Chapter 21
Motion Capture 3D Sign Language Resources

Zdeněk Krňoul, Pavel Jedlička, Miloš Železný, and Luděk Müller

Abstract The new 3D motion capture data corpus expands the portfolio of existing language resources by a corpus of 18 hours of Czech sign language. This helps alleviate the current problem, which is a critical lack of quality data necessary for research and subsequent deployment of machine learning techniques in this area. We currently provide the largest collection of annotated sign language recordings acquired by state-of-the-art 3D human body recording technology for the successful future deployment of communication technologies, especially machine translation and sign language synthesis.

1 Overview and Objectives of the Pilot Project

Sign language (SL) is a natural means of communication for deaf people. About 70 million people use SL as their first language and there are more than 100 different dialects used around the world. Although significant progress has been made in recent years in the field of language machine learning techniques, the field of SL processing struggles with a critical lack of quality data needed for the successful application of these techniques. SL resources are scarce – they consist of small SL corpora usually designed for a specific domain such as linguistics or computer science. There are some motion capture datasets for American Sign Language (ASL) and French Sign Language (Lu and Huenerfauth 2010; Naert et al. 2017) with a total recorded time of motion of up to 60 minutes. The situation is even worse for "small" languages.

The 3D reconstruction of human body motion using images and depth cameras is a common approach for capturing the movement of the human body (MMPose Contributors 2020). Current large SL datasets are mostly based on 2D RGB videos (Vaezi Joze and Koller 2019; Zelinka and Kanis 2020). The main goal of our project is to deliver a large 3D motion dataset collected using high precision optical marker-based motion capture and to extend the existing ELG portfolio of language resources

Zdeněk Krňoul · Pavel Jedlička · Miloš Železný · Luděk Müller
University of West Bohemia, Czech Republic, zdkrnoul@ntis.zcu.cz, jedlicka@ntis.zcu.cz, zeleny@ntis.zcu.cz, muller@kky.zcu.cz

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_21

by Czech sign language (CSE) data. For comparison SIGNUM, one of the largest video-based SL datasets, contains approximately 55 hours of SL recordings (Koller et al. 2015) and one of the largest 3D motion capture datasets contains only 60 minutes of SL recordings (Naert et al. 2017).

Motion capture technology guarantees precise recording of the signer's movements in 3D space at the cost of a more complex preparation phase compared to standard video recording. Optical marker-based motion capture has become the industry standard for capturing movement of the human body. In Jedlička et al. (2020), we collected the first 3D motion capture dataset for CSE, covering the weather forecast domain. It has a rather limited size and contains recordings of one signer only.

Our contribution can be summarised as follows:

- Proof of concept of large-scale motion capture recording of multiple SL speakers;
- Provide 3D motion capture data to cover wider domains, grammatical context and more signers. We perform proper data post-processing, annotate glosses, and develop tools for data extraction from the collected dataset;
- The largest SL motion capture dataset consisting of recordings of continuous SL phrases and a vocabulary of six native SL speakers from carefully selected domains, in total more than 18 hours;
- Tools that allow searching for individual glosses, phrases, or small movement sub-units (e. g., given hand shape/action) in the dataset.

2 Methodology and Experiment

A new recording procedure for a large amount of 3D motion capturing of SL was investigated to ensure sufficient diversity of SL speakers, grammar, and sign contexts. This makes the new language resource more versatile and useful in many different research fields such as further linguistic and SL motion analyses. The integral part of the experiment is data processing.

In Jedlička et al. (2020), the experimental recording setup with VICON 18 cameras was used as proof of the intended concept. The negative aspect of this setup was its high complexity; the setup was very time demanding and not suitable for large-scale data and multiple speakers.

The new procedure simplifies the process by dividing the setup into two separate parts: large-scale body movement and small-scale, highly detailed finger movement are recorded with two separate motion capture camera setups, each of which uses a reduced number of capture cameras and is adjusted slightly for different speakers.

2.1 Recording Setup

We used our laboratory equipment, i. e., the VICON motion capture system with eight cameras. We extended it with a standard color video camera for a reference video. The frame rate was 100 frames per second (fps) for the motion capture and 25 fps for the reference video. The VICON system records movement using passive retro-reflexive markers attached to the human body. Movement is modeled as a set of movements of the rigid parts connected by the skeleton; the marks are placed on the poles of the rotation axis of the main skeleton joints. Each body part is defined by at least four markers, except fingertips, see Figure 1.

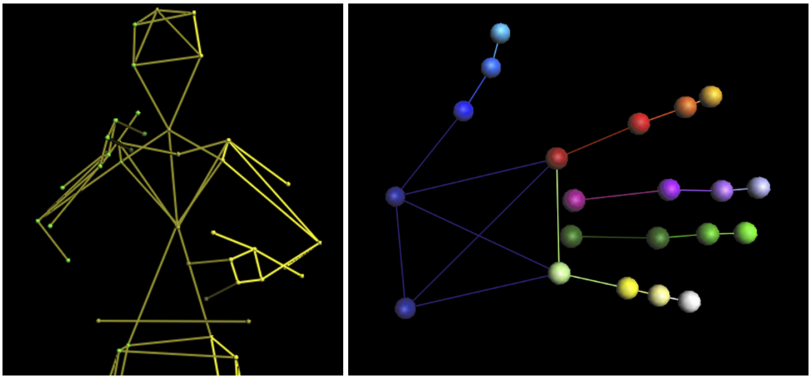

Fig. 1 Visualisation of SL body marker setup (left) and SL hand-shape marker setup (right)

The SL body marker setup is based on marker positions defined by the VICON three-finger standard. It uses a total of 43 markers for tracking upper body, head, arms, and palms movement. A simple hand pose is provided at the same time and incorporates tracking of thumb, index, and little fingertips. Moreover, this setup includes face tracking providing a non-manual component of SL, that is reduced to seven facial markers. The SL hand-shape marker setup is designed for detailed hand-shapes recording. Each hand-shape is recorded separately. Data is recorded for the right hand only. The movement starts from the relaxed hand-shape, then changes to the given hand-shape and back to the relaxed hand-shape. For both setups, data capturing was supervised by CSE linguists.

2.2 Data Annotation

An essential step is the annotation of captured SL utterances. We use time-synchronised reference video, the ELAN tool (Figure 2) and SL experts. The annotation of a sign is done by giving the information of the sign's meaning (gloss), and the right and the left hand-shape. If the sign consists of more than one defined hand-shape, the hand-shapes are annotated as a set of hand-shapes. Both the activities are very laborious and time-consuming. To successfully complete this task, we involved several trained annotators who worked in parallel.

2.3 Data Post-processing

Post-processing consists of data-cleaning, whole-body motion reconstruction, and data-solving. Data-cleaning removes noise and fills gaps in the raw 3D data caused by frequent mutual occlusions of markers during signing, and other noise caused by the environment. Motion reconstruction and data-solving recalculate marker positions into the movement of the skeletal model.

The data of both setups was post-processed. We reconstructed small gaps by the interpolation standard technique as long as the trajectory was simple enough. Note, that the recording speed is 100 fps, which is fast enough to contain minimal changes in trajectory between frames. We used semi-automatic 3D reconstruction of marker trajectories and labeling, and manual cleaning of swaps and gaps. For the body parts defined by at least four markers, filling in the trajectories of the marker is well automatised because at least three points are enough to define the missing position.

The body marker setup uses only one marker per fingertip and some larger gaps caused by more complex self-occlusions of body parts can obscure three or more markers in one rigid segment. Post-processing in those cases is more complicated and gaps must be filled in manually.

The full SL body movement is achieved as a composition of the body movement and corresponding data of the hand-shapes setup. For this purpose, the annotation of hand-shapes provides us temporal segmentation of the recordings. Thus the fingertip motion segments can provide information about dynamic changes during the performance of a particular SL hand-shape in a particular data frame.

The middle part of a given segment is always completed according to the hand-shape(s) assigned by the annotation. We captured full fingers motion only for the transition of the given hand-shape from and to the neutral hand-shape. Thus, for the other frames of the segment, the nearest hand pose with the smallest reconstruction

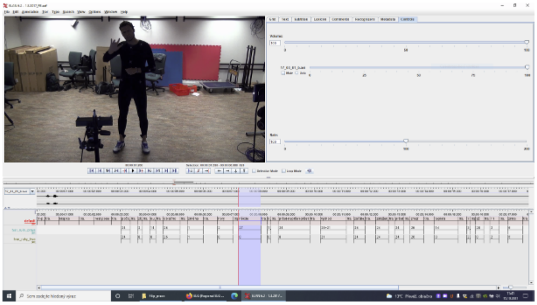

Fig. 2 Example of annotation work in ELAN, specifically designed software for the analysis of sign languages, and gestures

error can be used. We consider only those frames that have an alignment error below a given threshold. The remaining frames will have gaps in the final trajectories.

We solved the above problem as point-set alignment via Procrustes analysis that arises especially in tasks like 3D point cloud data registration. The rigid transformation of two sets of points on top of each other minimises the total distance in 3D between the corresponding markers (Arun et al. 1987). Since the data is noisy, it minimises the least-squares error:

$$err = \sum_{i=1}^{N} ||RM_f^i + t - M_{rf}^i||, \tag{1}$$

where M_f and M_{rf} are current and reference frame(s) respectively as a set of 3D points with known correspondences, R is the rotation matrix and t the translation vector. We define $N = 7$ as three fingertips (thumb, index, little finger), two wrist markers, and two knuckles of the index and little fingers. We aligned just the rotation and translation because the 3D transformation preserves the shape and size (same hand-shape and SL speaker). For the left hand, we mirrored the reference frame(s).

The last step is data-solving. It is a process of reconstruction of the 3D motion of the skeleton from the marker trajectories. For this purpose, we use the VICON software. The skeleton is well defined to directly control the SL avatar animation or handle animation retargeting.

2.4 Dataset Parameters

We limited the linguistic domain to two specific fields to reduce the number of unique signs. Weather forecasts and animal descriptions from the zoological garden domain were selected by CSE linguists. We were also given a list of all hand-shapes which occur in these domains. The dataset is collected from six SL speakers, who differ in their body size, age, and gender.

3 Conclusions and Results of the Pilot Project

SLs are not sufficiently supported through technologies and have only fragmented, weak, or no support at all. Our ELG pilot project offers a new SL resource designed for the development of language technologies (LTs) and multilingual services for Czech. The results contribute to the establishment of the Digital Single Market as one of ELG's objectives. In contrast to the all-in-one recording setup, the body movement is recorded separately from the highly detailed recording of hand poses. This separation reduces the camera setup complexity and the complexity of data during post-processing, which makes SL recording more flexible and adjustments for new SL speakers or data easier.

The project delivered a professionally created SL dataset via state-of-the-art 3D motion capture technology. The project provides data for the wider research community through ELG. We have recorded 18 hours of sign language and recorded six different speakers for two different domains.

We assume our results will be beneficial for other applications such as next generation SL synthesis that uses a 3D animated avatar for natural human movement reproduction or SL analysis or gesture recognition and classification in general.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects.

References

Arun, K. S., T. S. Huang, and S. D. Blostein (1987). "Least-Squares Fitting of Two 3-D Point Sets". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9.5, pp. 698–700. DOI: 10.1109/TPAMI.1987.4767965.

Jedlička, Pavel, Zdeněk Krňoul, Jakub Kanis, and Miloš Železný (2020). "Sign Language Motion Capture Dataset for Data-driven Synthesis". In: *Proceedings of the LREC2020*. Marseille, France: ELRA, pp. 101–106.

Koller, Oscar, Jens Forster, and Hermann Ney (2015). "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers". In: *Computer Vision and Image Understanding* 141. Pose & Gesture, pp. 108–125.

Lu, Pengfei and Matt Huenerfauth (2010). "Collecting a motion-capture corpus of American Sign Language for data-driven generation research". In: *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*. ACL, pp. 89–97.

MMPose Contributors (2020). *OpenMMLab Pose Estimation Toolbox and Benchmark*. URL: https://github.com/open-mmlab/mmpose.

Naert, Lucie, Caroline Larboulette, and Sylvie Gibet (2017). "Coarticulation Analysis for Sign Language Synthesis". In: *Universal Access in Human – Computer Interaction. Designing Novel Interactions*. Cham: Springer, pp. 55–75.

Vaezi Joze, Hamid and Oscar Koller (2019). "MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language". In: *BMVC*.

Zelinka, Jan and Jakub Kanis (2020). "Neural Sign Language Synthesis: Words Are Our Glosses". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3384–3392. DOI: 10.1109/WACV45572.2020.9093516.