Design and Recording of Czech Sign Language Corpus for Automatic Sign Language Recognition

Pavel Campr, Marek Hrúz, Miloš Železný

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

Abstract

We describe the design, recording and content of a Czech Sign Language database in this paper. The database is intended for training and testing of sign language recognition (SLR) systems. The UWB-06-SLR-A database contains video data of 15 signers recorded from 3 different views, two of them capture whole body and provide 3D motion data, and third one is focused on signer's face and provide data for face expression feature extraction and for lipreading.

The corpus consists of nearly 5 hours of processed and annotated video files which were recorded in laboratory conditions using static illumination. The whole corpus is annotated and pre-processed to be ready to use in SLR experiments. It is composed of 25 selected signs from Czech Sign Language. Each signer performed all of these signs with 5 repetitions. Altogether the database contains more than 5500 video files where each file contains one isolated sign.

The purpose of the corpus is to provide data for evaluation of visual parameterizations and sign language recognition techniques. The corpus is pre-processed and each video file is supplemented with a XML data file. It provides information about performed sign (name of sign, type of sign), signer (identification, left or right-handed person), scene (camera position, calibration matrices) and pre-processed data (regions of interests, hands and head trajectories in 3D space).

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

Index Terms: Sign language recognition, gesture recognition, sign language corpus, speech corpus, Czech

1. Introduction

Sign language is the main form of communication for deaf people, just as speech is for hearing 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. Main task of automatic sign language recognition is to recognize sign performed by a signer.

In speech recognition it is obvious that microphone is used as an input device. In sign language recognition more kinds of input devices can be used. Mechanical devices, which measure location of various parts of body, such as data gloves and haptic devices, have advantage in accuracy of measurements. But the disadvantage is that the signer is forced to wear a cumbersome device. This problem can be solved by using camera or multiple cameras as input devices and using computer vision algorithms for feature extraction from acquired images.

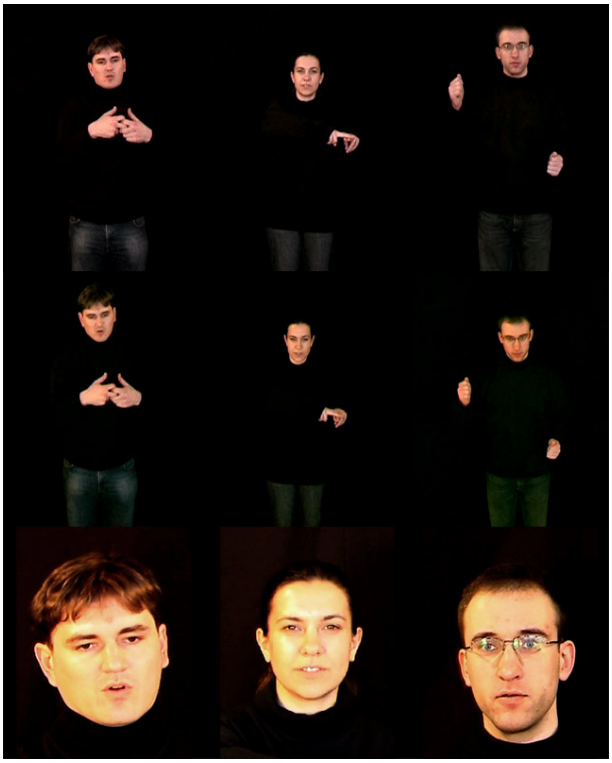

Figure 1: *Sample frames from the corpus from 3 different signers, row 1: front view, row 2: top view, row 3: face.*

Current state of the art in SLR is still far behind 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 [1] present survey of SLR and comparison of many gesture and sign language recognition systems. Chen [2] achieved a 92% success rate for a 5113 Chinese signs vocabulary using gloves and magnetic trackers as input device and whole-word HMMs for recognition of isolated signs. Zieren and Kraiss [3] reported 99.3% success rate for visual recognition of 232 isolated person-dependent signs in controlled environment and 87.8% average performance for six signers using a reduced vocabulary of 18 signs.

Our effort is focused on creation of information kiosk for

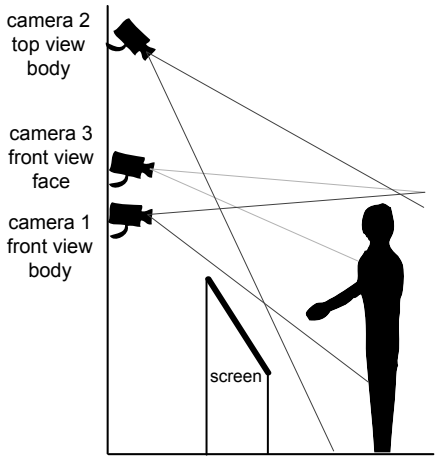

Figure 2: *Design of information kiosk for deaf people, proposed camera arrangement.*

deaf people providing timetable information on railway and bus stations. The kiosk will consist of one touchscreen displaying signing avatar [4] and three cameras providing input for SLR system. Topic of dialogue between signer and kiosk will be narrowed to travel information sentences.

Czech Sign Language is main communication language among deaf people in Czech Republic. Another form is Signed Czech which is used for communication between deaf person and hearing person. It uses grammatical resources of the spoken Czech language and is simultaneously loudly or quietly articulated during signing. Signed Czech will be main communication language used both for input and output in the information kiosk. Because signs in corpus are the same or very similar in both languages, we use the term Czech Sign Language in this paper.

Two of three cameras will be used for hands and head tracking in 3D space. Third camera will be aimed at the signer's face and used as a source for face feature extraction, where lipreading [6] is essential to distinguish signs differing only in articulation.

We recorded the corpus in laboratory conditions, containing 25 signs from 15 different signers. Purpose of this corpus is to verify that creation of information kiosk is feasible. We performed tracking experiments on this corpus which showed that our effort is suitable for next experiments with isolated sign recognition.

2. Database specification

Primary purpose of the UWB-06-SLR-A (UWB stands for University of West Bohemia, 06 for year of recording, SLR for sign language recognition, A for first version) corpus is to have experimental data for verification of SLR algorithms, that will be used in recognition part of the information kiosk. Recording conditions were set for easy feature retrieval. We retain constant illumination, a signer does not change his position in the scene, and there are large contrast differences between the objects of interest and background.

The corpus consists of 15 signer recordings. Each signer performed 25 signs with 5 repetitions. The database consists of several types of signs. Generally gestures can be divided into:

- one hand movement
- two hands movement
- movement with occlusion(s) between objects
- with specific hand shape and finger movement
- types that differ only in head movement and/or face expression

The average duration of one recording session was 14 minutes, which makes together more than 210 minutes of raw data for one camera. Since we used three cameras there are over 10 hours of raw material. After we cut this raw material into separated signs there are nearly 5 hours of processed recordings.

The corpus is not divided into training and testing section. We intend to use a larger part of the data for training purposes and the rest for testing.

3. Database recording

The database 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. That way we get an overview of the scene. In both camera 1 and camera 2 the actor's body is seen from head to knees. Before each sign the actor puts his arms near the lower body. This state corresponds to silence in spoken language and thus separates the individual signs. That way we are able to detect hands in the first frame of video, assuming it begins with silence. Camera 3 is used for capturing of the face in high resolution. It is intended for additional feature extraction. This camera is set as described in [5].

Cameras 1 and 2 recorded the scene in the resolution of 720x576 pixels, 25 frames per second. The shutter speed was set to 1/500 second. This way moving hand isn't blurred even at high velocity. The output video is interlaced. Also, manual focus was used to maintain constant parameters of the cameras. 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.

Recording of one speaker was done during one session. The signer was asked not to change position during recording. An assistant was signing the signs in advance to help the signer memorize the correct order of the signs.

A clapperboard was used at the beginning of recording. The time of 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 maximal difference is 1 cm between two body parts viewed from different cameras. This error is low enough for our purpose.

At last we recorded a box with chessboard tiles on each side. These data are 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. However this condition is not met precisely, that's why the actual angle must be estimated. To synchronize the recordings from different cameras we recorded a clapperboard between a series of signs.

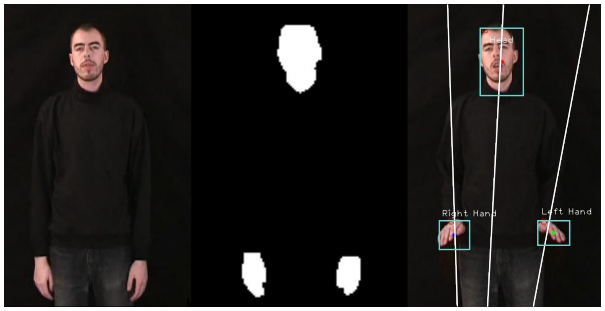

Figure 3: *(a) Source frame from front camera, (b) segmented image, (c) hands and head tracking.*

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 85 GB to 8 GB preserving good quality of recordings.

4. Data Processing

4.1. Calibration

First we make use of the part of the database that contains the calibration data. These data are acquired from frames containing box with a chessboard on each side. Next we must extract two images, each image from one camera capture, referring to the same time moment. In each image we find the corners of chessboard tiles (this is done automatically). Thus we get several points, which are then passed to the 8-point algorithm. The output of the algorithm is a fundamental matrix. The Fundamental matrix is essential for 3D representation of a scene. It is the algebraical representation of epipolar geometry. Using it we can find corresponding pixels in different views 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. 5). In our case we chose the metric to get the output in centimeters with an orthogonal base.

4.2. Feature extraction

The next step is to process each video file so that we get corresponding points representing position of head and hands in each view. We use a set of digital image processing tools 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 will be the initial state for our tracking algorithm. The algorithm itself consists of several steps:

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

Some signs can be observed with hands or head occlusions. First of all, thanks to the position prediction, we can predict

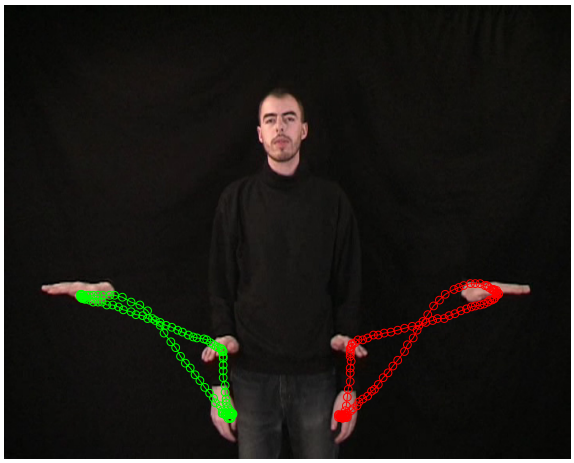

Figure 4: *Tracked trajectory of left and right hand projected in the front camera plane.*

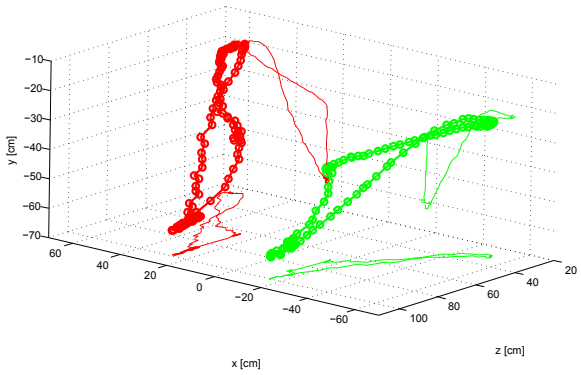

Figure 5: *Tracked trajectory of left and right hand in 3D space.*

these occlusions and then we can use the information from the other camera to resolve this occlusion. In the end we obtain three matrices containing the position of right hand, left hand, and head throughout the whole capture of the sign. Each 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 view.

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.

Position of hands and head and their derivations such as speed and acceleration are the most important features used in recognition of sign. Many signs have such a unique trajectory that these features are sufficient for successful classification. The rest of signs have the same or very similar trajectories. In these cases it is necessary to use another features for sign classification. We have performed experiments with hand shape feature extraction and liptracking. 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 cover all of the most important features for sign language recognition: hands and head position, hand shapes, articulation and face expression.

4.3. Data file

Recordings in our corpus are supplemented with description of each sign and preprocessing data. The description includes sign name, sign type (one or two handed), signer identification, and orientation of signer (right or left-handed person). The preprocessing data consist of static and dynamic data. Static data consist of calibration data (fundamental and projection matrices) and segmented regions, which correspond to skin color. Then each video file is processed as a whole and segmented regions from previous step are identified as head or hand depending on blob position and dynamics of movements. The data are stored in one XML file which is attached to each of video files.

5. Conclusion

The UWB-06-SLR-A Czech Sign Language database offers possibilities for testing of various feature extraction methods and recognition techniques. It was recorded using three cameras to provide 3D information about the head and hands position. By maintainig the parameters of the framework at the same level we are able to compare the results of different sign languages recognition approaches. This database is being used for design and evaluation of sign language recognition system, which will be used in information kiosk for deaf people providing traffic information at railway and bus stations.

6. Acknowledgement

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

7. References

[1] S.C.W. Ong and S. Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 873-891, June 2005.

[2] Chen, Y., et al., "CSLDS: Chinese Sign Language Dialog System", in Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and G.A. ten Holt Automatic Recognition of Dutch Sign Language 8 Gestures AMFG'03. 2003, IEEE Computer Society: Nice, France. p. 236-238.

[3] J. Zieren and K.-F. Kraiss, "Robust Person-Independent Visual Sign Language Recognition", in Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis, volume Lecture Notes in Computer Science, 2005.

[4] Krňoul, Z.; Kanis, J.; Železný, M.; Müller, L.; Císař, P., "3D symbol base translation and synthesis of Czech sign speech", in Proceedings of the 11th international conference Speech and computer SPECOM'2006 . St.Petersburg: Anatolya Publisher, 2006. s. 530-535. ISBN 5-7452-0074-X.

[5] Císař, P.; Železný, M.; Krňoul, Z.; Kanis, J.; Zelinka, J.; Müller, L., "Design and recording of Czech speech corpus for audio-visual continuous speech recognition", in Proceedings of the Auditory-Visual Speech Processing International Conference 2005. Vancouver Island : AVSP2005, 2005. s. 1-4. ISBN 1 876346 53 1.

[6] Císař, P.; Železný, M.; Krňoul, Z., "3D lip-tracking for audio-visual speech recognition in real applications", in Prodeedings of the Interspeech 2004 - ICSLP. Jeju : Sunjin Printing Co., 2004. s. 2521-2524. ISSN 1225-441x.