Towards Continuous Recognition of Illustrative and Spatial Structures in Sign Language

Valentin Belissen[1,2], Annelies Braffort[1], and Michèle Gouiffès[1,2]

[1] CNRS, LIMSI, Orsay, France
[2] Université Paris-Saclay, Orsay, France
{valentin.belissen,annelies.braffort,michele.gouiffes}@limsi.fr

Abstract. In recent research, attention has been drawn to recognizing lexical signs in continuous Sign Language (SL) corpora, often artificial. Because SLs are singularly structured by the use of space and iconicity, this focus does not allow for an easy transition towards SL understanding and translation. In this paper, we discuss the necessity and realizability of recognizing higher-level linguistic structures in SL videos, like classifier constructions, using natural corpora such as Dicta-Sign–LSF–v2.

Keywords: Continuous Sign Language Recognition, Iconicity

1 Introduction

Most research on Continuous Sign Language Recognition (CSLR) has focused on the recognition of lexical signs, which are *"highly conventionalised signs in both form and meaning [...] consistent across contexts"* [7]. However, SL discourse also makes extensive use of space and illustrative structures – *i.e.* depicting signs (DSs) – that, conversely, can not be listed in a dictionary.

In this paper, we present experiments for the continuous recognition of DSs on the finely annotated dialogue corpus Dicta-Sign–LSF–v2 [1]. In this regard, a compact signer representation and recognition model are detailed here.

2 The current paradigm and limits

Because the common acceptation of CSLR focuses on lexical signs, popular corpora – which are few – do not include higher-level linguistic annotations. Furthermore, the observed SL is often made of artificial sequences, or interpreted SL – like in RWTH Phoenix Weather [6] – which can be somewhat different from natural SL. In this approach (see for instance [4]), the usual metric is the Word Error Rate, that measures the normalized discrepancy between annotated and recognized sequences of signs.

However, this perspective is bound to be limited to very simple SL utterances. Indeed, natural SL exploits some key properties:

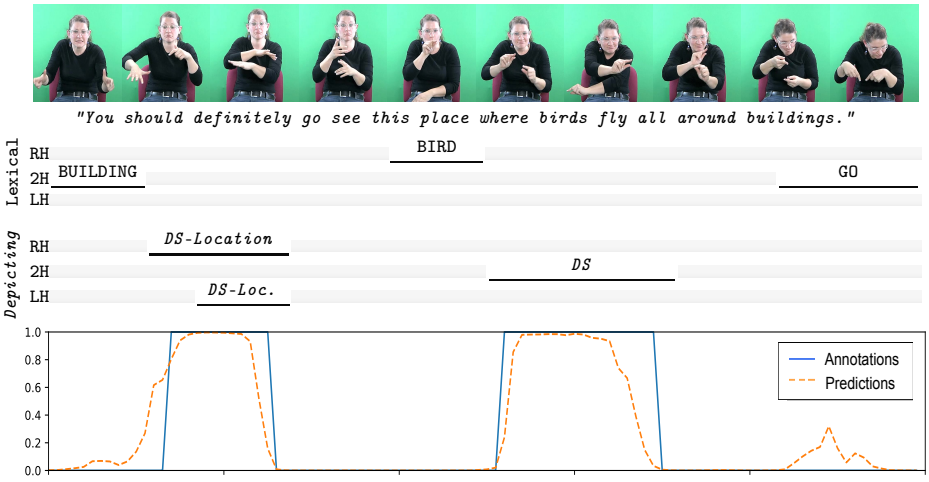

Fig. 1: Utterance from Dicta-Sign–LSF–v2 [1], with a strong use of space and iconicity (video reference: `S7_T2_A10` – duration: 4 seconds).
From top to bottom: thumbnails, proposed translation, expert annotation for the manual activity – Lexical signs and Depicting signs, each on three tracks (right handed (RH), two handed (2H), left handed (LH)) – and recognition results for depicting signs (F1-score: 86%).

Simultaneity: a great number of manual and non-manual articulators make it possible to convey information simultaneously, on different time scales.

Iconicity and use of space: building on the visual modality, iconicity enables to *show while saying*. Using the signing space in a visual way to structure discourse is also fundamental, and forms the core of the visual grammar.

A newer acceptation of CSLR, not restricted to the recognition of lexical signs and with more attention to iconicity, is thus needed.

3 A newer acceptation for CSLR

Natural SL corpora, annotated in detail, are an interesting material to get closer to SL translation. With this objective, we propose that CSLR should include the recognition of elements other than the conventional lexical signs.

Dicta-Sign–LSF–v2 [1], stemming from [9], is a French SL (LSF) corpus based on dialogue with very loose elicitation guidelines, thus highly representative of natural SL. With a total length of more than 11 hours, the annotated manual activity, inspired from the convention of [7], covers lexical signs, as well as DSs – with sub-categories –, pointing signs, buoys and more.

DSs, that amount to about 20% of the annotated frames of this corpus, appear to a major category of non-lexical signs, using the iconic modality. Some-

times referred to as classifier construction or illustrative signs, sometimes building on purely lexical signs, they are used to visually describe the location, motion, size, shape or the action of referents. In the following, we focus on the recognition of DSs. In terms of performance metrics the temporal – and possibly spatial – localization should be accounted for. Frame-wise F1-score, for instance, appears to be a good starting point.

4 Signer representation and learning model

A compact and generalizable representation of signers in videos is obtained by separately processing the upper body, hands and face.

The OpenPose library [3] is used to get a reliable 2D estimate on the upper body pose, that is later turned into a 3D estimate by training a similar model as that of [10]. Another Neural Network (NN) model [2] makes it possible to get a 3D estimate on the face pose, and a SL-specific model [8] yields hand shape probabilities for each hand. Then, meaningful SL features like relative joint positions, speeds and accelerations, angles etc. are derived to form a compact signer representation vector (about 400 features).

A convolutional-recurrent NN (CRNN) can thus be trained to take these features as input, and recognize different kinds of manual unit types, like lexical signs, DSs, pointing signs, etc.

5 Recognition results

Using the previously described signer representation, we trained our learning model for the recognition of DSs on Dicta-Sign–LSF–v2, in a signer-independent fashion, which is known to make learning more difficult as well as more generalizable. Fig. 1 illustrates the value and effectivity of such a recognition model. Indeed, the selected sequence clearly can not be reduced to its three lexical signs. Two long DSs are used to depict birds, their important number, the location and extent of their flight, the form of their trajectory, etc. The frame-wise F1-score for this particular sequence reaches 86%, while it averages 59% on the whole corpus.

6 Conclusion and perspectives

In this paper, we have insisted on the central role of iconicity and spatial structure in SL discourse, highlighting the fact that Lexical Sign Recognition is only a part of the CSLR task. Since prevalent SL corpora have intrinsic limits in terms of generalizability and do not include annotations outside lexicon, we felt it was important to point out that richer corpora do exist, with fine temporal annotations.

As a first attempt on the Dicta-Sign–LSF–v2 corpus, we have trained a CRNN to recognize depicting signs. Interesting recognition scores are met, especially when considering the unclear boundary between lexical and depicting signs.

Indeed, this frontier is dependent upon the chosen linguistic model, with no clear consensus on the matter [5]. Using finely annotated datasets, like Dicta-Sign–LSF–v2 or other corpora initially intended for linguistic studies, the relevance of prevalent linguistic descriptions of SLs could be further questioned and our work be extended. Conversely, the usual CSLR setting, with lexical annotations and WER metric prevents one from conducting this type of research. It implicitly uses the hypothesis that SL discourse can be described with sequences of lexical signs, which we have shown is far from sufficient.

Beside more analysis on the performance metric and linguistic model, future work will include further reflection on the ways spatial information can be annotated and included in automatic recognition models. On a long-term basis, we will also reflect on how to go from the detection of important discourse elements like illustrative structures to global SL understanding.

References

1. Belissen, V., Gouiffès, M., Braffort, A.: Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020) (2020)
2. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7291–7299 (2017)
4. Cui, R., Liu, H., Zhang, C.: A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. IEEE Transactions on Multimedia (2019)
5. Cuxac, C.: La langue des signes française (LSF): les voies de l'iconicité. No. 15-16, Ophrys (2000)
6. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). pp. 1911–1916 (2014)
7. Johnston, T., De Beuzeville, L.: Auslan Corpus Annotation Guidelines. Centre for Language Sciences, Department of Linguistics, Macquarie University (2016)
8. Koller, O., Ney, H., Bowden, R.: Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data is Continuous and Weakly Labelled. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3793–3802 (2016)
9. Matthes, S., Hanke, T., Regen, A., Storz, J., Worseck, S., Efthimiou, E., Dimou, N., Braffort, A., Glauert, J., Safar, E.: Dicta-Sign – Building a Multilingual Sign Language Corpus. In: Proceedings of the 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. pp. 117–122 (2012)
10. Zhao, R., Wang, Y., Martinez, A.M.: A Simple, Fast and Highly-Accurate Algorithm to Recover 3D Shape from 2D Landmarks on a Single Image. IEEE Transactions on Pattern Analysis & Machine Intelligence **40**(12), 3059–3066 (2018)