Sign Language Video Analysis For Automatic Recognition and Detection

Valentin Belissen

Sign Language Video Analysis For Automatic Recognition and Detection

Valentin Belissen
LIMSI, CNRS, Univ. Paris-Sud, Orsay, France

***Abstract*— This Research Project Summary presents ongoing work on automatic Sign Language Recognition (SLR) and detection, carried out as a PhD research project at the LIMSI, CNRS, France. It uses the French part of the Dicta-Sign corpus with 8 hours of annotated dialogue from 18 native speakers [1].**

This work tackles the issue of extracting relevant information from a Sign Language (SL) video, and developing systems to detect lexical signs or even high-level linguistic features.

***Index Terms*— Sign Language, Convolutional Neural Networks, Recurrent Neural Networks, Sign Language Recognition**

I. Addressed problem and motivation

A. Specificities of SL

Sign Languages (SL) are undoubtedly the most natural language for Deaf people. However, SL do not have a written equivalent, which makes it impossible to search amongst SL videos, like one would do for text documents. However, with the development of the Internet, there are more and more SL videos available to the Deaf, on public platforms, social networks, etc. It is thus a strong need inside the Deaf community to develop tools that would enable one to do such searches or queries, possibly without the use of any written language. Another motivation to this work is that SL linguistics are not as well described as those of other common spoken languages. Being able to detect high-level linguistic features would then be useful to better understand them.

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 (see for instance [2] and [3]). Although it might be a step towards true SLR, it has strong limitations, if only that signs are not achieved similarly in continuous discourse compared to when isolated. SLR is actually a much more difficult task, with more variability and continuous transitions between successive signs. It has been addressed on specific corpus and in a sequential way [4]–[6], although SL actually have strong characteristics that make them fundamentally different from unidimensional sequential languages:

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

• 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;

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

Last, but not least, is the fact that due to their inherent flexibility, SL show an important variability between different signers in terms of style, dynamics.

B. Dicta-Sign and SL linguistics

The Dicta-Sign corpus contains dialogue in four different languages: British Sign Language (BSL), German Sign Language (DGS), Greek Sign Language (GSL) and French Sign Language (LSF) [1]. For this work, only the French part was retained, containing about five hours of annotated dialogue with the following annotations according to [7]:

• *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 [8]). Their definition is close to what is called *iconic signs* in [9].

- *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.
- *Buoys*: they are handshapes held in the signing space (usually on the weak hand) while signing activity continues on the other hand [10]. They can be seen as a referent, and can be used for specific linguistic functions, like what was called qualification/naming structures in [11].

• *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).

C. Scope of this work

FLS are the most studied elements in the literature in terms of automatic recognition (see for instance [2]–[5], [12]–[15]). Conversely, pointing and buoys have not been dealt with to our knowledge, except in [16] which focuses on pointing signs but without any published result. However, these linguistic features have been shown critical in order to accurately describe SL speech [9]–[11], [17]. Our aim is to

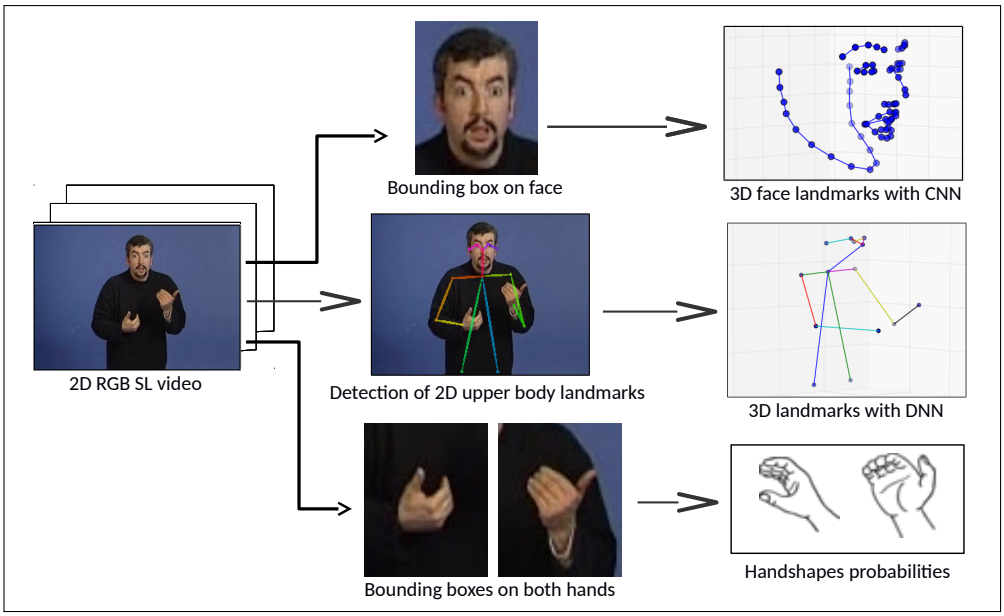

Fig. 1: Relevant data extraction pipeline for a 2D RGB SL video.

design a global SL recognition system that will be enriched step by step: first, we focus on detecting lexical signs, then we try to detect higher-level linguistic features.

II. Proposed solution and methodology of current work

A. Data processing pipeline

The input data of the present model is RGB videos (one video per signer), with a frequency of 25 fps and a resolution of 720x576.

We have decided to handle this data through three different channels: the upper body pose, the face and the hands. The whole pipeline is summarized on Fig. 1.

3D facial pose estimation. In this work, it was considered that facial pose in SL was not fundamentally different from that of 3D facial pose estimation in common situations datasets. Therefore, a pretrained Convolutional Neural Network (CNN) [18] is used to directly obtain this estimate.

3D upper body pose estimation. First, a 2D estimate of the upper body pose is obtained from Convolutional Neural Networks (CNN) and open source libraries like OpenPose [19]. Then, since SL are 3-dimensional, a Deep Neural Network (DNN) was built and trained following [20]. A corpus of 3D motion capture data [21] was used to train the network on LSF.

Hand pose. Ideally, one would like to estimate 3D hand pose as has been done for the body pose. Unfortunately, it has not yet been possible to find or develop a model able to get a reliable estimate on the 3D hand pose. Indeed, because of the speed of the hands in SL speech, the motion blur has prevented us from using 3D hand pose estimators like [22]. From our point of view, those models only work well for slow motion (or high frequency videos) associated with high resolutions.

It was then decided – for now – to use a trained CNN model [23] that classifies hand images into 60 SL handshapes. With a 85.5% top-1 accuracy, this CNN model is to our knowledge today's best SL handshape classifier. It is to be noted that while handshape can be accurately retrieved, information on hand orientation is somewhat lost in this classification although it conveys meaningful information.

B. Summary of processed SL data

After each image is processed by the three channels of the data extraction pipeline, final adjustments are made, 3D data of the head and the upper body are bound together and rescaled with respect to a shoulder width of length 1, and more meaningful features are calculated:

- every joint position with respect to its parent joint (e.g.: hand position is calculated w.r.t. elbow position, elbow position w.r.t. shoulder position, etc.);
- the position of one hand relatively to the other hand, as well as the Euclidean distance between them;
- head orientation.

Finally, a feature vector of size $F = 298$ is obtained and renormalized so that each feature has a mean of zero and a standard deviation of one. This will be the input of the learning model.

C. The learning model

Taking inspiration from state-of-the-art models for lexical SLR in continuous SL (see [4] and [5]), it was decided to build and train a Recurrent Neural Network (RNN) in order to predict Fully-Lexical Signs (FLS), Pointing Signs (PT), Depicting Signs (DS), Buoys and Non-Lexical Signs (NLS). The RNN model was built with Keras [24] on top of Tensorflow [25].

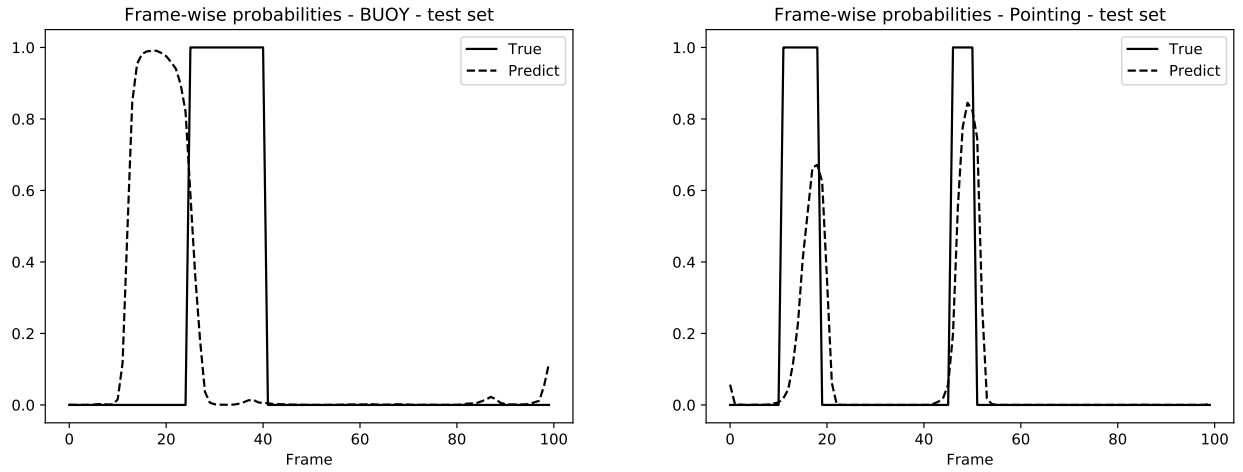

Fig. 2: Frame-wise probabilities on an example, after training, for both buoy (left) and pointing (right)

.

Fig. 3: Buoy (top) and pointing (bottom) sequence examples in Fig. 2 (respectively at left and right).

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 [26], which in the case of high frequency data like ours is critical. The number of recurrent layers is currently 2. Dropout is used to prevent overfitting in the LSTM layers [27].

D. Results

For both buoys and pointing signs, frame-wise predictions of the trained model compared to expected values, on two sequences from the test set, are presented on Fig. 2. The model correctly outputs its probabilities, with one buoy and two pointing signs correctly detected. The buoy prediction is slightly shifted to the left, which may be explained by the fact that buoys tend to have unclear limits, so the annotations are more subjective for them than for pointing signs that tend to be more accurately located. Fig. 3 shows a few of the images where a buoy (top) or two pointing signs (bottom) occur.

III. Future enrichments and challenges

This work has shown that, with image processing tools like CNNs and sequence processing models like LSTMs, it is now possible to detect high-level linguistic features in continuous SL videos, in an automated way. In the future, we will:

- Add more linguistic features to the model.
- Further improve current model performance.
- Improve our data processing pipeline, through the use of state-of-the-art hand pose estimators.
- Apply methods presented in this paper to other SL corpus.
- Propose a RNN-based SLR system that will make it possible to search through many SL videos, like one would do to search through web search engines with a text entry.

The long-term applications could be to integrate this kind of model into a bigger one and go towards realistic Sign Language Translation – including the understanding of spatial relationship and iconic structures.

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1583–1597, 2016.

[16] Julie Rinfret, Anne-Marie Parisot, Karl Szymoniak, Suzanne Villeneuve, and Thierry L. Chevalier. Methodological issues in automatic recognition of pointing signs in Langue des Signes Québécoise using a 3D motion capture system. In *Theoretical Issues in Sign Language Research 11*, 2013.

[17] Annelies Braffort. *Reconnaissance et compréhension de gestes, application à la langue des signes*. PhD thesis, Université de Paris XI - Orsay, 1996.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

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

[27] 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.