An Evaluation of RGB-D Skeleton Tracking for Use in Large Vocabulary Complex Gesture Recognition

Christopher Conly, Zhong Zhang, and Vassilis Athitsos
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, Texas, USA
cconly@uta.edu, zhong.zhang@mavs.uta.edu, athitsos@uta.edu

ABSTRACT

An essential component of any hand gesture recognition system is the hand detector and tracker. While a system with a small vocabulary of sufficiently dissimilar gestures may work well with approximate estimations of hand locations, more accurate hand position information is needed for the best results with a large vocabulary of complex two-handed gestures, such as those found in sign languages. In this paper we assess the feasibility of using a popular commercial skeleton tracking software solution in a large vocabulary gesture recognition system using an RGB-D gesture dataset. We also provide a discussion of where improvements in existing methods utilizing the advantages of depth-sensing technology can be made in order to achieve the best possible results in complex gesture recognition.

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

1. INTRODUCTION

RGB-D technology has applications that extend beyond the often-referenced Kinect gaming industry and into assistive technology. Zhang, et al., for example, use a Kinect to detect falls, which could be useful for home monitoring of elderly or injured individuals [20]. RGB-D cameras are also useful in gesture recognition since they provide multiple data modalities that can be used to interpret the scene. Hand gestures are a convenient form of human-computer interaction with a broad range of applications in various areas. They can be used, for example, to give commands to computers or assistive robots when traditional input methods may be impractical or entirely unusable. The ability to recognize gestures can not only enrich one's computing experience but potentially his or her life.

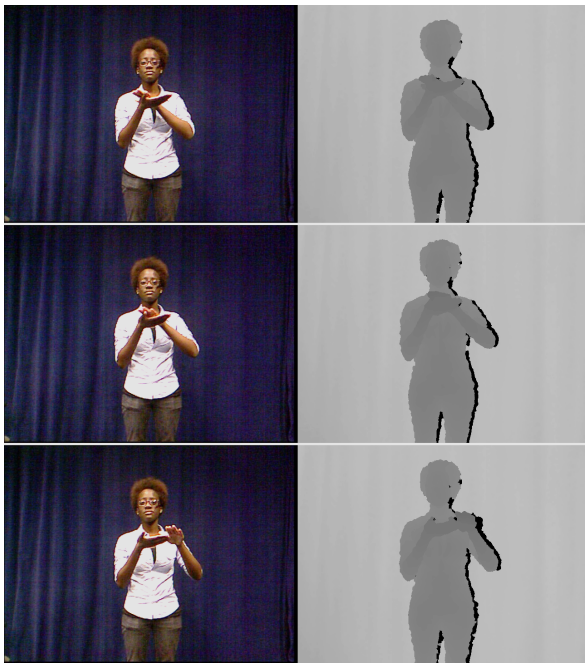

Figure 1: Sample depth and color frames from the beginning, middle, and end of a gesture.

A gesture recognition system such as that used in sign language video dictionaries [19, 5] needs a fast, reliable, and accurate hand locator and tracker. To minimize the work that the user must perform to match a sign or gesture, the system should automatically detect and track the hands without user intervention. Some existing systems, such as the ASL video dictionary system described by Wang, et al. [19], require several user-performed steps to match a sign. One of the steps is to provide a bounding box around each hand in the first frame of the sign to initialize the hand tracker for the remaining frames. The Kinect offers the potential to

PETRA'14, May 27 - 30 2014, Island of Rhodes, Greece
Copyright 2014 ACM 978-1-4503-2746-6/14/05 ... $15.00
http://dx.doi.org/10.1145/2674396.2674426

automatically locate and track the user's hands throughout the gesture, thus eliminating some steps, and it is with this motivation that we examine the potential of using a readily available RGB-D skeleton tracker for use with a vocabulary comprised of a large number of potentially complex gestures.

Rather than deliberately creating a limited gesture set that works well with Kinect-based gesture recognition, we wanted to apply RGB-D technology to an existing difficult vocabulary. Thus, for the testing set in our experiments, we used the 3D body part detection and gesture recognition dataset introduced in [4], comprised of Kinect recordings of a large vocabulary of American Sign Language (ASL) signs. Figure 1 shows example corresponding color and depth frames from the ASL sign for the adjective *clean*. At the time of writing, the dataset consists of a 1,113 sign vocabulary recorded with two signers, so that there are two examples of each sign. Eventually, we would like to extend the set to the roughly 3,000 signs found in the Gallaudet Dictionary of American Sign Language [18] and have three or four signers perform the set, as each individual performs the signs with some variation in hand placement, limb angles, and expressiveness of motion.

With an estimated 500,000 to 2,000,000 users of ASL alone in the just United States [10, 15], sign languages offer a realistic scenario of human-computer interaction. Instead of exaggerated movements in a small vocabulary of deliberately dissimilar one-handed gestures, as often comprises the gesture set in recognition projects, ASL consists of intricate one-handed and two-handed gestures that form a large expressive vocabulary, much like a written language. While this affords the opportunity for a richer interaction experience with computer systems beyond, for example, simple commands to an assistive robot to move to the right or left, it makes recognition more difficult, since so many gestures can share similar characteristics and since occlusions and the proximity of the hands to each other and the body can impede accurate hand tracking.

We use the OpenNI and NiTE development kits from Primesense, LTD to access the Kinect and generate skeleton joint data [13, 12]. They comprise a popular software platform for the development of Natural Interaction (NI) applications and are thus a common choice for gesture driven projects. This paper assesses how well these particular technologies perform in recognizing a large vocabulary of gestures.

2. RELATED WORK

There is an abundance of computer vision-based gesture recognition research employing RGB cameras. Sandjaja et al. achieve 85.52% accuracy in a Filipino Sign Language number recognition system but require the user to wear a multi-colored glove [14] to automate hand and finger location and tracking. Our system does not require such measures, and instead uses depth information from the Kinect to locate the hands.

Much of the research involves vocabularies of limited size. Zieren et al. achieve 99.3% accuracy in user-dependent sign language recognition experiments using a 232 sign vocabulary; the accuracy, however, decreases to 44.1% in user-independent experiments with a vocabulary of 221 signs [22]. Similarly, Kadir et al. achieve high accuracy with a vocabulary of 164 signs, but also use the same signer for the training and testing sets [8]. In our experiments, we employ a much larger vocabulary of 1,113 gestures and ensure user-independence.

Athitsos et al. [3] and Wang et al. [19] use vocabularies of comparable size to ours, but require the user to provide hand locations either for each frame or for the first frame to initialize a hand tracker. We automate this process by using a Kinect and skeleton tracking algorithms to estimate hand positions.

There is also a body of research using the Kinect or similar RGB-D cameras for gesture and sign language recognition, but these studies also tend to use limited vocabulary size and gesture complexity. In early Kinect research, Doliotis et al. reach 95% recognition accuracy in cluttered scenes but only employ a simple vocabulary of 10 digits drawn in space with the hand and make the assumption that the hand will be the closest body part to the camera [6]. This is often not the case with ASL. More recently, Agarwal and Thakur also achieve good results using a similarly sized static hand gesture vocabulary, consisting of Chinese Sign Language signs for digits [1].

Zafrulla et al. conduct a Kinect-based ASL recognition feasibility study in which they recognize 60 distinct, simple phrases of 3 to 5 signs using a 19 word vocabulary [21]. The authors conducted both seated and standing tests. They achieve word and sentence recognition accuracy of 74.48% and 36.2%, respectively, for seated tests and 73.62% and 36.3% for standing tests. Pedersoli et al. explore real-time gesture recognition using a vocabulary of 16 relatively simple one-handed gestures and achieve better than 70% accuracy [11]. However, it requires an open palm, forward-facing orientation for hand segmentation and assumes the hand is the closest object to the camera for hand pixel clustering to work.

Several datasets have been created that are useful in gesture recognition research. One RGB dataset is the American Sign Language Lexicon Video Dataset [2]—a large dataset that contains annotated recordings of multiple signers from several different camera views—that is used for training in our experiments. The ChaLearn gesture dataset is a large RGB-D dataset of 50,000 hand and arm gestures of varying complexity [7]. The lack of complete annotations, however, makes it difficult to use in our research. We have thus created a Kinect gesture dataset of 1,113 ASL signs, each recorded by two signers, to be used in our experiments for testing [4].

3. EXPERIMENTAL METHOD

In our recognition experiments, we extract scale and translation invariant features and use a similarity measure that accounts for temporal differences in gestures; all experiments are performed in a user-independent manner.

3.1 Feature Extraction

To make both the training and test gestures translation and scale invariant, we express the joint positions in a head-centric coordinate system using the position of the head in

the first frame as the origin and then scale the signs so that the diagonal of the face bounding box is equal to 1. This effectively aligns the orientation of the signs and normalizes the distance of the signer to the camera. Each sign is normalized to 20 frames using linear interpolation on joint positions as needed.

After the positions of the hands are extracted and expressed in the new coordinate system and the sign is normalized to 20 frames, a modified version of the trajectory-based feature vector described in [19] is generated. We do not include hand appearance in the feature vector since we cannot properly compare the test set 3D depth video handshapes to those of the color training videos. We package the following components into feature vector X_t for each frame t of sign video X.

1. $L_d(X,t)$ and $L_{nd}(X,t)$: The pixel position of the dominant and non-dominant hands, respectively, in frame t of sign video X

2. $L_\delta(X,t) = L_d(X,t) - L_{nd}(X,t)$: The dominant hand's position relative to the non-dominant hand.

3. $O_d(X,t)$ and $O_{nd}(X,t)$: The direction of motion from frame $t-1$ to frame $t+1$ for the dominant and non-dominant hands, respectively, expressed as unit vectors.

4. $O_\delta(X,t)$: The direction of motion for L_δ from frame $t-1$ to frame $t+1$, expressed as a unit vector.

The dominant hand is the hand that will be moving in signs for which only one hand moves.

3.2 Similarity Measure

The Dynamic Time Warping (DTW) [9] time series analysis method is useful in gesture recognition and was chosen to provide a similarity measure between the trajectories of signs in our experiments. For every sign in the test set, we use DTW to compare the feature vectors from its frames to those of each training example, creating a warping path W, or alignment between the frames of the query and model. The cost C of each warping path is the sum of the costs of matching the aligned frames. We define this local cost $c(Q_{q_i}, M_{m_i})$ of matching frame q in query sign Q to frame m of model sign M as the Euclidean distance between their feature vectors. Thus, for warping path W of length $|W|$ aligning query sign Q with model sign M:

$$C(W,Q,M) = \sum_{i=1}^{|W|} c(Q_{q_i}, M_{m_i}), \quad (1)$$

so that the DTW score D between query Q and model M is provided by the lowest cost warping path:

$$D_{DTW}(Q,M) = \min_W C(W,Q,M). \quad (2)$$

The lowest DTW score of the training examples is taken to be the score for that sign class.

3.3 Training Set

Due to the lack of publicly available 3D ASL datasets, we chose to use a standard 2D RGB dataset for training. In the experiments, we used a 1113 gesture vocabulary training set to which we matched our smaller subset of RGB-D signs. To ensure user independence, no videos from the test set signer appear in any of the training sets.

Three examples each, from different signers, of the 1113 gesture vocabulary were taken from the dataset described in [2] to be used as training examples. Since the videos are standard 2D RGB videos and there is no real-world distance information for the hand and head positions, we used their pixel locations in our training data.

3.4 Test Set

We selected 606 signs of varying complexity, 400 two-handed and 206 one-handed, from our 3D dataset recorded with OpenNI [13] and used the NiTE skeleton tracker [12] to determine the hand positions in each frame and the head position in the first frame of each sign. As the NiTE tracker provides positions for joints in a 3D Kinect-centric coordinate system, we used the projections of those positions onto the 2D depth image plane, so that instead of the real-world distance measures for the joints, we were using their pixel coordinates. This allowed for proper comparison with the pixel coordinates used in the training set. For comparison to a best possible scenario, we ran the experiments using the manual annotations of the hand positions in each frame in addition to the skeleton tracker-generated positions.

4. RESULTS

We first analyzed the accuracy of the hand locations provided by the skeleton tracker as described in [4]. To do so, we recorded the Euclidean pixel distance between the centroids of the manually annotated hand bounding boxes and their respective pixel positions output by the skeleton tracker in each frame. Figure 2 shows the percentage of depth video frames with a maximum Euclidean distance pixel error e. For example, the right hand in 80% the frames of 2-handed signs had a error of 27 pixels or less.

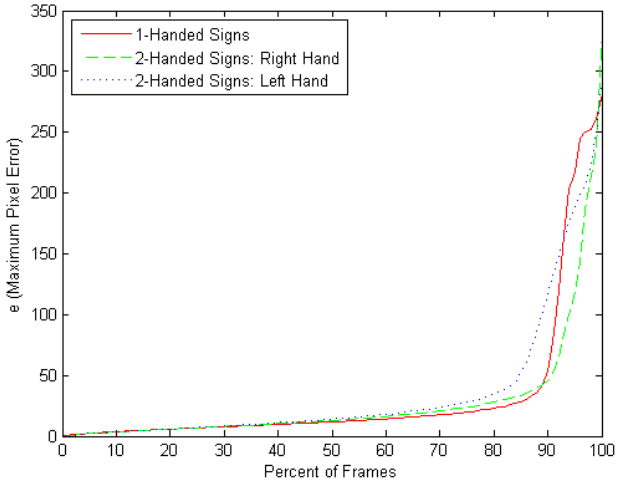

Figure 2: Hand location accuracy of the skeleton tracker using the method of [4].

We then ran the gesture recognition experiments using the skeleton tracker data and and calculated accuracy as a percentage of signs for which the correct sign was ranked in the top k matches. Figure 3 shows the results. For example, 66% of the two-handed signs ranked in the top 20 matches.

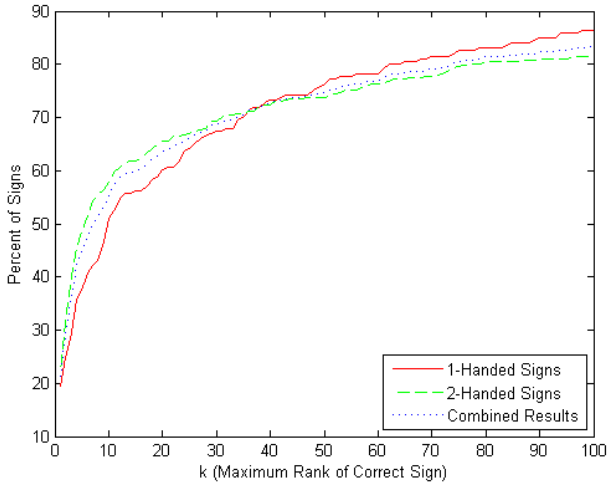

Figure 3: Gesture match accuracy using hand positions from the skeleton tracker.

We also ran the experiments using the manual annotations of the hand locations to establish a best-case scenario for this particular gesture recognition method. Figure 4 shows that 88% of the same two-handed signs rank in the top 20 matches.

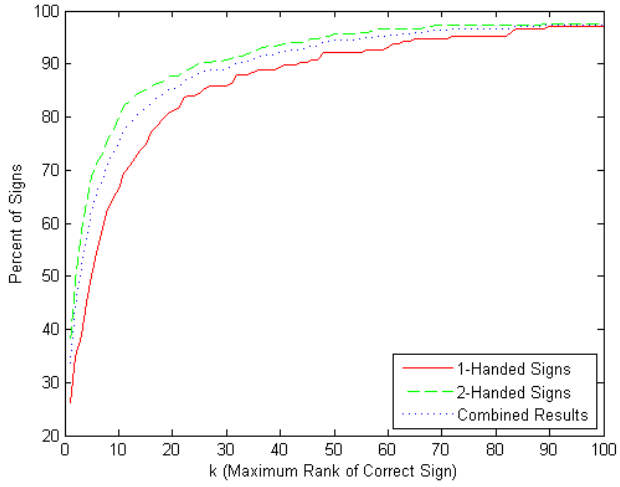

Figure 4: Gesture match accuracy using hand positions from the manual annotations.

Finally, we provide a comparison of gesture match accuracy on all signs using the skeleton tracker and manual annotations in figure 5. It is clear that while the skeleton tracker provides hand position information sufficient to achieve significant accuracy in complex gesture recognition, it is far from accurate enough to approach the best case results of using manual annotations.

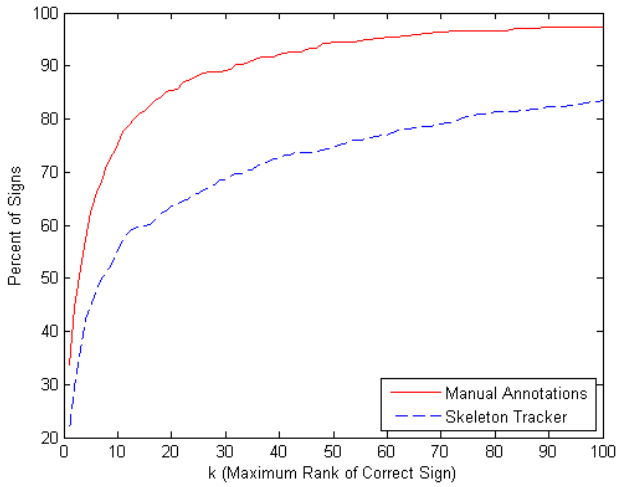

Figure 5: Comparison of gesture match accuracy using hand positions from the manual annotations and the skeleton tracker output.

5. DISCUSSION

When we examine the results and take note of the signs with poor recognition results, two general causes of problems become apparent: large vocabulary gesture similarity and skeleton tracker inaccuracy.

5.1 Gesture Similarity

Many gestures, particularly one-handed gestures, share a similar trajectory, and it can be seen in figures 3 and 4 that the one-handed signs are matched in the top 10 signs at a much lower rate than the two-handed gestures. Many of these signs are stationary gestures, in which the position is approximately the same across signs and only hand shape differs. Figure 6 shows frames from 4 such signs.

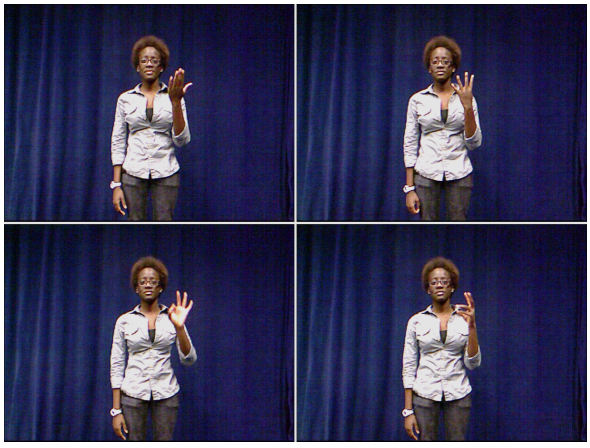

Figure 6: Examples of sign similarity. The position is roughly the same, but the hand shape differs. A skeleton tracker alone is insufficient to distinguish between these signs.

It is clear that the skeleton tracker alone does not output enough information to distinguish between the signs, since

it does not estimate the structure or finger configuration of the hand itself. The incorporation of hand shape or appearance comparison can improve the results, and the use of RGB video from the Kinect may provide that opportunity. However, since there is a disparity in the viewpoints of the depth and RGB cameras due to the physical separation of the sensors, much like with our own eyes, the two frames are not aligned, and there is not a one-to-one pixel correspondence between them. The scale portion of this disparity can easily be seen in the images of figure 1. Registration—the task of aligning the two frames—is not trivial, and the quality of alignment tends to vary with depth. Once aligned, however, one can take advantage of both color and depth information to improve recognition results by incorporating hand appearance into the similarity scores.

5.2 Skeleton Tracker Inaccuracies

It is clear that existing skeleton trackers are not designed for tracking complex and intricate skeletal joint movement. Joint proximity to the body can cause problems. The current depth-based trackers sometimes fail in instances when the hands and arms come into contact with the body, likely due to the limited depth resolution of the Kinect. Signs for which there is no clear separation and obvious distance between the limbs and the body cause the tracker to lose the joints.

In our dataset, when the signer lowers her hands between signs and places her arms at her sides, the tracker often loses lock of the joints as they blend into the mass of the body. When she lifts her arms to perform the next sign, the tracker can take a significant portion of the sign to relocate the joints. Such is the case in figure 7. The green shows the centroid of the manually annotated bounding box, while the red shows the skeleton tracker hand estimate before it relocated the hand position.

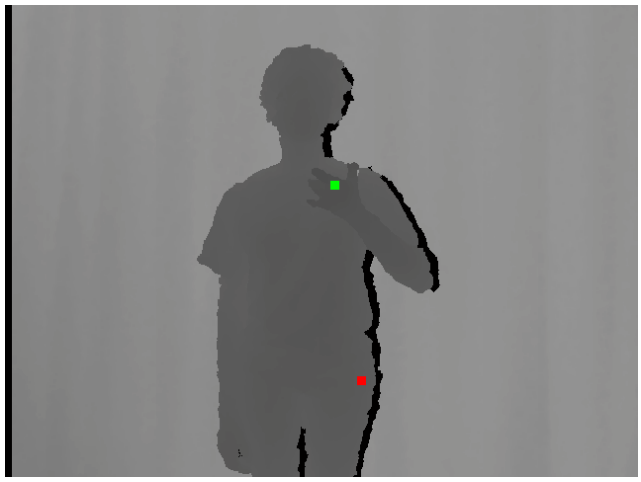

Figure 7: Failure of the skeleton tracker after the signer's arms were at her side. The red square is the tracker hand position estimate. The green square is the centroid of the hand bounding box.

This may not be an issue in a sign language video dictionary system when the user can ensure that the tracker is properly tracking movements before performing the sign.

Gestures in which the arms cross or are oriented toward the camera can also provide considerable difficulty for the skeleton tracker. When arms are oriented along the optical axis of the camera, much of them is self-occluded, and the tracker sometimes has difficulty determining arm joint positions. In the case of crossed arms, the tracker struggles to distinguishing between the arms, the joint position estimates begin to destabilize, and the tracker loses lock on the joints.

There are also joint estimate stabilization issues between frames. Even when the skeleton tracker does not lose track of the joints, the hand position, for example, can jump around the hand from frame to frame, even in a static gesture in which the hand does not move. When part of the feature vector extracted from hand position information includes various directions of motion and changes in those directions from frame to frame, this instability can have a significant effect on scores and recognition accuracy. Though you can apply a smoothing factor to the NiTE skeleton tracker, the smoothing can cause the tracker to be slow to react to changes in motion, thus losing information about joint movement in intricate gestures. Work is clearly needed on the stability of joint position estimates and the responsiveness of tracking to movement.

It is evident that the existing trackers are geared more towards whole body pose estimation and do well in recognizing action poses that use large deliberate movement, such as kicking, jumping, large arm movements, etc [16]. This makes sense, as the Kinect was designed to be part of a gaming system. Besides these full body poses, the only hand gestures it was designed to handle are for simple menu navigation.

Of these issues, improvements in joint tracking when arms and hands are in close proximity to the body or each other could perhaps prove most fruitful.

6. FUTURE WORK

In order to improve gesture recognition accuracy, new hand locating and skeleton tracking methods are clearly needed. Existing trackers provide insufficient hand location accuracy for such a large vocabulary of complex and often subtly different gestures as a sign language. Our lab continues to work toward this goal.

The addition of other joint information beyond the hands and head may also prove useful. With position information about the shoulders and elbows, we can ascertain limb arrangement and orientation that may lead to improved results.

Furthermore, some trackers only provide single candidates for joint locations, along with a confidence level. It can be useful to have a choice of multiple candidates as was shown in [17]. We are currently exploring methods of providing multiple joint position hypotheses.

By making use of color information, we can narrow the number of hand candidates by rejecting those that are not the color of the signer's skin. After proper RGB-D calibration, we can align the depth image to the RGB image and make use of both data modalities. We are currently assessing a number of RGB-D calibration methods.

We are also working on integrating hand shape analysis into our gesture similarity scores. While the current Kinect may not provide sufficient depth resolution at certain distances to extract 3D hand shape information, the new generation of Kinect promises a significantly higher resolution and the possibility of articulated hand tracking. With proper RGB-D calibration, however, we can currently make use of traditional RGB hand shape and appearance methods.

Finally, a promising area of research is recognition in 3D space, rather than on a 2D image plane. Using 3D features and trajectory information can help distinguish between signs with similar 2D trajectory projections but distinct 3D real-world trajectories. This will require scale and spatial orientation normalization and alignment for gestures, since they may be recorded from different angles and heights. We are currently performing research in this area.

7. ACKNOWLEDGEMENTS

This work was partially supported by National Science Foundation grants IIS-1055062, CNS-1059235, CNS-1035913, and CNS-1338118.

8. REFERENCES

[1] A. Agarwal and M. Thakur. Sign language recognition using microsoft kinect. In *Contemporary Computing (IC3), 2013 Sixth International Conference on*, pages 181–185, Aug 2013.

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

[3] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, A. Thangali, H. Wang, and Q. Yuan. Large Lexicon Project: {American Sign Language} Video Corpus and Sign Language Indexing/Retrieval Algorithms. In *Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, pages 11–14, 2010.

[4] C. Conly, P. Doliotis, P. Jangyodsuk, R. Alonzo, and V. Athitsos. Toward a 3d body part detection video dataset and hand tracking benchmark. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, pages 2:1–2:6, New York, NY, USA, 2013. ACM.

[5] H. Cooper, N. Pugeault, and R. Bowden. Reading the signs: A video based sign dictionary. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 914–919. Ieee, Nov. 2011.

[6] 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.

[7] 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.

[8] T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *In Proceedings of the 15th British Machine Vision Conference*, 2004.

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

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

[11] F. Pedersoli, S. Benini, N. Adami, and R. Leonardi. Xkin: an open source framework for hand pose and gesture recognition using kinect. *The Visual Computer*, pages 1–16, 2014.

[12] PrimeSense, LTD. NiTE 2.2.0.11 | OpenNI.

[13] PrimeSense, LTD. OpenNI SDK | OpenNI. http://www.openni.org/openni-sdk/.

[14] I. N. Sandjaja and N. Marcos. Sign Language Number Recognition. In *2009 Fifth International Joint Conference on INC, IMS and IDC*, pages 1503–1508. Ieee, 2009.

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

[16] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2821–2840, 2013.

[17] 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.

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

[19] 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.

[20] V. M. Z. Zhang, W.H. Liu and V. Athitsos. A viewpoint-independent statistical method for fall detection. In *International Conference on Pattern Recognition*, pages 3626–3630, Nov 2012.

[21] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 279–286, New York, NY, USA, 2011. ACM.

[22] J. Zieren and K.-F. Kraiss. Robust person-independent visual sign language recognition. In *Proceedings of the Second Iberian Conference on Pattern Recognition and Image Analysis - Volume Part I*, IbPRIA'05, pages 520–528, Berlin, Heidelberg, 2005. Springer-Verlag.