# A Kinect Corpus of Swedish Sign Language Signs

Kalin Stefanov and Jonas Beskow

KTH Speech, Music and Hearing
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden
{kalins,beskow}@kth.se

**Abstract.** We describe a corpus of Swedish sign language signs, recorded for the purpose of an educational "signing game". The primary target group of the game is children with communicative disabilities, and the goal is to offer a playful and interactive way of learning and practicing sign language signs to these children, as well as to their friends and family. As a first step, a dataset consisting of 51 signs has been recorded for a total of 10 adult signers. The signers performed all of the signs five times and were captured with an RGB-D (Microsoft Kinect) sensor, via a purpose-built recording application.

**Keywords:** multimodal, corpus, Kinect, RGB-D, sign language, Swedish, Tivoli

## 1 Introduction

Sign Language and different forms of sign based communication is important to large groups in society. In addition to members of the deaf community, that often have sign language as their first language, there is a large group of people who use verbal communication but rely on signing as a complement. A child born with hearing impairment or some form of communication disability such as developmental disorder, language disorder, cerebral palsy or autism, frequently have the need for this type of communication, in Sweden known as TSS (Signs as Support), and is a form of augmented and alternative communication (AAC).

Sign-based AAC systems borrow individual signs from sign language (e.g., TSS borrows from Swedish sing language, SSL). These signs support and enhance the verbal communication. As such, these communication support schemes do away with the grammatical constructs in sign language and keep only parts of the vocabulary. One important difference between SSL and TSS is that the latter is poorly formalized and described, and the extent and manner in which it is taught differ widely between different parts of the country. While many deaf children have sign language as their first language and are able to pick it up in a natural way from the environment, children that need signs for other reasons do not have the same rights and opportunities to be introduced to signs and signing.

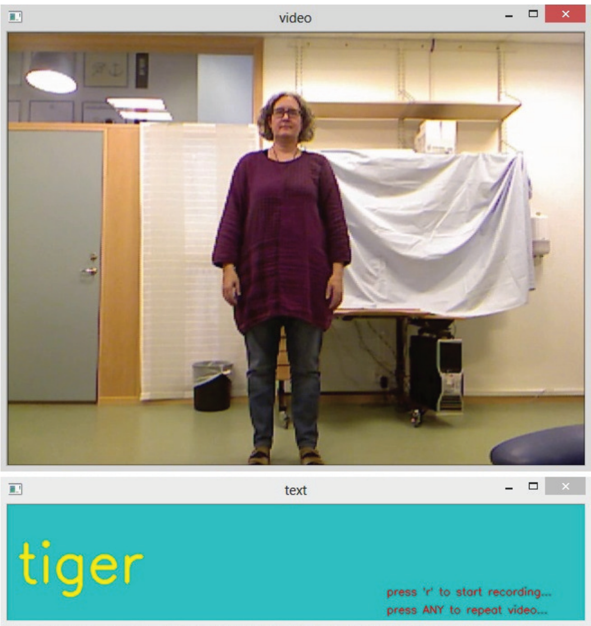
The dataset described in this manuscript is a product of the Swedish TIVOLI project, that aims at creating a learning environment where children can pick up signs in a game setting; an on-screen avatar presents the signs and gives the child certain tasks to accomplish, and in doing so the child gets to practice the signs. The presented dataset is an essential part of the development of an isolated sign recognizer for use in the Tivoli game. Automatic sign language recognition based on color only information is a difficult task, which motivates the use of RGB-D (Microsoft Kinect) sensor in the project - the capability of the sensor to produce simultaneously color, depth and skeleton data is exploited in order to build a robust automatic sign language recognizer.

Previous Swedish sign language resources include the The Swedish Sign Language Dictionary with approx. 8000 video recorded signs [2, 3] and [1] which is an online dictionary of Swedish sign languages videos. The main reason for recording our own corpus insted of using one of the existing resources is that we wanted RGB-D (Kinect) recordings rather than plain video. In addition, in order to capture enough variability for speaker dependent as well as speaker independent recognition experiments, we needed each sign to be repeated many times by different signers.

## 2 Data Collection

The input device for the data collection is the Microsoft Kinect sensor. For the purpose of data collection we created a recording tool that prompts the participant to perform a certain sign. The tool has two main parts - a window for playback of video that demonstrates the particular sign and a window for text where the name/meaning of the sign is displayed together with some information for the control of the tool. Screenshot of the tool is shown in Figure 1. The current implementation requires second person to control the recording flow with the means of pushing keyboard buttons. The logic of the tool offers control over the start and the end of recording, replying videos, saving data to disk, and skipping sings.

The process of recording one sign follows a certain path. First the meaning of the sign is displayed and the video is played. At this point the participant has an option of replaying the video as many times as he/she wants. Once the participant feels confident enough the recording of the sign is started. All participants were instructed to start at rest position (both hands relaxed), then, perform the sign and go back to rest position. At this point the recording of the sign is stopped. The participant can choose whether to save the recording and move forward to the next sign or record the sign again. The participant can stay in this state until he/she is satisfied by the sign performance. The upside of this set-up is that the signs are automatically annotated at the end of the recording session.



tiger

press 'r' to start recording...
press ANY to repeat video...

**Fig. 1.** Screenshot of the recording application

## 3 Corpus

To date, ten participants have been recorded performing a set of game related signs, where, none of the participants had prior experience in Swedish sign language. They were instructed on how to perform the signs by the means of video clips played in the data recording application. The goal of the participant then was to mimic the played sign (the videos capture signing of expert signer).

The size of the recorded vocabulary for the Tivoli game is 51 signs. The vocabulary is composed of four subsets - objects, colors, animals, and attributes. Each of the participants performed all 51 signs in one session, and after a short brake, new session was started. In this way 5 instances of each sign were recorded for each of the ten participants. Table 1 summarizes the used vocabulary.

**Table 1.** Recorded vocabulary

| **Objects (17)** | **Colors (10)** | **Animals (13)** | **Attributes (11)** |
|---|---|---|---|
| *car, tractor, boat,<br>book, ball, doll, frame,<br>computer, flashlight,<br>guitar, necklace, hat,<br>watch, key, patch,<br>drums, scissors* | *blue, brown,<br>green, yellow,<br>purple, pink,<br>red, orange,<br>black, white* | *moose, monkey,<br>dolphin, elephant,<br>horse, rabbit,<br>cat, crocodile,<br>mouse, tiger,<br>bear, camel, lion* | *angry, happy,<br>striped, small,<br>sad, checkered,<br>narrow, large,<br>thick, tired,<br>dotted* |

### 3.1 Data Streams

During recording, color, depth and skeleton data captured with the Kinect sensor were saved to disk. Furthermore, the color and depth streams were software synchronized and the data can be used for different purpose later, e.g. Microsoft Face Tracking SDK can be run on that data.

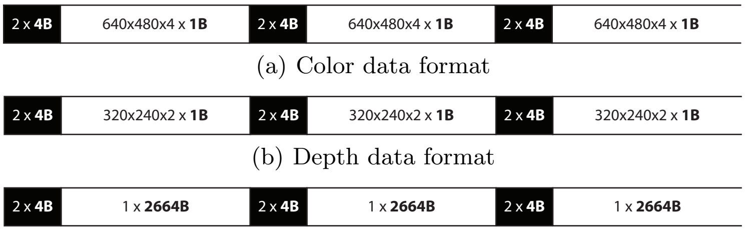
Figure 3 (at the end of the text) illustrates the superimposed trajectories over time of the right wrists of all ten participants (each plot represents an unique participant) while performing the sign for ‘dolphin’ five times. The trajectories are drawn by taking the Y-coordinate of the wrist joint generated by the Microsoft skeleton tracker over time. The horizontal axis in all figures is time (frames) and the vertical axis is the position in Y with respect to the origin of the Kinect (meters). The figure suggests that the chosen recording procedure paired with untrained signers already produces considerable variability within signers, in both space and time. Since we observed considerable variation in signing across different children, and we cannot expect that the kids will be experts in signing, we believe that the data produced in such way fairly reflects the difficult conditions under which the recognition system should work. The sign for ‘dolphin’ itself can be described as slow movement of the hand sideways (in horizontal direction) while following up-down undulating pattern in vertical direction.

### 3.2 Data Format

The color data is encoded in 32-bit, linear X8R8G8B8-formatted color bitmaps, in sRGB color space. The resolution is 640x480. Figure 2(a) illustrates how the color data is saved to disk (first three recorded frames). The color data for a certain sign is a continuous array of bytes where each frame starts with timestamp and is followed by the bytes of the color data. The depth data merges two separate types of data - depth data, in millimeters and player segmentation data (connected regions in the depth stream that are regarded as parts of the same object). The resolution is 320x240. Figure 2(b) illustrates how the depth data is saved to disk (first three recorded frames). The skeleton data contains the raw data streamed out of the sensor during recording. It also provides access to the floor clipping plane and the number of skeletons tracked. Figure 2(c) illustrates how the skeleton data is saved to disk (first three recorded frames).

### 3.3 Recognition Experiments

For the purpose of isolated sign recognition, we trained 8 state Hidden Markov Model for each of the 51 signs with simple set of spatial features generated from the skeleton data. The accuracy of the recognizer was tested for two cases - signer dependent and signer independent. The final results of both tests are based on leave-one-out cross-validation procedure. Signer dependent recognition rate was 96.5% for the most consistent signer, and 89.7% on the average, where the models were trained on 4 instances of each sign for each participant and



(a) Color data format

(b) Depth data format

(c) Skeleton data format

**Fig. 2.** Data formats

tested on the fifth instance of the sign performed by that participant. Signer independent recognition rates varied between 44% and 73%, with an average of 63.1%, where the models were trained on 45 instances of each sign (nine different signers) and tested on the 5 instances of each sign performed by the tenth participant.

## 4 Future Work

We plan to extend the dataset with seven more adult participants and recruit children to participate in recordings. We are adding new logic to the recording application so the person that controls the application will become redundant, and the participant can choose all discussed modes with the means of gesturing. The 'old' logic will be kept in case future participants find the gesturing control difficult.

## 5 Acknowledgments

The project Tivoli (Sign learning through games and playful interaction) is supported by the Swedish Post and Telecom Authority (PTS).

## References

1. Spreadthesign. `http://www.spreadthesign.com`, 2013. [Online; accessed 09-June-2013].
2. Johanna Mesch and Lars Wallin. From meaning to signs and back:lexicography and the swedish sign language corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon.*, pages 123–126, 2012.
3. Johanna Mesch, Lars Wallin, and Thomas Björkstrand. Sign language resources in sweden: Dictionary and corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon.*, pages 127–130, 2012.

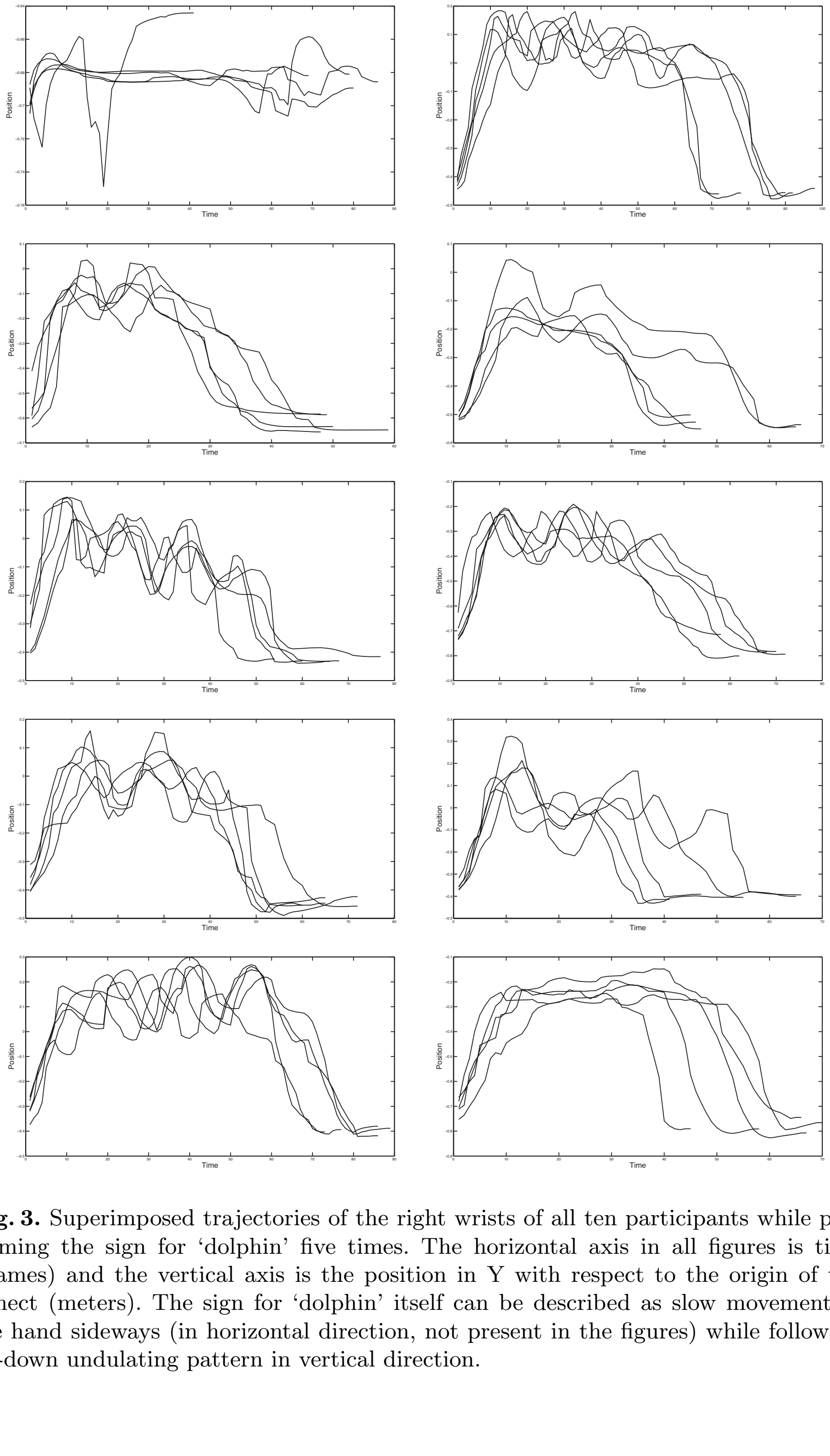


**Fig. 3.** Superimposed trajectories of the right wrists of all ten participants while performing the sign for ‘dolphin’ five times. The horizontal axis in all figures is time (frames) and the vertical axis is the position in Y with respect to the origin of the Kinect (meters). The sign for ‘dolphin’ itself can be described as slow movement of the hand sideways (in horizontal direction, not present in the figures) while following up-down undulating pattern in vertical direction.