2014-01-01

A new framework for sign language recognition based on 3D handshape identification ...

A New Framework for Sign Language Recognition based on 3D Handshape Identification and Linguistic Modeling

Mark Dilsizian*, Polina Yanovich*, Shu Wang*, Carol Neidle, Dimitris Metaxas***

*Rutgers University , **Boston University

*110 Frelinghuysen Road, Piscataway, NJ 08854 ,

**Boston University Linguistics Program, 621 Commonwealth Ave., Boston, MA 02215

mdil@cs.rutgers.edu, yanovich@cs.rutgers.edu, shuwang.cbim@cs.rutgers.edu, carol@bu.edu, dnm@rutgers.edu

Abstract

Current approaches to sign recognition by computer generally have at least some of the following limitations: they rely on laboratory conditions for sign production, are limited to a small vocabulary, rely on 2D modeling (and therefore cannot deal with occlusions and off-plane rotations), and/or achieve limited success. Here we propose a new framework that (1) provides a new tracking method less dependent than others on laboratory conditions and able to deal with variations in background and skin regions (such as the face, forearms, or other hands); (2) allows for identification of 3D hand configurations that are linguistically important in American Sign Language (ASL); and (3) incorporates statistical information reflecting linguistic constraints in sign production. For purposes of large-scale computer-based sign language recognition from video, the ability to distinguish hand configurations accurately is critical. Our current method estimates the 3D hand configuration to distinguish among 77 hand configurations linguistically relevant for ASL. Constraining the problem in this way makes recognition of 3D hand configuration more tractable and provides the information specifically needed for sign recognition. Further improvements are obtained by incorporation of statistical information about linguistic dependencies among handshapes within a sign derived from an annotated corpus of almost 10,000 sign tokens.

Keywords: ASL, hand tracking, structured prediction, 3D hand configurations, handshape dependencies

1. Introduction

Despite the fact that monocular 3D hand pose reconstruction from 2D video is an insoluble problem because of the large number of degrees of freedom and high occurrence of self-occlusion of hands and fingers, full 3D reconstruction of hand configuration is critical to the larger task of computer-based sign recognition. For purposes of sign recognition, we can, however, reduce the hand pose space to a finite set of linguistically relevant hand configurations using knowledge of American Sign Language (ASL) phonology. For a large class of ASL "lexical" signs, handshape identification can be further improved by leveraging statistical information about dependencies between start and end handshapes for a given sign and between the handshapes on the two hands (Thangali et al., 2011; Thangali, 2013). By combining the linguistic constraints of ASL with the geometric and kinematic constraints of a 3D model, we achieve tractability.

For 3D hand pose reconstruction, we first generate a synthetic training dataset comprised of a discrete set of hand configurations spanning the set of linguistically important ASL handshapes. We propose a Bayesian mixture of experts (BME) approach that trains a model of each hand configuration based on a set of 2D image features from varying poses. After tracking and segmenting the hand, we map from the 2D features directly to the 3D pose based on our trained model. After initialization, we use phonological and geometric/kinematic priors to inform weights applied to each predictor of the mixture model. This framework makes it feasible to achieve the 3D hand pose reconstruction that provides the information necessary for sign language recognition.

The accuracy of the identification of start and end handshapes for individual ASL signs and, consequently, of sign recognition enabled by handshape recognition, is further improved by utilization of the American Sign Language Lexicon Video Dataset (ASLLVD) (Neidle et al., 2012), which provides statistical information, from a corpus of nearly 10,000 sign tokens, reflecting the linguistic dependencies among handshapes within a given sign for the largest morphological class of ASL signs.

2. Previous work

Some success in sign recognition has been achieved using a variety of methods (e.g., Vogler and Metaxas, 1998; Vogler and Metaxas, 2004; Potamias and Athitsos, 2008; Thangali et al., 2011; Alon et al., 2009; Buehler et al., 2009; Yang et al., 2010; see also Ong and Ranganath, 2005; Von Agris et al., 2008; Vogler and Goldenstein, 2008). In many of the reported studies, the problem has been simplified in various ways, e.g., through restricting attention to limited vocabularies. With the aim of accomplishing large-scale sign language recognition, some researchers have focused on the essential linguistic parameters involved in sign production, including hand configuration, orientation, location in the signing space, and movement trajectory (Ding and Martinez, 2009; Ding and Martinez, 2007; Yuntao and Weng, 2000; Vogler, 2003). See Thangali (2013, chapter 3) for an overview of research focused specifically on recognition of handshapes that are critical to the composition and discrimination of signs.

Hand pose reconstruction methods in 3D have been either generative (model-based) (Heap and Hogg, 1996; Ding and Martinez, 2009; Ding and Martinez, 2007; Lu et al., 2003) or discriminative (appearance-based) (Athitsos and Sclaroff, 2003; Athitsos and Sclaroff, 2001; Yuntao and

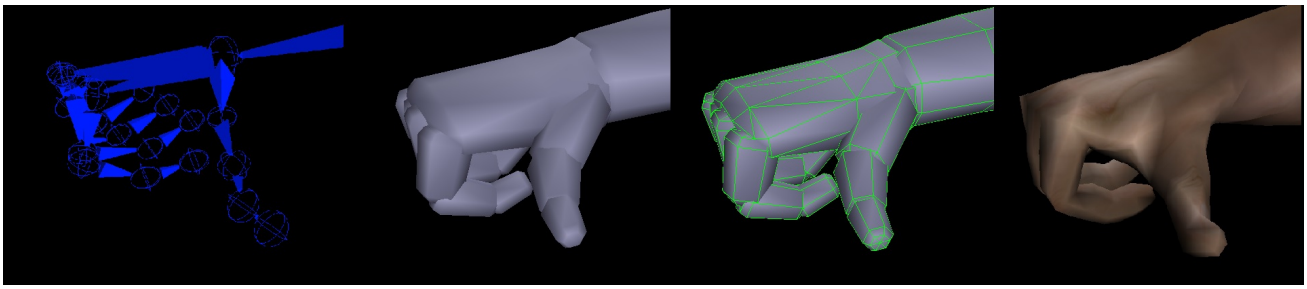

Figure 1: Synthetic Dataset: 3D reconstructions of hand configurations from Cyberglove joint angles.

Weng, 2000). However, both types of methods fail to constrain the problem adequately to deal with hands that have noisy segmentation. In Thangali et al. (2011) handshape recognition performed in 2D is enhanced by exploitation of linguistic constraints. However, accuracy is severely limited by the inability of a 2D model to capture the entire manifold of the 3D pose. Several papers (Potamias and Athitsos, 2008; Athitsos and Sclaroff, 2003; Rosales et al., 2001) attempt to reconstruct 3D hand pose. However, they are limited by using chamfer distance matching, which fails to generalize across different subjects. Such methods are also limited to a relatively small set of hand configurations.

3. Hand Tracking

Accurately locating the hands is a critical preprocessing step for 3D hand pose reconstruction. However, this task itself is challenging, since hands can vary greatly in size, shape, and viewpoint, can be closed or open, can be partially occluded, can have different articulations of the fingers, can be grasping other objects or other hands, etc. Methods based on detecting hands independently using skin detection (Wu and Huang, 2000; Zhu et al., 2000) or Haar-like features (Kolsch and Turk, 2004; Ong and Bowden, 2004; Viola and Jones, 2001) have shown limited success in unconstrained environments, which may be due to lack of training data and insufficient usage of shape information. Some success is achieved by detecting the hand as a part of human pictorial structure (Buehler et al., 2008; Karlinsky et al., 2010; Kumar et al., 2009). However, this method requires that several parts of the human (e.g., head and arms) also be visible in the image. Furthermore, this method can be used to detect only the hand poses for which they have been trained (e.g., they cannot deal with self-occlusion). Furthermore, the computational cost of these methods is considerable, which in turn constrains their application in online tasks. Additionally, in Mittal et al. (2011), a robust hand detection method is proposed that combines skin color detection and shape detectors for both hands and their context. However, this method trains 3 detectors off-line, with high computational cost, and is also not adaptive enough for handshapes with a large number of degrees of freedom.

In this paper, we present a novel tracking framework that decomposes the hand tracking task into tracking, detection, and learning. The tracker makes use of the continuous target information from frame to frame, and the detectors localize all previously known appearances. The learning process updates the online detector with missed detections and false alarms, thereby reducing the detector's confusion on difficult cases in following frames.

4. Hand configuration Synthetic dataset

In order to make it possible to learn the mapping from 2D image features to 3D hand configurations, the training data must contain a large number of diverse examples that include variations in viewing angle, hand size, skin color, and lighting. In order to meet these demands, a synthetic dataset has been generated based on motion capture. Two native signers wearing cybergloves demonstrated 87 handshapes used by the American Sign Language Linguistic Research Project (Neidle, 2011) (leaving out the 10 handshapes that occur rarely, if ever, in our dataset). A comprehensive set of hand orientations was recorded; subjects demonstrated the handshapes across a full range of typical sign language movements.

Next, a 3D hand mesh and texture were fitted to the joint angle data (see Figure 1). The hand was rotated across a discretized set of viewing angles with variations in lighting. For generation of additional data, the mesh was varied for different hand sizes and proportions, and the texture was varied for different skin color and appearance.

5. Structured Pose Prediction

To enable recognition of hand pose from a given image frame, we attempt to map from a vector of image features to the 3D pose. As seen in Figure 2, this pose is represented as a set of 5 finger tip and 15 joint positions in 3D space: 5 Metacarpophalangeal (MCP) joints that connect metacapal bones to the proximal phalanges, 5 proximal interphalangeal joints (PIP) that connect proximal and intermediate phalanges, 4 distal interphalangeal joints (DIP) that connect intermediate and distal phalanges, and a carpometacarpal joint (TCP) that connects the thumb metacarpal bone to the wrist. If we can find a mapping from 2D image features to their corresponding 3D hand structure, we will be able to find a 3D pose for any arbitrary set of 2D image features represented in our training dataset.

In order to find such a mapping, we must use an appropriate regression method. First we must consider that image features are interdependent, because features from an image region tend to be related to other nearby regions. Gaussian Processes have been shown to be effective methods of modeling non-linear dependencies among features to an

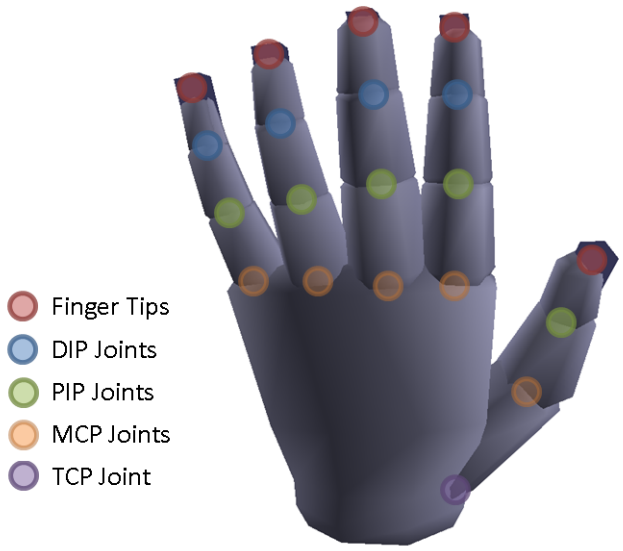

Figure 2: Joints Used to Model the Hand

output (Bo and Sminchisescu, 2010). However, the mapping from features to the manifold of hand poses is non-linear. Fitting a single model to each hand configuration would fail to provide a successful mapping. Furthermore, we know that the multiple outputs (the different joints of the hand pose) are not independent: they have a structure in 3D space, which means that the problem should be posed as a structured prediction problem. Therefore, we use Twin Gaussian Processes (TGP), an approach that has been used successfully on full body tracking, and we apply it to the hands. TGP not only models dependencies among image features, but also captures correlations within the 3D structure (Bo and Sminchisescu, 2010). It uses two Gaussian Processes, one from inputs to outputs, and one representing the reverse mapping. The divergence between the Gaussian Processes are minimized, so that similar observations map to similar poses.

To train our model, we use the synthetic dataset to extract, from each handshape, features and corresponding 3D hand poses. Image features are computed based on block normalized HOG (histogram of oriented gradient) features over multiple scales. These features allow us to capture contours critical to the overall shape and the interior contours of the fingers. We run TGP over multiple examples of each handshape and learn the non-linear mapping function.

6. Assigning Handshape Probabilities

For the purpose of testing our new model, hand locations are first acquired by our tracker from continuous ASL sequences. We segment the hands using skin color and compute the block normalized HOG features. We use our learned TGP function to map these features directly to a pose in 3D Euclidean space. In order to assign this arbitrary 3D hand pose to an ASL handshape, we must compute the distance from our learned pose to clusters that represent each shape. However, in order to attain a more normalized distance measurement that maximizes the distances among clusters, we first transform the pose space into a lower dimensional space.

We use the Spectral Latent Variable Model (SLVM) (Kanaujia et al., 2007) to learn the low-dimensional subspace representing the set of plausible 3D poses. The bi-directional mapping learned using SLVM allows projection of out-of-sample data points from Euclidean space to a lower dimensional latent space, and the back-projection of latent points to the Euclidean space. Because SLVM preserves global and local geometric properties of the modeled data, it is ideally suited to modeling the hand pose space. The lower dimensional latent space of all 87 ASL handshapes can be seen in Figure 3. The convexity of the manifold shows that learning a mapping to the 3D hand pose space is a realistic goal.

Next, we compute the distances from our learned hand pose to each handshape cluster in the lower dimensional latent space. While these distances would allow us to assign a handshape for each frame, it would be helpful to be able to relate shape detections of different frames to one another and to other probabilities, such as statistical distributions that come from language knowledge. Therefore, for each frame, we normalize the distances to each handshape cluster and convert to a probability distribution $P(S|x)$, where S represents the discrete set of 87 possible ASL handshapes for each frame x.

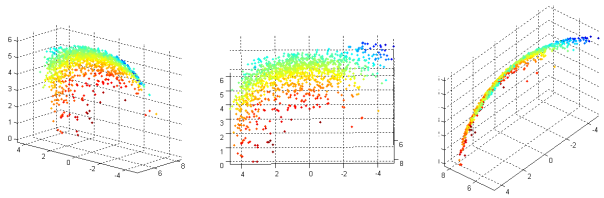

Figure 3: Multiple viewpoint visualization of the non-linear lower dimensional subspace of 87 ASL handshapes of the synthetic dataset

7. Linguistic Constraints

We achieve further improvements by leveraging linguistic constraints that hold on sign production. For the class of signs under consideration in this research, "lexical signs," the relationships between the start and end handshape for both the dominant and non-dominant hands, are highly constrained, as is the relationship between the two hands, in two-handed signs. Note that for sign identification, the handshapes are the most linguistically informative at the beginning and end of such lexical signs.

To leverage these linguistically based dependencies to improve the accuracy of handshape recognition for the start and end handshapes of each of the citation-form signs in the data set, we set out to identify the handshape (on the dominant hand) specifically at the start and end of each sign and to exploit the statistics from the dataset reflecting the dependencies.

For a given signing sequence we must first determine which frames represent the start handshape and which represent the end shape. For signs where the shape does not change,

$$P(S_i|x_1,...,x_{\lfloor \frac{n}{2} \rfloor}) = \prod_{k=1}^{\lfloor \frac{n}{2} \rfloor}(P(S_i|x_k)) \qquad \text{and} \qquad P(S_j|x_{(\lfloor \frac{n}{2} \rfloor+1)},...,x_n) = \prod_{k=\lfloor \frac{n}{2} \rfloor+1}^{n}(P(S_j|x_k)) \tag{1}$$

$$P(S_i, S_j|x_i,...,x_n) = P(S_i|x_1,...,x_{\lfloor \frac{n}{2} \rfloor}) \times P(S_j|x_{\lfloor \frac{n}{2} \rfloor+1},...,x_n) \tag{2}$$

$$\hat{P}(S_i, S_j|x_i,...,x_n, P_o) = \alpha P(S_i, S_j|x_i,...,x_n) + (1-\alpha)P_o(S_i, S_j) \tag{3}$$

this division should not affect the ultimate accuracy computation. Although we do not currently explicitly detect when shape changes occur, we follow a method that allows us to identify the best frames to analyze, for a given sign, in order to determine the start and end hanshapes of that sign. We start by splitting each signing sequence in half. This division is crude because it may not represent an optimal division between start and end handshape frames and it potentially includes intermediary frames where the handshape may be changing. However, we are able to overcome this lack of precision in two ways.

First, we apply a weighting scheme that applies larger weights to frames at the start and end of the sequence and smaller weights to frames in the middle. For signs with shape changes, this is effective, because the most representative observations of handshapes are likely to occur at the start and end of the sequence while the frames in the middle are more likely to include intermediary shapes. We apply an inverse Gaussian membership function where $\sigma = \lfloor \frac{n}{2} \rfloor$, and where n is the number of frames in the sequence.

Second, a filter is used to threshold out frames that have poor observation confidence. The reliability of the observation α is determined based on the normalized distance of a prediction to the nearest handshape cluster in the SLVM reduced space. This confidence metric is often low during handshape changes because during handshape transitions, the 3D configuration is likely to match none of the 87 canonical configurations, and because of motion blur that often accompanies quick movements. This filter has the added benefit of removing frames that may include noisy image data due to motion blur or poor hand tracking. We use a threshold of 80%, meaning we remove the bottom 20% of observations for each sequence.

Once the frames have been split and filtered, we define joint probability distributions over the total of n frames contained in each of the sequences being used to analyze the start frames $P(S_i|x_1,...,x_{\lfloor \frac{n}{2} \rfloor})$ and end frames $P(S_j|x_{(\lfloor \frac{n}{2} \rfloor+1)},...,x_n)$, by computing the joint probability distribution between the start and end segments over each frame x_k (See Equation 1). In Equation 2, we compute the most likely start and end shapes for the sequence by taking the joint probability of the two distributions.

In addition to computing the joint probability between start and end handshapes, we also have prior information about the distributions of start and end handshape co-occurrence in ASL. A matrix of co-occurrence likelihoods $P_o(S_i, S_j)$, based on language statistics, represents probabilities based on the frequencies of all start and end handshape pairings. This prior distribution should be utilized in a way that is proportional to our observation confidence: the better our observation, the less we depend on the prior. Therefore, in Equation 3 we apply the prior using a convex combination of our joint probability $P(S_i, S_j|x_i,...,x_n)$ and our prior $P_o(S_i, S_j)$, with the observation certainty coefficient α.

8. Experimental Results

Initial results are promising. After training on our synthetic dataset of 87 handshapes, we test on the dominant hand of 100 ASL signs from the largest morphological type of signs in ASL: so-called "lexical" signs; these signs are subject to specific types of constraints that hold between handshapes on the left and right hands, as well as between start and end handshape for a given sign. The sample includes a total of 77 handshapes linguistically important for ASL. The sequences cover 5194 frames and four subjects. About 40% of the signs in the test sample involved a handshape change on the dominant hand between the start and the end of the sign, which is approximately representative of the percentage of such signs among the nearly 10,000 tokens in the ASLLVD dataset.

TGP returns a hand pose, and probability distributions are acquired for each frame. However, there is some noise in the data because of frames that include poor hand tracking and/or motion blur in the image. To minimize the effects of such noise, we use a sliding window to smooth over a span of 5 frames. First, each window is assigned a shape based on majority voting of the most probable handshape in each frame of the window. Accuracy (percent of correctly identified handshapes) across all instances was 64.03%. Next, using our probabilistic framework, we compute the joint probabilities over all frames in the window; using this approach, we reach an accuracy of 71.02%. Tracking, segmentation, and recognition results for example frames can be seen in Figure 4. This is a substantial improvement over previous work using the same dataset that reported 32.1% accuracy in handshape recognition (Thangali et al., 2011; Thangali, 2013).

Finally, we test for improvement based on our application of language knowledge. The joint probability distribution from Eqation 2 over the start and end windows leads to an improved overall accuracy of 74.71%. When the start and end co-occurrence prior is applied to the joint distributions, we achieve a final accuracy of 81.76%. Our overall accuracy was already higher, on a per-frame/window basis, than Thangali (2013) achieved for handshape recognition of data from this same dataset (the ASLLVD) even after incorporation of linguistic knowledge. When we combine our handshape recognition with start and end handshape dependency knowledge, our new overall accuracy far exceeds

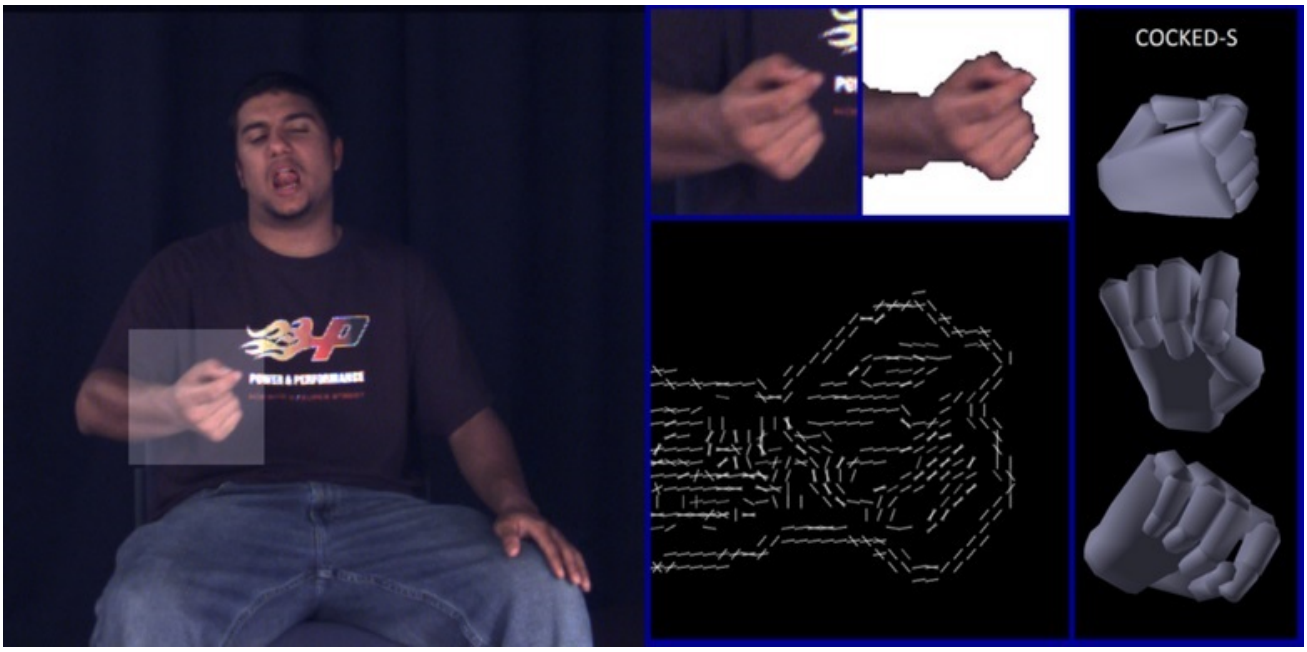

Figure 4: Hand tracking, segmentation, HOG features, and 3D hand configuration result

what has previously been reported. In the work reported by Thangali, the use of linguistic knowledge resulted in an increase in the handshape recognition rate from 30.4% to 44.4% (Neidle et al., 2012). In our approach, accuracy is improved by 15.1%, from 71.02% to 81.76%. Overall, our final accuracy represents an 84.1% improvement over those previously reported results.

9. Conclusion

We have proposed a novel hand tracker and handshape recognition method that enable us to predict 3D handshapes from monocular imagery. Recognition is achievable because we limit the subspace of handshapes to those that are part of the ASL inventory. Accuracy is improved further by leveraging linguistically constrained handshape dependencies. Our results already show significant improvements over previous attempts to recognize handshapes in ASL. In the future, additional linguistic and kinematic constraints can be leveraged for further refinement of the handshape recognition.

10. Acknowledgments

For the handshape synthetic dataset, we thank Matt Huenerfauth and the signing participants. The ASLLVD was constructed in collaboration with Ashwin Thangali and Stan Sclaroff (Neidle et al., 2012). Thangali also developed an excellent interface to facilitate verifications of the linguistic annotations and viewing and searching the dataset, which has greatly facilitated our work. We also thank Vassilis Athitsos, Ben Bahan, Christian Vogler, Iryna Zhuravlova, and the many signers, students, and annotators who have contributed to that project, including especially Naomi Berlove, Elizabeth Cassidy, Lana Cook, Braden Painter, Tyler Richard, Dana Schlang, Robert G. Lee, Joan Nash, Tory Sampson, Donna Riggle, Amelia Wisniewski-Barker, and Rachel Benedict, and annotators who have contributed to that project. The work reported here has been partially funded by grants from the National Science Foundation (#IIS-1065013, #IIS-0964385, #EIA-9528985, #CNS-1059218).

11. References

Alon, J., Athitsos, V., Yuan, Q., and Sclaroff, S. (2009). A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1685–1699.

Athitsos, V. and Sclaroff, S. (2001). 3D Hand Pose Estimation by Finding Appearance-based Matches in a Large Database of Training Views. Technical report, Boston University Computer Science Department.

Athitsos, V. and Sclaroff, S. (2003). Estimating 3D Hand Pose from a Cluttered Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–432. IEEE.

Bo, L. and Sminchisescu, C. (2010). Twin Gaussian Processes for Structured Prediction. In *International Conference on Computer Vision (ICCV)*, volume 87, pages 28–52.

Buehler, P., Everingham, M., Huttenlocher, D., and Zisserman, A. (2008). Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts. In *Proceedings of the 19th British Machine Vision Conference*, pages 1105–1114. BMVA Press.

Buehler, P., Everingham, M., and Zisserman, A. (2009). Learning Sign Language by Watching TV (using Weakly Aligned Subtitles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ding, L. and Martinez, A. (2007). Recovering the Linguistic Components of the Manual Signs in American Sign Language. In *Conference on Advanced Video and Signal*

Based Surveillance, 2007. AVSS 2007, pages 447–452. IEEE.

Ding, L. and Martinez, A. (2009). Modelling and Recognition of the Linguistic Components in American Sign Language. *Image and Vision Computing*, 27(12):1826–1844.

Heap, T. and Hogg, D. (1996). Towards 3D Hand Tracking using a Deformable Model. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition.*, pages 140–145. IEEE.

Kanaujia, A., Sminchisescu, C., and Metaxas, D. (2007). Spectral Latent Variable Models for Perceptual Inference. In *International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE.

Karlinsky, L., Dinerstein, M., Harari, D., and Ullman, S. (2010). The Chains Model for Detecting Parts by their Context. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25–32. IEEE.

Kolsch, M. and Turk, M. (2004). Robust Hand Detection. In *International Conference on Automatic Face and Gesture Recognition*, pages 614–619.

Kumar, M., Zisserman, A., and Torr, P. (2009). Efficient Discriminative Learning of Parts-based Models. In *International Conference on Computer Vision (ICCV)*, pages 552–559. IEEE.

Lu, S., Metaxas, D., Samaras, D., and Oliensis, J. (2003). Using Multiple Cues for Hand Tracking and Model Refinement. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–443. IEEE.

Mittal, A., Zisserman, A., and Torr, P. (2011). Hand Detection using Multiple Proposals. In *British Machine Vision Conference*.

Neidle, C., Thangali, A., and Sclaroff, S. (2012). Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, Istanbul, Turkey. LREC.

Neidle, C. (2011). Movies of Handshapes used in American Sign Language from Different Views.

Ong, E. and Bowden, R. (2004). A Boosted Classifier Tree for Hand Shape Detection. In *Proceedings of the Sixth International Conference on Automatic Face and Gesture Recognition*, pages 889–894. IEEE.

Ong, S. and Ranganath, S. (2005). Automatic Sign Language Analysis: A Survey and the Future Beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891.

Potamias, M. and Athitsos, V. (2008). Nearest Neighbor Search Methods for Handshape Recognition. In *Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments*, page 30. ACM.

Rosales, R., Athitsos, V., Sigal, L., and Sclaroff, S. (2001). 3D Hand Pose Reconstruction using Specialized Mappings. In *International Conference on Computer Vision (ICCV)*, volume 1, pages 378–385. IEEE.

Thangali, A., Nash, J., Sclaroff, S., and Neidle, C. (2011). Exploiting Phonological Constraints for Handshape Inference in ASL Video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 521–528. IEEE.

Thangali, A. (2013). *Exploiting Phonological Constraints for Handshape Inference in ASL Video*. Ph.D. thesis, Boston University.

Viola, P. and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511. IEEE.

Vogler, C. and Goldenstein, S. (2008). Toward Computational Understanding of Sign Language. *Technology and Disability*, 20(2):109–119.

Vogler, C. and Metaxas, D. (1998). ASL Recognition Based on a Coupling between HMMs and 3D Motion Analysis. In *International Conference on Computer Vision (ICCV)*, pages 363–369. IEEE.

Vogler, C. and Metaxas, D. (2004). Handshapes and Movements: Multiple-channel American Sign Language Recognition. In *Gesture-Based Communication in Human-Computer Interaction*, pages 247–258. Springer.

Vogler, C. (2003). *American Sign Language Recognition: Reducing the Complexity of the Task with Phoneme-based Modeling and Parallel Hidden Markov Models*. Ph.D. thesis, University of Pennsylvania.

Von Agris, U., Zieren, J., Canzler, U., Bauer, B., and Kraiss, K.-F. (2008). Recent Developments in Visual Sign Language Recognition. *Universal Access in the Information Society*, 6(4):323–362.

Wu, Y. and Huang, T. (2000). View-independent Recognition of Hand Postures. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 88–94. IEEE.

Yang, R., Sarkar, S., and Loeding, B. (2010). Handling Movement Epenthesis and Hand Segmentation Ambiguities in Continuous Sign Language Recognition Using Nested Dynamic Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):462–477.

Yuntao, C. and Weng, J. (2000). Appearance-based Hand Sign Recognition from Intensity Image Sequences. *Computer Vision and Image Understanding*, 78(2):157–176.

Zhu, X., Yang, J., and Waibel, A. (2000). Segmenting Hands of Arbitrary Color. In *Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition*, pages 446–453. IEEE.