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

Authors: Robson Souza[1], José Mario[1], Janice Gonçalves[2], Ivani Rodrigues[2]

[1]School of Electrical and Computer Engineering. University of Campinas
[2]Faculty of Medical Sciences, University of Campinas

The Sixth International Conference on Informatics and Assistive Technologies for Heath-Care, Medical Support and Wellbeing – HEALTHINFO 2021

Barcelona, Spain, October 03-07, 2021

UNICAMP

IARIA

Robson Souza

- Robson Souza received the master's degree in electrical engineering from the Federal University of Amazonas, Brazil in 2015. He is currently a doctoral student majoring in engineering at the School of Electrical and Computer Engineering, University of Campinas.

- His research interest lies in the areas of Computer Vision and Machine Learning, working mainly in the development of projects in assistive technologies.

UNICAMP

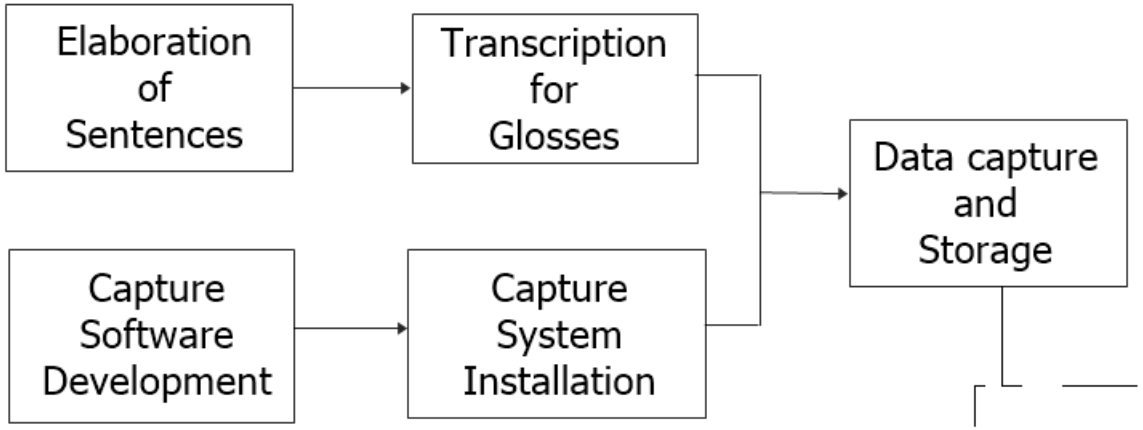
IARIA

Aims and contributions of our paper

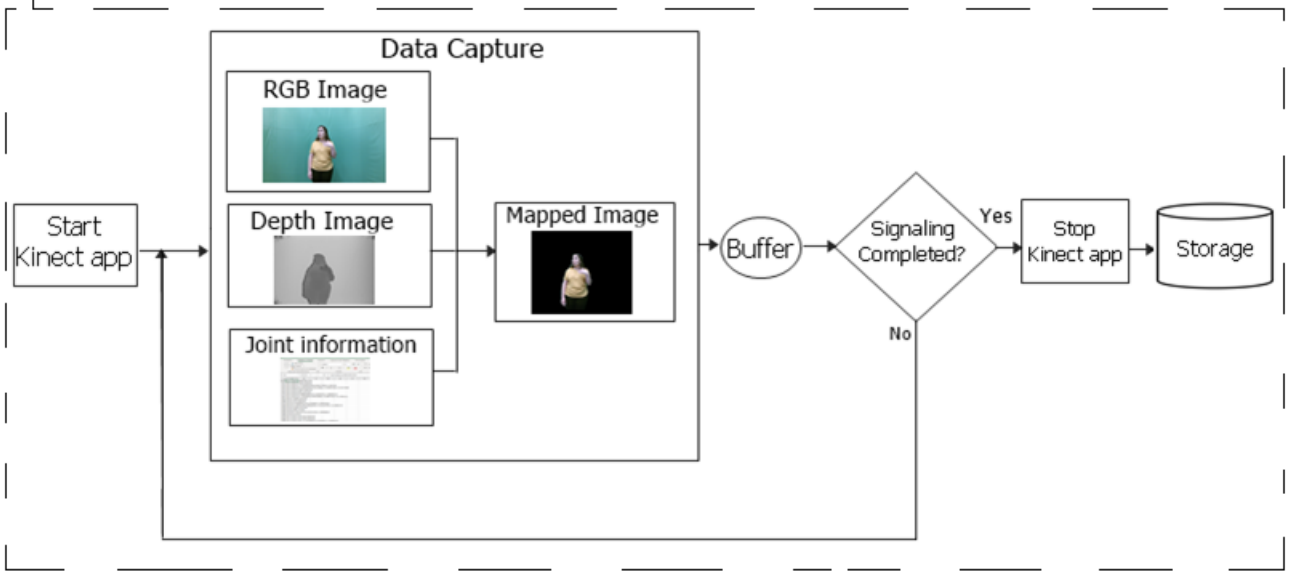
- **In our paper, we aimed at:**
- The main objective of this work is to develop an approach for the automatic recognition of Libras continuous signing, in the context of a medical interview, using data captured by a depth camera.

- **Contributions of our study are:**
- Implementation of a computer vision based translation system that aligns Libras video frames to Portuguese word sequences.
- Development of an optimization algorithm for the translation system in order to maximize the recognition of frame-by-frame signing and reduce the computational cost.
- Construction of a dataset composed of RGB and depth information of Libras content in order to support research in the area.

Execution flow for the construction of the dataset

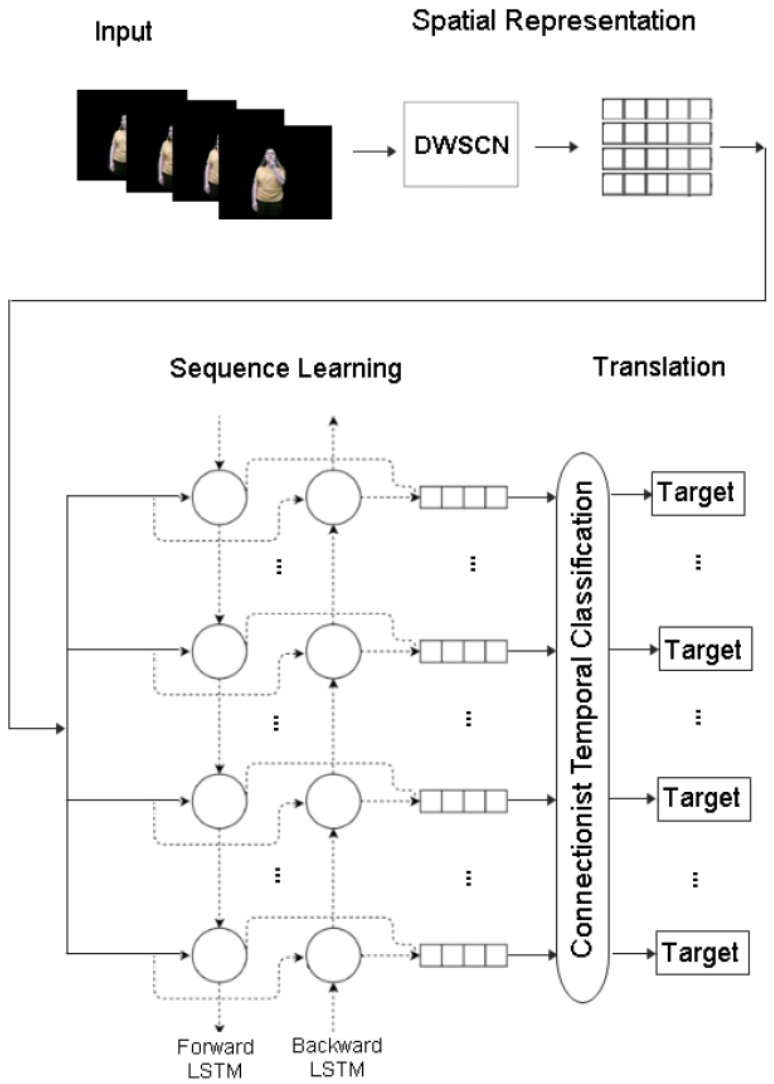


A gloss is a word, in our case a Portuguese word, that is consistently used to label a sign within the corpus.

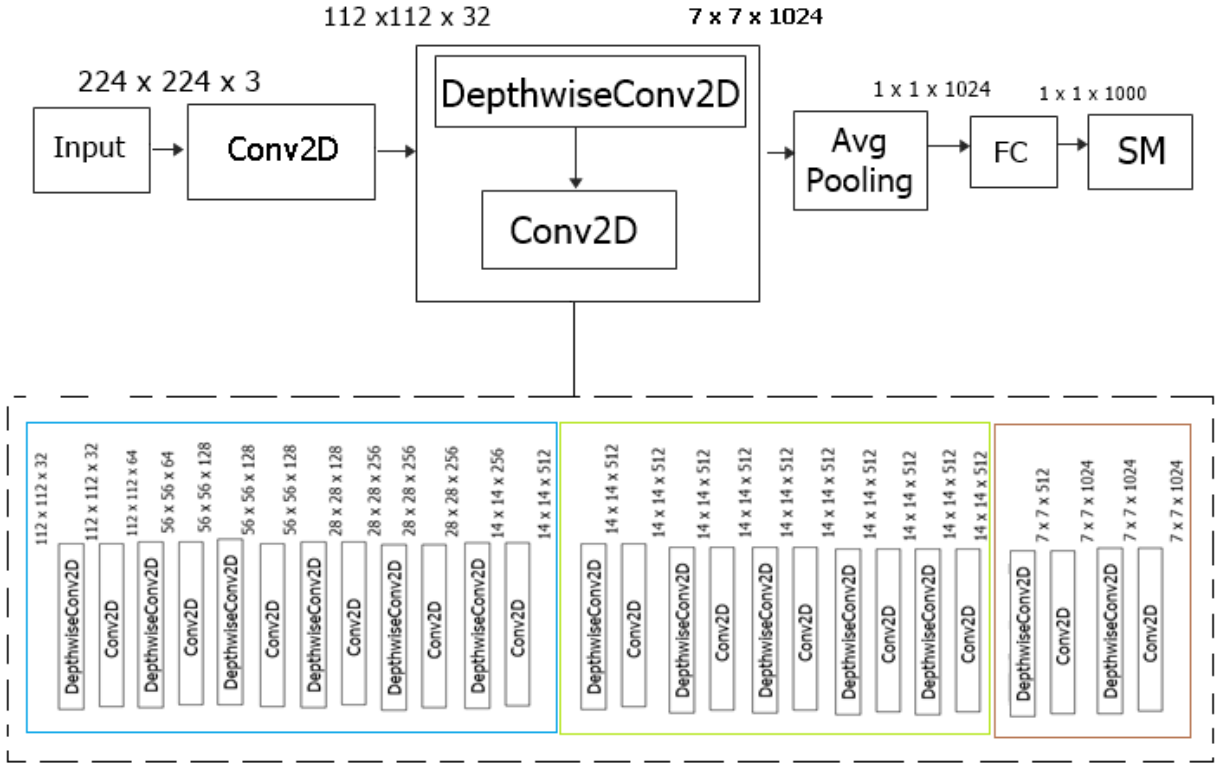


Overview of our continuous sign language recognition approach

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



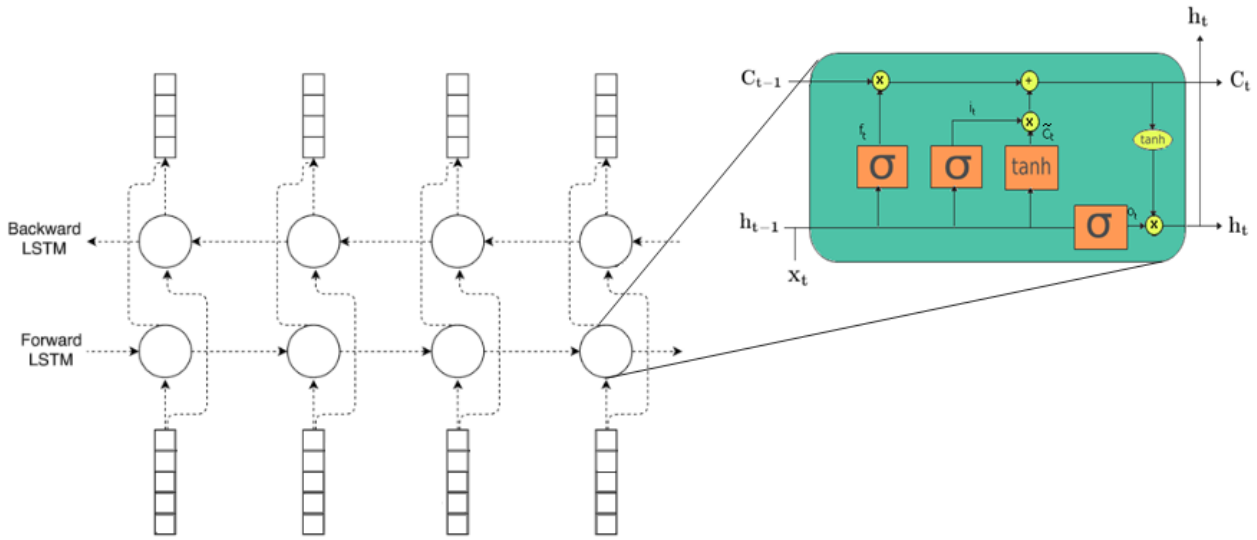
Features extraction: Execution of a Depth-Wise Separable Convolutional Network (DWSCN) based architecture



The pre-trained MobileNetV1 operational model is among the models based on the DWSCN.

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 and in Connectionist Temporal Classification (CTC).



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

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

Our approach uses BLSTM to model the correspondences between the input sequences and output glosses.

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

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.

EXPERIMENTS

- 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 - Training and Evaluation

Arguments	Value
initialization for the recurrent weights	orthogonal
initialization for the non-recurring weights	glorot uniform
initialization for the vector bias	zeros
optimizer	Root Mean Square Propagation (RMSprop)
learning rate	0.01
batch size	82
Beam width CTC	10

Results

- 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.

- 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.

Sentences with prediction errors

#	Target	Prediction
1	START BEFORE-YESTERDAY [COMEÇAR ANTEONTEM]	START YESTERDAY [COMEÇAR ONTEM]
2	START THURSDAY PAST [COMEÇAR QUINTA-FEIRA PASSADA]	START TUESDAY PAST [COMEÇAR TERÇA-FEIRA PASSADA]
3	START MONDAY PAST [COMEÇAR SEGUNDA-FEIRA PASSADA]	START WEDNESDAY PAST [COMEÇAR QUARTA-FEIRA PASSADA]
4	START TUESDAY PAST [COMEÇAR TERÇA-FEIRA PASSADA]	START THURSDAY PAST [COMEÇAR QUINTA-FEIRA PASSADA]
5	BAD-BREATH BAD-SMELL HAVE [MAU-HÁLITO FEDOR TER]	BAD-BREATH BAD-SMELL RED HAVE [MAU-HÁLITO FEDOR VERMELHO TER]
6	MY TOOTH PAIN [MEU DENTE DOR]	MY BACK HAVE [MEU COSTAS TER]
7	MY NOSE PAIN [MEU NARIZ DOR]	MY LEFT-EYE SWOLLEN [MEU OLHO-ESQUERDO INCHADO]
8	RIGHT-EYE POINT RED HAVE [OLHO-DIREITO APONTAR VERMELHO TER]	RIGHT-EYE POINT RED FLAVOR DO-NOT-HAVE [OLHO-DIREITO APONTAR VERMELHO SABOR NÃO-TER]
9	MY RIGHT-EYE PAIN [MEU OLHO-DIREITO DOR]	MY RIGHT-EYE SWOLLEN [MEU OLHO-DIREITO INCHADO]

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

$$acc = 1 - WER$$

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.

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%.**

UNICAMP

Conclusions

- 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.

- 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.