Toward Sign Language Motion Capture Dataset Building

Zdeněk Krňoul[1(✉)], Pavel Jedlička[2], Jakub Kanis[1], and Miloš Železný[1]

[1] Faculty of Applied Sciences, NTIS - New Technologies for the Information Society,
University of West Bohemia, Univerzitní 8, 306 14 Pilsen, Czech Republic
{zdkrnoul,jkanis,zelezny}@ntis.zcu.cz

[2] Faculty of Applied Sciences, Department of Cybernetics,
University of West Bohemia, Univerzitní 8, 306 14 Pilsen, Czech Republic
jedlicka@kky.zcu.cz

Abstract. The article deals with a recording procedure for motion dataset building mainly for sign language synthesis systems. Data gloves and two types of optical motion capture techniques are considered such as one source of sign language speech data for advanced training of more natural and acceptable body movements of signing avatars. A summary of the state-of-the-art technologies provides an overview of possibilities, and even limiting factors in relation to the sign language recording. The combination of the motion capture technologies overcomes the existing difficulties of such a complex task of recording both manual and non-manual component of the sign language. A result is the recording procedure for simultaneous motion capture of signing subject towards further research yet unexplored phenomenon of sign language production by a human.

Keywords: Corpus building · Sign language · Motion capture

1 Introduction

In these days sign language (SL) translation or TV broadcast is provided by humans. SL synthesis is considered as supplementary communication means of the deaf individuals. One perspective technique is virtual 3D character animation in the form of the signing avatar [5]. However, there is still poor realism of the character animation compared to the standard video of the signing subject causing overall rejection of the signing avatars by the deaf community.

One reason for the rejection is that artificial signing avatars are not able to sign fluently and naturally and, therefore, it is difficult or uncomfortable to understand them. Integration of high-quality motion capture data is essential for any further research and gives certain assumptions to provide accessible SL synthesis [4]. The full body motion capture (mocap) including hand, finger, facial expression, and eye gaze movements may provide spatial-temporally synchronous records of all the channels [2].

A. Ronzhin et al. (Eds.): SPECOM 2016, LNAI 9811, pp. 706–713, 2016.
DOI: 10.1007/978-3-319-43958-7_86

There are different approaches using different technology for body mocap [3]. These approaches are optical, gyroscopic, mechanical, etc. and designed for mocap of different body parts. There are more specialized techniques based on markers fixed on speaker's face or marker-less techniques tracing the face by image processing gray, color and/or depth data[1]. The optical mocap systems are based on special cameras to track active or passive markers in 3D space (e.g. VICON, VICON Cara[2], Qualisys, OptiTrack, Optotrak). Whilst data processing provided by the VICON and VICON Cara systems are very beneficial for SL corpora building, there is a limited functionality of tools for CyberGlove3[3] data glove [6] using for precise finger mocap. The both facial and body capturing by one mocap system may result in noisy positions of the facial markers. Moreover, simultaneous capturing of the body, fingers, and facial data at high-frequency rates cause technical difficulties.

In the paper, we present SL recording procedure allowing the simultaneous body, finger, and facial mocap; flexible setting of the data parameter; spatial-temporally synchronous record; the data glove calibration; and mocap data interpretation by the 3D character model.

2 Combined SL Motion Capturing Method

We consider three mocap systems for the SL data acquisition task: VICON, VICON Cara, and CyberGlove3. The VICON and VICON Cara systems are a marker-based optical system. The optical capture principle was chosen because the signing subject is not wearing any special suit that limits his or her natural movement and the marker-based principle was chosen for its higher precision compared to non-marker approaches. Two CyberGlove3 data gloves provide robust finger mocap using bent sensor principle.

2.1 Body Motion Capturing

In our case, the VICON motion capture system consists of eight T-series cameras measuring a motion of passive spherical retroreflective markers in the infrared spectrum. The T-20 is a high-frequency camera with 2 Mpx resolution and is capable of frame rate 1200 fps (690 fps in full resolution). The system includes also VICON Blade software used for a camera set-up, calibration, and motion capturing itself. There are some limiting factors in consequence of the capturing principles and recording of complex body movement.

The main limiting factor is the tracking of the finger movements. Since the finger markers are close to each other, there is a significant number of overlapping situations (frames with marker swaps), especially the hand contacts have to be resolved during data post-processing. Moreover, such mocap setup requires at least 30 additional finger markers. We observed also negative effects of fixation

[1] www.faceshift.com.

[2] https://www.vicon.com/products/camera-systems/cara.

[3] http://www.cyberglovesystems.com/cyberglove-iii/.

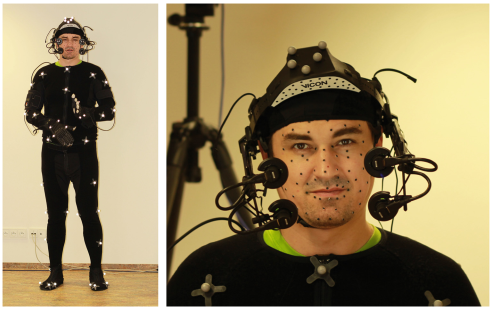


Fig. 1. The marker setups for the optical mocap systems. On the left: body mocap consisting of 53 14 mm retroreflective markers for the VICON system, on the right: 54 black passive markers for the VICON Cara system.

of the finger markers when they were not rigid to the particular finger segment during its bending. It causes an inaccuracy in the identification of the skeleton model internally used by the VICON system. On the other hand, there are unwanted losses of the finger markers attached directly to the skin caused by frequent touches of the hands during the signing. We have observed also the higher motion speed of the finger markers mainly for fingertips, which requires higher camera frame rate compared to capturing remaining body parts.

According to our experience, a standard set of 53 passive 14 mm markers fixed on the body of the signing subject is optimal to capture a head, shoulders, arms and wrist including hand/body contacts. The consider marker setup contains 10 markers on each arm and 15 markers on the torso and head providing mocap of any general movement of the whole upper body, see Fig. 1 on the left.

2.2 Hand Motion Capturing

The CyberGlove3 data glove is based on the resistive sensors of finger bending that provide robust measurements of hand shape especially during finger contacts on one or mutually between hands, see Fig. 2. In addition, the data glove measures also palm flex and wrist rotation like pitch and yaw. On the other hand, the reading of one sensor is relative to the preceding finger segment or the wrist and thus does not capture absolute 3D position.

The calibration is needed to found the conversion relationship between the sensor raw data and the actual finger bending. The manual calibration controlled by a protocol is a preferred option for SL mocap corpus building [8]. However, once identified calibration parameters does not provide precise conversion of hand shapes after re-dressing of the gloves by the same subject. This data inconsistency must be taking into account while creating of the SL mocap corpora. The standard CyberGlove3 tool enables the calibration only for common simple hand shapes. The very laborious and time-consuming process is calibration of the thumb touch with the rest of fingers. But the thumb-pinky finger touch was not achieved anyway. There are, moreover, also reported problems of glove sensor cross-coupling [9].

2.3 Facial Motion Capturing

The VICON Cara is a motion capture system considered for the marker-based facial motion capturing. The system consists of a headgear (HeadRig) with four cameras, a processing unit, a storage device, and a battery pack, see Fig. 1 on the right. An integral part is the operating and the post-processing software tool enabling calibration, triggering the Cara device and 3D reconstruction. The HeadRig is equipped with four 720p HD high-speed cameras with the framerate up to 60 fps. It is possible to natively synchronize time-code with the VICON system. Constant light conditions are provided by a custom designed controllable rig of four lights. The cameras have 3 mm F2.0 IR filtered lens. Noise caused by the T-series camera strobes is reduced by the IR filtered lens. The storage device has a recording capacity of 64 GB and two hours battery time.

The markers can be placed on the subject's face in the form of glued circles, drawn by an ink marker, or drawn by a make-up. White or black passive facial markers can be used. The certain positions of the marker are not required and there is also no default SL facial marker set. For example, the MPEG4 standard defines 53 markers as the set of Facial Feature Points, in the SignCom project, there were 41 markers used for mocap of the French SL [1] and the set of 60 markers was used in the Sign3D project [7]. The software tool detects these markers in the record as circular blobs and then finds it's centroids. After that, the 3-D position of each centroid is computed and the set of (x, y and z) positions over time forms the facial mocap data.

2.4 Combining Optical and Data Glove Recording

It is necessary to determine which body parts will be included in the motion capture data. There is a defined connection of the body mocap and the hand mocap and a connection of the body mocap and the facial mocap. The first option for the hand and body mocap connection is to use a mapping of the wrist pitch and jaw sensor to the target model and the VICON system determines only the wrist position and the forearm twist. The preferred option is tracking a full/global wrist rotation by the VICON system and only fingers and palm flex by the data glove. In this case, at least two optical markers on the back of the hand have to be added to the two markers placed on the wrist joint, see Fig. 2. The connection body part of the body mocap and the facial mocap is subjects head. The global position of the head is tracked by the VICON system through the HeadRig of the Cara system and the facial mocap data are relative to the transformation.

2.5 Character Model

The objective of the mocap data for SL synthesis is its proper interpretation by the 3D model. We assume 3D character model created by Autodesk character generator[4], see Fig. 3 on the right. The model overcomes a limitation of the

[4] Available at https://charactergenerator.autodesk.com/.

built-in hand model internally used by the CyberGlove3 tool. In addition, the model is appropriate also for the body and the facial mocap data.

The model includes the standard three bones per finger, moreover, index and pinky metacarpal bone, and 21 auxiliary bones of the facial rig. There is a support for the characterization of the body mocap data and it also allows the animation retargeting to the different body proportions (SL speaker/model). The standard skinning method is based on a weighted transformation of each bone and affects vertices of the surface mesh.

The bone-ends of the facial rig are fixed to predefined 3D positions in the model mesh surface. In general, the positions differ from the positions of chosen facial marker set. We consider manual retargeting of the facial mocap data to the character's face. For this purpose *position constraints* were defined by the professional 3D character animation software Autodesk MotionBuilder (MB). One constraint defines the affected bone-end as a constrained object and one or more of the facial markers as source objects. As a result, all the facial markers are transformed by weighting interpolation to the facial rig of the model.

3 SL Recording Procedure

The sign language recording procedure determines suitable steps for feasible and functional simultaneous recording of SL mocap data. The procedure divides the data acquisition to a capturing session and data post-processing.

3.1 Capturing Session

Facial Mocap. First of all, it is necessary to prepare and adjust the VICON Cara system for capturing a particular subject. The position of each camera has to be adjusted and focused on a target part of the subject's face. The next step is a standard calibration of the system. Markers are placed on the subject's face according to the desired model while the HeadRig is removed. After that, the HeadRig is returned on the subject and recording of the range of movement follows.

Hand Mocap. We consider capturing of a raw glove motion data without predefined glove calibration. We assume recently developed tools for the control and the communication with the CyberGlove3 gloves [6]. The tools provide an interface for recording with one or two (left and right) gloves at the time and also enables necessary time synchronization between the gloves and the VICON system.

The glove recording session starts by a launching of the above-mentioned tool for the simultaneous recording of both gloves. First, the time-synchronization stamp from the VICON system is set to the gloves. The particular commands for the time set are sent to the gloves at the same time. However, it can be executed by each glove with a slightly different delay depending on the processing unit of each glove. To time-synchronize the data recording we set the same internal time

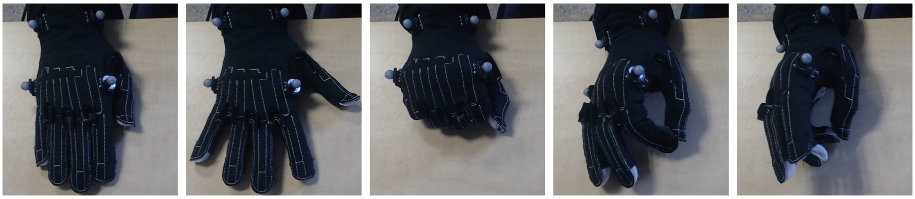


Fig. 2. Calibration take hand shapes.

for both gloves by one command and then start the recording simultaneously with another one command. This procedure allows us to reach the time difference between gloves in a range of a one data frame, i.e. 33.3 ms because there are 30 frames per second. The time difference can be greater and it is recommended to keep the internal time setting until the difference is acceptable. As soon as it is acceptable, we can start the recording.

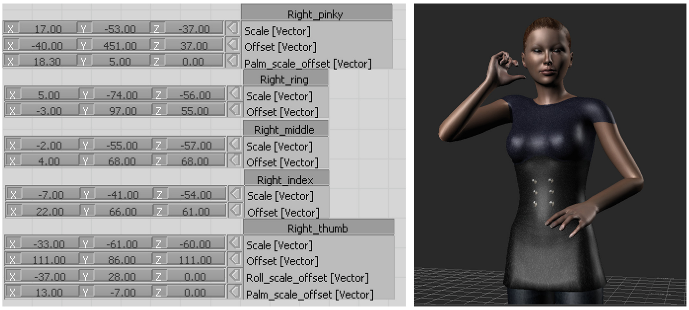
The first step (and it is beneficial to be the last one too) of the glove recording session is capturing of a calibration take which is essential for the successful glove raw data interpretation. The calibration take consists of five hand shapes: a flat hand, a stretching of all fingers, a fist and two "o" hand shapes, the one with thumb – index touch and the second with thumb – pinky touch respectively, see Fig. 2. The most important feature of the calibration take is to cover the full range of all finger movements. But a researcher can define its own calibration take which better suits his needs. Next, we can launch the standard recording.

Body Mocap. The T-20 cameras are situated and aimed at the captured subject. The camera layout depends on the subject's body proportion and on the complexity (range) of the SL recording material. The next step is the calibration of the system. The markers are placed on the subject according to the body model. Each capturing session starts with the standard recording of the range of movements (ROM).

3.2 Data Post-processing

The data acquired during the session have to be post-processed to get the standard motion capture data. The VICON Cara Post is a software tool used for post-processing data acquired by VICON Cara. The centroids of the markers placed on the subject's face are identified and cleared from an incidental noise. The 3-D reconstruction and the labeling of the final motion capture data are made after that.

The data from T-20 cameras are post-processed in the VICON Blade software tool. The reconstruction of 3-D data is made and necessary manual denoising is needed. The noise can be caused e.g. by body marker occlusions. The labeling of the data and export as the final mocap data follows.



Right_pinky
X 17.00 Y -53.00 Z -37.00 Scale [Vector]
X -40.00 Y 451.00 Z 37.00 Offset [Vector]
X 18.30 Y 5.00 Z 0.00 Palm_scale_offset [Vector]

Right_ring
X 5.00 Y -74.00 Z -56.00 Scale [Vector]
X -3.00 Y 97.00 Z 55.00 Offset [Vector]

Right_middle
X -2.00 Y -55.00 Z -57.00 Scale [Vector]
X 4.00 Y 68.00 Z 68.00 Offset [Vector]

Right_index
X -7.00 Y -41.00 Z -54.00 Scale [Vector]
X 22.00 Y 66.00 Z 61.00 Offset [Vector]

Right_thumb
X -33.00 Y -61.00 Z -60.00 Scale [Vector]
X 111.00 Y 86.00 Z 111.00 Offset [Vector]
X -37.00 Y 28.00 Z 0.00 Roll_scale_offset [Vector]
X 13.00 Y -7.00 Z 0.00 Palm_scale_offset [Vector]

Fig. 3. On the left: right hand manual calibration interface, on the right: the 3D character model.

The glove data post-processing phase allowing interpretation of the raw data and include the glove data calibration. The post-processing starts by the downloading of the recorded data files from the glove internal memory cards (calibration and data takes). The time-corresponding records for the left and right hand are then converted to the XMLTRC format (newly designed XML version of the TRC (Track Row Column[5]) format) and merged to the one corresponding XMLTRC file. An arbitrary XMLTRC file can be anytime later converted to the standard TRC format which is suitable for the processing of the 3D motion data by a standard animation software. The glove data calibration can be done in automatic and/or manual manner. For this purpose, we used the MB with a calibration template integrating the graphical user interface, see Fig. 3. This template allows a manual adjusting of all necessary calibration parameters (all scale and offset linear equation parameters). The TRC file of the merged calibration take is loaded into the MB with the active calibration template. The researcher can then adjust the template parameters until the finger motions of the given 3D model appropriately match the calibration take finger motions. The provided automatic calibration method can be optionally used as a starting point for the manual calibration. To be able to use the automatic calibration tool, the user only needs to identify calibration gesture keyframes in the calibration take by the supplied tool.

4 Conclusion

The recording procedure for motion dataset building is a crucial step to research new methods for the sign language synthesis systems. We combine the data gloves and optical motion capture techniques to collect source data of the sign language. The state-of-the-art technologies VICON, VICON Cara, and CyberGlove3 are discussed to summary advantage and also limiting factors in relation to motion capturing of the sign languages. The time-consuming and laborious

[5] http://simtk-confluence.stanford.edu:8080/display/OpenSim/ Marker+(.trc)+Files.

calibration of the two gloves is moved from a recording session to the phase of an off-line data post-processing when the presence of the signer is not required. The combination of the motion capture technologies overcomes the existing difficulties of such a complex task. The recording procedure provides instruction for researchers dealing with simultaneous recording both the manual and the non-manual component of the sign language. In this context, further research will be aimed to uncover naturalness of movements provided by the signing human.

Acknowledgments. This work was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by the UWB grant, project No. SGS-2016-039.

References

1. Gibet, S., Courty, N., Duarte, K., Naour, T.L.: The signcom system for data-driven animation of interactive virtual signers: methodology and evaluation. ACM Trans. Interact. Intell. Syst. **1**(1), 6:1–6:23 (2011). http://doi.acm.org/10.1145/2030365.2030371
2. Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., Turki, A.: Interactive editing in French Sign Language dedicated to virtual signers: requirementsand challenges. Universal Access in the Information Society, September 2015. https://hal.archives-ouvertes.fr/hal-01205742
3. Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: Computer Vision and Pattern Recognition, CVPR 2009, pp. 224–231 (2009)
4. Huenerfauth, M., Lu, P., Kacorri, H.: Synthesizing and evaluating animations ofamerican sign language verbs modeled from motion-capture data. In: SLPAT 2015, pp. 22–28. ACL, Dresden (2015). http://www.aclweb.org/anthology/W15-5105
5. Krňoul, Z., Kanis, J., Železný, M., Müller, L.: Czech text-to-sign speech synthesizer. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 180–191. Springer, Heidelberg (2008)
6. Krňoul, Z., Kanis, J., Železný, M., Müller, L.: Semiautomatic data glove calibration for sign language corpora building. In: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, LREC, May 2016, in press
7. Lefebvre-Albaret, F., Gibet, S., Turki, A., Hamon, L., Brun, R.: Overview ofthe Sign3D project high-fidelity 3D recording, indexing and editing of French Sign Language content. In: SLTAT 2013, Chicago, United States (2013). https://hal.archives-ouvertes.fr/hal-00914661
8. Lu, P., Huenerfauth, M.: Accessible motion-capture glove calibration protocolfor recording sign language data from deaf subjects. In: Proceedings of the11th International ACM SIGACCESS Conference on Computers and Accessibility, Assets 2009, pp. 83–90. ACM, New York (2009). http://doi.acm.org/10.1145/1639642.1639658
9. Wang, Y., Neff, M.: Data-driven glove calibration for hand motion capture. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2013, pp. 15–24. ACM, New York (2013). http://doi.acm.org/10.1145/2485895.2485901