

Collection and Preprocessing of Czech Sign Language Corpus for Sign Language Recognition

Pavel Campr, Marek Hruží, Jana Trojanová

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia in Pilsen
Univerzitní 8, 306 14 Pilsen, Czech Republic
campr@kky.zcu.cz, mhruz@kky.zcu.cz, trojana@kky.zcu.cz

Abstract

This paper discusses the design, recording and preprocessing of a Czech sign language corpus. The corpus is intended for training and testing of sign language recognition (SLR) systems. The UWB-07-SLR-P corpus contains video data of 4 signers recorded from 3 different perspectives. Two of the perspectives contain whole body and provide 3D motion data, the third one is focused on signer's face and provide data for face expression and lip feature extraction. Each signer performed 378 signs with 5 repetitions. The corpus consists of several types of signs: numbers (35 signs), one and two-handed finger alphabet (64), town names (35) and other signs (244). Each sign is stored in a separate AVI file. In total the corpus consists of 21853 video files in total length of 11.1 hours. Additionally each sign is preprocessed and basic features such as 3D hand and head trajectories are available. The corpus is mainly focused on feature extraction and isolated SLR rather than continuous SLR experiments.

1. Introduction

The corpus consists of more 11.1 hours of processed video files which were recorded in laboratory conditions using static illumination. The whole corpus is annotated and pre-processed and is ready for use in SLR experiments. It is composed of 378 selected signs from Czech sign language. Nearly every sign was repeated 5 times by each signer. Altogether the corpus contains 21853 video files where each file contains one isolated sign captured from one perspective.

The purpose of the corpus is to provide data for evaluation of visual parameterizations and sign language recognition techniques. The corpus is preprocessed and each video file is supplemented with a data file which contains information about performed sign, signer, scene (camera position, calibration matrices) and preprocessed data (regions of interests, hands and head trajectories in 3D space).

The presented corpus is collected, preprocessed and is ready to use for subsequent experiments on sign language recognition.

2. Related Work

Several institutions concerned with sign language recorded their own corpora, but only few of them are suitable for automatic sign language recognition due to poor image quality and other conditions. Such corpora usually focus on linguistic analysis and educational purposes.

The European Cultural Heritage Online organization (ECHO)¹ published a corpora for British sign language (Woll et al., 2004), Swedish sign language (Bergman and Mesch, 2004) and the sign language of the Netherlands (Crasborn et al., 2004). All of the corpora include several stories signed by a single signer.

The American Sign Language Linguistic Research group at Boston University² published a corpus in American sign language. The corpus was recorded from three different

perspectives. The Human Language Technology and Pattern Recognition group at Aachen University processed this corpus and published new corpora for automatic sign language recognition (Dreuw et al., 2007).

TV broadcast news for the hearing impaired are another source of sign language recordings. Aachen University published a German Sign Language Corpus of the Domain Weather Report (Bungeroth et al., 2006). Automatical creation of similar corpus is described in (Aran et al., 2007) where automatical speech recognition (ASR) is used to find borders of performed signs in Turkish news.

3. Sign Language in View of Human-Computer Interaction

Sign language is the main form of communication for deaf people. Inspired by speech recognition, where the rate of progress has been enormous in the past decade, new ways of communication between deaf people and computers or hearing people are being developed. The main task of automatic sign language recognition is to recognize one isolated or multiple signs performed by a signer.

The current state-of-the-art in SLR is still far behind that of speech recognition. There are many reasons for this disparity: research in SLR started later, usage of advanced input devices, higher computational requirements, sign language uses simultaneous events for expressing terms, unavailability of large number of training data etc. Despite of all this there are many successful achievements. Ong and Ranganath (2005) present a survey of SLR and comparison of many gesture and sign language recognition systems.

Our effort is focused on the creation of assistive HCI systems for hearing impaired. The first goal is to create a Czech sign language to Czech language dictionary and Czech sign language tutoring tool similar to the one proposed by Aran et al. (2006). The second goal is to create an information kiosk for deaf people providing timetable information on railway and bus stations (Železný et al., 2007) where both sign language recognition and sign language

¹<http://www.let.kun.nl/sign-lang/echo/>

²<http://www.bu.edu/asllrp/>



Figure 1: Sample frames from the corpus, 4 different signers, row 1: front perspective, row 2: face and top perspective

synthesis (Kanis et al., 2008) will be used. The integration of SL recognition and synthesis in one system avoids using text modality which can cause problems for many hearing impaired who have difficulties with reading.

4. Corpus Specification

Primary purpose of the UWB-07-SLR-P (UWB stands for University of West Bohemia, 07 for year of recording, SLR for sign language recognition, P for professional speakers) corpus is to have experimental data for verification of SLR algorithms. Recording conditions were set for easy feature retrieval. We retain constant illumination, the signer does not change her position in the scene and there is a large contrast differences between the objects of interest and the background.

The corpus was recorded by 4 different signers from which two are deaf. Each signer performed 378 signs with mostly 5 repetitions. The corpus consists of several types of signs:

- numbers (35 signs)
- day and month names (19 signs)
- finger alphabet, both one and two-handed (64 signs)
- town names (35 signs)
- selected signs (225 signs)

Each sign is stored in a separate AVI file. For 378 signs, 4 speakers, 5 repetitions and 3 perspectives the corpus consists of 21853 avi files in total length of 11.1 hours. Some of them were removed due to errors in signing.

We selected 225 signs from the *Train timetable dialogue corpus* (Kanis et al., 2006). These 225 signs are the most frequent words which are used in spoken train information service dialogues. With those signs we are able to create a dialogue system which will be used in the train station information kiosk for the deaf.

The camera setup and recording conditions are similar to our previous database UWB-06-SLR-A (Campr et al., 2007), but we have extended the number of signs and the signers were deaf persons or persons who are familiar with sign language in everyday use.

4.1. Data File

All recordings are supplemented with a description of each sign and preprocessed data. The description includes sign name and signer identification. The preprocessed data consist of calibration data (fundamental and projection matrices) and segmented regions of interest such as head and hands. The data is stored in a single file which is attached to each video file.

5. Corpus Recording

The corpus was recorded with three cameras (see Fig. 2). Camera 1 is situated in front of the actor. Camera 2 is about a meter above camera 1 and looks downwards. In both camera 1 and camera 2 the actor's body is seen from the head to the knees. The actor puts her arms near the lower body before performing each sign. This state corresponds to silence in spoken language. Camera 3 captured the face in high resolution in the same setup as described in (Císař et al., 2005). It is intended for additional feature extraction from the face, e.g. face expression extraction or lip-reading (Císař et al., 2006).

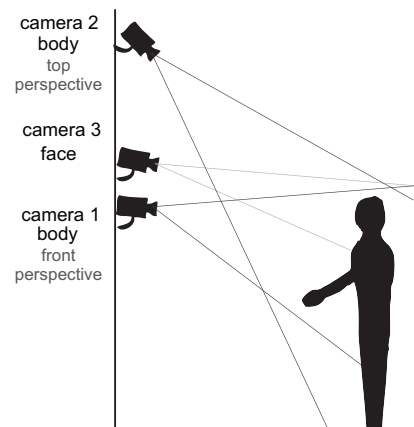


Figure 2: Camera arrangement. Two cameras (top and front perspective) capture the actor's body from knees to the head, one camera captures only the face.

Cameras recorded the scene in the resolution of 720x576 pixels, 25 frames per second. The shutter speed was set to 1/500 second so that moving hands are not blurred even at high velocities. The signer was dressed in black or dark clothing with visible hands and head. There is a black sheet in the background so that we eliminate the undesirable effect of background noise in the image. The scene is well illuminated to minimize the presence of shadows cast on the actor. There are two lights in front of the scene and another two lights, each from one side of the scene.

A clapperboard was used at the beginning of the recording. The time of a clap can be measured with precision of one frame. This information is sufficient for camera synchronization. The maximum time shift between two synchronized videos is 10 ms (one half of duration of one frame, one frame lasts 20 ms after deinterlacing). If we assume maximum speed of hand movement 1 m/s then the maximum difference is 1 cm between two body parts observed from two cameras. This error is small enough for our purpose.

At last we recorded a box with chessboard tiles on every side. This data is used for calibration. The box is rotated towards the camera so that each side of the box forms an angle of 45 degrees with the camera plane. Because this condition is not met precisely, the actual angle must be estimated.

Raw visual data were stored on a camcorder tape and acquired later using the IEEE1394 (Firewire) interface. We preprocessed recorded data by disabling audio channels, deinterlacing and compressing using Xvid codec. Thus we reduced required space for storing data from 230 GB to 19 GB preserving high quality of the recordings.

6. Data Preprocessing

6.1. Annotation

Camera recordings were annotated with ELAN annotation tool. The annotator marked every sign in the recordings. Afterwards each marked part was extracted into a single avi file. For each single avi file the information about signer, sign group, calibration and defects of recording (e.g. wrong face expression of a signer etc.) is available.

6.2. Calibration

Calibration data were acquired from frames containing box with a chessboard on every side. We find the corners of chessboard tiles in every image. Thus we get several points which are passed to the 8-point algorithm. The output of the algorithm is a fundamental matrix. The fundamental matrix is essential for 3D representation of the scene. It is the algebraical representation of epipolar geometry. Using it we can find corresponding pixels in different perspectives of the same scene.

Knowing the position of the box in 3D space we are able to create a projection matrix. We get two projection matrices, one matrix for each camera. These matrices are used for representing two 2D corresponding points as one 3D point. By choosing the right metric the output can be visualised for comparison with the observed trajectory of the sign (see Fig. 4). In our case we chose the metric to get the output in centimeters with an orthogonal base.

6.3. Feature Extraction

We use a set of image processing algorithms to separate the objects of interest from the rest of the image (see Fig. 3). In the first image we find hands and head via the skin color model. This is the initial state for our tracking algorithm. The algorithm itself consists of several steps:

- detection of the object in the actual frame, using information about it's predicted position
- prediction of the position of the object in the next frame
- assigning this object to a real world object

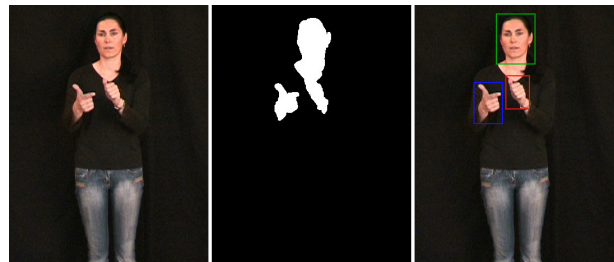


Figure 3: (a) Source frame from the front camera, (b) segmented image, (c) hands and head tracking.

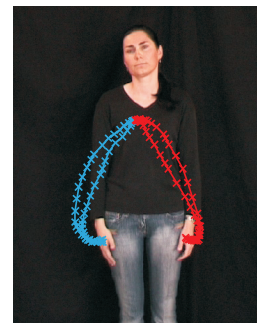


Figure 4: Example of hand tracking for one sign from the front camera

We obtain three matrices containing the position of the right hand, the left hand and the head. Every matrix has four columns. The columns represent the horizontal and vertical position of the object. There are two pairs of these columns, each pair for one perspective.

Using the projection matrices and the output matrices we compute the 3D trajectory of the sign. Because of the orthogonality of the base we can easily visualize the output (see Fig. 5).

The position of hands and head (manual sign components) and their derivations such as speed and acceleration are important features which are used in sign language recognition. Many signs have such a unique trajectory that these features are sufficient for successful classification. The rest of the signs have the same or very similar trajectories. In these cases it is necessary to use another features such as hand shape and lip shape (non-manual sign components).

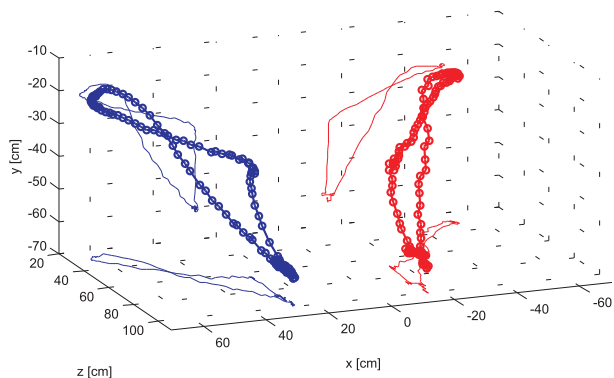


Figure 5: Trajectory tracking of the left (red curve) and right (blue curve) hand in 3D space. The coordinate origin is located in the mean position of the head.

We are preparing new experiments where all of these features will be used in recognition process. Afterwards, features of face expressions will be added to include all of the most important features for sign language recognition: hands and head position, hand shapes, articulation and face expression.

Extracted features have to be evaluated. For the trajectory features we have developed a semiautomatic annotation tool for tracking accuracy evaluation. The results for our tracking method will be available soon.

7. Conclusion

The UWB-07-SLR-P Czech sign language corpus offers possibilities for testing various feature extraction methods and recognition techniques. It was recorded by using three cameras to provide 3D information about the head and hands positions. By maintaining the parameters of the framework at the same level we are able to compare the results of different sign language recognition approaches. This corpus is being used for design and evaluation of the sign language recognition systems.

8. Acknowledgement

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416 and by the Ministry of Education of the Czech Republic, project No. ME08106.

9. References

Oya Aran, Ismail Ari, Alexandre Benoit, Ana Huerta Carrillo, François-Xavier Fanard, Pavel Campr, Lale Akarun, Alice Caplier, Michele Rombaut, and Bulent Sankur. 2006. Sign Language Tutoring Tool. *Proceedings of eINTERFACE 2006, Summer Workshop on Multimodal Interfaces, Dubrovnik, Croatia*.

Oya Aran, Ismail Ari, Pavel Campr, Erinc Dikici, Marek Hruz, Deniz Kahramaner, Siddika Parlak, Lale Akarun, and Murat Saraclar. 2007. Speech and Sliding Text Aided Sign Retrieval from Hearing Impaired Sign News Videos. *eINTERFACE 2007 Proceedings*.

B. Bergman and J. Mesch. 2004. ECHO Data Set for Swedish Sign Language (SSL). Department of Linguistics, University of Stockholm. <http://www.let.ru.nl/sign-lang/echo>.

J. Bungeroth, D. Stein, P. Dreuw, M. Zahedi, and H. Ney. 2006. A German Sign Language Corpus of the Domain Weather Report. In *Fifth International Conference on Language Resources and Evaluation*, pages 2000–2003, Genoa, Italy, May.

P. Campr, M. Hruz, and M. Železný. 2007. Design and Recording of Czech Sign Language Corpus for Automatic Sign Language Recognition. *Proceedings of the Interspeech 2007, Antwerp, Belgium*.

P. Císař, M. Železný, Z. Krňoul, J. Kanis, J. Zelinka, and L. Müller. 2005. Design and Recording of Czech Speech Corpus for Audio-visual Continuous Speech Recognition. In *Proceedings of the Auditory-Visual Speech Processing International Conference AVSP2005*, pages 1–4, Vancouver Island.

P. Císař, J. Zelinka, M. Železný, A. Karpov, and Ronzhin A. 2006. Audio-visual Speech Recognition for Slavonic Languages (Czech and Russian). In *Proceedings of the 11th international conference Speech and computer SPECOM'2006*, pages 493–498, St.Petersburg. Anatolya Publishers.

O. Crasborn, E. van der Kooij, A. Nonhebel, and W. Emmerik. 2004. ECHO Data Set for Sign Language of the Netherlands (NGT), Department of Linguistics, Radboud University Nijmegen, <http://www.let.ru.nl/sign-lang/echo>.

P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. 2007. Speech Recognition Techniques for a Sign Language Recognition System. In *Interspeech 2007*, pages 2513–2516, Antwerp, Belgium, August.

Jakub Kanis, Jirí Zahradil, Filip Jurčiček, and Luděk Müller. 2006. Czech-Sign Speech Corpus for Semantic Based Machine Translation. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, *Proceedings of TSD 2006*, volume 4188 of *Lecture Notes in Computer Science*, pages 613–620. Springer.

Jakub Kanis, Pavel Campr, Marek Hruz, Zdeněk Krňoul, and Miloš Železný. 2008. Interactive HamNoSys Notation Editor for Sign Language Synthesis. In *LREC 2008. Workshop proceedings: Construction and Exploitation of Sign Language Corpora, ELRA*.

S. C. W. Ong and S. Ranganath. 2005. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):873–891.

M. Železný, P. Campr, Z. Krňoul, and M. Hruz. 2007. Design of a Multi-Modal Information Kiosk for Aurally Handicapped People. *Proceedings of SPECOM 2007, Moscow, Russia*.

B. Woll, R. Sutton-Spence, and D. Waters. 2004. ECHO Data Set for British Sign Language (BSL). Department of Language and Communication Science, City University (London). <http://www.let.ru.nl/sign-lang/echo>.