The Importance of 3D Motion Trajectories for Computer-based Sign Recognition

Mark Dilsizian*, Zhiqiang Tang*, Dimitris Metaxas*, Matt Huenerfauth, and Carol Neidle*****

*Rutgers University , **Rochester Institute of Technology, ***Boston University,
110 Frelinghuysen Road, Piscataway, NJ 08854 ,
*Golisano College of Computing and Information Sciences, 152 Lomb Memorial Drive, Rochester, NY 14623
**Boston University Linguistics Program, 621 Commonwealth Ave., Boston, MA 02215
mdil@cs.rutgers.edu, zt53@cs.rutgers.edu, dnm@rutgers.edu, matt.huenerfauth@rit.edu, carol@bu.edu

Abstract

Computer-based sign language recognition from video is a challenging problem because of the spatiotemporal complexities inherent in sign production and the variations within and across signers. However, linguistic information can help constrain sign recognition to make it a more feasible classification problem. We have previously explored recognition of linguistically significant 3D hand configurations, as start and end handshapes represent one major component of signs; others include hand orientation, place of articulation in space, and movement. Thus, although recognition of handshapes (on one or both hands) at the start and end of a sign is essential for sign identification, it is not sufficient. Analysis of hand and arm movement trajectories can provide additional information critical for sign identification. In order to test the discriminative potential of the hand motion analysis, we performed sign recognition based exclusively on hand trajectories while holding the handshape constant. To facilitate this evaluation, we captured a collection of videos involving signs with a constant handshape produced by multiple subjects; and we automatically annotated the 3D motion trajectories. 3D hand locations are normalized in accordance with invariant properties of ASL movements. We trained time-series learning-based models for different signs of constant handshape in our dataset using the normalized 3D motion trajectories. Results show significant computer-based sign recognition accuracy across subjects and across a diverse set of signs. Our framework demonstrates the discriminative power and importance of 3D hand motion trajectories for sign recognition, given known handshapes.

Keywords: ASL, hand tracking, sign recognition, sign motion trajectory estimation

1. Introduction

Recognizing a large set of ASL signs is a difficult challenge when posed strictly as a computer vision classification problem. Classification would require vast amounts of training data representing a range of subject-specific signing variations. However, top-down linguistic knowledge imposed on the data analysis can help constrain the problem in order to make learning and sign recognition more feasible.

We have previously achieved high accuracy with respect to handshape recognition from video (Dilsizian et al., 2014). However, for frequently occurring combinations of start and end handshapes, there are large numbers of signs that have those handshapes in common. In the current study, the set of 3D hand configurations has been limited to a set of linguistically important ASL handshapes appropriate for sign recognition.

We demonstrate here that analysis of movement trajectories allows us to achieve high rates of accuracy in discriminating among signs, holding the start and end handshape constant. Thus we expect that combining the techniques reported here with our prior work on handshape recognition will allow us to achieve high accuracy in identification of specific signs.

2. Related Work

Sign recognition has been approached by Vogler et al. (Vogler and Metaxas, 1998; Vogler and Metaxas, 2003) as a time-series modeling problem using Hidden Markov Models (HMMs) over 3D hand models. However, this work is limited to a small vocabulary and laboratory conditions

Other works attempt to recognize signs from real world video. The work in (Ding and Martinez, 2007; Ding and Martinez, 2009) attempts to incorporate modeling of motion trajectories with face and hand configuration recognition. However these works are limited to 2D trajectories and fail to build a stochastic model of the sign that can leverage phonological constraints or inter-subject variation.

Cui et al. (Cui and Weng, 2000) monitor changes in hand observations over time in an attempt to capture *spatiotemporal events*; signs are then classified with respect to these events. In addition, (Buehler et al., 2009) recognize signs by matching windowed video sequences. Although, some success has been achieved, sign recognition research to date has failed to model different components of signs in order to fully leverage important linguistic information.

Some works focus entirely on handshape recognition as an intermediate step to sign recognition. Handshapes are recognized in 2D using nearest neighbor classification

in (Potamias and Athitsos, 2008). (Thangali et al., 2011) achieve improvements in handshape recognition by modeling phonological constraints between start and end handshapes, but handshape estimation is limited to 2 dimensions. The handshape model is extended to 3 dimensions with significant improvement in handshape recognition accuracy in (Dilsizian et al., 2014). While these works show good recognition accuracy for handshape, this research has not yet been extended to full sign recognition/identification because of the existence of potentially large numbers of signs with the same start and end handshape pairs.

Although handshape-dependent upper body trajectories have not been previously explored in the literature, 3D human pose and upper-body estimation has been studied extensively. Several generative (Isard and Blake, 1998; Deutscher et al., 2000; Sigal et al., 2004; Bălan et al., 2007) as well as discriminative (Rosales and Sclaroff, 2001; Sminchisescu et al., 2007; Agarwal and Triggs, 2004; Sigal et al., 2007) methods exist for 3D human pose prediction. These works attempt to model multi-valuedness (ambiguities) in mappings from 2D images to 3D poses (Rosales and Sclaroff, 2001; Sminchisescu et al., 2007; Sigal et al., 2007) and employ coarser, global features (Agarwal and Triggs, 2004; Sminchisescu et al., 2007; Sigal et al., 2007) such as silhouette shapes, to generalize the trained models to different scenarios.

Alternatively, (Ferrari et al., 2008) proposed an algorithm to localize upper body parts in 2D images using a coarse-to-fine approach. Humans are coarsely detected using current human detectors. Foreground is extracted within the bounding box using grabcut. The work uses edge-based soft detectors (Yang and Ramanan, 2013) to first detect the torso and head and other parts. The appearance is learned from the detected parts and used to detect further parts using a MAP optimization. The method is extended to spatiotemporal parsing. Anthropometric priors have been extensively applied to constrain this problem.

However, both discriminative methods and the 2D-part based approaches are highly dependent on the use of training data. Because very little 3D upper body trajectory data of ASL signing exists, we are unable to sufficiently train state-of-the-art pose estimation methods.

3. 3D Hand Tracking Dataset

As is well known, along with handshapes, orientation, and place of articulation in space, movement trajectories are an essential component of signs, and thus computer-based recognition of motion patterns is essential for automatic sign recognition. In order to test the ability of a computer vision system to access this discriminative information, we recorded a dataset of 3D upper body motion trajectories across multiple signs and subjects, holding handshapes constant.

3.1. Data Collection

ASL signers

Five ASL signers were recruited on the campus of the Rochester Institute of Technology (home of the National Technical Institute of the Deaf) and from the surrounding community in Rochester, NY, using social media advertising. The participants included 2 men and 3 women, ages 21-32 (mean 24.2). Participants were recorded in a video studio space using a Kinect™ v2 camera system and custom recording software developed at Rutgers University, as described below. A total of 3,627 sign productions were recorded (about 25 tokens each of 139 distinct signs). Because of time limitations, however, for this paper data from 2 signers were prioritized for processing and analysis. The entire set of subjects will be analyzed and discussed in the LREC presentation.

Recording of Motion Trajectories

The Microsoft Kinect™ v2 provides a robust platform for recording 3D upper body joint configurations combined with calibrated 2D color video data. We developed a tool for recording and automatic annotation of joint locations for different ASL signs (see Figure 1).

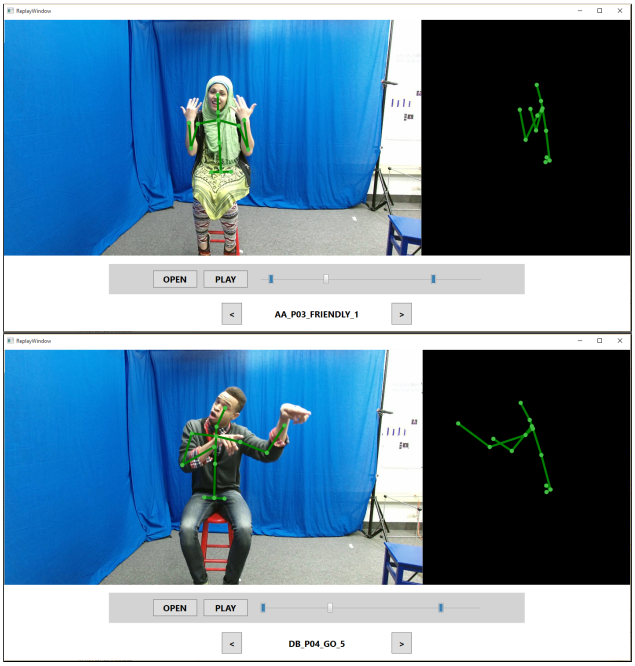


Figure 1: ASL trajectory recording software developed to capture a dataset of 3D ASL movements.

Stimuli

We considered the most common handshapes for 2-handed signs with the same handshapes on both hands throughout the sign production. The signers were recorded as they reproduced two-handed ASL signs shown to them in a video recording of signs from the ASLLVD data set (Nei-

dle et al., 2012) (http://secrets.rutgers.edu/dai/queryPages/) with one of three common handshapes used at both the beginning and end of the sign (the B-L, 1, and 5) varying in their motion trajectories.[1] The B-L, 1, and 5 handshapes are illustrated in Figure 2.

3.2. 3D Tracking and Refinement

Because tracking from the Kinect™ v2 sensor is based on a trained discriminative model (Shotton et al., 2013), it is optimized for average case performance. In order to capture subtle discriminative cues in the motion, we refine the output of the camera by taking a cloud of neighboring depth points around each predicted joint location. We constrain each joint to lie near the center-of-mass of its neighborhood. We also smooth these predictions using a Kalman filter.

4. Sign Classification

In order to train a model for the trajectories of different signs, we must ensure that our modeling is invariant to several factors: (1) variation in sign production (*signing style*); (2) variations in body proportion between different subjects; and (3) noise in 3D tracking data.

4.1. Normalizing Motion Trajectories

Improved invariance to different anthropomorphic proportions and ranges of movement can be achieved by normalizing the 3D motion trajectories. First, trajectories are transformed to a common world coordinate system by computing joint locations as the relative distance from the *root* position (located approximately between the hips).

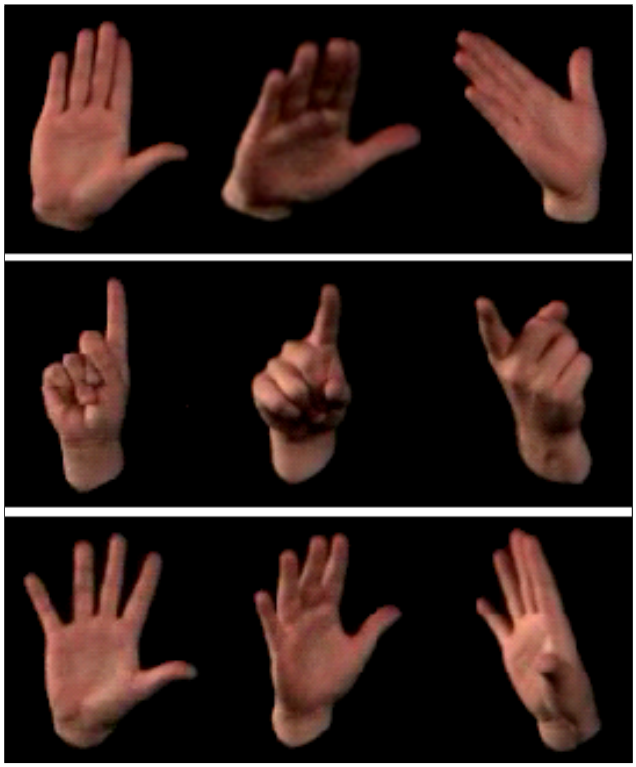


Figure 2: The B-L *(top)*, 1 *(middle)*, and 5 *(bottom)* handshapes.

Second, since it is also important that our model be invariant to differences in the ranges of movement across different subjects. Rather than normalizing according to the overall movement of each trajectory, we normalize over the average range of both left and right hands per subject; this ensures that we preserve the relative range of movement between the left and right hands. An example is shown in Figure 3 for the sign DARK. The bottom row shows the significant reduction of the variance between 2 subjects that results from use of our normalization methodology.

4.2. Training Sign Trajectory Models

In order to overcome noise in 3D hand tracking and variations in signing style, we must learn a robust model that avoids over-fitting to noise or insignificant variation.

The Hidden Markov Model (HMM) has been very popular in the machine learning community and widely applied to speech recognition, part-of-speech tagging, handwriting recognition, bioinformatics, and ASL recognition (Vogler and Metaxas, 1998). HMMs assume that a signing sequence is a Markov process describing how hand locations change through the sign production. A number of states are used to represent different parts of the signing action. These states are not directly visible. Instead, they are perceived

[1] Signs were generated by 5 subjects performing approximately 5 examples of each of the ASL signs glossed in the BU ASLLRP corpora (Neidle et al., 2012) (http://secrets.rutgers.edu/dai/queryPages/) as follows: (1)GO-STEADY++, (1)WHEELCHAIR, (1h)HAPPY, (5)WEATHER, (Vulcan)FILE, ABSTRACT+, AFTERNOON, AGREE, ALLERGY, ALL-RIGHT, ALSO, ANSWER, APPLAUSE, ARRIVE, AVERAGE, BALANCE, BEACH, BECOME, BELOW, BETWEEN/SHARE, BLOOD, BOAT, BOIL, BOTHER++, BOX_2, BREAK-DOWN-BUILDING, BRING-1p, BUT, CALM-DOWN, CHEAP, CHILD, CLOSE-WINDOW, COME, CONFLICT/INTERSECTION, COOKING, COOL, CORRECT, CRACK, CYCLE, DEAF-APPLAUSE, DEPEND, DIE, DISAGREE, DIVE, DONT, DURING/WHILE, EASY+, EMBARRASS, END, EVERY-MONTH/RENT, FALL-INTO-PLACE, FAT, FINALLY, FINGERS, FIRE, FOCUS/NARROW, FOOTSTEP, FRESHMAN_3, FRIENDLY, GENERAL, GENERATIONS-AGO, GLORY, GLOVES_2, GO, GRAY, HALL, HANDS, HARP, HERE+, HUMBLE, INSPIRE, JUNIOR_3, KNIFE, LAPTOP, LEAVE-THERE, LIFT, LOUDSPEAKER, MARCHING, MAYBE, MERGE/MAINSTREAM, MOOSE, MOTIVATE, MUSIC, NECKLACE, NEXT-TO, NOISE, OBSCURE, OFTEN+++, ONE-MONTH, OPPOSITE, PANCAKE, PARALLEL, PERSON, PIMPLES, PLEASE/(1h)ENJOY, POPE, PREGNANT, PROGRESS++, PSYCHOLOGY, PUSH, RAIN, REFLECT, REJECT, REQUEST, ROAD, SAD, SCARE, SENIOR, SIGN, SKYSCRAPER, SLOW, SMILE, SOCKS, SPANK, SPIN, STAR, STEEP, STOP, SUCCEED, SUNDAY_2, SWIM, TAP-DANCE, THING, THROAT-HURT, TORNADO, TRAFFIC, TRAVEL, VACATION, VARY, WAIST, WALK, WASH-DISH, WASH-HANDS, WATER-RISE, WEAVE, WHAT, WHEN, WIND, WRAP.

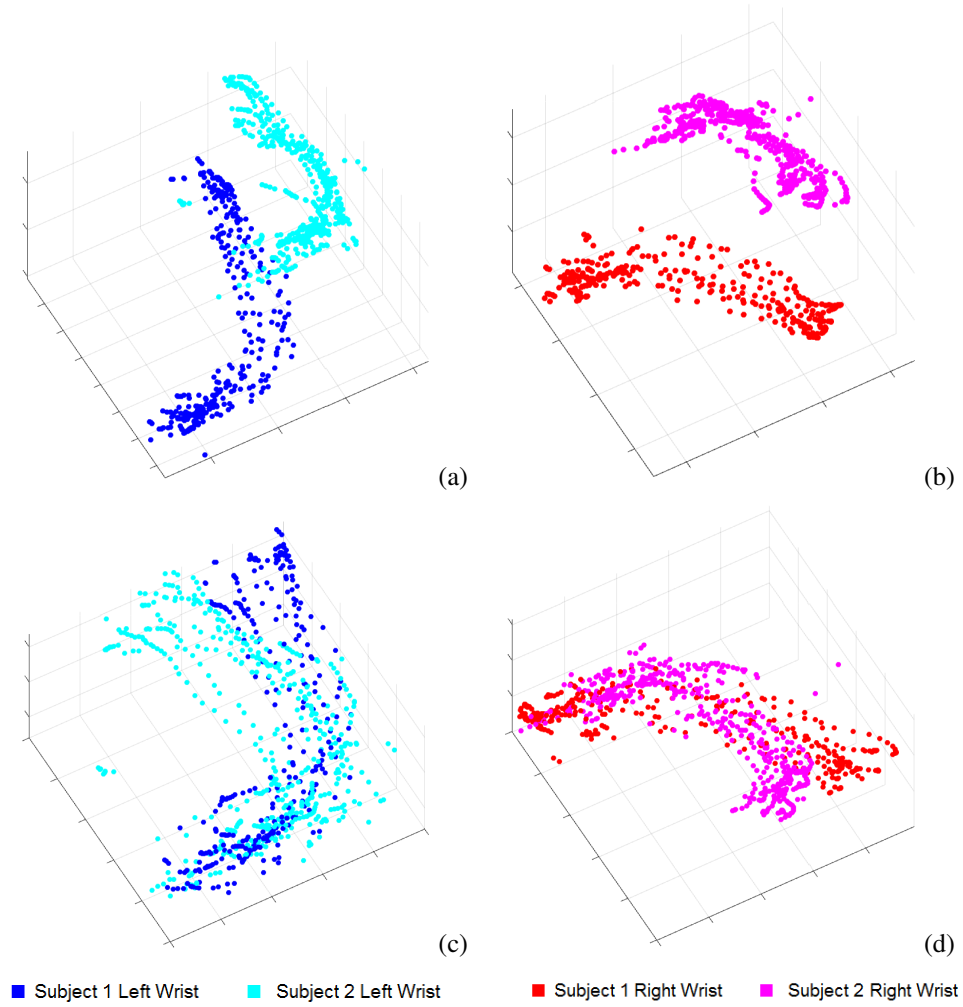


Figure 3: 3D wrist trajectories ({X,Y,Z} Euclidean locations) comparing multiple productions of the ASL sign glossed as DARK by each of two signers. The *top row*, (a) and (b), shows the original data space with evident variations between subjects with respect to sign production and anthropomorphic proportions. The *Bottom Row*, (c) and (d), shows the normalized data space which maximizes inter-subject overlap of trajectories. *Note: the PDF file of this paper contains interactive 3D content accessible by clicking on the figure.*

indirectly through depth image observations. An observation likelihood distribution models the relationship between the states and the observation.This likelihood distribution is represented by a mixture-of-Gaussian(MoG) density function, which is a combination of several Gaussian distribution components. Based on the previous state and the current observation, the HMM may switch from one state to another. During training, the number of states and the number of components in the mixture-of-Gaussian likelihood distribution are chosen using a model selection method known as the Bayesian Information Criterion (BIC). This BIC technique selects the optimal model that best describes the statistics of the training data while avoiding over-fitting. Therefore, using using the BIC technique allows for improved generalization to previously unseen test data.

In order to classify a given sign, we train a Support Vector Machine Hidden Markov Model (SVM-HMM) (Altun et al., 2003). The SVM-HMM is a discriminative sequence labeling model that combines the advantages of HMM and SVM by assuming Markov chain dependency structure between labels and using dynamic programming for optimal inference and learning. At the same time, the learning is based on a discriminative, maximum margin principle that can account for overlapping features. Moreover, unlike HMMs, it can learn the non-linear discriminant functions using kernel-based inputs. An SVM-HMM is trained for each sign which can best be discriminated from all other motion trajectories. This model implicitly captures properties of the motion that are invariant across different examples of the same sign.

5. Results

We train an SVM-HMM for each sign (with constant handshape) and use cross-validation and a two-tailed significance test to determine the parameters (states and Gaussian mixture components) of our SVM-HMMs. Sign labels are

assigned to each test sequence according to the SVM-HMM that returns the minimum log-likelihood indicating that the sequence belongs to a trained sign trajectory.

Despite the fact that the sample included some signs with relatively similar motion patterns, we were able to discriminate among these signs with an average of 78.0% accuracy (with cross-validated 50/50 training/testing split). Accuracy by handshape is shown in Table 1.

Handshape	Signs Trained/Tested	Accuracy (%)
B-L	67	75.7
1	35	80.2
5	37	80.3

Table 1: Percent accuracy and number of signs trained and tested (5-10 examples per subject)

While initial results leave some room for improvement, the correct sign classification is located in the top 3 ranked estimations in 96.1% of test examples. We have only used data thus far from two of the subjects. As additional subjects are incorporated into the SVM-HMM model, more general and robust discrimination should be possible. Moreover, additional information from the upper body tracking (i.e. limb locations, body leaning, etc.) can be integrated to improve recognition rates. Overall, the trajectory classification results suggest that a complete sign language recognition framework is feasible when this approach is combined with previously demonstrated handshape recognition.

In order to test the robustness of our modeling, we also tested with different percentages of training and testing splits ($10-90\%$) using cross-validation on 30 common B-L signs. Results across different sized training sets are shown in Figure 4. The stability in sign recognition accuracy even for low percentages of training data suggests that our approach is scalable and can discriminate among signs even when trained on a small set of examples. This is a necessary and critical property to any framework that seeks to scale to a significantly large set of signs and variations.

6. Conclusion

We show here that modeling movement trajectories of the hands provides important information that can be combined with previously demonstrated handshape recognition for purposes of discriminating among ASL signs. We chose a sample of 139 signs that have the most common combination of start and end handshape for 2-handed signs (i.e., signs that use the so-called B-L, 1, and 5 handshapes) throughout the articulation of the sign. We demonstrate a framework and methodology for classifying signs according to 3D motion trajectories.

The next step is to extend this method to construct a full

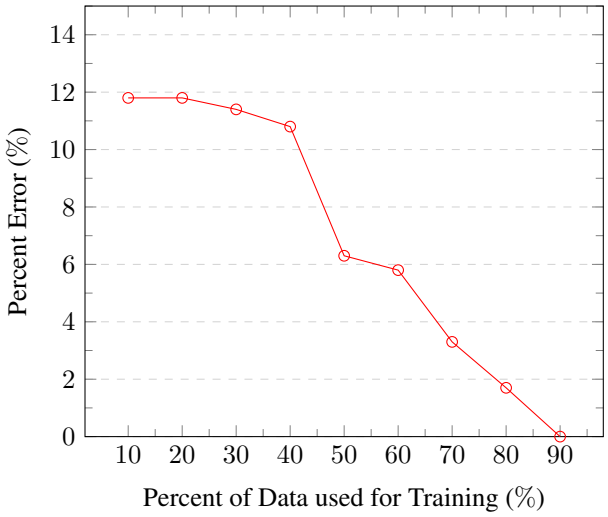


Figure 4: Sign recognition error rates across 30 signs (2-handed B-L handshape) and 2 subjects for different sized training and testing sets.

system for sign recognition/identification from video based on a combination of the methods that we have developed for (1) handshape recognition and (2) analysis of motion trajectories. We plan to report on the extension of these preliminary results to larger sets of signs with varying handshapes and motion trajectories, and larger numbers of signers, in the LREC presentation.

7. Bibliographical References

Agarwal, A. and Triggs, B. (2004). 3D Human Pose from Silhouettes by Relevance Vector Regression. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–882. IEEE.

Altun, Y., Tsochantaridis, I., Hofmann, T., et al. (2003). Hidden Markov Support Vector Machines. In *ICML*, volume 3, pages 3–10.

Bălan, A. O., Sigal, L., Black, M. J., Davis, J. E., and Haussecker, H. W. (2007). Detailed Human Shape and Pose from Images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE.

Buehler, P., Zisserman, A., and Everingham, M. (2009). Learning Sign Language by Watching TV (using Weakly Aligned Subtitles). In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2961–2968. IEEE.

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

Deutscher, J., Blake, A., and Reid, I. (2000). Articulated Body Motion Capture by Annealed Particle Filtering. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 126–133. IEEE.

Dilsizian, M., Yanovich, P., Wang, S., Neidle, C., and Metaxas, D. N. (2014). A New Framework for Sign

Language Recognition based on 3D Handshape Identification and Linguistic Modeling. In *LREC*, pages 1924–1929.

Ding, L. and Martinez, A. M. (2007). Recovering the Linguistic Components of the Manual Signs in American Sign Language. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 447–452. IEEE.

Ding, L. and Martinez, A. M. (2009). Modeling and Recognition of the Linguistic Components in American Sign Language. *Image and vision computing*, 27(12):1826–1844.

Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008). Progressive Search Space Reduction for Human Pose Estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE.

Isard, M. and Blake, A. (1998). Condensation - conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 29(1):5–28.

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, LREC 2012, Istanbul, Turkey*.

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. and Sclaroff, S. (2001). Learning Body Pose via Specialized Maps. In *Advances in Neural Information Processing Systems*, pages 1263–1270.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time Human Pose Recognition in Parts from Single Depth Images. *Communications of the ACM*, 56(1):116–124.

Sigal, L., Bhatia, S., Roth, S., Black, M. J., and Isard, M. (2004). Tracking Loose-limbed People. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–421. IEEE.

Sigal, L., Balan, A., and Black, M. J. (2007). Combined Discriminative and Generative Articulated Pose and Non-rigid Shape Estimation. In *Advances in Neural Information Processing Systems*, pages 1337–1344.

Sminchisescu, C., Kanaujia, A., and Metaxas, D. N. (2007). BM^3E: Discriminative Density Propagation for Visual Tracking. *Pattern Analysis and Machine Intelligence (PAMI)*, 29(11):2030–2044.

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

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

Vogler, C. and Metaxas, D. (2003). Handshapes and Movements: Multiple-channel American Sign Language Recognition. In *Gesture workshop*, volume 2915, pages 247–258. Springer.

Yang, Y. and Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2878–2890.