# How to Use Depth Sensors in Sign Language Corpus Recordings

**Rekha Jayaprakash, Thomas Hanke**
Institute of German Sign Language and Communication of the Deaf, University of Hamburg
{rekha.jayaprakash, thomas.hanke}@sign-lang.uni-hamburg.de

### Abstract

We describe the experimental setup of positioning two depth sensors in the existing DGS corpus studio configuration. This includes investigation of the challenges of including depth sensors in the setup that already consists of other cameras. We also discuss about how these sensors can be helpful in automatic analysis of non-manuals like facial expression recognition for corpus recordings with our experimental configuration.

**Keywords:** Kinect, Carmine 1.09, DGS corpus, studio setting, non-manuals recognition

## 1. Introduction

Recently, combined camera and depth sensor devices caused substantial advances in Computer Vision directly applicable to automatic coding a signer's use of head movement, eye gaze, and to some extent, facial expression. Automatic and even semi-automatic annotation of non-manuals would mean dramatic savings on annotation time and are therefore of high interest for anyone working on sign language corpora.

Optimally, these devices need to be placed directly in front of the signer's face at a rather short distance. While this might be ok for some experimental setups, it is not acceptable in a corpus setting for at least two reasons: (i) The signer looks at the device instead of into the eyes of an interlocutor. (ii) The device is in the field of view of other cameras used to record the signer's manual and non-manual behaviour.

We report on experiments determining the degradation in performance when moving the devices away from their optimal positions in order to achieve a recording setup acceptable in a corpus context. For these experiments, we used two different device types (Kinect and Carmine 1.09) in combination with one mature CV software package specialised on face recognition (Faceshift).

## 2. Setup

The experiment is located in an existing studio configuration adapted from the DGS corpus recording setup (Hanke et al., 2010). The major change is that the signers are standing instead of sitting. For the first round of experiments, only one signer is prsent, signing into an HD camera at face level. For the time being, we ignore the visibility of the two sensors in the total scene camera perspective, but concentrate on its visibility in the frontal view. (The bird's eye view turned out not to pose a problem.) In this context, the signer is located facing the HD camera a distance of about 2.8 meters distance.

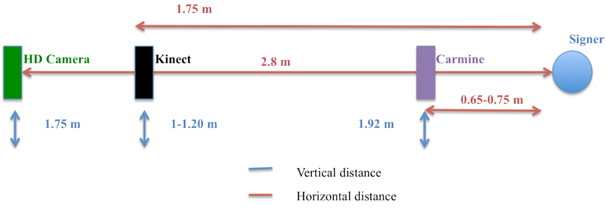

Figure 1: Diagrammatic representation of our final setup

Now we consider this length as our area of interest to position the sensors. In the experiment, a deaf colleague produced random signing, i.e. we did not make use of the monitors to provide elicitation materials which made things easier as we had to accommodate some improvised mounting devices (cf. fig. 2) for this experimental purpose. Finally we arrived at the configuration as shown in fig. 1 & 2.

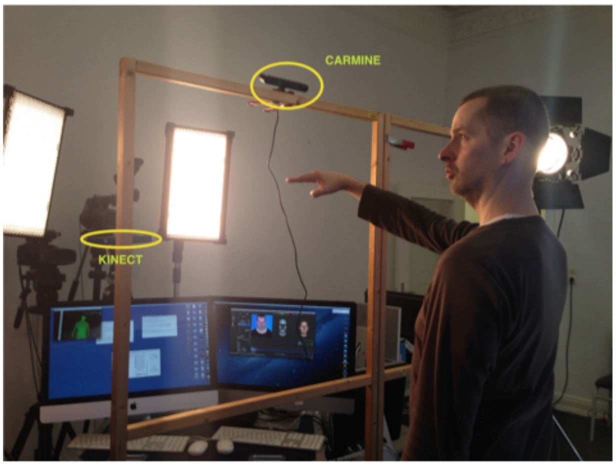

Figure 2: Setup with Carmine and Kinect sensor

## 3. Depth Sensor Positions

Placing these two depth sensors in the existing DGS corpus studio setup and arriving at the final optimal solution of the current studio setup are described in this section.

The Kinect is well known for its full body tracking capability if operated at a distance of more than half a meter. The Carmine 1.09 is a near mode depth sensor that can sense from less than half a meter, making it a good candidate for facial expression recognition software (e.g Faceshift). There are couple of important constraints to be considered about the performance range of the depth sensors and challenges to be resolved during recording.

*Constraints:*

1) The Carmine 1.09 can only be placed 0.65 meters maximum away from the face (front-facing) with permissible rotation along horizontal axis.

2) The Kinect should be placed between 1.5 to 1.75 meters away from the signer to get good skeleton

tracking.

*Challenges:*

1) The signer's eye gaze should not be distracted by the sensor.

2) These sensors can appear in other camera's fields of view.

Now we analyse different positions based on the combination of constraints and challenges within the length of interest i.e. between the signer and front facing HD camera.

### 3.1 Facial Expression Recognition

Faceshift is a facial motion capture software package that takes input from depth sensors like Carmine. Carmine 1.09 is a near-mode sensor recommended for Faceshift. Prior to the performance test, the system was trained to the signer's face for achieving a calibrated expression model for his/her face. For example, most common facial expressions like neutral, smile, frown etc., are considered as sample data for training and classification. The facial expression recognition highly relies on good training data of each individual signer. Another important point is that we are interested in estimating an optimal orientation of the Carmine device such that tracking and recognition are consistent and independent of different signers' physiognomies.

We tested performance of this setup varying the parameters distance (ranging from 0.35 to 0.8 meters) and rotation. As we can see from fig. 4, good lighting and the face close enough to the sensor result in a good accuracy of expression recognition, even with some rotation. By analyzing several test data we observed that the optimal distance is 0.70m. After resting the base of Carmine on a mounting surface (in our case a wooden frame and a stand), we rotated the head manually in 'yaw' direction in order to find best orientation. (The reason why there seems to be different lighting in the samples is sensor rotation.) The vertical field of view of the Carmine device is 45 degrees. It required a lot of trial and error experiments to adjust the sensor head rotation, based on the following rule: (1) Forehead and chin area should be visible prominently in the field of view (see fig. 3) as they are relevant clues in tracking the signer's face.

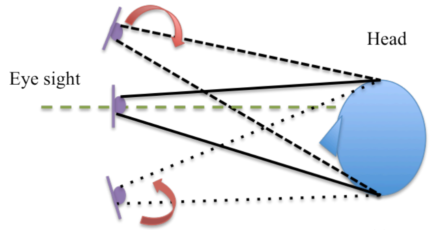

Figure 3: Sensor head rotation in 'yaw' direction

Once we had the optimal distance as well as rules for determining the best possible orientation and position of the sensor for one particular signer, the next step was to find a solution for achieving an optimal orientation of the sensor which is acceptable for signers of varying height. Of course one could adjust the height of the sensor but the setup should be tolerant enough not to require time-intensive calibration In order not to touch the sensor at all in our experiments, we simply asked shorter signers to stand on some pedestal-like boxes.

We also varied the sensor horizontally to the left and right of the signer's face within the range of 30 cm as shown in fig. 4. Within this range, the movement does not have an impact on face tracking (given a distance of 70cm).

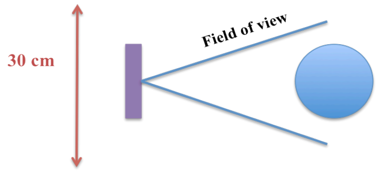

Figure 4: Sensor horizontal sliding

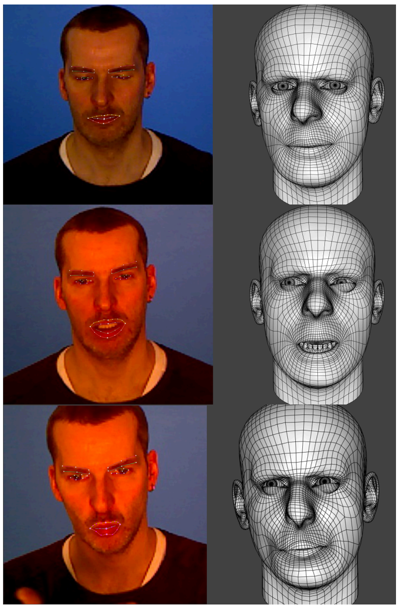

Figure 5: Sample data (1-3) showing accuracy of expression recognition based on table 1

| Sample | Distance in meters | Performance |
|---|---|---|
| 1 | 0.80 | Unreliable |
| 2 | 0.75 | Fewer false positives |
| 3 | 0.70 | Better tracking |

Table 1: Comparison of performance with varying distance

Figures 5, 6 and table 1 explain the dependency between the distance from the sensor, the orientation and the recognition quality. Considering this fact, we observed that the optimal height of mounting the Carmine is above the signer' head level with reasonable rotation, the upper option in fig. 3.

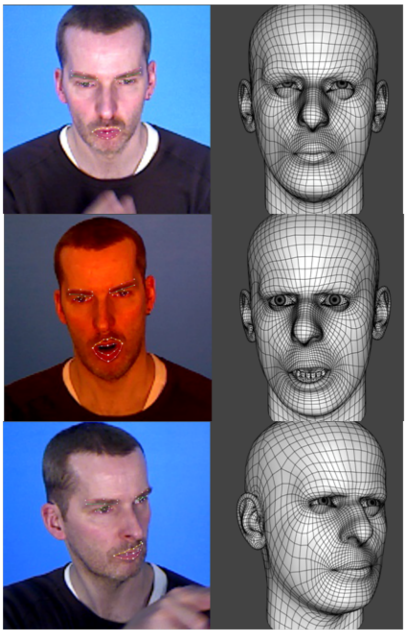

Figure 6: Sample (1-3) data showing sensor rotation

| Sample | Rotation range in degrees | Performance |
|---|---|---|
| 1 | 25 -30 | Unreliable eye tracking |
| 2 | 20-25 | Good |
| 3 | 10-20 | Better tracking |

Table 2: Comparison of tracking performance with varying rotation

Once the optimal rotation is set, the system gets trained for that particular signer's face and the recording begins. After initiating the recording, the Carmine sensor should not be rotated as that will result in inconsistent tracking for that signer. Prediction of rotation variation of the sensor head is not necessary. We have the option of extracting the head rotation in the 'yaw' direction. From the normal case (see fig. 3), i.e when the sensor is frontal to the signer's face, the rotation values can be even higher than this. But in our case where the sensor is not exactly facing the signer's face, but slightly from above. So we have restrictions to have minimum rotation as shown in Table 2. Another important issue is inconsistency in the lip movement recognition, which occurs due to head movement and tracking failures after occlusion of the lips. This issue was rectified to some extent in the refinement process. Fig. 5 shows the lip expression recognition of sample (2) and (3) from fig. 7. There is a possibility of analysing 48 built-in facial expressions in Faceshift.

After achieving a satisfying outcome from performance tests and height adjustment of the Carmine sensor, we filmed some sample data to check the visibility of the sensor in the front-facing HD camera. What we observed is that the Carmine sensor remains invisible in the field of view when the camera focuses on the signer's signing space below the head as shown in fig. 6. However, if the camera is set to also capture signs above the head, a little part of the sensor mounting became visible.

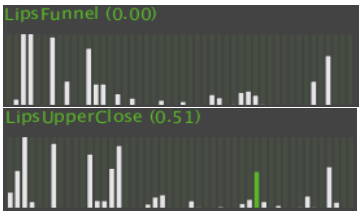

Figure 7: Lip expression accuracy (in green) of sample (2) and (3) from fig. 6

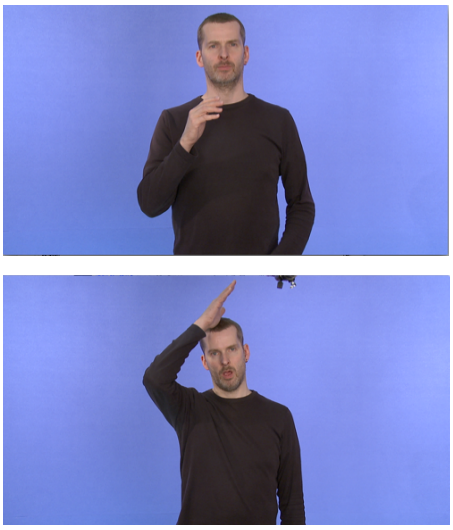

Figure 8: Visibility of Carmine from front facing HD camera when signing occurs below head and above head

Feedback from our transcription team suggests that a bit of appearance of Carmine will not disturb their further work with the film. For production videos, the mounting could later be removed automatically from the movie footage as long as there is no overlap of the signer's hands and the mounting device.

### 3.2 Kinect Positioning for Skeleton Tracking

The Kinect (xbox) is placed in front of the frontal HD camera as shown in fig. 1 and at a distance of 1.75 meters away from the signer and 1.0 meter above the ground. This is the final position where we could get satisfying results.

Before deciding the best position for the Kinect, we tried to explore the various pitfalls with different heights of placing the sensor as given in table 3. Since the motion is mainly happening in the upper part of the body (Torso), there were heights at which the tracking started collapsing by dropping too large an amount of frames initially. This is crucial because initial frame drops cannot be afforded in our case (for sign language corpus analysis later).

We show a couple of test cases to prove the dependency between distance nearer to the signer and skeleton tracking performance. We placed the Kinect sensor at:

a) Test case 1: a distance of 2.10 meters and a height of 1.70 meters to make sure it fits as close as possible to front facing HD camera. Tracking failed due to calibration failure.
b) Test case 2: a distance of 1.50 meters and a height of 1.40 meters. Tracking starts only after dropping frames, but it is unreliable. As you can see from fig. 9 (2) the green color of the tracker indicates second

user being detected in the scene, which is not true in our case.

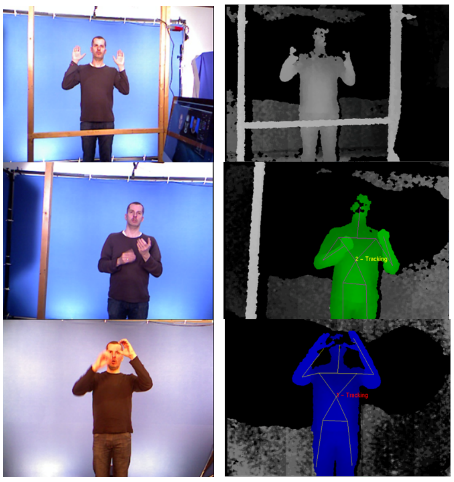

Figure 9: Different test cases (1-3) of Kinect positions showing the colour frames and tracking performance

c) Test case 3: a distance of 1.75 meters and height of 1.00 meters. Tracking and calibration are good.

As shown in table 3, we also tested the tracking performance with different heights. The possibility of moving the Kinect away from the signer was restricted due to the space constraint in the current studio setup. In a more regular setup, there will be enough space to test different other positions and heights.

| Kinect at a Distance of 1.75 meters away from the signer | Kinect's Height in meters | Tracking performance |
|---|---|---|
| | 1.60 | Calibration failed & no tracking |
| | 1.30 | Tracking started only in the middle of the film |
| | 1.00 | Tracks well |

Table 3: Comparison of Skeleton tracking performance with varying height at fixed distance from the signer

## 4. Future work

When trying to apply the current approach to the studio setting with two informants (whether seated or standing), The current solution for the Carmine devices can simply be doubled. However, one degree of freedom for positioning the Kinect devices is lost: Following the results obtained so far, the only reasonable position for the Kinect devices is directly above the screens used for elicitation material in order to minimize distraction to the informants. The experiments carried out so far suggest that a setup like that shown in fig. 10 will be possible. Fig. 11 shows possible configurations how to place two Kinect devices (one for each signer) relative to each other in order to minimize the space needed. Another problem to be researched is synchronization issues involved in non-manuals recognition resulting from the use of two different sensors requiring different recording software.

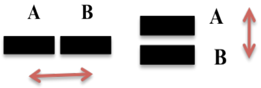

Figure 10: DGS corpus studio setup - Two signers interacting in sitting position, Kinect devices mounted between the two screens

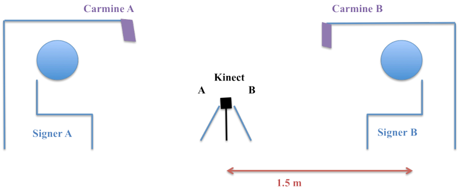

Figure 11: Possible configurations of placing two Kinect devices

## 5. Technical details

The depth sensors we use are Kinect xbox 360 for body tracking and Carmine 1.09 for facial feature tracking. These two sensors are operated using two different software packages. Data recording from Kinect xbox and Carmine are achieved by OpenNI & OpenCV program and Faceshift software at 640x480p30 respectively (for both depth and RGB channel). The recording with Kinect can be done automatically (continuously) or manually for each user.

## 6. Conclusion

Although the current studio setup has limited space to accommodate extra sensors (and their stands!), our additional sensors positions do not make the informants feel uncomfortable or the images more difficult to process by human annotators. The positioning of the sensors for the current corpus studio configuration increases our confidence that it will be possible to use these two depth sensors in corpus recordings resulting in valuable automatic annotation of non-manuals. As a by-product, we might be able to annotate emotional facial expressions.

## 7. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

## 8. References

Hanke, T., König, L., Wagner S., Matthes S., (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. $4^{th}$ *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pp. 106–109.

http://www.faceshift.com/product/

http://www.openni.org/

http://www.primesense.com/