Deep Learning Applied to Sign Language[⋆]

Jérôme Fink[1], Anthony Clève[1], and Benoît Frénay[1]

NADI Institute - PReCISE Research Center
University of Namur - Faculty of Computer Science
Rue Grandgagnage 21, Namur - Belgium.

1 Context

The French Belgian Sign Language laboratory (LSFB-lab) recorded 150 hours of conversations to create a corpus useful to analyze the Belgian sign language and its usage [3]. The videos are captured in RGB without depth information. 15 hours of these video have been annotated. They picture 52 different speakers and contains 670 signs with, at least, 10 occurrences. The LSFB-lab is interested in the development of an algorithm able to translate sign language to french. It could help the deaf community during medical or administrative appointments as finding the right word is difficult in those specific contexts.

This thesis aims to make a first step in that direction by exploring existing methods developed in video recognition and analyzing the LSFB dataset.

2 Related Work

Deep learning is successfully applied to image recognition challenges. To tackle the problem of video recognition, methods for handling the temporal dimension have been developed. The mains methods identified are[1] :

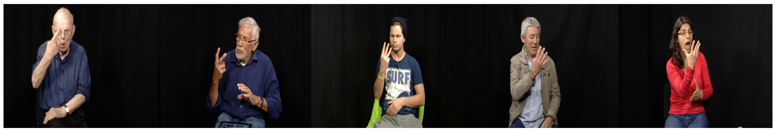
- *LSTM networks*: Each frame is presented to a network able to retain information about the evaluation of the previous frame. The retained information captures the temporal dimension of the video.
- *Two stream networks*: Two parallel networks focus on different aspects of the frames constituting a video. Usually, one network focuses on the temporal while the other extracts 2D information from one particular frame. Their results are then fused.
- *3D Convolution layer*: An extension of the 2D convolution able to handle the temporal dimension induced by video.

The state of the art architectures mix those methods.

Other datasets have been considered to find the best for deep learning training. Some identified datasets provide more examples per signs but the LSFB dataset has the advantage to capture more realistic signs as they are performed during an active conversation. It is also one of the biggest dataset available in terms of number of signs available. Each region has its sign language, therefore it is impossible to fuse two datasets from different region of the world.

3 Preliminary Experimental Results

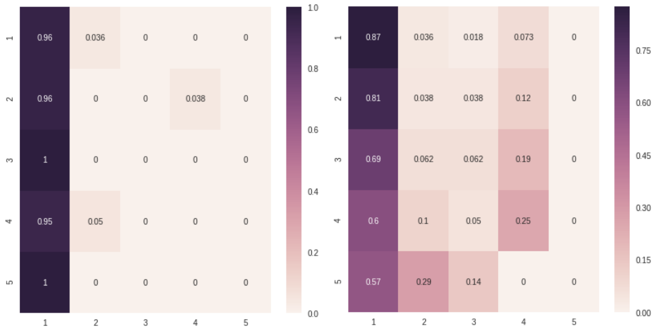
To reduce the scope of the problem, the thesis proposes and introduces the *Sign Language MNIST dataset* with five classes shown by Fig 1.



(a) Sign for 1 (b) Sign for 2 (c) Sign for 3 (d) Sign for 4 (e) Sign for 5

Fig. 1: Example from the Sign Language MNIST dataset picturing digit from 1 to 5. The angle of the camera hides one finger in the example for the number 5.

Preliminary experiments have only considered one frame for each video. The VGG16 and InceptionV3 models pre-trained on Imagenet have been tested. They have been retrained using a triplet loss [2]. Fig 2 shows the result of this first experiment. VGG16 is able to capture more useful information than InceptionV3 who is the state of the art in visual recognition.



(a) InceptionV3 (b) VGG16

Fig. 2: Confusion matrix for the Sign Language MNIST dataset

Even if the results obtained in this master thesis are still preliminary, they show that deep learning is an interesting option for sign language recognition. This master thesis will be the starting point of new research in the context of an interdisciplinary project, whose aim is to create a bilingual and context-sensitive dictionary for the french belgian sign language.

References

1. Joao Carreira and Andrew Zisserman, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR (2017)
2. Herve Bredin, TristouNet: Triplet loss for speaker turn embedding. ICASSP (2017)
3. Laurence Meurant, Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB). (2015)