# Toward a 3D Body Part Detection Video Dataset and Hand Tracking Benchmark

Christopher Conly[1], Paul Doliotis[1,2], Pat Jangyodsuk[1], Rommel Alonzo[1], Vassilis Athitsos[1]

[1]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
[2]Institute of Informatics and Telecommunications, N.C.S.R. "Demokritos", Athens, Greece

cconly@uta.edu, doliotis@uta.edu, pat.jangyodsuk@mavs.uta.edu, ralonzo@uta.edu, athitsos@uta.edu

## ABSTRACT

The purpose of this paper is twofold. First, we introduce our *Microsoft Kinect*–based video dataset of American Sign Language (ASL) signs designed for body part detection and tracking research. This dataset allows researchers to experiment with using more than 2-dimensional (2D) color video information in gesture recognition projects, as it gives them access to scene depth information. Not only can this make it easier to locate body parts like hands, but without this additional information, two completely different gestures that share a similar 2D trajectory projection can be difficult to distinguish from one another. Second, as an accurate hand locator is a critical element in any automated gesture or sign language recognition tool, this paper assesses the efficacy of one popular open source user skeleton tracker by examining its performance on random signs from the above dataset. We compare the hand positions as determined by the skeleton tracker to ground truth positions, which come from manual hand annotations of each video frame. The purpose of this study is to establish a benchmark for the assessment of more advanced detection and tracking methods that utilize scene depth data. For illustrative purposes, we compare the results of one of the methods previously developed in our lab for detecting a single hand to this benchmark.

## Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: 3D/Stereo Scene Analysis, Motion, Video Analysis;
I.4.8 [**Scene Analysis**]: Depth Cues, Motion, Time Varying Imagery, Tracking

## General Terms

Experimentation, Measurement

## Keywords

gesture recognition, Kinect, hand location, tracking


*PETRA* '13 May 29 - 31, Island of Rhodes, Greece


## 1. INTRODUCTION

With the advent of the Microsoft Kinect in 2010, computer vision researchers were presented with a new opportunity to utilize scene depth information, a capability previously only available with more expensive or cumbersome systems, such as laser depth sensors, stereo cameras, or multi-camera systems. The Kinect and its kin are thus usable in products that are more approachable by the average consumer. Soon after the release, open source Kinect libraries were created, and algorithms were written to detect and track a user's skeleton and joint positions. The ability to determine the user's hand positions, in particular, presents researchers interesting human-computer interaction possibilities.

A gesture recognition system or sign language video dictionary system like that described in [12, 15] necessitates a certain level of visual human-computer interaction. More specifically, they require a vision system that is able to reliably detect and track a user's hands, so that information about them—for example position, appearance, and movement—can be used to accomplish a certain task or perform an action, like looking up the definition of a sign. Traditional methods based on 2D color or grayscale images abound, but the Kinect offers the potential to more reliably detect and track the hands using scene depth information.

Incorporating information about the third dimension into gesture recognition tasks affords us a more accurate representation of what is actually occurring in the scene. A gesture is not merely a 2D, planar event. It has a 3D trajectory and thus, for the utmost accuracy in its representation, demands the third dimension information for trajectory matching using such time-series comparison algorithms as Dynamic Time Warping (DTW) [7]. There is a lack, however, of publicly available datasets that include such depth information.

ASL is a method of communication estimated to be used by 500,000 to 2,000,000 users in the United States alone [8, 10] that often involves complex movements during which the hands are close to each other or the body. Reliable detection of the hands can be difficult in these real world scenarios, and ASL was thus chosen to build the dataset presented in this paper. The dataset allows researchers to develop new hand detection and tracking algorithms and experiment with 3D gesture recognition methods. The benchmark presented in this paper for hand location accuracy provides a baseline measurement to which they can compare their own hand detection and tracking methods.

## 2. RELATED WORK

One of the highest quality video datasets useful for gesture recognition research is the American Sign Language Lexicon Video Dataset (ASLLVD) [3]. It consists of a large set of recordings from multiple camera angles of the signs contained in the Gallaudet Dictionary of American Sign Language [13], performed by native signers. Each sign is annotated with the gloss label (approximate English translation), start and end frames, hand shapes at the start and end frames, and position of the hands and face, with multiple examples per sign. Such datasets, while useful, lack any information about scene depth, since they were recorded with standard color video cameras. Thus, when using them, researchers suffer from the limitations of having to use conventional 2D hand detection and tracking algorithms based on, for example, skin color and motion. Hand tracking using standard video is particularly challenging because of occlusions, shading variations, and the high dimensionality of the motion.

Guyon, et al., present a Microsoft Kinect–based 3D gesture dataset for the ChaLearn gesture recognition competition in [5] that contains 50,000 gestures recorded by 20 different users organized into 500 batches of 100 gestures. Compared to the ChaLearn dataset, the dataset we describe in this paper has certain advantages. First, our dataset is recorded at a resolution of 640 x 480 pixels and a framerate of 25 frames per second, as opposed to 320 x 240 pixels at 10 fps. Secondly, only 400 frames are manually annotated with any skeletal information in the ChaLearn dataset, which makes it difficult to quantify the efficacy of any body part locator/tracker being developed. Our dataset will contain skeletal information for every frame of every gesture. Finally, as the ChaLearn videos are offered only as AVI files, we cannot translate the pixels into x,y,z coordinates in a 3D world reference frame. Our dataset provides access to the raw scene depth information and allows us to determine the x,y,x coordinates with respect to the Kinect reference frame.

One of the earliest hand location/tracking methods utilizing scene depth information was introduced by Nanda, et al. [9]. They employed a depth-based capturing system that relied on the time-of-flight (ToF) principle [6]. Depth and color information can simultaneously be acquired by using the same optical axis in real-time. By using depth data, they proposed a method that was able to track hands in highly cluttered environments. The potential fields of possible hands or face contour were computed by three algorithms: 1) by using distance transform, 2) k-components-based potential fields with weights, and 3) basin of attraction. The system was tested in head tracking and hand tracking on ten people with good results.

Van den Bergh, et al., in [14], used a ToF camera with a low resolution (176 x 144 pixels) to get depth images for segmentation, which was combined with a VGA resolution RGB camera for hand detection. Both cameras were calibrated and an initial background subtraction was performed based on simple thresholding. The remaining pixels were passed through a skin color detection module in order to get hand data. The skin color model employed by the authors used pre-trained skin color histograms combined with an adaptive skin color model, which was updated with color information taken from the face. Three situations were evaluated in hand detection: the hand was next to the face, the hand overlapped with the face, and a second person was behind the tester. Depth-based detection achieved more than 98% accuracy in all three situations, while the accuracy of color-based detection decreased dramatically from 92% in the first situation to 19.8% in the third situation.

In 2010, when Microsoft released the Kinect, an inexpensive tool was made available to the public that offered both color video and depth information. One of the earlier methods using the Kinect, proposed by Doliotis, et al., uses a combination of depth video motion analysis and scene distance information [4]. The algorithm isolates the person performing the gesture using segmentation on the depth data, and a score is calculated for each pixel belonging to the person based on distance to the camera and motion. The higher scoring pixels are more likely to belong to the hand, since it is assumed that the hand will be closest to the camera and will exhibit the greatest amount of motion. The method was tested in single-handed gesture recognition, including cases in which there was a person standing behind the gesturer.

In [11], the authors proposed an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. A large and highly varied training dataset allows the classifier to estimate body parts invariant to pose, body shape, clothing, etc. Randomized decision trees (RDT) are trained on this synthetic dataset, which are then used to classify each pixel of the retrieved depth image. Each pixel gets assigned a certain body part. Finally, a mean shift algorithm is used in order to estimate the joint centers of the body parts. The resulting framework can estimate the full body poses in real time.

Clearly, depth information is useful in computer vision research, and datasets that include this information are needed in order to develop and test the most accurate hand location and tracking methods possible.

## 3. DATASET

Our goal is to create a structured motion dataset that will enable researchers to explore body part (i.e. hands) detection and tracking methods, as well as gesture recognition algorithms not possible with such datasets as the ASLLVD [9] by including scene depth information. The dataset is being recorded with a Microsoft Kinect, which allows us to capture both color video and frames that include scene depth information. Figure 1 shows an example from one of the recording sessions. In this particular representation, the darker gray areas of the image are located closer to the camera. The black regions are portions of the scene for which depth information was not available.

### 3.1 Size and Scope

We hope that the final dataset will contain most of the 3,000 signs found in The Gallaudet Dictionary of American Sign Language [13], which will offer an abundance of complex movements of the hands and arms. Currently, 1113 signs, both one-handed and two-handed, have been recorded with one fluent signer and 200 with another fluent signer, but in the future, we may add more signers, so that there are multiple examples of each sign.

As with [3], fingerspelled signs, loan signs, and classifiers are not included in the dataset. A fingerspelled sign is a word that is spelled out by using the manual alphabet. When a signer has to use a letter that is part of the overall sign, that letter is known as a *loan sign.* Classifiers provide addi-

**Figure 1: Sample dataset sign frame. Top: color video frame; Bottom: depth video frame.**

tional information about the object being signed, but since there are infinite variations of them, they are excluded. The ASLLVD paper [9] contains more information the about motivations for excluding certain types of signs.

## 3.2 Technical Specifications

Both the Kinect color frames and depth frames have a resolution of 640 x 480 pixels and are recorded at frame rate of 25 frames per second. The signers perform groups of ten signs per video in front of a neutral backdrop in a lab with consistent lighting. The signs are performed while standing, and the scene is framed so as to include the region from about the knees to a few inches above the signer's head. This ensures that the entire signing space is included in the video frame. Each video begins with a calibration pose that can be used to detect the signer and initialize tracking. After the pose, between each sign, and after the last sign, the signer returns her hands to her side, creating a clear separation of the signs in the video.

We currently use the OpenNI framework [2] to record the signs in the ONI format, but we may rerecord them in the future with the Microsoft Kinect SDK [1] so that researchers can experiment with both platforms. OpenNI is an open source sensing development framework used in many third party APIs. Its purpose is to standardize compatibility and interoperability of Natural Interactive devices and applications. It and third party software developed around it are useful to researchers that want to develop their own detection and tracking tools. Compressed and uncompressed AVI videos of the recordings are also available.

## 3.3 Annotations

Each video in the dataset is annotated with the start and end frames of each sign so that any sign can be quickly accessed. The first depth video frame of each sign is annotated with a bounding box around the signer's face to give an idea of the scale of the individual in the video. With this information, the researcher has an idea of how to scale the query sign to which it is being compared. Furthermore, every depth frame belonging to a sign is also annotated with bounding boxes around the hands, which give an indication of the hand's location when the box's centroid is calculated. In the future, we will annotate the frames with additional skeletal information such as elbow and shoulder locations to extend the usefulness of the dataset beyond hand location and tracking to multiple body part detection. Furthermore, we may annotate the color video in the same manner.

The annotations also include information about the signs themselves, such as signer ID, file locations, sign type (two-handed or one-handed), and gloss, or rough English translation. The hand and face annotations for an example sign frame are shown overlain on the depth frame image in figure 2.

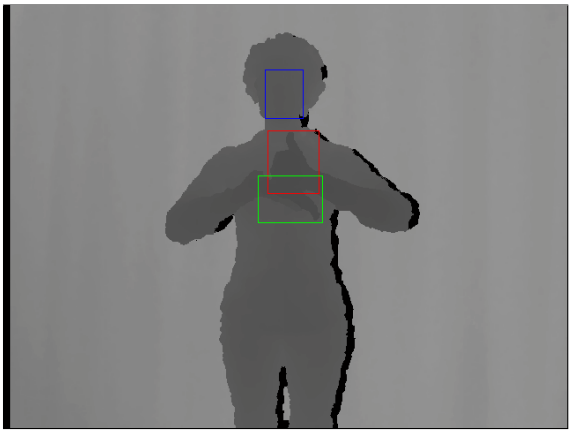

**Figure 2: Sample hands and face annotations of a single depth video frame.**

# 4. EXPERIMENTS

In order to establish the benchmark, we chose to use the hand location capabilities of the user skeleton tracker included in the OpenNI 1.5 NiUserTracker sample program [2], since it is open source and available to anyone. Once it has found the signer via the standard psi calibration pose, the program creates a skeletal model of the person and tracks joint position movement. In particular, we were interested in the arms and defined the hand locations to be the hand endpoints of the elbow-hand portions of the arms.

To evaluate the efficacy of using the skeleton tracker to approximate the positions of the hands, we used 70 randomly selected signs from the dataset described in section 3—35 one-handed and 35 two-handed—and processed them with the tracker. For one-handed signs, only the signing hand was considered. Once the hand positions were obtained, they were compared to the ground truth positions from the manual annotations, and the pixel Euclidean distance between

them was recorded as a score, so that a lower score would indicate a closer estimation of the hand's actual location. This operation was performed on each frame of the signs, and the accuracy was calculated to serve as the benchmark for the evaluation of future methods. As an example comparison, we processed the one-handed signs with the single hand locator described in [4]—a method based on movement and depth alone—and calculated the results using the same pixel Euclidean distance similarity measure.

## 5. RESULTS

We calculated overall accuracy as a percentage of frames in which the automatically generated hand locations fell within in various pixel distances (termed pixel error) of the manual hand annotations. The OpenNI skeletal tracker was used on both the one-handed and two-handed signs, while the comparison hand locator was only used on the one-handed signs, as it was not designed to detect multiple hands.

Figure 3 shows the accuracy of the skeletal tracker in locating the signer's hands in both one-handed and two-handed signs. For example, in 90% of the frames, the skeletal tracker had a pixel error of 37 pixels or less for the right hand.

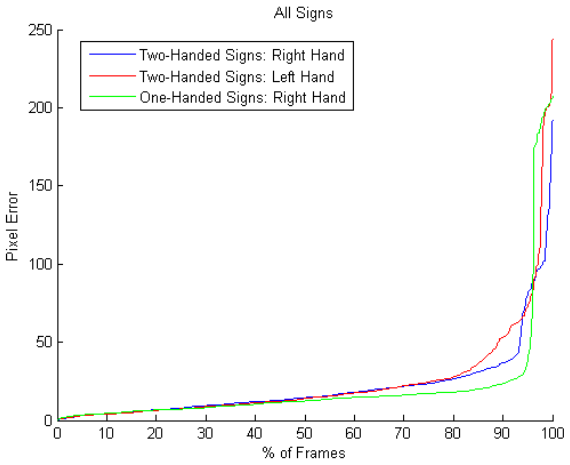


**Figure 3: Skeletal tracker pixel error for both one-handed and two-handed signs.**

It is clear that the skeletal tracker performs notably better on one-handed than on two-handed signs. For the right hand, 90% of the frames have a pixel error of about 24 pixels or less in one-handed signs, as opposed to 37 pixels or less in two-handed signs. This increased pixel error in two-handed signs is perhaps due to the close proximity of the hands to each other, which can make it difficult for the tracker to discriminate between the hands.

Figure 4 shows the accuracy of the skeletal tracker and the comparison hand locator on one-handed signs. It can be seen that the comparison hand locator does not perform as well on this dataset. It is understandable when we consider that it was written for use with simple hand gestures in which the hand will likely be the closest part of the body to the camera. Indeed, after examination of the results, we determined that it tends to fail when other body parts that are also in movement, such as the elbow, are closer to the camera.

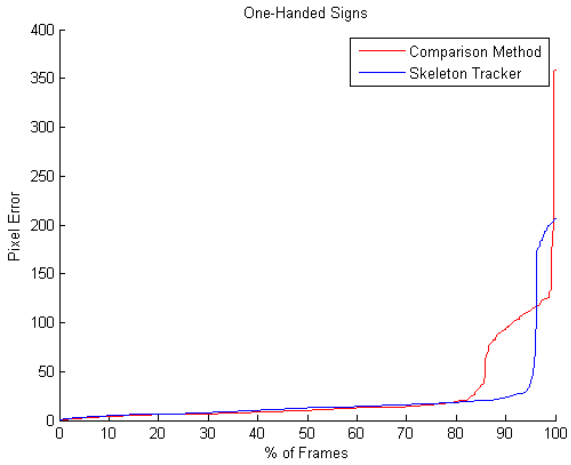


**Figure 4: Comparison of the skeletal tracker and the method from [4] on one-handed signs.**

We also calculated the maximum pixel error for each sign, separated into one-handed and two-handed signs. Figures 5 and 6 show the results for the skeletal tracker and its comparison to the single hand detector, respectively. We can see in figure 6, for example, that 50% of the signs had a maximum pixel error of about 22 pixels or less when the comparison method of [4] was used to detect the hands.

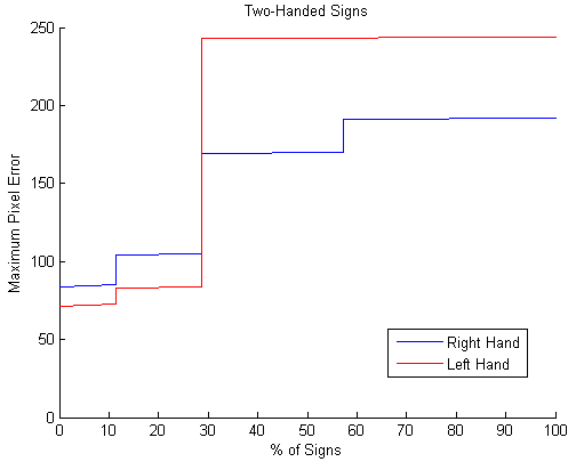


**Figure 5: Maximum hand location pixel error on a per sign basis for the skeletal tracker.**

Figure 7 shows an example frame with good accuracy using the skeletal tracker on both hands in a two-handed sign. In this example, the error on both hands is only a few pixels. Figure 8 shows a visualization of three levels of skeleton tracker accuracy from good to poor on a single-handed sign, with the pixel error ranging from a few pixels to a few hundred pixels. In both these figures, the manual annotations are shown in green and the skeleton tracker hand locations in red.

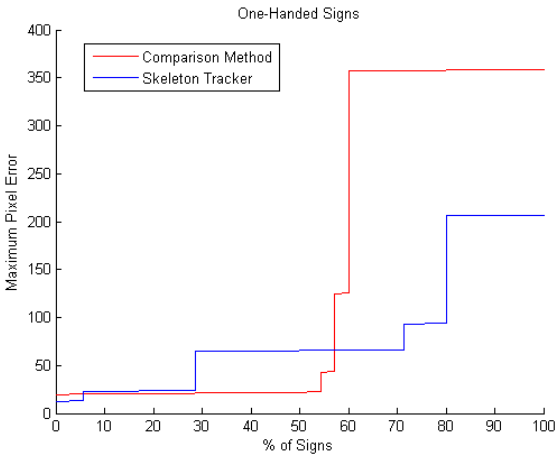


**Figure 6: Skeletal tracker and one hand gesture method maximum pixel error on a per sign basis.**

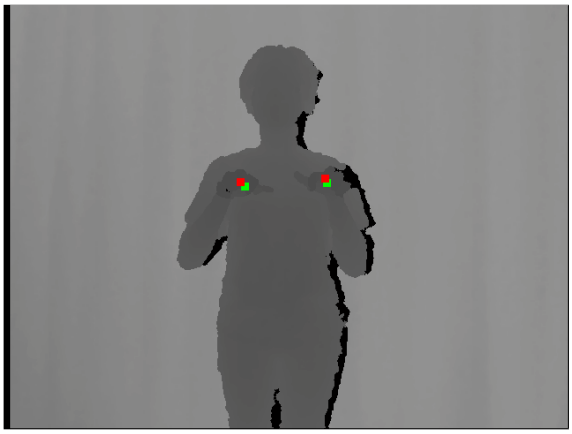

**Figure 7: Example accuracy of the skeletal tracker on a two-handed sign.**

## 6. CONCLUSION AND FUTURE WORK

We have presented a video dataset based on ASL recordings that provides both color video and valuable scene depth information for use in the development and testing of hand detection and tracking methods, as well as in 3D gesture recognition and natural interaction projects. The dataset provides a large number of gestures that involve one or both hands with varying levels of movement and hand shape complexity and presents an opportunity to develop algorithms that are viable in real world scenarios.

We have used this dataset and a readily available skeleton tracker to develop a benchmark for analysis of future detection and tracking algorithms. We provided a sample comparison of another hand location method that uses this scene depth information to the benchmark as an illustration to assess its potential usability in gesture recognition.

We will continue expanding the dataset described in section 3 until we have multiple examples of most of the signs found in [13] and will manually annotate the frames with skeletal and joint information. Finally, we will continue to develop more accurate hand location and tracking algorithms using the new dataset and the benchmark for comparisons and apply them, along with handshape analysis in our gesture recognition research.

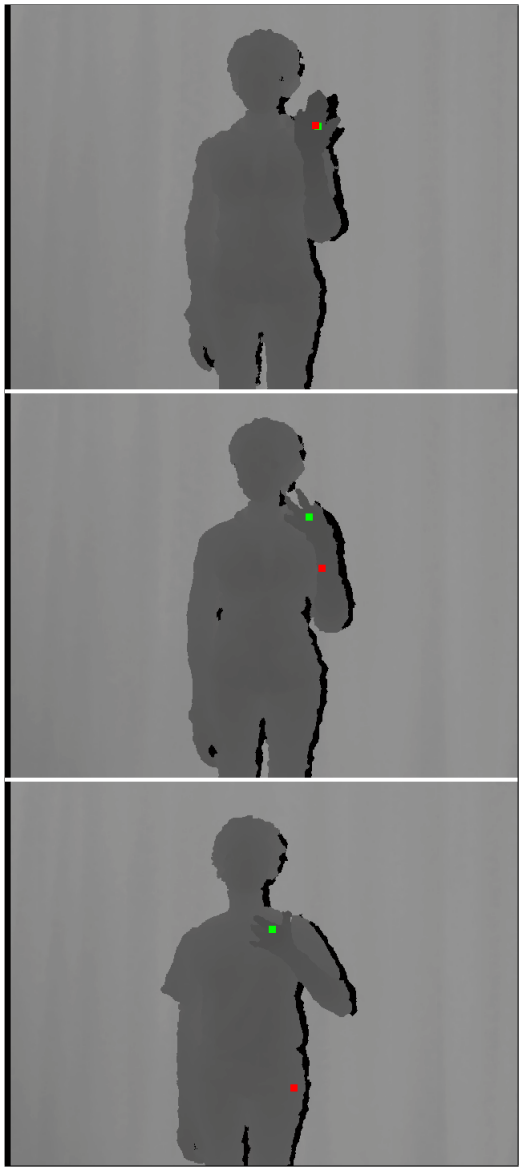

**Figure 8: Varying accuracy on one-handed signs.**

## 7. ACKNOWLEDGEMENTS

This work was partially supported by grants from the National Science Foundation, IIS-0812601, IIS-1055062, CNS-1059235, and CNS-1035913. The authors would like to thank Joan Bempong and Carolyn Stem for their contributions to the creation of this dataset.

## 8. REFERENCES

[1] Developer SDK, toolkit & documentation | kinect for windows. http://www.microsoft.com/en-us/kinectforwindows/develop/.

[2] OpenNI SDK | OpenNI. http://www.openni.org/openni-sdk/.

[3] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, and A. Thangali. The American Sign Language Lexicon Video Dataset, June 2008.

[4] P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos. Comparing gesture recognition accuracy using color and depth information. In *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '11*, page 1, New York, New York, USA, 2011. ACM Press.

[5] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. Escalante. Chalearn gesture challenge: Design and first results. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–6, 2012.

[6] G. J. Iddan and G. Yahav. G.: 3d imaging in the studio (and elsewhere. *In: SPIE*, pages 48–55, 2001.

[7] J. B. Kruskal and M. Liberman. The symmetric time warping algorithm: From continuous to discrete. In *Time Warps*. Addison-Wesley, 1983.

[8] H. Lane, R. J. Hoffmeister, and B. Bahan. *A Journey into the Deaf-World*. DawnSign Press, San Diego, CA, 1996.

[9] H. Nanda and K. Fujimura. Visual tracking using depth data. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, page 37, june 2004.

[10] J. Schein. *At home among strangers*. Gallaudet U. Press, Washington, DC, 1989.

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297 –1304, june 2011.

[12] A. Stefan, H. Wang, and V. Athitsos. Towards automated large vocabulary gesture search. *Proceedings of the 2nd International Conference on PErvsive Technologies Related to Assistive Environments - PETRA '09*, pages 1–8, 2009.

[13] C. Valli, editor. *The Gallaudet Dictionary of American Sign Language*. Gallaudet U. Press, Washington, DC, 2006.

[14] M. Van den Bergh and L. Van Gool. Combining rgb and tof cameras for real-time 3d hand gesture interaction. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 66 –72, jan. 2011.

[15] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, and F. Kamangar. A system for large vocabulary sign search. In *Proceedings of the 11th European conference on Trends and Topics in Computer Vision - Volume Part I*, ECCV'10, pages 342–353, Berlin, Heidelberg, 2012. Springer-Verlag.