A Multistream Model for Continuous Recognition of Lexical Units in French Sign Language

Yanis OUAKRIM[1, 2] Denis BEAUTEMPS[1] Michèle GOUIFFÈS[2] Thomas HUEBER[1]
Frédéric BERTHOMMIER[1] Annelies BRAFFORT[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

[2] LISN, Univ. Paris-Saclay, CNRS, 91405 Orsay, France

Résumé – Les langues des signes constituent le premier moyen de communication des personnes sourdes, pourtant elles restent peu dotées du point de vue du traitement automatique des langues, et les outils linguistiques (traducteurs, concordanciers) qui leur sont dédiés sont rares. Dans ce papier, partant d'un corpus de dialogues en LSF de taille limitée, nous proposons un tout premier modèle d'apprentissage pour la reconnaissance de signes lexicaux en Langue des Signes Française (LSF). Les signeurs sont représentés par leur pose 3D, qui est fournie à un réseau de neurone récurrent bi-directionnel, entrainé à l'aide de l'approche CTC (Connectionist Temporal Classification). L'apport des principaux articulateurs est évalué dans la reconnaissance de la LSF et des approches mono et multiflux (un modèle par articulateur) sont comparées.

Abstract – Sign languages are the primary means of communication for deaf people, yet they remain poorly endowed from a natural language processing perspective, thus linguistic tools (translators, concordancers) dedicated to them are rare. In this paper, starting from a corpus limited size, we propose a very first learning model for the recognition of lexical signs in French Sign Language (LSF). Signers are represented by sequences of 3D pose, which are decoded into sign sequences thanks to a bidirectionnal recurrent neural network trained with a CTC loss (Connectionist Temporal Classification). Different strategies for combining the different articulators of LSF within the neural decoder (single vs. multistream architecture), as well as the contribution of main articulators to the overall performance of the system, are investigated.

1 Introduction

It is estimated that there are 169,000 French Sign Language (LSF) signers in the world today [17] [1], and although deaf people are increasingly being taken into account, many services are still difficult for them to access. Sign language processing aims to provide tools to make content and services available to deaf people.

Sign languages are visual languages; signers communicate by representing their ideas iconically in a three-dimensional signing space. Iconicity exists at all levels of sign languages (from sub-lexical to speech). There are lexical units stabilized in form and meaning that are often called signs and that can be listed in a dictionary. There are also illustrative structures that allow us to show while saying without using stabilized signs. These structures are created as needed and cannot be listed in a dictionary. There are other types of units such as pointing, fbuoys, or dactylology [3].

Sign languages do not rely solely on the signers' hands but also on the configuration of other 'articulators', such as facial expressions, mouth opening, arm configurations, shoulder movement, gaze, frowning, etc. The frequency and amplitude of use of each articulator can vary depending on the signer or the sign language [5].

As concerns automatic recognition of Sign Language, there are two distinct tasks : classification of isolated signs and recognition of signs in a continuous context. In the first case, the signs are signed independently and most often in a very standard way. In the second case, the signs to be recognized are co-articulated and undergo deformations of form or speed that depend on the context. In this article, we place ourselves in the framework of this second task.

Early work on Sign language recognition was based on glove-based system (e.g. [6] for LSF) but most recent studies have moved to deep learning approaches using large video corpora, some of them being annotated at the gloss level. [2] Among the few available corpora, we can cite the widely used RWTH-Phoenix corpus [14] in German Sign Language (DGS).It consists of 16 hours of DGS interpreted from a weather broadcast by 9 hearing signers. Because of its origin, the signs used in the corpus are mainly about the weather, which has the advantage of limiting the scope of the vocabulary (1082 signs).More recently,[15] proposed a dense annotation at the sign level of the British Sign Language (BSL) corpus BOBSL [1]. It results in a corpus of 1,467 hours with 24,800 annotated signs.

As concern French Sign Language (LSF), there is only one densely annotated video corpus : Dicta-Sign-LSF-v2 [4]. DictaSign-LSF-v2 is the first video corpus of continuous and natural LSF of conversations with 16 deaf signers (i.e. not interpreters) annotated in glosses. It consists of about 4h of spontaneous dialogues on the theme of travel and transportation, its vocabulary is made of 2,252 unique glosses, i.e. twice the size of the RWTH-Phoenix corpus, for a duration 4 times shorter.

It has been used in [3] for sign categorization, that is automatically predicting whether a sign is a lexical sign, a depicting sign, a pointing or a fbuoy.

1. Sign languages are not universal.

2. Glosses are lexical sign form identifiers, one gloss can correspond to multiple meanings.

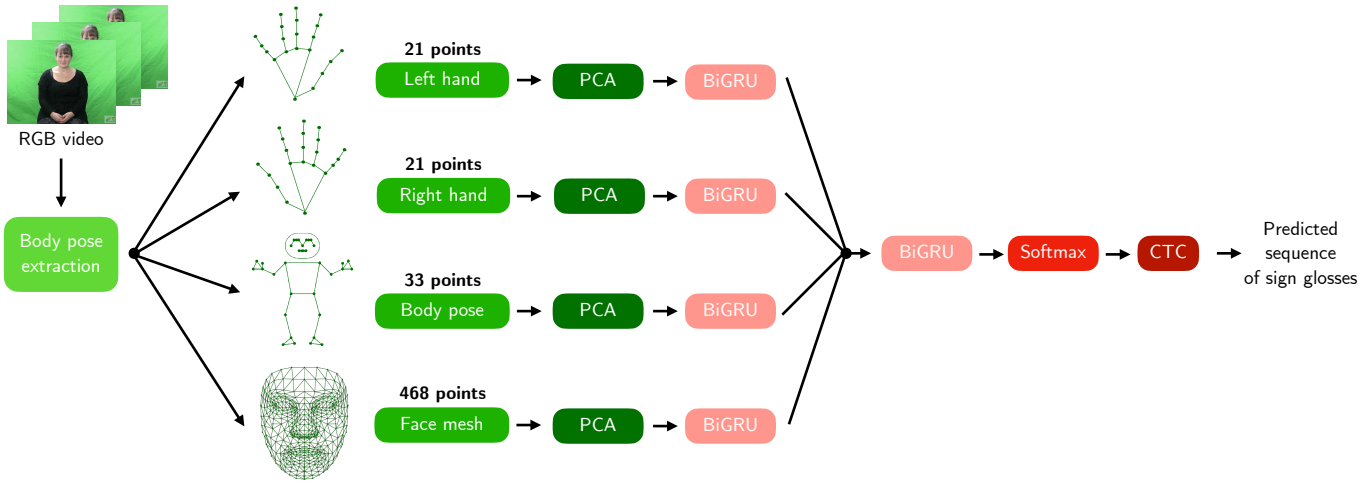

Figure 1 – Overview of the proposed multri-stream architecture for sign language recognition.

In recent studies on Sign Language processing, several techniques have been proposed to extract high-level features from raw videos, among which CNN, I3D [11], video transformers [10] and pose-based technique (a comparison of convolution vs. posed-based technique for isolated sign language recognition was proposed in [16]). When used for recognition, these techniques are typically combined with encoder-decoder models (RNN or Transformer-based models as in [8, 9]). Concerning the recognition performances, [8] which presents an architecture based on LSTM/CTC has a global performance of 43.1% WER in the limited linguistic context of speech to sign weather broadcast translation. Such models require a large amount of training data which is unfortunately not yet available in LSF.

The present study addresses the problem of (co-articulated) lexical signs recognition in LSF. To deal with the limited size of Dicta-Sign-LSF-v2, we proposed a lightweight model based on (i) a posed-based feature extraction technique, (ii) a bi-directional RNN (Bi-GRU) trained with a CTC loss (connectionist temporal classification). We also investigated different strategies for combining the different articulators of LSF within this neural decoder (single vs. multistream architecture). Finally, we assessed the contribution of the main articulators (hands, face and body pose) to the overall recognition performance.

2 Material and methods

2.1 Dataset

This work relies on the Dicta-Sign-LSF–v2 [4] corpus which has been annotated with glosses in a dense manner. The study is limited to the hundred most frequent signs in order to have a sufficient number of occurences per sign (> 47). The corpus is subsampled to have excerpts containing at least one sign belonging to the hundred most frequent classes. We only use the videos with one signer in the frame (the corpus includes a view with the two signers on top of individual recordings). The resulting corpus is made available to the community [3]. We obtain 5,737 sequences that have a maximum duration of 254 frames (11s). We reserve 10% of the sequences for testing, 10% for validation and the rest for training.

2.2 Feature extraction

The first step of the proposed pipeline consists in extracting a representation of the signers in the form of a pose for each video frame. For this purpose, we rely on the pre-trained MediaPipe Holistic model[2], which estimates a pose of the subjects from a raw video and is able to run in real time (> 20 fps). Obtained poses consist of a set of 3D points, grouped by articulator: 21 points per hand, 33 points for the body in general and 468 points for the face. Each signer is represented by 543 points, thus 1629 values. To limit the size of the input vectors, we apply a principal component analysis (PCA) on each articulator and reduce each articulator to a 20-dimension feature vector (explained variance > 99%). The obtained representation is thus particularly light (80 values with 4 articulators) in comparison with the raw list of coordinates (1,629 values).

2.3 Network architecture

Figure 1 provides an overview of the architecture of our neural network model. Similarly to [9] and [18], our model is composed of one input stream per articulator. Thus, with four articulators: two hands, face, and general body pose, we get 4 streams. Each of the streams has its own bidirectional GRU layer, capable of processing the pose data sequentially. The different streams are then concatenated into a single stream to pass through a second layer of bidirectional GRUs. The model ends with a *softmax* layer providing the posterior probability of each class at each timestep. The model is trained with a CTC loss [13] which eliminates the need for a temporal segmentation of the training set at the gloss-level (as opposed to a sliding window approach). At test time, the sequence of posterior probabilities is converted into a sequence of labels (i.e the glosses) using a beamsearch decoding algorithm [12]. This algorithm is used to consider several hypotheses when

3. Will be made available before the publication of the paper. Existing Dicta-Sign-LSF-V2 corpus can be found at https://www.ortolang.fr/market/corpora/dicta-sign-lsf-v2.

evaluating the performance. Also, in section 3.1, we also test a mono-stream alternative with an early-fusion of the 4 input streams as a single input stream.

Implementation details: We used the Keras and CTCKeras [19] python libraries for model development. Each layer of Bi-GRU is composed of 128 hidden units. We conducted several experiments to find optimal hyperparameters, as a result, we set the learning rate to 10^{-3}, the batch size to 32 and we train the model for 50 epochs. The resulting model has 1,373,029 parameters in total (the training time is about 30 minutes on a machine equipped with a Nvidia Quadro RTX 8000 GPU). With the same hardware, inference takes on average 55ms per sequence of 255 timestep (≈11s of video).

2.4 Metrics

To evaluate the performance of the recognition system, we rely on the Word Error Rate (WER) defined as: $WER = (S + D + I)/N$, where S, D and I correspond respectively to the number of substitutions, deletions and insertions and, N is the total number of glosses in the test set. We also report the WER when considering the top-2 hypotheses given by the beamsearch decoding algorithm. We call this metric Top-2 WER. The 95% confidence interval is derived from the Wilson method with continuity correction.

3 Results

3.1 Mono-stream or multistream ?

First, we compared the use of a mono-stream architecture with early concatenation (fusion) of the features of the 4 articulators with the use of multi-stream architecture with later concatenation where the 4 articulators are fed to 4 different input streams (see Figure 1). With the multistream architecture, we obtain a 7.7% improvement of the WER compared to the mono-stream case (WER of 53.1% in multi-stream vs WER of 60.8% in monostream). One hypothesis is that the multi-stream model allows the model to take into account the asynchrony between the articulators but also eases the ability of the model to learn articulator-specific characteristics (e.g. handshapes).

3.2 Articulators

Second, we assessed the contribution of each articulator w.r.t the overall performance. Table 1 shows a detailed comparison of the WER obtained depending on the articulators provided as input. Without surprise, the right hand (which is the dominant hand of most of the subjects) seems to be the articulator providing the most information. Ranked by performance level, we then obtained - right hand, left hand, body pose and face - which seems consistent with the structure of LSF.

3.3 Number of classes

Third, we evaluated the model performance w.r.t the number of glosses considered at training time. Table 2 presents both the WER and the Top-2 WER obtained when considering the 10, 25, 50 and 100 most frequent signs of the corpus. As expected, the performance significantly improves when we reduce the number of classes, likely because it increases the number of occurences per sign. The Top-2 WER shows that when a sign is badly predicted, the correct prediction often lies is the second hypothesis, the fewer the number of class, the bigger the gap between original WER and top-2 WER.

Table 1 – Ablation study of input articulators. Right hand, left hand, general body pose, and face mesh.

RH	LH	Body	Face	WER	95% Conf. Int
✘	✘	✘	✓	95.5	[93.9;96.6]
✘	✘	✓	✘	94.1	[92.4;95.5]
✘	✓	✘	✘	85.4	[83.1;87.6]
✓	✘	✘	✘	72.0	[69.1;74.8]
✓	✓	✘	✘	59.9	[56.8;63.0]
✓	✓	✓	✘	55.0	[51.8;58.1]
✓	✓	✘	✓	54.6	[51.5;57.7]
✓	✓	✓	✓	**53.1**	[49.9;56.2]

Table 2 – Performance of the model depending on the number of glosses considered at training.

# classes	Min occ./sign	WER	Top-2 WER
10	166	**23.9**	13.9
25	114	37.7	28.3
50	72	44.0	38.0
100	47	53.1	47.3

3.4 Qualitative evaluation

Table 3 – Examples of predicted sequences of glosses and the corresponding ground truth.The first two sequences obtained a WER of 50%, the last one obtained a WER of 75%.

GT	DEUX CONSEQUENCE CONSEQUENCE LIGNE
Pred	DEUX CONSEQUENCE JUSTE IL-N'Y-A-PAS
GT	GARE AUTREFOIS UN PERSONNE
Pred	GARE GARE AUTREFOIS PERSONNE
GT	AJOUTER VRAI AJOUTER AUTREFOIS
Pred	AJOUTER REGARDER AJOUTER AUTREFOIS

Table 3 shows some examples of predicted gloss sequences. The first two sequences correpond to a WER of 50% (thus representative of the overall system performance). We can see that they are not too far from the ground truth. The first two examples allow us to see that our model has difficulty in handling sign repetitions in the corpus. In the first example, it does not

take into account the repetition of the gloss CONSEQUENCE and in the second example it adds a repetition of the gloss GARE. It should be noted that repetition has a particular role in sign language, it sometimes indicates a quantity or a degree of intensity. Finally, some standardized lexical signs are based on movement repetition like the sign in LSF that corresponds to *work*. According to the third example, the model can confuse similar signs, indeed, the signs corresponding to the glosses VRAI and REGARDER correspond to the same handshape of the dominant hand (fingers in the shape of a V). This shows that the model takes into account the position of the joints and fingers to discriminate the signs.

4 Conclusion

In this study, we presented a first model for lexical sign recognition for French Sign Language in a continuous context. To deal with the relative small size of existing corpora in LSF, we proposed a lightweight model combining pose extraction and a RNN-based decoder trained with CTC loss (alleviating the need of a temporal segmentation of the training corpus at the gloss level). Finally, we precisely evaluated the relative and absolute contribution of main articulators (right hand, left hand, face and body) in the decoding performance and showed the advantage of a multi-stream approach compared to mono-stream approach. This first work on sign recognition in LSF opens many perspectives. One of the interesting points would be to apply our model on the RWTH-Phoenix corpus to compare ourselves with the models of the literature developed for this corpus of DGS. This could also allow us to test transfer learning: does a pre-learning on another sign language improve the results for the target language? Another perspective is to use the proposed model as a representation extractor for other tasks of automatic processing of LSF, such as automatic segmentation of subtitles of LSF videos [7].

5 Acknowledgement

We would like to thank Sanjana Sankar who provided us with part of the code of [18] that has been helpful for the implementation of the presented method. This work has been funded by the Bpifrance investment "Structuring Projects for Competitiveness" (PSPC), as part of the Serveur Gestuel project (IVès and 4Dviews Companies, LISN — University Paris-Saclay, and Gipsa-Lab — University Grenoble Alpes).

References

[1] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew Mcparland, and Andrew Zisserman. BBC-Oxford British Sign Language Dataset. working paper or preprint, January 2022.

[2] V. Bazarevsky, I. Grishchenko, K Raveendran, T. Zhu, F. Zhang, and M. Grundmann. Blazepose: On-device real-time body pose tracking, 2020.

[3] V. Belissen. *From Sign Recognition to Automatic Sign Language Understanding : Addressing the Non-Conventionalized Units*. Theses, Univ. Paris-Saclay, 2020.

[4] V. Belissen, A. Braffort, and M. Gouiffès. Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In *LREC*, 2020.

[5] F. Bigand, E. Prigent, and A. Braffort. Person identification based on sign language motion: Insights from human perception and computational modeling. In *MOCO*, 2020.

[6] A. Braffort. A gesture recognition architecture for sign language. In *ASSETS*, 1996.

[7] H. Bull, M. Gouiffès, and A. Braffort. Automatic segmentation of sign language into subtitle-units. In *ECCV Workshops*, 2020.

[8] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *ICCV 2017*.

[9] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *ECCV Workshops*, 2020.

[10] N. C. Camgöz, O. Koller, S. Hadfield, and R. Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *CVPR*, 2020.

[11] A. Duarte, S. Albanie, X. Giró-i Nieto, and G. Varol. Sign language video retrieval with free-form textual queries. In *CVPR*, 2022.

[12] A Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006.

[14] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 2015.

[15] L. Momeni, H. Bull, K. R. Prajwal, S. Albanie, G. Varol, and A. Zisserman. Automatic dense annotation of large-vocabulary sign language videos. In *ECCV*, 2022.

[16] A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgoz, R. Bowden, T. Jiang, A. Rios, M. Muller, and S. Ebling. Evaluating the immediate applicability of pose estimation for sign language recognition. In *CVPR Workshops*, 2021.

[17] L. M. Paul, G. F. Simons, and C. D. Fennig. *Ethnologue: Languages of the World, Eighteenth edition*. 2015.

[18] S. Sankar, D. Beautemps, and T. Hueber. Multistream neural architectures for cued-speech recognition using a pre-trained visual feature extractor and constrained CTC decoding. In *ICASSP*, 2022.

[19] Y. Soullard, C. Ruffino, and T. Paquet. CTCModel: Connectionist Temporal Classification in Keras, 2018.