Sign Language Recognition and Translation - A Review

Daniela Winter

November 30, 2022

1 Introduction

1.1 Sign Languages

Sign languages are visual languages that are usually used as the form of communication for deaf and many hard-of-hearing people. They are official languages with defined signs and grammar. Different signs are expressed by manual and non-manual components. Manual components include hand poses, orientation of those and location of the hands in space, while upper-body movements, lip movement and other facial expressions constitute the non-manual components [ASC+21].

1.2 Transcription with Glosses

Glosses are one way of representing signed languages in a written form. include orientation, location, repetition, classifiers [ASC+21]

1.3 Sign Language Translation as a Machine Learning Task

Sign Language Recognition (SLR) and Sign Language Translation (SLT) can be realised with Machine Learning methods. SLR describes the detection and identification of different signs in a video. This term is also often used for the automatic transcription of a sign language video into glosses or some other form of textual sign language representation. SLR can be a categorization task where one video is only constituted of one sign. This kind of task is called Word-Level Sign Language Recognition (WSLR) or Isolated Sign Language Recognition (ISLR).
In contrast to WSLR, there is Continuous Sign Language Recognition (CSLR). In CSLR, a complete utterance is used as an input. In this case, the CSLR process needs to determine where a sign ends and a new one begins and output an annotate each of them in sequence.
SLR is a visual-spatial recognition task, that poses typical challenges of Computer

Vision tasks: orientation and 3D understanding, occlusion difficulties, differentiating multiple components (e.g. different speakers in one video).
SLT on the other hand is more related to a sequence-to-sequence machine translation task that aims to translate representations of the signed video to a spoken language text form. This can either be done by directly using video frame embeddings as input or by using the output of the CSLR step such as glosses. Some authors argue that the gloss representation as an intermediate step pose a disadvantage to pipelines without any intermediate transcription as it can act as a bottleneck and make the performance depend solely on the quality of the glosses obtained by the CSLR process [CKHB20]. Others argue that this does not have to be the case and show results where the final performance with predicted glosses exceeds the performance with grount truth glosses [YR20]. While it is desirable to skip the step of gloss representations also for reasons of simplicity, the performance of such pipelines are often worse. This could be the case because the task of SLT is too complex to learn for the proposed models so that some form of intermediate supervision helps guiding towards the right direction [CKHB20].
While machine translation works quite efficiently for spoken languages already, SLT is still a very challenging task. In comparison to spoken language translation, in SLT, the mappings of the signed language and its respective spoken language are often one-to-many mappings since a lot of information can be contained in one sign. Moreover, the mappings are highly non-monotonic which means that the signed and the spoken sentence often have a completely different word order and grammar [CKHB20].
The opposite direction of the SLR and SLT pipeline from spoken language text or speech to sign language is called Sign Language Production (SLP). Since this task is very different from what is describes above, this review will not focus on SLP. But since it is a translation process as well, the term of SLT is sometimes used in a way that includes SLP. In the following, the term "SLT" describes the translation process starting with a video as an input and outputting the spoken language text. In addition to that, it will be used to differentiate the text-to-text translation task from the SLR step in pipelines where both steps are separate. This usage conforms with the literature and the exact meaning can be determined by context.

2 Feature Extraction from Sign Language Videos

-¿ loss of information (human oracle bad performance)

3 Architectures

3.1 Recurrent Neural Networks

3.2 Sequence-to-Sequence Models

3.2.1 Without Attention

3.2.2 Attention-Based

3.3 Transformers

3.4 Reinforcement Learning

4 Evaluation Metrics

The most commenly used evaluation metric for SLT are the different BLEU-Scores BLEU-Scores for translation Word Error Rate (WER) for recognition [CKHB20] ROUGE-Score Top-5 recognition rate meteor (Yin, STMC-Transformer)

5 Data Sets

lack of datasets Most studies until recent years used only weather forecast phoenix14T Recent years-¿ bigger datasets but not many studies on them yet

5.1 German Sign Language Data Sets

Figure 1: This frog was uploaded via the file-tree menu.

5.2 American Sign Language Data Sets

5.3 Chinese Sign Language Data Sets

5.4 Other Data Sets

6 Analysis of Current Performances

Item	Quantity
Widgets	42
Gadgets	13

Table 1: An example table.

7 Conclusion and Outlook

References

[ASC+21] Tejaswini Ananthanarayana, Priyanshu Srivastava, Akash Chintha, Akhil Santha, Brian Landy, Joseph Panaro, Andre Webster, Nikunj Kotecha, Shagan Sah, Thomastine Sarchet, et al. Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing (TACCESS)*, 14(4):1–30, 2021.

[CKHB20] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[YR20] Kayo Yin and Jesse Read. Attention is all you sign: sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, volume 4, 2020.