Methods for Recognizing Interesting Events within Sign Language Motion Capture Data

Pavel Jedlička, Zdeněk Krňoul, Miloš Železný
University of West Bohemia, Faculty of Applied Sciences, NTIS - New Technologies for the Information Society
Univerzitní 8, Pilsen, Czech Republic
jedlicka@kky.zcu.cz, zdkrnoul@kky.zcu.cz, zelezny@kky.zcu.cz

Abstract

Rising popularity of motion capture in movie-production makes this technology more robust and more accessible. Utilization of this technology for sign language capturing and analysis is evident. The article deals with the usability of the motion capture in creating sign language corpora. A large amount of the data acquired by the motion capture has to be processed to provide usable data for wide range of research areas: e.g. sign language recognition, translation, synthesis, linguistics, etc. The aim of this article is to explore possible methods to detect interesting events in data using machine learning techniques. The result is a method for detection of the beginning and the end of the sign, hand location, finger and palm orientation, whether the sign is one or two handed, and symmetry in the two-handed signs.

Keywords: motion capture, machine learning, creating sign language corpora

1. Introduction

In these days signing language translation or TV broadcast is provided by humans. Sign language (SL) synthesis is considered as supplementary communication means of the deaf individuals. There are SL approaches intended for creating sign language in an artificial way. One perspective technique is using virtual 3D character animation as a signing avatar (Krňoul et al., 2008). However, there is still poor realism of such produced character animation compared to the standard SL video of signing subject causing overall unacceptability of the signing avatars by the deaf community. A huge disadvantage of image processing is that computer vision is a very complex task for the SL videos. Image recognition of the position of body parts such as arms, hands, and handshapes is possible, but accuracy is far worse than in using motion capture.

One reason is that artificial signing avatars are not able to sign fluently and naturally and, therefore, it is difficult or uncomfortable to understand them. On the other hand, replaying an utterance in 3D animation generated from the motion capture of sign language speaker provides very natural outcome because the captured motion copies movements of the SL subject. Such continuous data reflects a certain number of still unidentified phenomena of SL production system. Therefore, integration of high-quality motion capture data is essential for any further research and gives certain assumptions to provide accessible sign language synthesis (Huenerfauth et al., 2015).

The full body capture including hand, finger, facial expression and eye gaze movements is a condition to collect spatial-temporally synchronous records of all the channels (Gibet et al., 2015). However, for such complex recording, the motion capture hinders movements of subject's body so it does not have to compose a natural move. Moreover, an interconnection of SL annotations and motion capture data seem to be a crucial issue (Lu and Huenerfauth, 2012), (Gibet et al., 2015). Hereby, analyses of the motion capture data are often taken into account in limited short time intervals.

An analysis of 50 minutes of videos combined with motion captured data from French sign language corpora was conducted to extract low-level or high-level motor schemes (Gibet et al., 2012). There is incorporation of an automatic segmentation technique of the short hand-shape sequence (Heloir et al., 2006), a statistic analysis of phasing between hand motion and handshapes, categorizing of hand motion velocity profiles within signs and during sign transitions.

In the paper, we present initial experiences in the full body motion capture of Czech Sign Language interpreter. Each lexical item from a dictionary is produced when the signer's hands are returned to a relax-pose between the items during recording. The new technique for the motion capture data processing is presented to explore capabilities of automatic identification of start and end pose of the signs. In the context of the SL recording scenario, the experiment is uncovering helpful aspects that can lead to further inter-sing segmenting techniques of the SL motion capture or video data.

2. Sign language motion capturing

The popularity of using motion capture systems in many different tasks causes this system to be more accessible for non-commercial subjects. It also causes this technology to improve more and becoming more precise. There are more different systems using different technology for motion capturing (Hasler et al., 2009). These different approaches are optical, gyroscopic, mechanical, etc.

The optical system was chosen because the signing subject is not wearing any special suit that limits his or her natural movement. The marker-based system was chosen for its higher precision compared to non-marker approaches.

2.1. Initial experiences

The VICON system was chosen as a main motion capture technology of the data acquisition for the sign language synthesis task. The VICON motion capture system is based on the principle of high-frequency cameras measuring a

motion of passive spherical retroreflective markers in the infrared spectrum. However, there are some limitation factors given by the capturing principle for the finger movements and the handshapes used in sign languages. The first factor is the number of cameras. We found that it is sufficient to use eight cameras for an accurate and robust motion capturing of the body torso, arms and head of a signing subject. In this case, according to our experience, a standard set of optical markers is sufficient to exactly capture overlapping arms as well as other hand/body contacts that widely occur in the sign languages.

However, the simultaneous capturing of the fingers of both hands and the rest of the body requires at least 30 additional markers. The markers have to be smaller compared to the different proportion of the fingers and the rest of the body and also they must be somehow rigidly attached. For example, they can be mounted on a conventional glove like the other body markers. But in this case, we observed negative effects of such fixation. Mainly for a smaller hand, the markers were not rigid to the particular finger segment during its bending. Although this is a relatively small movement, it causes an inaccuracy in the identification of a model internally used by the VICON system. The problem can be eliminated so that the markers are attached directly to the finger skin. In this case, however, their unwanted loss caused by frequent touches of the hands while signing is not excluded. It was also observed that there is higher speed of marker movements mainly for fingertips, which requires higher camera frame rate than that for tracking other parts of the body.

The main limitation factor is the tracking of the finger markers that are close to each other and significantly increase overlapping situations (frames with marker swaps), especially during the hand contacts. These problems can be partially solved by the good positioning of the cameras, but this leads to increasing the number of cameras to 20 or more which can be of course expensive and difficult solution with limited functionality in the case of the full marker occlusions.

2.2. Combining optical and data glove record

The combination of the aforementioned optical and the data glove motion capturing is an alternative recording technique. The measurement principle of the finger bending is based on the resistive sensors that provide robust measurements of finger contacts on one hand or mutually between hands. In addition, CyberGlove3 glove measures palm flex and wrist rotation (pitch and yaw). On the other hand, the reading of one sensor is relative to the reading of the preceding finger segment or the wrist and thus we do not get absolute 3D position. Thus, the CyberGlove3 motion capture data are relative to the 3D position of the forearm.

3. Dataset

Data were acquired by VICON motion capture system using 8 T-20 cameras. The T-20 camera has 2 Mpx resolution and it is possible to record at a speed up to 690 frames per second (fps). Recording, reconstruction, and data post-processing were made in Blade software from VICON. This software provided also a body model. Motion capture of the handshapes was recorded simultaneously using Cybergloves3 based on flex sensor technology. There was also the availability of facial motion data by VICON motion capture Cara. It is a marker-based motion capture system using 4 cameras aimed at the face of the signing subject. It is possible to track tens of markers placed on the face, lips, and even eyelids. But this was not involved due to higher demands on recording procedure and research purposes of the dataset.

The dataset used for this research contains two hours of signing. For motion-capturing were placed 53 passive 14 mm markers on the body of the signing subject. Used marker setup contains 10 markers on each arm and 15 on the torso and head providing the possibility to track any general movement of the whole upper body.

The subject signed about 1000 dictionary signs in Czech Sign Language and each individual sign was recorded separately starting and ending in the relax-pose. This restriction was chosen for more robust separating of single signs and it does not affect the quality of this particular research. Motion capture frame rate was set to 120 fps. This rate was accepted as sufficient because movements with faster changes were not observed. Higher frame rate is not requisite as the amount of data increases significantly. Part of the dataset was manually segmented by two different persons for further evaluation.

4. Automatic feature detection

The purpose of the first developed method is to automatically detect the relax-pose to separate individual signs. The sign segmented this way is surrounded by resting in the relax-pose and there is a characteristic movement of the signing subject when leaving the relax-pose and moving hands to start point and when returning back to the relax-pose from the end point. We developed the estimating method using this feature to determine the time stamp of the beginning and the end of the sign. The segment acquired by using this method was used for further analysis. In this article, we also focused on events in a starting point configuration (hand location). An important characteristic of sign language is the dominance of one the signing subject's hands. There is only one signing subject in used dataset and it is a priori known which of her hands is dominant. But it is also possible to recognize this information automatically simply by measuring the length of the trajectory of each hand. The dominant hand is apparently the one which moves the longer distance than the other.

4.1. Relax-pose detection

As it was mentioned, the first step of data processing was detecting the relax-poses to separate signs in the record. The relax-pose was defined as a position of hands freely hanging in front of the stomach. The beginning and the end of the relax-pose was detected by positioning the dominant hand in the expected area and by the decrease of the speed of this hand (particularly wrist joint) below the threshold. The speed v is measured as a difference of the position of

dominant hand's wrist joint in two following frames:

$$dx(n) = x(n) - x(n+1)$$
$$dy(n) = y(n) - y(n+1)$$
$$dz(n) = z(n) - z(n+1)$$
$$v(n) = \sqrt{dx^2(n) + dy^2(n) + dz^2(n)},$$

where $x(n)$, $y(n)$, $z(n)$ are positions in frame n. Origin is placed on the ground, positive y-axis leads upwards, positive z-axis leads forwards, and x-axis leads on the right, all from the subject's orientation.
Area boundaries for each axis and the speed threshold were determined by supervised learning on the part of the data. An example of manual and automatic segmentation is in Figure 1.

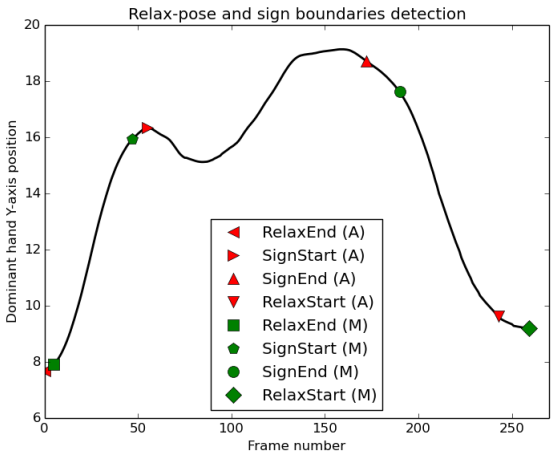

Figure 1: Automatic (A) and manual (M) detections of start and end points for both relax-pose and sign visualized for y-axis position of the dominant hand.

4.2. Sign-beginning detection

There is a specific movement from the end of the relax-pose to the beginning of sign. This movement seems to be more fluent than the movement during the sign. The automatic detection of the sign-beginning is based on measuring acceleration and deceleration of the dominant hand.
All data contains some low-level noise caused by the environment during recording. This noise doesn't affect detections based on position and speed but it causes problems in detection of the features in acceleration. It is necessary to filter acceleration signal before detecting points of interest with lowpass filter. Filter parameters were experimentally chosen corresponding to recording frame rate.
As the dominant hand is leaving the relax-pose acceleration increases. The hand is decelerating when approaching the hand location. Subject starts signing after that movement and, therefore, accelerates his or her hand again, in other words, the point of the second acceleration of the dominant hand is the hand location. Acceleration a is defined:

$$a(n) = v(n+1) - v(n).$$

Values of the speed and the acceleration for the same example as in Section 4.1. is shown in Figure 2. This acceleration-based approach was successful in most cases but there was a phenomenon in some signs which caused to trigger detection too early. It was caused by a sudden change of acceleration during the movement from the relax-pose to the sign location. This problem was solved by adding a maximum speed threshold as a parameter for the sign location detection.

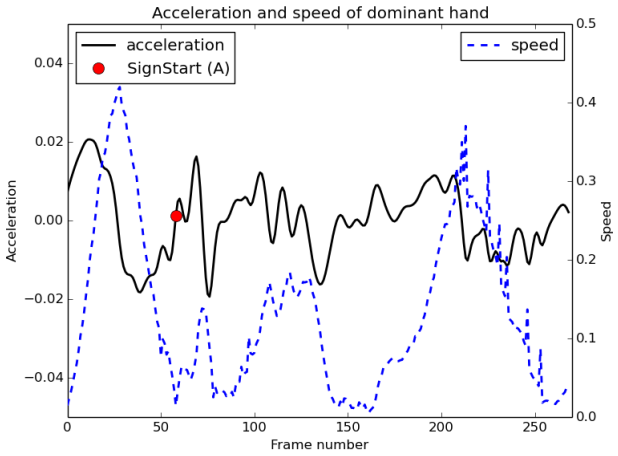

Figure 2: Speed and acceleration values.

4.3. Recognising features

There are some sign types which can be simply recognised e.g. one or two handed sign. Other observable specifications are different types of symmetry, static and dynamic signs, etc.

4.3.1. One handed sign detection

It was already mentioned that the signing subject has his or her dominant hand. Detection of the dominant hand is described in Section 4. The next step in feature recognition is to decide whether the sign is one or two handed. The detection is based on the same principle as the relax-pose detector but in this case it is focused on the secondary hand assuming that the dominant hand is signing.

4.3.2. Sign location detection

Many sign language notification systems describing handshape and location at the beginning of the sign. While the beginning of the sign is detected, it is simple to describe hand location and wrist orientation because the motion capture data already contains this type of information. There is no difference in one or two-handed signs, because the moment of the beginning of the sign is same for both hands. It is simple to acquire data for both hands.

4.3.3. Symmetry detection

Another interesting feature is the symmetry of two-handed signs. There are two types of symmetry. The natural type of symmetry is mirror movement of both hands. Inverse symmetry is when both hands start in mirror symmetry but each hand moves in opposite direction. In our research, it makes no difference whether the handshapes are the same or not, but this information can be added using data from

finger motion capture provided by CyberGloves3. The initial experiment was made by correlating movements of both hands. The correlation alone is not robust enough for detection of symmetry and further research is needed.

5. Results and future work

The results were validated on two sets manually segmented by two persons. Each set contained 20 signs. Both manual segmentations were compared to each other for defining the cross-annotation difference. Only one of the sets (set 2) was used for supervised learning incorporated in detection techniques.

set	r-p end	sign start	sign end	r-p start
set 1	9.2	14.3	9.15	9.95
set 2	9.75	8.2	13.95	8.65

Table 1: Manual segmentations comparison.

Set 1	r-p end	sign start	sign end	r-p start
man 1	7.85	11.7	19.65	25.1
man 2	6.45	14.8	14.7	27.75

Table 2: Automatic segmentation validation on dataset 1.

Set 2	r-p end	sign start	sign end	r-p start
man 1	9.1	11.1	15.0	20.5
man 2	5.95	14.9	18.45	22.0

Table 3: Automatic segmentation validation on dataset 2.

The results are summarized in Tables 1, 2, and 3. The columns correspond to the relax-pose end, sign-beginning, sign-end, and relax-pose beginning. Values in the rows correspond to average absolute frame difference. It can be observed that the difference of two manual segmentations and the difference of automatic and manual segmentations are similar. It should be reminded that the frame rate of the record is 120 fps. This means that 1 frame difference equals 8.33 miliseconds. Human eye is not able to recognize framerate 24 fps which is framerate of video. Standard video frame length equals approximately 5 frames in used motion capture.

The results for recognizing important events such as an end of the relax-pose and sign beginning are very satisfactory because the difference between the automatic and the manual segmentation tends to be slightly lower compared to two manual segmentations. Worse results in the recognising end of the sign seem to be caused by not well-bounded signs at its end. Signing subjects tend to lose his or her hands during the end of the sign fluently. Human segmentation is more or less intuitive for this feature. On the other hand, automatic segmentation reflects more on distinct events in the data.

The relax-pose beginning detection results are not satisfactory. Automatic segmentation triggers when the dominant hand's speed decreases below threshold but manual segmentation tends to trigger earlier. This may be caused by the fact, that the human validator knows that the sign will end soon and he or she does not wait until hands stay still. The question is which segmentation is better and whether this difference means that automatic segmentation is better than human. Anyway, the beginning of the rest pose is the least important event of four evaluated features and does not cause any transferred inaccuracy.

In further work, we will focus on different approaches to sign segmentation as well as on sign location analysis. The next step is fluent sign speech analysis. The long term goal is data-driven sign language synthesis.

6. Acknowledgements

This research was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506 and by the grant of the University of West Bohemia, project No. SGS-2016-039.

7. Bibliographical References

Gibet, S., Marteau, P.-F., and Duarte, K. (2012). Toward a motor theory of sign language perception. In *Proceedings of the 9th International Conference on Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, GW'11, pages 161–172, Berlin, Heidelberg. Springer-Verlag.

Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., and Turki, A. (2015). Interactive editing in French Sign Language dedicated to virtual signers: requirements and challenges. September.

Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., and Seidel, H. P. (2009). Markerless motion capture with unsynchronized moving cameras. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 224–231, June.

Heloir, A., Gibet, S., Multon, F., and Courty, N. (2006). Captured motion data processing for real time synthesis of sign language. In *Proceedings of the 6th International Conference on Gesture in Human-Computer Interaction and Simulation*, GW'05, pages 168–171, Berlin, Heidelberg. Springer-Verlag.

Huenerfauth, M., Lu, P., and Kacorri, H. (2015). Synthesizing and evaluating animations of american sign language verbs modeled from motion-capture data. In *Proceedings of SLPAT 2015*, pages 22–28, Dresden, Germany, September. Association for Computational Linguistics.

Krňoul, Z., Kanis, J., Železný, M., and Müller, L. (2008). Czech text-to-sign speech synthesizer. *Machine Learning for Multimodal Interaction, Series Lecture Notes in Computer Science*, 4892:180–191.

Lu, P. and Huenerfauth, M. (2012). Learning a vector-based model of american sign language inflecting verbs from motion-capture data. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, SLPAT '12, pages 66–74, Stroudsburg, PA, USA. Association for Computational Linguistics.