Automatic recognition of Sign Language structures in RGB videos: the detection of pointing and lexical signs

Valentin Belissen, Michèle Gouiffès, Annelies Braffort

Automatic recognition of Sign Language structures in RGB videos: the detection of pointing and lexical signs

Valentin Belissen
LIMSI, CNRS, Université Paris Saclay, Orsay,France

Michèle Gouiffès
LIMSI, CNRS, Orsay,France

Annelies Braffort
LIMSI, CNRS, Orsay,France

Abstract

This work presents a generic approach to tackle continuous Sign Language Recognition (SLR) in ordinary RGB Sign Language videos.

While usual SLR systems only analyze SL through the lexical level, it is shown here that both lexical and SL-specific features can be accurately detected and localized. This paves the way to better understanding Sign Language linguistics, as well as benefiting the Deaf community in need of SL-specific tools like video querying.

Generic human motion features are first extracted from videos, so that a very compact modeling is obtained and future training time is limited. A Recurrent Neural Network is then trained on those generic features to recognize lexicon and SL-specific structures like pointing.

Applied to the French Sign Language corpus Dicta-Sign, pointing signs detection gets a 78% sequence-wise $F1$-score on 4 seconds chunks. The network also gets a 70% sequence-wise $F1$-score for the detection of lexical signs with less than 50 instances in the training set. These are very promising results for a first SLR trial on a French Sign Language corpus, given its relatively short total duration, its low image resolution and frame rate, and the unconstrained nature of the recorded dialogs.

1 Introduction

Sign Languages (SL) are clearly the most natural language for Deaf people. However, to this day, SL linguistics has not been as thoroughly described and understood as those of other common spoken languages [3].

As it will be described later in the paper, SL production includes lexical signs as well as other linguistic structures, such as pointing signs or classifiers [31] as illustrated on Fig. 1. These high-level structures are usually not considered in the automatic Sign Language Recognition (SLR) literature. Being able to detect high-level linguistic features as well as specific lexicon is needed to better understand SL, and there is a strong demand from language experts to develop such tools. Luckily, there is more and more SL videos available to the Deaf, on public platforms, social networks, etc.

In this paper, we present two major contributions to the field of continuous SLR:

1. A generalizable modeling of a signer, as well as a generic pipeline to process RGB videos and extract relevant motion features using open source software. This model does not require the use of a specific sensor, and can be applied to webcam videos. Contrary to most models, it is independent both from the appearance of the signer and from the environment.

2. A common neural architecture to detect both lexical signs and high-level linguistic structures – in this paper we illustrate the example of pointing signs.

This paper is organized as follows: after discussing the state-of-the-art in automatic SLR and its short-comings in Section 2, we introduce a relevant and generalizable signer modeling in Section 3. The details of the continuous French SL dialog corpus *Dicta-Sign* are given in Section 4. The Recurrent Neural Network that was built for this paper is outlined in Section 5 and the experiments we conducted are presented in Section 6. Finally, we end with conclusions and future work in Section 7.

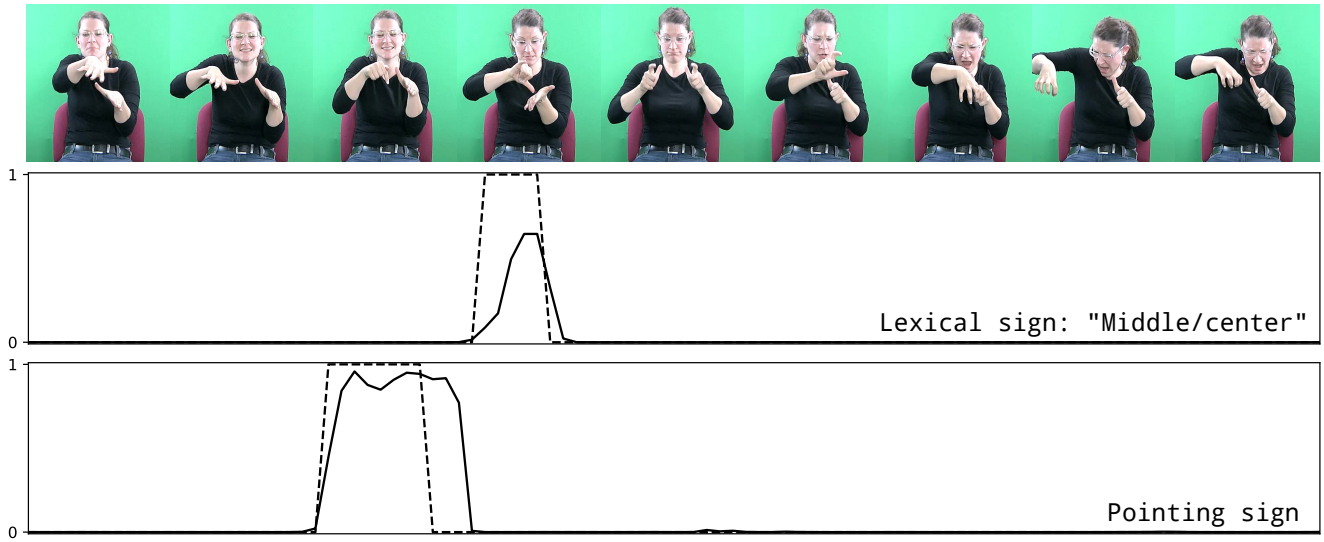

Figure 1: French Sign Language sequence example (duration: 4 seconds), with a lexical sign ("Middle/center") and a pointing sign accurately detected and localized by our model. Dashed lines are ground truth probabilities and solid lines are predictions. Possible translation: *At the very center of this area, there is a large building surrounded by restaurants.*

2 Related work and limitations

In this section, we discuss the usual SLR hypotheses, both in terms of input data characteristics and in terms of SL linguistics – the latter is developed in Section 2.1.

2.1 Sign Language Recognition and associated hypotheses

From [23], we use this commonly accepted definition: *fully-lexical signs are highly conventionalised signs in both form and meaning in the sense that both are relatively stable or consistent across contexts. Fully-lexical signs can easily be listed in a dictionary.*
Lexical signs can be used as verbs, nouns, adjectives, etc. As mentioned and explained further below, they only account for a fraction of SL production.

A lot of past and current work has focused on the problem of recognizing lexical signs that are realized in an isolated way, usually called *citation-form* lexical SLR [24, 39, 40, 44, 38, 10]. Although it might be seen as a step towards true SLR, it has strong limitations, if only that signs are not achieved similarly in continuous discourse compared to when isolated.

Continuous SLR is actually a much more challenging task, with a lot of variability and continuous transitions between successive signs. It has only been addressed from a lexical perspective and with the – sometimes unstated – hypothesis that SL production can be reduced to a sequence of lexical signs [28, 6, 12].

However, SL actually have – at least – three strong characteristics that make them fundamentally different from unidimensional sequential languages:

1. They are **multi-channel**: information is conveyed through hand motion, shape and orientation, body posture and motion, facial expression and gaze;

2. They are strongly **spatially organized**: events, objects, people and other entities are placed in the signing space and related to each other in a visual way. This is a fundamental point since all of the above-mentioned papers assume that a SL production can be analyzed through the sequence of detected lexical signs, which is wrong for natural SL discourse. Indeed, the grammar of SL is structured by the use of space;

3. They allow signers to generate new signs – that would not appear in a dictionary – on the go, in an iconic way, or even modify lexical signs. More generally, **SL do not only consist of lexical signs** but they also make use of more complex iconic structures.

In the authors' opinion, the above-mentioned papers actually deal with continuous *lexical sign* recognition, while continuous *sign language* recognition is still to be addressed.

2.2 Non-lexical Sign Language complexity

An interesting illustration to the complexity of SL and to the short-comings of lexical-only approaches is given in Fig. 1 (top part). This – only 4 seconds long – sequence (extracted from a longer video) can be decomposed as follows:

- Frame 1 and 2: Production of the lexical sign "Area", with a modification of the citation-form to highlight the fact that the area is wider than normal. The left hand seems to be static which is usually called a *fragment buoy*, thus helping the interlocutor understand that what is being said relates to an element introduced earlier in the talk (probably just before the beginning of this clip).
- Frame 3: Pointing sign to the middle of the area in question ; the pointing sign is not exactly in its standard form since the left hand still keeps its fragment buoy function.
- Frame 4: Lexical sign "Middle/center", insisting on the fact that what is going to be said is at the *very* center of the area.
- Frame 5: Lexical sign "Building", with a modification of the citation-form to highlight the fact that the building is larger than normal.
- Frame 6: the left hand is static, maintaining a fragment of the "building" sign. It has the function of providing a reference point to understand that what will be said is still in the same setting, with the large building at the center of the area. The right hand produces the lexical sign "Restaurant" (usually it is a two-handed sign, but the left hand is already being used as a fragment buoy).
- Frame 7, 8 and 9: a standard classifier that can be understood – among others – as a small building is successively placed all around the area. Three instances are placed, but the face expression suggest that there are many of them, probably more than just three. The left hand still maintains the reference point to the large building.

This random example shows that SL production should by all means not be seen and analyzed as a succession of citation-form lexical signs. Many other gestural units are used in SL, and the objective of this paper is to explore this direction, showing that they can be dealt with. In this paper, we initiate this exploration by the analysis of pointing signs (as well as lexical signs).

2.3 Data acquisition and processing

A lot of SL recording and subsequent analysis has been done in controlled environments, with specific conditions, like RGB-D setup [38, 40, 44, 48] or very high recording frame rate [17, 16, 15, 47].

In a less controlled environment with more general conditions (RGB images and 25 frames per second), a lot of research has been conducted on the *RWTH-PHOENIX-Weather 2014* dataset [25]. The usual approach on this corpus has been to start with a Convolutional Neural Network in order to derive features on the images [13, 29, 48, 12, 28, 5, 26, 27]. Apart from the limitations in terms of linguistics on this corpus – the only annotations are lexical signs, and the research on this corpus has only been addressed in a linear way –, the features derived from a CNN might be prone to a lack of generalizability, for instance if applied to videos where scale or appearance are changed.

All of these point have led us to consider and develop another type of signer modeling, which is described in the Section 3.

3 Generalizable signer modeling

Following-up the discussion of Section 2, a completely different modeling process is presented here, and summarized in Fig. 2. It was decided to make use of several open-source programs. The modeling of upper body pose, face and hands is first outlined, then the manufacturing of relevant features is described.

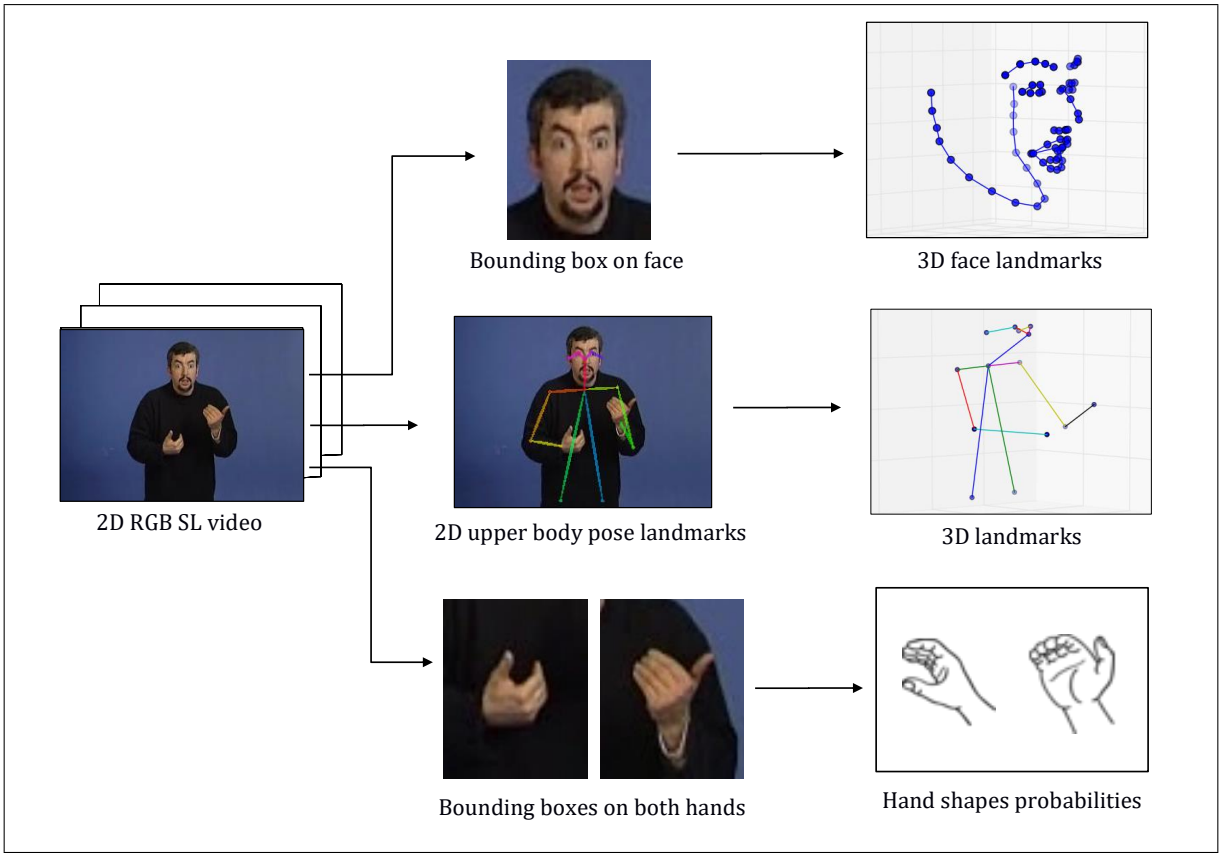

Figure 2: Modeling of a signer in a 2D RGB SL video

Figure 3: Mocap1 dataset: OpenPose landmarks no not exactly match sensors position

3.1 Body pose

While previous methods were usually based on optical flow and skin color detection [20, 9, 30], Convolutional Neural Networks (CNN) have emerged as a very effective tool to get relevant features from images. OpenPose [7, 41, 37, 45] is a powerful open source library, with real-time capability for estimating 2D body, face and hand pose. While body pose estimation appeared to be reliable, hand pose estimation did not perform well on 25 *fps* SL videos, because of motion blur around the hands. Indeed, hands move quite fast in SL production.

Since SL are 3-dimensional, we decided to train a deep neural network, reproducing the architecture from [49], in order to get an estimate on the 3D upper body pose from the 2D upper body pose. This 2D to 3D model was trained on motion capture data from the French Sign Language (LSF) corpus MOCAP1 [33], only on upper body pose. Finally, we mention that body size is normalized in order to increase the model generalizability.

In order to obtain a reliable "OpenPose 2D" to 3D model, we first had to estimate the upper body keypoints position from the sensors position that were positioned on the person's body during motion capture. Indeed, sensors position do not exactly match OpenPose keypoints, as can be seen on Fig. 3. Each OpenPose landmark position was calculated as a linear combination of the motion capture sensors position – for instance, OpenPose wrist is calculated as the average between the two sensors that were placed around the wrist during motion capture.

3.2 Facial landmarks

Although OpenPose outputs a 2D estimate on the facial pose, a 3D estimate is directly obtained from video frames thanks to a CNN model [4] trained on $230,000$ images. This model consists of 68 facial landmarks that can be seen at the upper right on Fig. 2.

3.3 Hand modeling

A lot of information in SL production, if not most of it, is conveyed through the hands. More specifically, the location, shape and orientation of both hands are critical, along with the dynamics of these three variables, that is: hand trajectory, shape deformation and hand rotation.

Ideally, one would thus greatly benefit from a frame-wise 3D hand pose estimate on RGB images. Although such algorithms have been developed [46, 21, 42, 50, 36, 37], we have not been able to get a reliable estimate on hand pose on real-life 25 *fps* SL videos with these algorithms. Indeed, because hands move constantly and relatively fast in SL videos, motion blur makes the frame-wise estimation of hand pose very difficult.

While 3D hand pose estimation on real-life RGB videos has not appeared to be reliable to this day, a SL-specific model was developed in [26]. This model was trained on about a million frames, including motion blurred images. This CNN model classifies cropped hand images into 61 predefined hand shapes classes. Whereas this model focuses on hand shape – which only accounts for a portion of the information conveyed through the hands –, we decided to use it as our hand modeling system. Thus, for each frame and each hand, we scale hand data down to a vector of 61 probabilities. In Section 6.3, we also investigate a degraded model with 21 2D hand keypoints from OpenPose.

It is to be noted that the dataset used in [26] comprises hand crop images from Danish Sign Language, New Zealand Sign Language and German Sign Language. Although French Sign Language (LSF) was not part of this dataset, we have made the assumption that this model could be transferred to LSF. Even though the hand shapes probability distribution is very likely to be different between different Sign Languages, most hand shapes seem to be shared between all SL, so that the assumption seems valid.

With this modeling, the trained models can be applied to any SL RGB video, even if scaling or appearance differ from the training data.

3.4 Pre-computed body & face features

In order to even reduce training time, some features can be pre-computed and thus replace the raw 3D position of body and facial landmarks:

- 9 head features:
  - Each rotation angle, speed and acceleration for axes X, Y and Z of the centroid of the head. The head or gaze generally follows the placement of objects in the signing space.
- 4 face features:
  - Horizontal and vertical mouth openness

  - Relative motion of the eyebrows to the eyes

- 222 body features:
  - Every joint angle, orientation and their dynamics (speed and acceleration)
  - Every joint position relative to the parent joint – for instance left elbow relative to left shoulder), and their dynamics
  - Relative position, speed, acceleration and distance of one hand to the other – according to [2], three categories of bi-manual signs can be distinguished with respect to their relative motion (and with respect to the difference of their configuration)

Most of these features were selected empirically, and should be thoroughly analyzed at some point of this research project. We are aware that there might be useless features, or features we did not think of, but in any case they reflect the fact that SL do not only rely on hands to convey information.

4 Dicta-Sign corpus: going beyond lexicon

In this section, we first discuss requirements to select a SL dataset that would fit our goals, then an overview of the Dicta-Sign corpus is given.

4.1 Choosing a relevant SL dataset

A number of criteria were taken into account in order to pick an appropriate dataset to train and test our model:

- French Sign Language (LSF) was preferred. Indeed, most research has focused on American Sign Language [15, 48, 16, 17, 47, 35], German Sign Language [13, 6, 29, 12, 5], some on Greek Sign Language [10, 11] and Chinese Sign Language [40, 44]. Our goal is to draw inspiration from these works, while producing a general enough model that could be applied to other Sign Languages than LSF;
- A continuous SL recording, with the lowest possible restriction in terms of language, in order to get realistic data;
- The video resolution and frequency was to be relatively low so that our model could be applied in most use cases;
- We wanted a corpus that was not annotated only on the lexical level.

We then decided to train the model on the Dicta-Sign corpus, which is presented and described in the next section.

4.2 Details of the corpus

Dicta-Sign corpus contains dialogs in four different languages: British Sign Language (BSL), German Sign Language (DGS), Greek Sign Language (GSL) and French Sign Language (LSF) [34]. Fig. 4 shows the setup of dialog recorded for the corpus. For this work, only the French part was retained, containing about five hours of annotated dialogs with the following annotations according to [22]:

- *Fully Lexical Signs (FLS)*: they form the basic lexicon of SL, as it was mentioned before. FLS only account for a fraction of what can be analyzed from SL discourse.
- *Partially Lexical Signs (PLS)*: they are also referred to as classifier signs or classifier predicates (see [31]). Their definition is close to what is called *iconic signs* in [14].
  - *Pointing (PT)*: Pointing signs, as mentioned in the name, are used to point towards an entity in the signing space, that is to link what is said to a spatial referent. Since SL are spatially organized, they are of prime importance to understand a discourse.
  - *Depicting Signs (DS)*: they form a broad category of signs, the structure of which is easily identified. They are used to describe the location, motion, size, shape or the action of an entity, along with trajectories in the signing space. They sometimes consist of the tweaking/enrichment of a lexical sign.

Figure 4: Dicta-Sign corpus: setup for the recording of two signers

  - *Fragment buoys*: they are hand shapes held in the signing space (usually on the weak hand) while signing activity continues on the other hand [32]. They can be seen as a referent, and can be used for specific linguistic functions, like what was called qualification/naming structures in [18].

- *Non Lexical Signs (NLS)*: Here NLS comprise fingerspelling (FS), numbering (N) and gestures that are not typically specific to SL and can be culturally shared with non SL signers (*i.e.* speakers).

The image resolution of this corpus is 480p, while the frame rate is 25 *fps*. Thus, it is consistent with the objectives in Section 4.1.

There are 16 native signers in the LSF corpus. Since each video consists of recording the conversation between two people, there are many blanks, they sometimes laugh, interrupt each other, they can sign very fast at times and slowly too. Different instructions were given to the signers, different topics were discussed, consequently different narrative and signing styles were observed. The whole corpus is thus very different from others with a much more controlled environment and constrained production.

Our goal is to be able to detect any type of lexical or non-lexical feature in natural continuous SL like the Dicta-Sign corpus. For that purpose, we present further below a recurrent architecture that was built, and the results it achieved.

5 Sign Language features detection network

In this section, we describe the architecture of the training model and explain the choices we made, along with the parameters we had to optimize.

Taking inspiration from state-of-the-art models for lexical sign recognition in continuous SL (see [28, 6]), it was decided to build and train a unique Recurrent Neural Network (RNN) in order to detect both Fully-Lexical Signs (FLS) and Pointing Signs (PT).

The RNN model was built with Keras [8] on top of Tensorflow [1].

Since this work does not aim at real-time applications, the recurrent layers have been chosen as bidirectional. The type of units is Long Short-Term Memory (LSTM) – they handle vanishing gradient issues [19], which in the case of high frequency data like ours is critical. Dropout is used to prevent overfitting in the LSTM layers [43].

A simplified representation of this network is given on Fig. 5, with two LSTM layers. The input data to this model is, for each frame t, a vector x_t which is the concatenation of the output of the 3D upper body pose model, the 3D facial landmarks model and the hand shapes probabilities model. The first operation is a 1D-convolution, then the data is processed by the first BLSTM layer, and concatenated. The operation is repeated on the second BLSTM layer. Finally, fully-connected (FC) layers directly output probabilities y_t of either one or several FLS or of a PT being executed.

Network optimization includes many parameters, including:

- Convolution kernel size, and number of filters

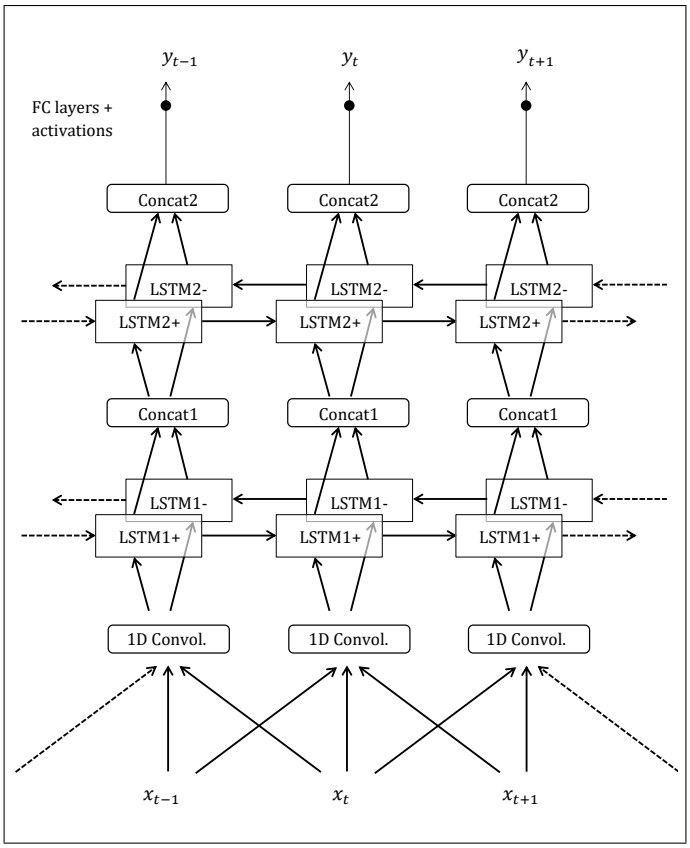

Figure 5: Simplified unfolded scheme of a 2-layer bidirectional LSTM network for SL linguistic features learning

Table 1: Sign count for the two FLS ("Same/Also" and "Line/Wire") and Pointing signs

Set	"Same/Also"	"Line/Wire"	Pointing
Training	208	47	2398
Validation/test	60	23	1154
Total	268	70	3552

- Number of LSTM layers, and their number of units
- Batch size, learning rate

In this work, our main objective is to validate the signer modeling that was described in Section 3. Therefore, we leave room for further RNN optimization in the future.

6 Experiments

In this section, we report and discuss our experimental results on the Dicta-Sign corpus. Specifically, we stress the importance of signer modeling to the model performance.

6.1 Pointing and lexical signs detection in the Dicta-Sign corpus

In order to evaluate the presented model, we applied it on the detection of:

- two lexical signs: a more frequent one ("Same/Also") and a less frequent one ("Line/Wire") ;
- pointing signs, that belong to non-lexical SL signs (see Sections 2.1 and 4).

The detailed sign count is given in Table 1.

According to the signer modeling and data processing pipeline presented in Section 3, the first operation is to run the 5 hours of LSF from the Dicta-Sign corpus through the pipeline, in order to get the 357 features detailed in Section 3.4. Then, the network is trained on these features along with Dicta-Sign corpus annotations.

6.2 Performance measure

Although the models we trained output frame-wise predictions, most frame labels are "blank", so that frame-wise accuracy is close to 100%. For sake of clarity, it was then decided to present in this paper sequence-wise precision P, recall R and $F1$-score, defined as

$$F1 = 2\frac{PR}{P+R}$$

with sequence length set at 100 frames (that is 4 seconds).

Here, we only look at global precision/recall/$F1$-score, even though – as can be seen on Fig. 1 – the model enables one to accurately localize sign production. We are aware that this performance measure does not exactly match the performance goal we are targeting – that is: precision in terms of detection *and* localization – but it enables one to relatively compare different modelings as can be seen in the next section.

Sequences are chopped with a sliding window, so that every possible sequence is seen in training, but the sequences may start or finish at the middle of a sign – which obviously worsens the model apparent performance.

$F1$-score is also used to look for convergence during training. Indeed, the training loss is defined frame-wise, but as it was just stated, we do not focus on frame-wise accuracy to inspect model performance. Fig. 6 is an illustration of $F1$-score improvement on pointing signs during training. This figure shows that the network architecture makes it possible to properly learn how to detect and localize pointing. From epoch 0 to epoch ˜40, the output of the model is always "blank", so that $R = 0$, $P \simeq 1$ and $F1 = 0$. Proper performance improvement starts at epoch ˜50. Then, from epoch ˜60 to epoch ˜160, precision and recall increase concurrently. Starting from epoch ˜160, precision decreases while recall increases, with $F1$ still improving until convergence.

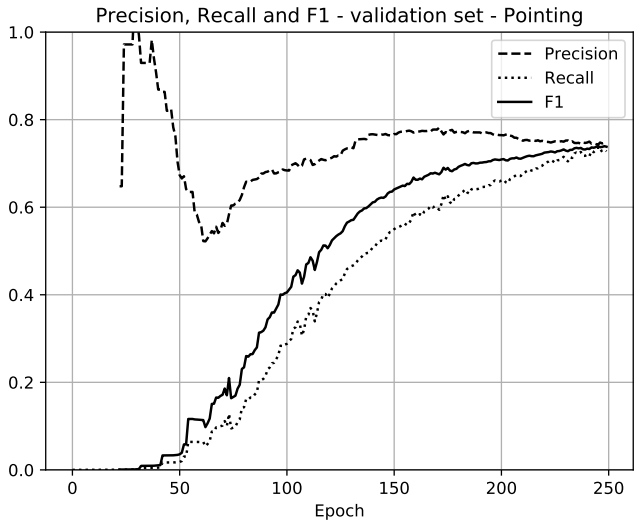

Figure 6: $F1$-score learning curve on pointing signs

Table 2: Impact of signer modeling on performance (best sequence-wise F1-score on test set). The number of LSTM layers to achieve best performance is also indicated.

Signer modeling	N_{layers}	FLS ("Same/Also")	FLS ("Line/Wire")	Pointing
OP(2D)	4	0.478	0.510	0.549
OP(2D)+3D+HS	4	0.586	0.523	0.605
OP(2D)+3D-features+HS	1	0.646	0.700	0.781

6.3 Impact of signer modeling on computed performance

In this section, we analyze the impact of signer modeling on the overall FLS/PT prediction performance.

It is to be noted that particular attention is paid to using a signer independent performance measure: the signers in test set do not appear in training set, which makes the model much better at generalizing to unseen data. In [28], the signer-dependent Word Error Rate (WER) is 26.8%, while signer-independent WER is 44.1%, which is – relatively – 65% worse.

Three modelings are considered:

1. *OP(2D)*: raw OpenPose output is used (with 2D upper body pose, 2D facial keypoints and 2D hand keypoints). The frame-wise vector size is 248;
2. *OP(2D)+3D+HS*: raw OpenPose output is used, with 3D upper body pose estimation from our trained model, and with the hand shapes probabilites from [26]. The frame-wise vector size is 452;
3. *OP(2D)+3D-features+HS*: same as above, but instead of raw data, features from 3.4 are pre-computed. The frame-wise vector size is 357.

Table 2 presents the performance difference in terms of $F1$-score of the three considered modeling, applied to the three chosen SL examples. The benefit of the signer modeling described in Section 3 is clear. Indeed, it can be seen that the 3D upper body estimation, the hand shape classification and the manufactured body and face features all play a role in improving model performance.

Furthermore, although this paper does not aim at presenting the optimization of the network architecture, one should note that best model performance was obtained with fewer LSTM layers for the third model than for the first one and the second one. This might be understood by the fact that the manufactured features were somehow correctly built, and the first LSTM layers for the first two models might be used by the network to learn similar features (including speed and acceleration of joints).

The discrepancies between our model best results and a 100% $F1$-score can be related to several factors, amongst which:

- The chopping of sequences at the middle of certain signs;
- A non fully optimized network architecture. Indeed, as stated earlier (see Section 5), network optimization includes many parameters that were not completely investigated;
- Annotation errors and bias or subjectivity;
- A great variability between signers;
- A lot of variability between signs because of the continuous nature of the corpus.
- More generally, the continuous nature of Sign Language in itself that makes it very difficult to classify its parameters into a finite number of categories.

7 Conclusions and future works

In this paper, a French Sign Language dialog corpus was studied for the first time, from the Sign Language Recognition perspective. Dicta-Sign is a very natural corpus, with very few constraints on the style and content. Furthermore, the recording quality is very common. Our work is thus easily generalizable to unseen data.

We have developed a generalizable model of the signer, independent from the appearance, size, context and that can be computed from standard videos. We have also built a network for detecting lexical and non lexical structures, which is general but can be specialized for a given application. Although tested on French Sign Language videos, it could easily be extended to other SL. Among others, it would be interesting to test whether non-lexical structures could be detected in other SL without retraining the model.

Despite the low video resolution and frame rate, the continuous and unconstrained nature of the French Sign Language corpus dialogs, our model was able to accurately detect and localize the queried structures.

Fine-tuning the network, adding new pre-computed features and new preprocessing algorithms – in particular hand pose recognition –, we are optimistic that performance can be very much improved. Also, gradually adding other non-lexical features to the network, we see this work as a stepping stone towards Sign Language understanding and analysis as a whole, in an automated way.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Robbin Battison. Phonological Deletion in American Sign Language. *Sign language studies*, 5(1):1–19, 1974.

[3] A. Braffort. *La Langue des Signes Française (LSF): Modélisations, Ressources et Applications*. Collection Sciences cognitives. ISTE/Hermes Science Publishing, 2016.

[4] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*, 2017.

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*, 2017.

[8] François Chollet et al. Keras, 2015.

[9] Helen Cooper, Brian Holt, and Richard Bowden. Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer, 2011.

[10] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13(Jul):2205–2231, 2012.

[11] Helen Cooper, Nicolas Pugeault, and Richard Bowden. Reading the signs: A video based sign dictionary. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 914–919. IEEE, 2011.

[12] Runpeng Cui, Hu Liu, and Changshui 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*, pages 7361–7369, 2017.

[13] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 2019.

[14] Christian Cuxac. *La langue des signes française (LSF): les voies de l'iconocité*. Number 15-16. Ophrys, 2000.

[15] Mark Dilsizian, Dimitris Metaxas, and Carol Neidle. Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition. *LREC*, 2018.

[16] Mark Dilsizian, Zhiqiang Tang, Dimitris Metaxas, Matt Huenerfauth, and Carol Neidle. The Importance of 3D Motion Trajectories for Computer-based Sign Recognition. *7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, The 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[17] Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris N Metaxas. A New Framework for Sign Language Recognition based on 3D Handshape Identification and Linguistic Modeling. In *LREC*, 2014.

[18] Michael Filhol and Annelies Braffort. A study on qualification/naming structures in Sign Languages. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the 8th International Conference on Language Resources and Evaluation (LREC 2012), ELRA*, pages 63–66, 2012.

[19] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *IET*, 1999.

[20] Matilde Gonzalez Preciado. *Computer vision methods for unconstrained gesture recognition in the context of sign language annotation*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2012.

[21] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand Pose Estimation via Latent 2.5 D Heatmap Regression. *arXiv preprint arXiv:1804.09534*, 2018.

[22] Trevor Johnston and L De Beuzeville. Auslan corpus annotation guidelines. *Centre for Language Sciences, Department of Linguistics, Macquarie University*, 2014.

[23] Trevor Johnston and Adam Schembri. *Australian Sign Language (Auslan): An introduction to sign language linguistics*. Cambridge University Press, 2007.

[24] Hamid Reza Vaezi Joze and Oscar Koller. MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language. *arXiv preprint arXiv:1812.01053*, 2018.

[25] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, December 2015.

[26] Oscar Koller, Hermann Ney, and Richard 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*, pages 3793–3802, June 2016.

[27] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference 2016*, September 2016.

[28] Oscar Koller, Sepehr Zargaran, and Hermann Ney. Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs. In *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[29] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs. *International Journal of Computer Vision*, 126(12):1311–1325, 2018.

[30] François Lefebvre-Albaret. *Traitement automatique de vidéos en LSF Modélisation et exploitation des contraintes phonologiques du mouvement*. PhD thesis, Université Paul Sabatier-Toulouse III, 2010.

[31] Scott K Liddell. *An investigation into the syntactic structure of American Sign Language*. University of California, San Diego, 1977.

[32] Scott K Liddell. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003.

[33] Limsi and CIAMS. MOCAP1, 2017. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[34] S. Matthes, T. Hanke, Anja Regen, J. Storz, Satu Worseck, E. Efthimiou, N. Dimou, Annelies Braffort, J. Glauert, and E. Safar. Dicta-Sign – Building a Multilingual Sign Language Corpus. In , pages 117–122, Istanbul, Turkey, 2012.

[35] Dimitris N Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. In *LREC*, pages 2414–2420, 2012.

[36] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. GANerated Hands for Real-time 3D Hand Tracking from Monocular RGB. *arXiv preprint arXiv:1712.01057*, 2017. Modèle récupéré mais pas testé Performances a priori moins bonnes que Spurr.

[37] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single RGB frame for real time 3D hand pose estimation in the wild. *arXiv preprint arXiv:1712.03866*, 2017.

[38] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. Sign language recognition using convolutional neural networks. In *Workshop at the European Conference on Computer Vision*, pages 572–578. Springer, 2014.

[39] Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, 126(2-4):430–439, 2018.

[40] Junfu Pu, Wengang Zhou, Jihai Zhang, and Houqiang Li. Sign language recognition based on trajectory modeling with HMMs. In *International Conference on Multimedia Modeling*, pages 686–697. Springer, 2016.

[41] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. *arXiv preprint arXiv:1704.07809*, 2017.

[42] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal Deep Variational Hand Pose Estimation. In *CVPR*, 2018.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[44] Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. Isolated Sign Language Recognition with Grassmann Covariance Matrices. *ACM Transactions on Accessible Computing (TACCESS)*, 8(4):14, 2016.

[45] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[46] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. *arXiv preprint arXiv:1812.01598*, 2018. https://github.com/xiangdonglai Pas de modèle dispo pour l'instant https://www.youtube.com/watch?v=-7rQSPYZRNw.

[47] Hee-Deok Yang and Seong-Whan Lee. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters*, 34(16):2051–2056, 2013.

[48] Yuancheng Ye, Yingli Tian, Matt Huenerfauth, and Jingya Liu. Recognizing American Sign Language Gestures from within Continuous Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2064–2073, 2018.

[49] Ruiqi Zhao, Yan Wang, C Fabian Benitez-Quiroz, Yaojie Liu, and Aleix M Martinez. Fast and precise face alignment and 3D shape reconstruction from a single 2D image. In *European Conference on Computer Vision*, pages 590–603. Springer, 2016.

[50] Christian Zimmermann and Thomas Brox. Learning to Estimate 3D Hand Pose from Single RGB Images. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. https://arxiv.org/abs/1705.01389.