Automatic Recognition of Continuous Signing of Brazilian Sign Language for Medical Interview

Robson Silva de Souza
Dept. of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
Campinas, SP, Brazil
robsonnddesouza@gmail.com

José Mario De Martino
Dept. of Computer Engineering and Industrial Automation
School of Electrical and Computer Engineering
Campinas, SP, Brazil
martino@unicamp.br

Janice Gonçalves Temoteo Marques
Dept. of Human Development and Rehabilitation
Faculty of Medical Sciences, University of Campinas
Campinas, SP, Brazil
janicetm@unicamp.br

Ivani Rodrigues Silva
Dept. of Human Development and Rehabilitation
Faculty of Medical Sciences, University of Campinas
Campinas, SP, Brazil
ivanirs@unicamp.br

***Abstract*—In this article, we present an automatic image recognition approach for assisting the communication between deaf patients, speakers of the Brazilian Sign Language (Libras), and hearing physicians. The aim of the approach is to help the interaction and exchange of information during medical interviews. Its scope is the automatic recognition of the continuous signing of Libras through the analysis of traditional video and depth data (RGB-D data). Recognition is performed by a cascade of two neural networks. The first, a convolutional neural network, encodes the visual input and extracts relevant features. The second, a recurrent neural network, learns the mapping of the extracted features into Brazilian Portuguese words. To train the recurrent network with videos of different lengths and word sequences, we use the Connectionist Temporal Classification approach. Experiments using a dataset of 280 videos encompassing 56 sentences composed of 67 different signs results in an accuracy of round 91%.**

***Index Terms*—Libras, Sign language recognition, Continuous signing, long short term memory, connectionist temporal classification**

I. Introduction

Anamnesis and clinical examination are the standard procedures of physicians to diagnose diseases and health problems of their patients. Anamnesis is a process of interviewing the patient to collect information about his/her current health complaints and medical history. The precise disclosure, correct understanding, and assessment of this information are preconditions for an effective diagnosis and the identification of the appropriate therapy. However, the effectiveness of the medical interview is jeopardized if the physician and the patient do not have a common language for communication. That is usually the case when we consider a deaf patient who has sign language as his/her first language and does not master the written language of the physician who, by his/her side, does not understand sign language. A common solution to overcome this problem is to have a sign language interpreter assisting the deaf patient during the interview. Besides the operational difficulties of organizing an interpreter, another important drawback is the uncomfortable situation created by the introduction of a third party in the medical interview. During a medical interview, the patient should feel comfortable enough to share very personal and sensitive information, providing any and all relevant information to help the doctor to make a correct diagnosis. A solution to overcome this potential breach of patient-doctor confidentiality is to provide a robust computer-based solution to support the communication between physicians and deaf patients. Although the interaction between doctor and patient is a two-way process, in this article, we focus only on the issue of automatic recognition of continuous signing based on computer-based recognition of video imagery. Our study deals specifically with the Brazilian Sign Language. However, the findings can be extended and applied to other sign languages.

Sign languages convey information by the movement of the hands, body, and face. They are perceived by vision. There is not a single, universal sign language used worldwide by deaf people. Each country has its own sign language [1]. The sign language of a country is independent of its oral language. For example, Deaf Americans speak the American Sign Language (ASL), the Deaf in the UK use the British Sign Language (BSL), and Deaf Australians speak the Australian Sign Language (Auslan). Deaf Brazilians use the Brazilian Sign Language (Libras).

There is an increasing research interest in automatic sign language recognition in recent years. Automatic sign language recognition applies computer vision combined with machine learning techniques to analyze and translate, into a written form, videos with sign language content.

The development of robust automatic sign language recognition systems is challenging. Several techniques have been proposed for automatic sign language recognition for a variety of sign languages, including the Brazilian Sign Language (Libras). Most efforts, however, have been limited to the study of isolated sign recognition, postures representative of cardinal

 ISBN: 978-1-61208-916-4

numbers (0 to 10), and the manual alphabet or fingerspelling. Research on continuous signing recognition is still rare.

In this article, we present a method for automatic continuous sign language recognition of Libras during medical interviews. Applying the method, we implement an approach based on Deep Learning that is capable of finding and using extracted data from signing from full-frame sequences. Therefore, it aligns sequences of video frames displaying Libras content to sequence glosses. A gloss is a word, in our case a Portuguese word, that is consistently used to label a sign within the corpus, regardless of the meaning of that sign in a particular context or whether it has been systematically modified in some way [2]. As pointed out in [3], glosses are a convenient way to write down the meaning of a sign, as they use another language to represent the signs.

The main contributions of this article are:

- The construction of a robust and representative dataset, composed of RGB information and depth of signage in Libras in order to contribute to the advancement of the research in this area.
- Execution of a Depth-Wise Separable Convolutional Network (DWSCN) based architecture, as feature extractor preprocessor. Insofar as we know, we are the first to employ this type of architecture in continuous sign language recognition systems.
- The development of a new architecture of sequential learning, based on recurrent neural networks and Connectionist Temporal Classification (CTC), which learn to find and store relevant data in its memory cells from the full-frame sequences, without importing in its subsystems structures that process image patches.

The remainder of the paper is organized as follows: Section II contains a review of relevant related work. Section III presents our approach. Section IV describes the experiments performed, and Section V presents the conclusions.

II. Related Work

The recognition of continuous signing is a far more complex task than the recognition of isolated signs, requiring more sophisticated methods to deal with the dynamics of production and the transition between signs. On the other hand, continuous signing recognition systems are more appropriate for real-world scenarios of interpersonal communication. However, it is observed that there is still little research that seeks to solve this problem. In the following paragraphs, we present approaches aimed at recognizing continuous signing based on computer vision.

Research using deep learning models has increased considerably in recent years. The work of [4] proposes an approach that breaks down the problem of recognizing signs into a series of expert systems called subunits. Each subunit consists of three layers of neural networks; Convolutional Neural Network (CNN) for extraction of spatial features, Bidirectional Long Short-Term Memory (BLSTM) [5], an extension of LSTM [6] that temporarily models the features and a loss layer based on the CTC. A recent work, [7], also uses CNN and LSTM but encapsulated it in an Hidden Markov Model (HMM) model following the hybrid approach used in his previous work, this time exploring sequential parallelism to learn sign language, mouth shapes, and hand shape classifiers.

The works [8]–[12] use CNNs as feature extractors, a 3D CNN model, or a 3D residual convolutional network (3D-ResNet). For modeling and sequential learning, they use dilated convolutional networks or RNNs such as LSTM, Gated Recurrent Unit (GRU) [13] and their variants in combination with the CTC algorithm. Among these approaches, [14] is the one that achieved the best performance in the RWTH-PHOENIX-Weather dataset and also in a set of images captured by the Kinect called CSL-25K, which covers 100 daily life sentences expressed in Chinese Sign Language (CSL).

In our proposal, we also use recurrent neural networks with CTC, but differently from the other approaches, we apply depth-wise separable convolutional network that contains far fewer parameters and is computationally cheaper than the state-of-the-art convolutional neural networks, as for example VGG16 [15], ResNet50 [16] and InceptionV3 [17].

III. Method

In this section, we present the protocol used to build a sign language dataset in the context of a medical interview and our approach for recognizing continuous signing in Libras.

A. Dataset construction

The existence of a dataset composed of Libras sentences related to medical interviews is fundamental to develop and test our approach. No publicly available image databases of continuous signing in Libras have been found.

Through the study of existing datasets of other sign languages [18] and with the intent of meeting our objectives, we developed specifications to be followed for the construction of our dataset. The proposal is to develop a robust dataset that simulates the internal environment of a clinic with artificial lighting, in which the deaf volunteer or interpreter performs the sign naturally.

Fig. 1 shows the execution flow. Thereafter, each module will be described in details.

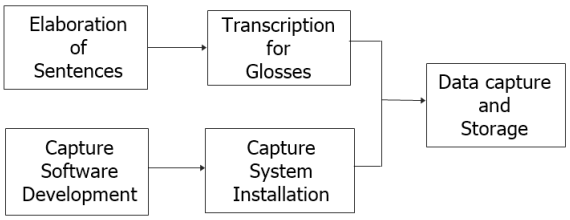

Fig. 1. Execution flow for the construction of the dataset

Sentences elaboration. It comprises the elaboration of sentences in the Portuguese language related to the answers of a patient in the context of a medical consultation (general practitioner). The sentences are established through the study of signs and manifested individual symptoms accordingly to the anamnesis medical procedure described in [19] and [20].

Transcription of the sentences in Portuguese to glosses. It is created through the assistance of a fluent sign language specialist. The right columns on Tables I and II present the transcriptions of the sentences from the previous stage.

TABLE I
EXAMPLES OF SENTENCES DEVISED IN PORTUGUESE LANGUAGE AND ITS TRANSCRIPTIONS TO GLOSSES.

#	Target	Prediction
1	Eu tenho febre	EU FEBRE
2	Eu estou fraco	EU FRACO
3	Eu estou com diarréia	EU TER DIARRÉIA
4	Meu braço esquerdo dói	MEU BRAÇO-ESQUERDO DOR
5	Minha urina está marrom	MEU XIXI COR MARROM

TABLE II
EXAMPLES OF SENTENCES DEVISED IN PORTUGUESE LANGUAGE AND ITS TRANSCRIPTIONS TO GLOSSES - VERSION IN ENGLISH

#	Target	Prediction
1	I have fever	ME FEVER
2	I am weak	ME WEAK
3	I have diarrhea	ME HAVE DIARRHEA
4	My left arm hurts	MY LEFT-ARM PAIN
5	My urine is brown	MY PEE BROWN COLOR

Capture device and development of the capture software. The data recording is made through the Kinect device v2 for Windows. The capture application is developed using Kinect's own software development kit (SDK). This application captures and stores RGB images, depth images, and mapped images (RGB images mapped on the depth images), in which all pixels not belonging to the signer are converted to black.

Installation of the capture system. The Libras signing recordings executed by the volunteer are made in a laboratory, with artificial illuminations and homogeneous scene background.

Capture and data storage. A Libras interpreter teacher, member of the research team helps with the video acquisition. During the signing the images are captured and stored on the computer.

B. Our approach

The approach that recognizes continuous Libras signing includes a CNN-based model for features extraction and an RNN architecture for learning the spatial-temporal dependencies that exist between the sentence signs. To solve the alignment problem between the probability sequences in the RNN outputs with the sequences of glosses, we used CTC.

Fig. 2 presents a general view of our approach composed of three main models. The first comprises spatial modeling, while the others encompass sequential learning and a CTC loss layer to decode categorical probabilities in sequences of glosses.

Features extraction. DWSCN is used for representations of spatial features of the frame sequences. The pre-trained MobileNetV1 [21] operational model is among the models

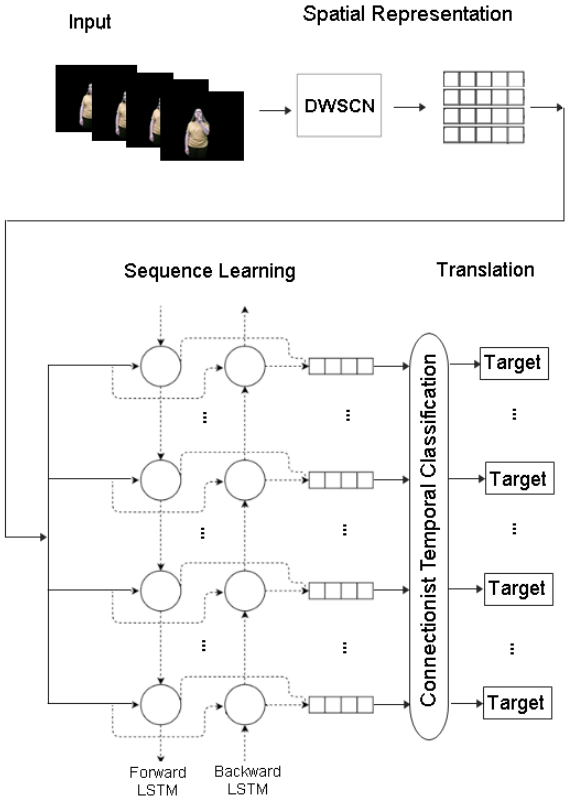

Fig. 2. Overview of our continuous sign language recognition approach

based on the DWSCN. The use of pre-trained models enables developing efficient models in situations of limited data availability, in addition to reducing processing time [22].

MobileNetV1 was pre-trained on ImageNet [23] and has characteristics of having reduced size (17MB) and reduced number of parameters (4,2 million) when compared to other state-of-the-art models.

To use MobileNet as a feature extractor preprocessor, the softmax classification layer (SM) and the completely connected layer (FC) have been removed, keeping all the depthwise separable convolution blocks and the Average Pooling layer.

All dataset images are processed by the resulting model. As the last layer has 1024 nodes, each image will be represented as a 1024 value vector. Each video sample results in a three-dimensional array of dimensions equal to 1 x number of frames x 1024 features. Since the number of frames are different between the videos, the padding in each array has been performed to allow the concatenating of all feature arrays. The data is used as the entry to train our model based on recurrent neural networks.

The glosses are coded in categorical variables and together with the feature arrays are used as input to train our model based on recurring neural networks. This is a weakly supervised learning problem, that is, the gloss sequences are available but not its time limits.

Sequential learning. Our approach uses BLSTM to model

 ISBN: 978-1-61208-916-4

the correspondences between the input sequences and output glosses. This architecture is capable of storing data for long periods of time and try to avoid the explosion of the gradient, a common problem of the Vanilla neural networks.

To implement a BLSTM network, it takes two parallels layers of LSTM cells, backward LSTM and forward LSTM, each of them being responsible for processing the information in the direction of time. The final hidden layer is given by the concatenation of the two networks.

The memory neurons of an LSTM are called cells. Fig. 3 presents the structure of a BLSTM network and highlights one single memory cell. The cells are capable of storing data in the course of a sequence through units called gates. According to [24], these units calculate the weights that connect them to avoid the gradient degradation through parameterized or manually chosen values.

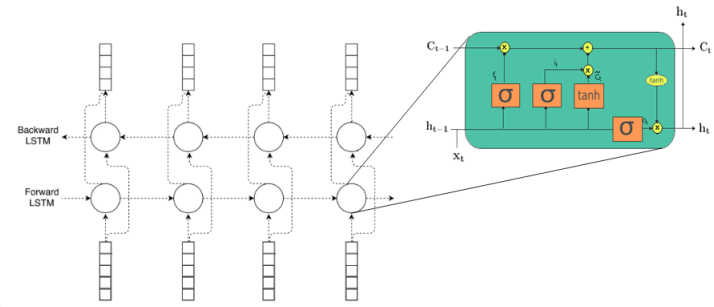

Fig. 3. BLSTM network structure, highlight to a single memory cell

A softmax activation function on a fully connected layer is used in the network output and is applied to each time frame.

Connectionist Temporal Classification. In the BLSTM training phase, CTC is used to calculate the cost value. During prediction, it decodes the probability matrices of the softmax function in gloss sequences.

To allow the CTC algorithm to decode the target sequence, one more unit is introduced to the total number of labels in the softmax output layer. This unit refers to a token named blank, that models the transitions between different labels.

Let us consider the mapping of the input frames sequence $X = [x_1, x_2, ..., x_T]$, for the sequences of output words $Y = [y_1, y_2, ..., y_T]$. The CTC cost function for a pair (X, Y) has the conditional probability $p(Y/X)$ equal to the sum of all the valid paths $A \in A_{XY}$, calculating the probability $p_t(a_t|X)$ to a single step-by-step alignment following (1).

$$p(Y/X) = \sum_{A \in A_{XY}} \prod_{a}^{b} p_t(a_t|X) \tag{1}$$

For a training set M, the model parameters are tuned to minimize the negative log-likelihood. That way, the CTC objective function is given by (2).

$$Loss_{CTC} = \sum_{(X,Y) \in M} -\log p(Y/X) \tag{2}$$

To calculate the CTC loss efficiently, the Forward-Backward algorithm given in [25] is used.

IV. Experiments

This section reports on the experiments performed and the performance of our architecture in continuous Libras signing recognition.

A. Dataset

In order to develop and test our approach, 280 sentences signed by a professional interpreter were captured, corresponding to 5 repetitions of 56 sentences. 42663 frames are obtained at a rate of 30 fps. The number of glosses is equal to 67. The number of glosses per sentence varies from 2 to 6 and the number of frames per sentence varies from 124 to 277.

B. Evaluation Metrics

The Word Error Rate (WER) is the metric widely [8], [9], [26], [27], [12], [10], [11], [14], [4], [28], [29], [30] used in continuous sign language recognition work and, therefore, will be the metric used in this paper. The WER is given by (3).

$$WER = \frac{I + D + S}{N} \tag{3}$$

Where I is the number of errors entered, D is the number of deletion errors, S is the number of substitution errors, and N is the total number of glosses in the reference sentence.

The accuracy is given by (4).

$$acc = 1 - WER \tag{4}$$

C. Training and Evaluation

We performed experiments on an Nvidia RTX 2080Ti, and the model is implemented in the Keras framework [31], using tensorflow [32] as a backend. In our experiments, we used 80 percent of the data (224 sentences) for the training set and 20 percent for the test set (56 sentences).

The simulations performed processed the images mapped. Initially, these images were resized to 224 X 224 pixels, dimensions expected by the MobileNetV1 network.

After spatial modeling with our structure based on DWSCN, the resulting feature matrix has a dimension equal to the number of samples x time steps x features.

According [33], training small datasets has some challenges, as the network effectively memorizes the training dataset. The author recommends that adding noise is an approach to improve the generalization error and to enhance the structure of the mapping problem during learning. Thus, we applied Gaussian noise, at the entrance of the BLSTM network, with a standard deviation of 0.5 during the training phase.

The training of our BLSTM architecture is performed by implementing the backpropagation algorithm through time, [34]. The initialization for the recurrent weights matrix is the orthogonal [35], for non-recurring weights the glorot uniform [36] and the vector bias is initialized with zeros. The optimizer used is the Root Mean Square Propagation (RMSprop) [37] with a learning rate of 0.01, a discounting factor of 0.9, a

 ISBN: 978-1-61208-916-4

momentum of zero, (default values in the framework), and a batch size of 82. Then, we use the CTC beam decoder described in the work of [38] to decode sentences with beam width 10.

For the aforementioned configurations, dozens of experiments were carried out using different network topologies, with a maximum of 4 layers (1 to 3 recurring layers and a completely connected layer) and the number of neurons equal to powers of 2 in the range of 2 to 512. The last layer is fixed with 68 neurons (one for each vocabulary label plus the blank label). Given the stochastic nature of the algorithms used, repetitions of the tests are performed in order to determine the most promising models.

In order to detect overfitting and determine the most promising models, a validation set is adopted, based on the training set, consisting of 60 sentences. During the training, at the end of each epoch, the value of the loss CTC is calculated in the validation set, and the best model in each training is determined according to the lowest value of the loss in that set.

Also, to identify and soften the effect of overfitting, we used the method of regularization called dropout, presented in [39]. Dropout values equal to 0.5 were applied for both recurrent and non-recurrent connections.

Among the best models that fit the data, the simplest model, that is, with the least hyperparameters, is considered the most plausible to be used in the test set.

D. Results

Our best result is achieved by configuring two recurrent layers with 32 and 64 neurons, respectively. At the end of 30000 epochs, it was determined that the best model corresponds to epoch 21422. The values of the initial weights and the settings referring to that model are saved and stored for reproducibility, as well as for use in the unseen data set during the training.

Of the 56 sentences in the test set, 11 obtained some kind of error in the model prediction. The average WER was 8.92% and therefore, an accuracy of 91.07%. In Table III we can observe some errors found, comparing the results of the model with the ground-truth sentences. Bold words are associated with errors in prediction. Table IV presents the equivalent results in English.

Therefore, the errors found were: 13 substitutions, 2 insertions, and no deletions. Low values in relation to the total amount of glosses existing in the dataset demonstrating the effectiveness of our architecture.

TABLE III
SENTENCES WITH PREDICTION ERRORS

#	Target	Prediction
1	COMEÇAR ANTEONTEM	COMEÇAR **ONTEM**
2	COMEÇAR QUINTA-FEIRA PASSADA	COMEÇAR **TERÇA-FEIRA** PASSADA
3	COMEÇAR SEGUNDA-FEIRA PASSADA	COMEÇAR **QUARTA-FEIRA** PASSADA
4	COMEÇAR TERÇA-FEIRA PASSADA	COMEÇAR **QUINTA-FEIRA** PASSADA
5	MAU-HÁLITO FEDOR TER	MAU-HÁLITO FEDOR **VERMELHO** TER
6	MEU DENTE DOR	MEU **COSTAS TER**
7	MEU NARIZ DOR	MEU **OLHO-ESQUERDO INCHADO**
8	OLHO-DIREITO APONTAR VERMELHO TER	OLHO-DIREITO APONTAR VERMELHO **SABOR NÃO-TER**
9	MEU OLHO-DIREITO DOR	MEU OLHO-DIREITO **INCHADO**

TABLE IV
SENTENCES WITH PREDICTION ERRORS - VERSION IN ENGLISH

#	Target	Prediction
1	START BEFORE-YESTERDAY	START **YESTERDAY**
2	START THURSDAY PAST	START **TUESDAY** PAST
3	START MONDAY PAST	START **WEDNESDAY** PAST
4	START TUESDAY PAST	START **THURSDAY** PAST
5	BAD-BREATH BAD-SMELL HAVE	BAD-BREATH BAD-SMELL **RED** HAVE
6	MY TOOTH PAIN	MY **BACK HAVE**
7	MY NOSE PAIN	MY **LEFT-EYE SWOLLEN**
8	RIGHT-EYE POINT RED HAVE	RIGHT-EYE POINT RED **FLAVOR DO-NOT-HAVE**
9	MY RIGHT-EYE PAIN	MY RIGHT-EYE **SWOLLEN**

V. CONCLUSIONS

In this article, we presented an approach for recognition of continuous signing of Libras. This approach receives sequences of images of a person communicating in Libras and translates signs to the Portuguese language. The efficacy of our proposed methods was proven by state-of-the-art results.

In general, when compared to other approaches in the literature, our approach demonstrates a series of advantages:

i) It does not depend on the extraction of manual features, specifically designed for a domain and laboriously calculated from the geometry of the hands and arms.

ii) It takes into account characteristics related to non-manual expressions, such as movements of the face, eyes, head, and torso, instead of using only continuous sequences of the hands.

iii) Contrary to other studies' continuous signing recognition, which performs the feature extraction process in video segments related to isolated signs, our spatial representation module is processed on the entire video. Our choice is due to the fact that video representation based on fixed-length signs can compromise the continuous recognition of signing in real situations since the same sign varies in length in a video, even when performed by the same person in different situations

iv) Our spatial modeling, which is based on depthwise separable convolutions, reduces the latency and favors the development of real-time sign recognition because of the accuracy and the number of parameters and demanded calculations. This is a great advantage when compared to other convolutional neural networks.

 ISBN: 978-1-61208-916-4

v) Our architecture based in BLSTM with CTC learns to find and store information relevant memory cells from the data channels included in full-frame sequences. This is done without injecting subsystems in its structure that process image patches. Consequently, our approach presents a greater capacity for temporal learning compared to studies that import extra data in its system to ease the learning.

Our approach demonstrates the potential to be applied in signing recognition on heterogeneous backgrounds due to the use of Kinect, which performs the segmentation of the individual while capturing the depth and color of images. In our upcoming work, we intend to include more signage and diversify the recording scenarios of our dataset images, as well as increase the vocabulary in order to maximize the robustness of our recognition approach.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

[1] N. Timmermans *et al.*, *The status of sign languages in Europe*. Council of Europe, 2005.

[2] T. Johnston, "From archive to corpus: transcription and annotation in the creation of signed language corpora," in *Proceedings of the 22nd Pacific Asian Conference on Language, Information, and Computation*, pp. 16–29, 2008.

[3] A. Baker, B. van den Bogaerde, R. Pfau, and T. Schermer, *The linguistics of sign languages: An introduction*. John Benjamins Publishing Company, 2016.

[4] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3075–3084, IEEE, 2017.

[5] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] O. Koller, C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[8] J. Pu, W. Zhou, and H. Li, "Dilated convolutional network with iterative optimization for continuous sign language recognition.," in *IJCAI*, pp. 885–891, 2018.

[9] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for continuous sign language recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4165–4174, 2019.

[10] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7361–7369, 2017.

[11] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Transactions on Multimedia*, 2019.

[12] H. Zhou, W. Zhou, and H. Li, "Dynamic pseudo label decoding for continuous sign language recognition," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1282–1287, IEEE, 2019.

[13] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[14] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-temporal multi-cue network for continuous sign language recognition," *arXiv preprint arXiv:2002.03187*, 2020.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision. 2015," *arXiv preprint arXiv:1512.00567*, 2015.

[18] N. B. Ibrahim, H. H. Zayed, and M. M. Selim, "Advances, challenges and opportunities in continuous sign language recognition," *Journal of Engineering and Applied Sciences*, vol. 15, no. 5, pp. 1205–1227, 2020.

[19] F. Veiga and A. B. Souza, *Physical Exam Manual*. Elsevier Brasil, 2019.

[20] M. H. Swartz, *medical semiology treatise*. Elsevier Brasil, 2015.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[22] J. Brownlee, *Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python*. Machine Learning Mastery, 2019.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, ACM, 2006.

[26] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical lstm for sign language translation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[27] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, "Video-based sign language recognition without temporal segmentation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] O. Koller, R. Bowden, and H. Ney, "Automatic alignment of hamnosys subunits for continuous sign language recognition," *LREC 2016 Proceedings*, pp. 121–128, 2016.

[29] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3793–3802, 2016.

[30] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: Hybrid cnn-hmm for continuous sign language recognition," in *Proceedings of the British Machine Vision Conference 2016*, 2016.

[31] F. Chollet *et al.*, "Keras." https://keras.io, last accessed on 02/08/21, 2015.

[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[33] J. Brownlee, *Better Deep Learning: Train Faster, Reduce Overfitting, and Make Better Predictions*. Machine Learning Mastery, 2018.

[34] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[35] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[36] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[37] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, p. 8, 2012.

[38] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[39] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

 ISBN: 978-1-61208-916-4