A data augmentation approach for sign-language-to-text translation in-the-wild

Fabrizio Nunnari
German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus D3 2, Saarbrücken, Germany

Cristina España-Bonet
German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus D3 2, Saarbrücken, Germany

Eleftherios Avramidis
German Research Center for Artificial Intelligence (DFKI), Alt Moabit 91c, 10559 Berlin, Germany

Abstract

In this paper, we describe the current main approaches to sign language translation which use deep neural networks with videos as input and text as output. We highlight that, under our point of view, their main weakness is the lack of generalization in daily life contexts. Our goal is to build a state-of-the-art system for the automatic interpretation of sign language in unpredictable video framing conditions. Our main contribution is the shift from image features to landmark positions in order to diminish the size of the input data and facilitate the combination of data augmentation techniques for landmarks. We describe the set of hypotheses to build such a system and the list of experiments that will lead us to their verification.

2012 ACM Subject Classification Computing methodologies → Machine learning; Human-centered computing → Accessibility technologies; Computing methodologies → Computer graphics

Keywords and phrases sing language, video recognition, end-to-end translation, data augmentation

Digital Object Identifier 10.4230/OASIcs.LDK.2021.4

Category Short Paper

Acknowledgements The research reported in this paper was supported by the BMBF (German Federal Ministry of Education and Research) in the project SOCIALWEAR (Socially Interactive Smart Fashion, DFKI Kst 22132).

1 Introduction

During the last years (multimodal) language technology has seen immense progress due to the great performance of deep neural networks working on large amounts of text, image or video data [26, 7, 1, 16, 15, 21]. This progress has enabled solutions and products which serve the majority of the consumer basis, which has the ability to speak and hear, but has comparatively neglected a considerable part of the population which is deaf or hearing impaired. In our effort to intensify research towards supporting deaf people with tools for their integration in the society, we focus on Sign Language (SL).

Sign Language is the main communication language for deaf people and used by more than 10 million people in the world [8]. People who are deaf from birth, also due to the lack of exposure to corresponding vocal signals, are not proficient in reading text translations and not comfortable with writing, as the spoken language is for them a foreign language. Hence, the only effective mean of communication are motion videos, either captured or played-back.

The research community has been investigating the machine-driven translation of SL for more than 20 years; at the beginning, with the introduction of text-to-SL tools to render sign language videos through the use of virtual characters [9, 12, 17]. More recently, the

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 4; pp. 4:1–4:8
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

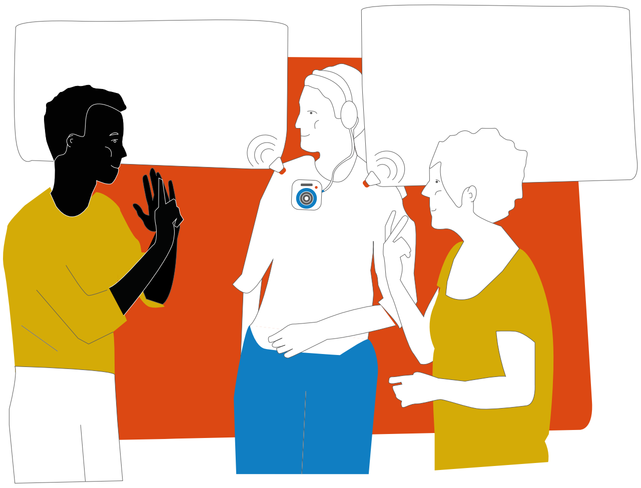

Figure 1 An illustration of a possible application scenario: a hearing person, wearing a technologically augmented jacket embedding a camera, can follow the discussion between two sign language speakers. *Illustration by Mia Grote.*

focus moved towards the more challenging SL video-to-text direction [25, 3], requiring a more computational intensive analysis of video streams.

Within the SocialWear project,[1] we are conceiving solutions supporting an easier integration between the deaf and the speaking communities. One of the scenarios considered within the project aims at supporting the integration of a speaking person within a group of deaf speakers by translating, in real-time, sign language videos taken from wearable cameras, into voice (see Figure 1). The most challenging technological aspect being the need to recognize motion of people with diverse body proportions and clothing, framed from cameras positioned at different height, unpredictable framing angles, various lighting conditions and any background.

On the other hand, systems to estimate body and facial configurations "in the wild" do exist, see for instance [2, 5, 18], and could be trained due to the availability of big datasets. However, such training data for sign language translation of many national sign languages is missing (and will likely be missing for many years ahead), thus preventing the development of end-to-end translation systems in uncontrolled scenarios.

Therefore, we are proposing an approach to recognize sign language in the wild through a training pipeline that includes an augmentation of sign language animation data via synthetic generation (see Figure 2). The main idea is to delegate the identification of 3D landmarks in sign language video streams to specialized software, which is trained on big corpora in diverse conditions. Then, a 3D software will augment the 3D landmarks corpus to simulate cameras with different lenses and framing angles. Finally, train a neural network able to translate 3D landmark information into text, also bypassing any intermediate symbolic representation of sign language.

After introducing some related work (Section 2), in Section 3, we describe our envisioned pipeline and how to implement it. Section 4 describes our hypotheses and the evaluation plan, and finally Section 5 concludes the paper.

[1] https://www.dfki.de/en/web/research/projects-and-publications/projects-overview/projekt/socialwear/

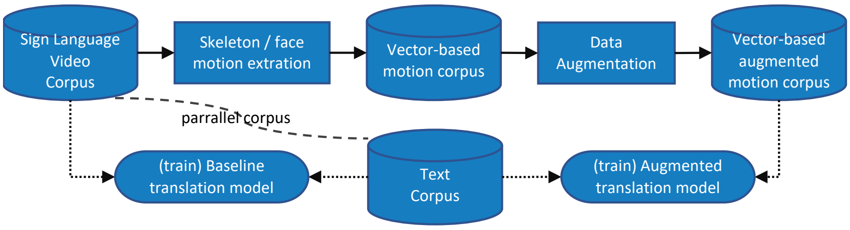
Sign Language Video Corpus
Skeleton / face motion extration
Vector-based motion corpus
Data Augmentation
Vector-based augmented motion corpus
parralel corpus
(train) Baseline translation model
Text Corpus
(train) Augmented translation model

Figure 2 A diagram of the proposed translation pipeline.

2 Background

Very recent works approach SL video-to-text as a translation task that in most cases uses intermediate sign glosses [3, 4, 27, 28, 22]. These works employ a variety of neural machine translation (NMT) architectures, which mainly differ on how the input (video) is encoded —CNN, STMC networks, etc. Traditionally, the 2-steps conversion video-to-gloss followed by gloss-to-text has performed better than the end-to-end task. However, as currently happens in several deep learning problems, the use of transformer architectures [26] starts favouring end-to-end learning. Camgoz et al. (2020) [4] and Yin et al. (2020) [27] achieve state-of-the-art results on the RWTH-PHOENIX14T corpus [10, 11] —a standard test set for SL interpretation— with transformer-based NMT systems.

As already introduced, recognizing sing language motion in the wild would require an amount of data that is not available as of today. To circumvent this problem, we propose splitting the translation pipeline into a first phase, recognizing 3D landmarks from videos. This would allow augmenting the data by applying transformation techniques (e.g. simulating various recording conditions, size of body parts etc.) or creating additional features given the landmarks. Then the augmented vector-based information can be used to train the translation from landmarks into text.

Within this approach lies the work done by Ko et al. [14, 13]. They use NMT architectures as in previous works, but they extract 2D coordinates of human keypoints from the input videos and use these coordinates to train the neural translation systems. Vector-based animation data allows for applying object 2D normalization whereas the authors apply random frame skip sampling to augment the video data. Unfortunately, [4, 27] and [13] cannot be compared directly because systems are applied to different sign languages, domains, and test sets. Contrary to our suggestion, Ko et al. apply augmentation techniques directly on the video frames (prior to the landmark recognition) but not at the recognized landmarks (after the landmark recognition and before the translation from landmarks to text), as suggested by us.

More elaborated synthetic data augmentation techniques have been already employed in the generation of synthetic data for the task of recognizing hand poses. For example, Malik et al. [20, 19] generated more than 5 million images of hand configurations by setting up a virtual human in front of a desktop environment and simulating the random movement of a hand in front of a webcam. Also, Mueller et al. [23] generated a synthetic dataset by first capturing the motion of real hands, retargeting the motion to a virtual hand, framing it from an egocentric point-of-view, and then augmenting the dataset by modulating hand shape and skin color, adding occluding objects, and imposing random real-world backgrounds. In Covre et al. [6], the authors recorded gestures executed by a single human and transferred it to a virtual human. Then, they augmented the motion of the virtual character in order to train a gesture classification model based on random forests. The augmentation concerned both modulating the gesture dynamics and moving the virtual camera. These techniques have not been applied to sign language, which is based on the movement of more body elements than just hands (e.g., posture, face).

3 Proposed methodology

State-of-the-art work from Camgoz et al. [4], achieves a BLEU score of more than 20 points by employing deep learning on an end-to-end translation approach from video to text. The input is processed within the neural network via a spatio-temporal embedding bound with a positional encoding. A major weakness of this approach is that the translation system relies on the raw input of a stream of video pixels, thus leading, from our point of view, to the following limitations:

- The resolution of the input video stream is forced by the architecture. Hence, videos at higher resolution must be scaled down even when a higher resolution, and thus more details, would be available. However, the higher the resolution of the video, the higher the computational power needed to train and run the neural network;
- The system is bound to the recording conditions of the corpus (RWTH-PHOENIX14T [10]) used for the experiments such as camera lenses (aperture and distortion), camera distance and angle, lighting conditions, background;
- The system is bound to the physical characteristics and the dress-code directives (black long-sleeved sweaters) present in the training corpus.

We assume that such a system would not be able to reach the same performance when tested on a video of a person with different clothing and skin colour, different light conditions or viewing distance or angle. This can be attributed to a known limitation of neural networks, which badly generalize for input coming from a different *distribution* than the one used for training and testing. An end-to-end neural architecture would be in principle able to learn to generalize the translation given different conditions, but this would require a vast amount of video-to-text parallel corpora, where the same phrases would be repeated under these different conditions. Unfortunately, the lack of sign language parallel corpora and the difficulty to obtain them is a major obstacle to building such a robust system.

Hence, we propose adding an intermediate level of indirection in the translation pipeline, by first extracting *motion data* of the SL speakers in terms of skeletal motion (for body and hands) and displacement of key points of the skin (for the face) from the video streams. Figure 2 shows a diagram of the proposed architecture. This way, the translation of the sign language into text is performed on the animation data of a 3D virtual human, rather than on the video of a real human. This would lead to several advantages:

1. There are sufficient data and very strong existing models for recognizing skeletal and facial motion in-the-wild, such as OpenFace [2], OpenPose [5], and MediaPipe [18];
2. the performance of the system would not be affected by the identity of the signer nor by their clothes;
3. the translation system would be independent from lighting conditions;
4. it would be possible to provide the network with additional data points, like the distances between joints (a feature often very useful in hand pose or gesture recognition);
5. the size of the skeleton and the face could be normalized to improve the sign recognition on bodies with very different proportions;
6. it opens the possibility to augment the animation information through the simulation of different camera position and lenses via geometrical transformations, thus training a system able to recognize signs from different distances and shooting angles;

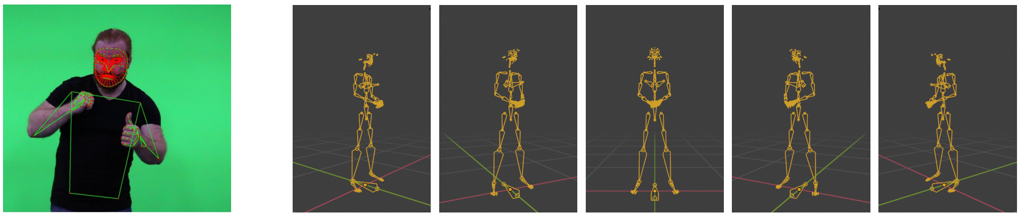

Figure 3 Preliminary work on (left) capturing the skeletal motion from a user and (right) augmenting skeletal motion as seen from multiple points of view. [24]

7. the neural architecture dedicated to the translation of motion data into text would be smaller, and hence faster and more energy efficient, as the quantity of information received as input would be two orders of magnitude inferior to video data.

Concerning the last point, as a rough estimation of the reduction in the quantity of information, consider that one second of RGB color video at resolution 210x260 pixels (as for RWTH-PHOENIX14T) at 25 FPS would require 4095 KBytes. In contrast, animation data, counting roughly 60 bones for the upper body (4-tuple quaternions for the rotations) and 468 face landmarks (3-tuple for each vertex in space), encoded as 4-byte floats, makes 6576 bytes per frame, which recorded at 25 FPS leads to approximately 164 KBytes per second. This is about 4% of the corresponding low-resolution video data.

Points 4-7 above demonstrate an advantage as compared to state-of-the-art Ko et al. [14, 13], as the augmentation will occur directly on the landmarks, providing additional data related to observed weaknesses of the existing models.

A drawback of this approach is that any error introduced by the skeletal and the facial recognition stage would propagate to the translation stage. Nevertheless, given the consistent improvement of the technologies specifically dedicated to the motion tracking of body, hands, and face we are confident that the tracking errors introduced by the motion analysis stage would be limited and well compensated by the advantages of a lighter architecture dedicated to the translation process. Additionally, if deemed necessary, deep learning offers the possibility of handling both the skeletal/face recognition and the translation through the same joint neural network, which would minimize the effects of error propagation.

4 Empirical Evaluation Plan

As already introduced, we plan to extract skeletal and facial motion data from videos, and use those to feed a MotionData-to-text (MD2Text) translation system. For the extraction of motion data, we plan to use the recent MediaPipe[2] [18] framework, which provides tools for the extraction of body, hands, and facial motion data. Figure 3 shows the result of initial tests. As for the corpus, we will use the RWTH-PHOENIX-14T dataset as it has been used in previous related research such as that of Camgoz et al. [4].

We can summarize our experiments with the following pipeline:

1. Retrieve the corpus `V` of videos from RWTH-PHOENIX-14T;
2. create a baseline model `V2Text`, which takes plain videos as input: train it and measure its performances on corpus `V`;

[2] `https://mediapipe.dev/`

3. create a corpus `MD` (motion data for body, hands, and face) by analysing the videos of `V` using MediaPipe;
4. define a model `MD2Text`, which takes motion data as input: train it and measure its performances on corpus `MD`;
5. augment the `MD` corpus with additional feature like mutual distances between joints (`MD+D`), camera settings and positions (`MD+C`), and by normalizing body proportions (`MD+B`);
6. train a model `MD+2Text` on the augmented `MD+D+C+B` corpus and measure its performances;
7. create a corpus of "videos in the wild" (`WV`) of new signers, with random clothes, diverse camera framing angles and lenses;
8. measure the performances of `V2Text` on `WV`;
9. extract the motion data `WMD` from `WV`; and
10. measure the performances of `MD2Text` and `MD+2Text` on `WMD`.

The goal of the set of experiments is to verify the following hypotheses:

- **H1**: `V2Text` performs worse on `WV` than on `V`;
- **H2**: `MD2Text` performs better than `V2Text`;
- **H3**: `MD2Text` and `MD+2Text` require much less computational resources than `V2Text` for both training and inference (for the latter, sum up the inference time of MediaPipe);
- **H4**: `MD+2Text` performs better than `MD2Text` when tested on `MD`;
- **H5**: `MD+2Text` performs better than `MD2Text` when tested on `WMD`;
- **H6**: finally, `MD+2Text` performs on `WMD` as good as on the original `MD`.

5 Summary

In this paper we have described and analyzed the current main approaches for the translation of sign language into text, and detected the main weaknesses for an application in real daily life. To overcome those limitations, we proposed an approach based on chaining a video-to-motion recognition system followed by an end-to-end translation approach from motion vectors into text.

Whereas, nowadays, researchers are proving the superiority of pure end-to-end architectures trained on huge quantities of "dirty" data, such approach cannot be yet applied in the context of sign language because of the scarcity of resources. In general, in this work we are exploring the possibility of training systems starting from a smaller quantity of "clean" data (recorded in controlled conditions) and improve performances through an artificial introduction of data variability. An alternative and complementary approach would be the exploitation of fine-tuning and domain adaptation techniques which would allow using models trained in richer settings (such as video captioning, video question answering or video and language inference) as initialisation of sign language translators. This is another possible research line left as future work and has been not considered in the discussion.

A reasonable limitation of our approach is that data augmentation is not really equivalent to increasing the sampling size, but rather a localized exploration of the neighbourhood of existing samples along some of the features characterizing the input domain. Still, data augmentation has proven to be effective in image classification, and it is part of the challenge to prove that it will be effective on motion analysis too.

In the presented proposal, we described the idea of augmenting the data mainly by generating different camera framing conditions. In future work, we could explore the benefits of augmentation applied to the human motion, too, by performing a modulation of the dynamics of the motion (e.g., time scaling and time warping) and by the manipulation of motion trajectories.

References

1 Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1219.

2 Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, Xi'an, May 2018. IEEE.

3 Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. doi:10.1109/CVPR.2018.00812.

4 Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. IEEE, 2020. doi:10.1109/CVPR42600.2020.01004.

5 Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

6 Nicola Covre, Fabrizio Nunnari, Alberto Fornaser, and Mariolino De Cecco. Generation of action recognition training data through rotoscoping and augmentation of synthetic animations. In *Augmented Reality, Virtual Reality, and Computer Graphics*, pages 23–42, Cham, 6 2019. Springer International Publishing. doi:10.1007/978-3-030-25999-0_3.

7 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

8 David M. Eberhard, Gary F. Simons, and Charles D. Fenning. *Ethnologue: Languages of the World. Twenty-third edition.* SIL International, Dallas, Texas, 2020.

9 R. Elliott, J. R. W. Glauert, J. R. Kennaway, and I. Marshall. The development of language processing support for the visicast project. In *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, Assets '00, page 101–108, New York, NY, USA, 2000. Association for Computing Machinery. doi:10.1145/354324.354349.

10 Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Language Resources and Evaluation*, pages 3785–3789, Istanbul, Turkey, May 2012.

11 Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *LREC*, pages 1911–1916, 2014.

12 Alexis Heloir and Michael Kipp. EMBR - A Realtime Animation Engine for Interactive Embodied Agents. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents (IVA-09)*, 2009.

13 Sang-Ki Ko, Kim Kim, Hyedong Jung, and Choong sang Cho. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9:2683, 2019.

14 Sang-Ki Ko, Jae Gi Son, and Hyedong Jung. Sign language recognition with recurrent neural network using human keypoint detection. In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems*, RACS '18, page 326–328, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3264746.3264805.

15 Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of*

the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2046–2065, Online, November 2020. Association for Computational Linguistics.

16 Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020.

17 Vincenzo Lombardo, Cristina Battaglino, Rossana Damiano, and Fabrizio Nunnari. An avatar-based interface for the italian sign language. In *Proceedings of the 2011 International Conference on Complex, Intelligent, and Software Intensive Systems*, CISIS '11, pages 589–594, Washington, DC, USA, June 2011. IEEE Computer Society. `doi:10.1109/CISIS.2011.97`.

18 Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines. *arXiv:1906.08172 [cs]*, June 2019. arXiv: 1906.08172.

19 Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, and Didier Stricker. Simple and effective deep hand shape and pose regression from a single depth image. *Computers & Graphics*, 85:85–91, October 2019. `doi:10.1016/j.cag.2019.10.002`.

20 Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In *2018 International Conference on 3D Vision (3DV)*, pages 110–119. IEEE, 9 2018. `doi:10.1109/3DV.2018.00023`.

21 Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

22 Taro Miyazaki, Yusuke Morita, and Masanori Sano. Machine translation from spoken language to sign language using pre-trained language model as encoder. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144, Marseille, France, May 2020. European Language Resources Association (ELRA).

23 Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE ICCV Workshops*, Oct 2017.

24 Florian Schicktanz, Lan Thao Nguyen, Aeneas Stankowski, and Eleftherios Avramidis. Evaluating the translation of speech to virtually-performed sign language on AR glasses. In IEEE, editor, *Proceedings of the Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)*, 2021.

25 Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *British Machine Vision Conference*, Northumbria, UK, 2018. British Machine Vision Association.

26 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

27 Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. `doi:10.18653/v1/2020.coling-main.525`.

28 Jiangbin Zheng, Zheng Zhao, Min Chen, Jing Chen, Chong Wu, Yidong Chen, Xiaodong Shi, and Yiqi Tong. An Improved Sign Language Translation Model with Explainable Adaptations for Processing Long Sign Sentences. *Computational Intelligence and Neuroscience*, 2020:11, 2020. `doi:10.1155/2020/8816125`.