Detection of Major ASL Sign Types in Continuous Signing for ASL Recognition

Polina Yanovich, Carol Neidle, Dimitris Metaxas
Rutgers, The State University of New Jersey, Computer Science
Boston University, Linguistics Program
110 Frelinghuysen Rd., Piscataway NJ, 08854
621 Commonwealth Ave., Boston, MA 02215
yanovich@cs.rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu

Abstract

In American Sign Language (ASL) as well as other signed languages, different classes of signs (e.g., lexical signs, fingerspelled signs, and classifier constructions) have different internal structural properties. Continuous sign recognition accuracy can be improved through use of distinct recognition strategies, as well as different training datasets, for each class of signs. For these strategies to be applied, continuous signing video needs to be segmented into parts corresponding to particular classes of signs. In this paper we present a multiple instance learning-based segmentation system that accurately labels 91.27% of the video frames of 500 continuous utterances (including 7 different subjects) from the publicly accessible NCSLGR corpus <http://secrets.rutgers.edu/dai/queryPages/> (Neidle and Vogler, 2012). The system uses novel feature descriptors derived from both motion and shape statistics of the regions of high local motion. The system does not require a hand tracker.

Keywords: continuous signing, ASL recognition, sign classification

1. Introduction

Computer-based ASL recognition often focuses on signs that have been pre-extracted from video (with known start and end frames) (Elons et al., 2013a; Mohandes et al., 2012, e.g.). Learning methods are typically employed for sign recognition; these are frequently based on Hidden Markov Models (HMMs) (Vogler and Metaxas, 2004) or Conditional Random Fields (CRFs) (Wang et al., 2006), trained for each sign in the vocabulary.

However, continuous signing presents a significantly greater challenge, not only because of co-articulation effects, but also because of the existence of multiple classes of signs with fundamentally different internal composition. Unlike spoken languages, in which phonemes are concatenated, in signed languages the components of words combine both sequentially and non-sequentially (simultaneously), subject to different linguistic constraints depending on the morphological class to which the signs belong, and these linguistic constraints can be leveraged to improve computer-based recognition (Athitsos et al., 2010; Thangali et al., 2011). The three most prevalent classes in signed languages are lexical signs, fingerspelled signs, and classifier constructions.

Lexical sign production involves combinations of specific hand configurations, orientations, locations in signing space, and movement trajectories. Strict linguistic constraints govern the relationships between start and end handshapes of a given sign, and between the two hands (in 2-handed signs) with respect to hand configuration and movement trajectory (Battison, 1978; Brentari, 1998; Thangali et al., 2011).

Fingerspelled signs consist of sequences of letter handshapes from the manual alphabet, produced with rapid finger movements at a fairly constant global hand location (Athitsos et al., 2010), potentially with relatively small left-to-right movement. Although some fingerspelled signs are in frequent usage, fingerspelling is often used for proper names and spoken language borrowings. Thus there is no fixed fingerspelled vocabulary: many fingerspelled productions would not be included in any ASL dictionary.

Classifier constructions incorporate substantial variability in their realizations (Emmorey, 2013, e.g.).

In current work, we omit from consideration so-called name signs, loan signs (which originated as fingerspelled, but have become more like lexical signs), and index signs (used for pronominal reference).

In light of the fundamentally different nature of these sign classes, distinct, class-specific, recognition strategies are needed.

2. Previous work

Techniques for recognition from continuous signing based on HMMs (Assaleh et al., 2008; Theodorakis et al., 2012; Vogler and Metaxas, 2001; Kong and Ranganath, 2014), CRFs (Yang and Lee, 2013), or intelligent search (Gao et al., 2004; Elons et al., 2013b; Sarkar et al., 2011) can be applied to entire sentences or individual signs (pre-segmented from sentences in a preprocessing step). Some approaches also exploit movement epenthesis (between signs) for sign recognition (Gao et al., 2004) or segmentation (Kong and Ranganath, 2014; Yang and Sarkar, 2006). Other researchers have attempted semi-supervised (Theodorakis et al., 2012; Madeo et al., 2012; Bowden et al., 2004) or unsupervised (Han et al., 2013; Cooper et al., 2012; Nayak et al., 2012) sign decomposition into subunits (a.k.a. signemes or atomic shapes or motions). Subunits are then used (instead of per-frame features) for further recognition using learning methods, such as HMMs or CRFs. Another semi-supervised approach involves learning specific signs from sentences containing them (Nayak et al., 2012; Sarkar et al., 2011), e.g., from subtitled TV programs (Pfister et al., 2013; Pfister et al., 2014).

Recognition rates for these semi-supervised and unsupervised approaches are in the 80-93% range for limited vo-

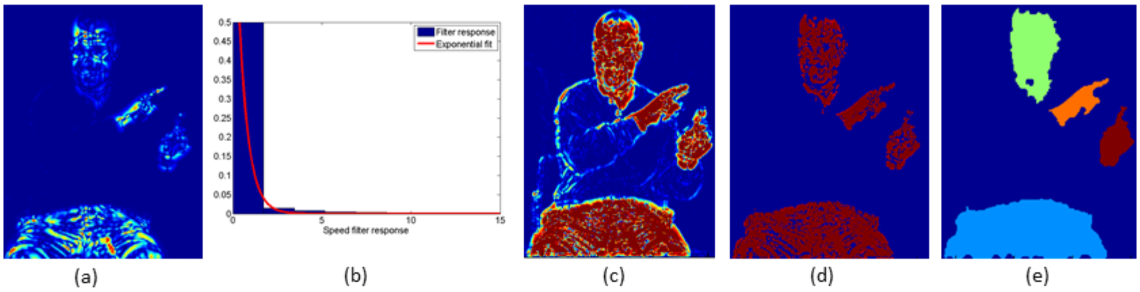

Figure 1: a) Non-zero optical flow locations; b) Exponential PDF fit to the product $|sX| \cdot |sY|$; c) Exponential PDF values over the product $|sX| \cdot |sY|$; d) Thresholded PDF; e) Extracted moving regions;

cabularies (20-142 signs). Existing supervised sign recognition methods achieve accuracy of 91-97% for a limited inventory (25-80 signs) (Assaleh et al., 2008; Vogler and Metaxas, 2001; Yang and Lee, 2013; Nayak et al., 2012; Sarkar et al., 2011; Yang and Sarkar, 2006), and often from single-subject data (Assaleh et al., 2008; Nayak et al., 2012; Sarkar et al., 2011; Yang and Sarkar, 2006; Gao et al., 2004). Accuracy quickly degrades to 85-90% with somewhat more extensive vocabulary (100-250 signs) (Elons et al., 2013b; Kong and Ranganath, 2014) and subject independence (Kong and Ranganath, 2014).

None of the work mentioned above takes into account the existence of sign types other than lexical and/or fingerspelled. Classifier constructions, which occur with high frequency in signed languages, are largely ignored. Furthermore, prior work on recognition from continuous signing does not apply different strategies based on differentiation of types of signs; and recognition research generally focuses on a single type of sign (e.g., there is some research on recognition of lexical signs, other research on recognition of fingerspelling).

3. Our contribution

In order to tailor recognition strategies to the distinct sign classes (Tsechpenakis et al., 2006; Tsechpenakis et al., 2008; Dilsizian et al., 2014), we must first be able to segment continuous signing video into subsequences corresponding to the distinct types of sign production (i.e., sequences of one or more signs of the same class). We introduce here a multiple instance learning (MIL) system for this task. Our contributions are three-fold:

- Unlike most previous research, we detect classifier signs.
- Our novel feature extraction method does not require pre-segmented hands of specific size and scale, thus eliminating the need for a hand tracker.
- We formulate sign classification as an intra-frame MIL problem, allowing us to capture indirectly important relationships among multiple moving regions in the image.

We test our system on 500 utterances containing 3,085 signs captured across 7 subjects from the National Center for Sign Language and Gesture Resources (NCSLGR <http://secrets.rutgers.edu/dai/queryPages/>) corpus of utterances collected and annotated at Boston University (Neidle and Vogler, 2012). Our system produces sign class labels that match those of human annotators for 91.27% of the video frames; the remaining 8.73% include marginal cases that are too short (e.g., fingerspelled "on"), cases with articulatory properties consistent with more than one sign class, as well as some cases where the human annotation turned out to be inaccurate (see section 7).

4. Problem formulation

In ASL, hand, arm, upper body, and head movement conveys important linguistic information of various kinds, as do facial expressions. However, since moving body parts are more relevant for recognition than stationary ones, we extract relevant features solely from image regions with significant motion, without restricting attention to specific body parts.

We formulate sign class recognition as a multiple instance learning (MIL) problem (Ben-Hur and others, 2012). In the MIL context a video frame is represented by a bag of multiple moving regions (e.g., hands, arms, face). The number of moving regions is not regulated. The task is then to find the frame type (sign class) given the set of moving regions, without attempting to find the sign class for each moving region. Thus, MIL indirectly captures the relationship between different moving body parts.

Not all moving regions inside the frame are relevant to the actual signing. For example, background motion or clothing with salient texture can produce significant local motion detector output. Given sufficient quantities of data, the MIL framework will filter out the moving regions (like legs in Figure 1a) that are inconsistent with the training data.

We use the Citation KNN (k-Nearest Neighbors) (Wang and Zucker, 2000) implementation of the MIL paradigm. This implementation builds a distance map between the training examples using ranked Hausdorff distance between two point sets (sets of moving regions in the case of our application). Hausdorff distance is not necessarily symmetric. Therefore, there is a difference between the nearest neighbors of the given frame and the frames that would consider the given frame as their nearest neighbor.

For each new frame, R references and C citers are computed using training data. References are the nearest neighbors of the new frame. Citers are the training frames that would consider the new frame their nearest neighbor within a certain rank. The rank is passed to the algorithm as parameter c. We used $R = 2$ and $c = 2$.

The resulting system captures local motion inside the frame, but does not capture global temporal changes. However, local motion within a frame can be consistent with multiple sign classes. Therefore, we used a one state per frame CRF (Lafferty et al., 2001) on top of the MIL-framework output to model the global interframe dynamics. For the MIL-framework input, we use a set of fixed-size feature descriptors for each of the moving parts. Popular feature descriptors, such as Shape Context (Belongie et al., 2002), SIFT (Lowe, 1999), or HOG (Dalal and Triggs, 2005), either do not describe regions or do not result in constant-size feature vectors. We, therefore, propose a new feature descriptor suitable for our application.

5. Feature Extraction

Our approach extracts two types of features:

1. features to detect the moving regions of interest, e.g., the hands; and
2. spatio-temporal features from the regions of interest, suitable for recognizing the type of sign

For (1), we use optical flow (Fleet and Weiss, 2006) to find the moving regions with significant speed (sX, sY) at location X, Y in the image. We compute the PDF (probability density function) of the distribution of the product $|sX| \cdot |sY|$ (Figure 1a-1c) and threshold the data on probability density <0.1 (Figure 1d) to include only highly noticeable motion.

We then collect regions with an area of >100 pixels (10×10 pixels, which is of the order of magnitude of a fingernail for the 640×480 frames in the NCSLGR dataset); see Figure 1e.

In step 2, we extract features related to the shape and velocity of the moving regions (e.g., hands). For implementation, we choose the set of image filters shown in Table 1, which allows us to capture predominant orientations and ridge strength signatures of both shape (intensity) and local motion (optical flow speeds). Each filter response is computed over the Gaussian pyramid of the input image to account for scale issues. We have observed that each filter response consistently follows a specific parametric distribution for all regions of interest (Figure 2). The parameters of that distribution become a part of the feature vector. The total feature vector length is 14; and each frame would have approximately from 1 to 6 feature vectors corresponding to the high local motion regions.

6. Experimental evaluation

We used the Weka (Hall et al., 2009) implementation of the Citation KNN MIL (Wang and Zucker, 2000) classifier, trained on the pre-segmented NCSLGR data. For training we used 570 randomly chosen instances of each sign class (lexical, fingerspelled, classifiers) for a total of 1,710 sign instances across 7 subjects. Using the trained MIL classifier, we obtained frame labels for 1,110 utterances. We removed movement epenthesis frames and frames belonging to sign classes not included in the current research, and used the rest to train a CRF.

We tested on 500 complete utterances across the same 7 subjects. These utterances consisted of 3,085 relevant signs (42,947 relevant frames plus 21,535 frames of movement epenthesis or sign classes not considered here). Therefore, our test and training datasets were significantly different in temporal structure.

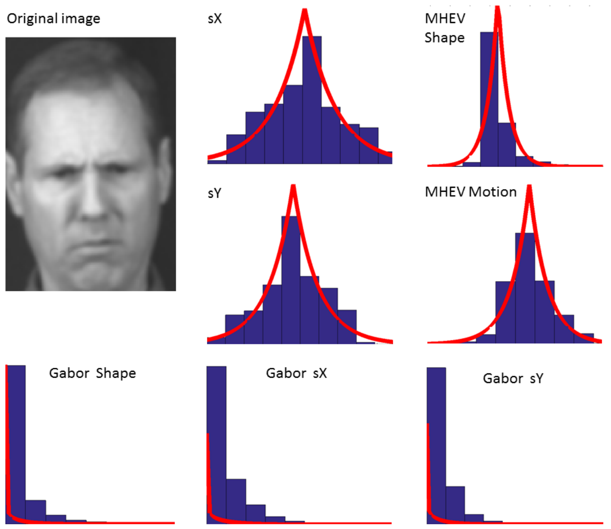

Figure 2: Filter responses and their parametric distributions. All fitting procedures converged to a 95% confidence interval.

Filter	Parametric distribution	Number of parameters
Gabor absolute value on shape	Gamma (Stacy, 1962)	2
Gabor absolute value on speeds sX, sY obtained from optical flow	Laplace (Kotz et al., 2001)	2, 2
Maximum Eigen value of the Hessian on speeds sX, sY 1	Laplace	2
Maximum Eigen value of the Hessian on shape	Laplace	2
sX, sY	Laplace	2, 2

Table 1: Image filter responses and their parametric distributions over the Gaussian pyramid. Total feature descriptor length is 14.

7. Results and Discussion

The system labeling matched that of the human annotators for 91.27% of the 42,947 relevant frames. Sample utterances are presented in Figures 3 and 4 (gloss labels include the prefix "fs-" for fingerspelled signs and "BCL-", "DCL-" or "SCL-" for (different types of) classifier constructions).

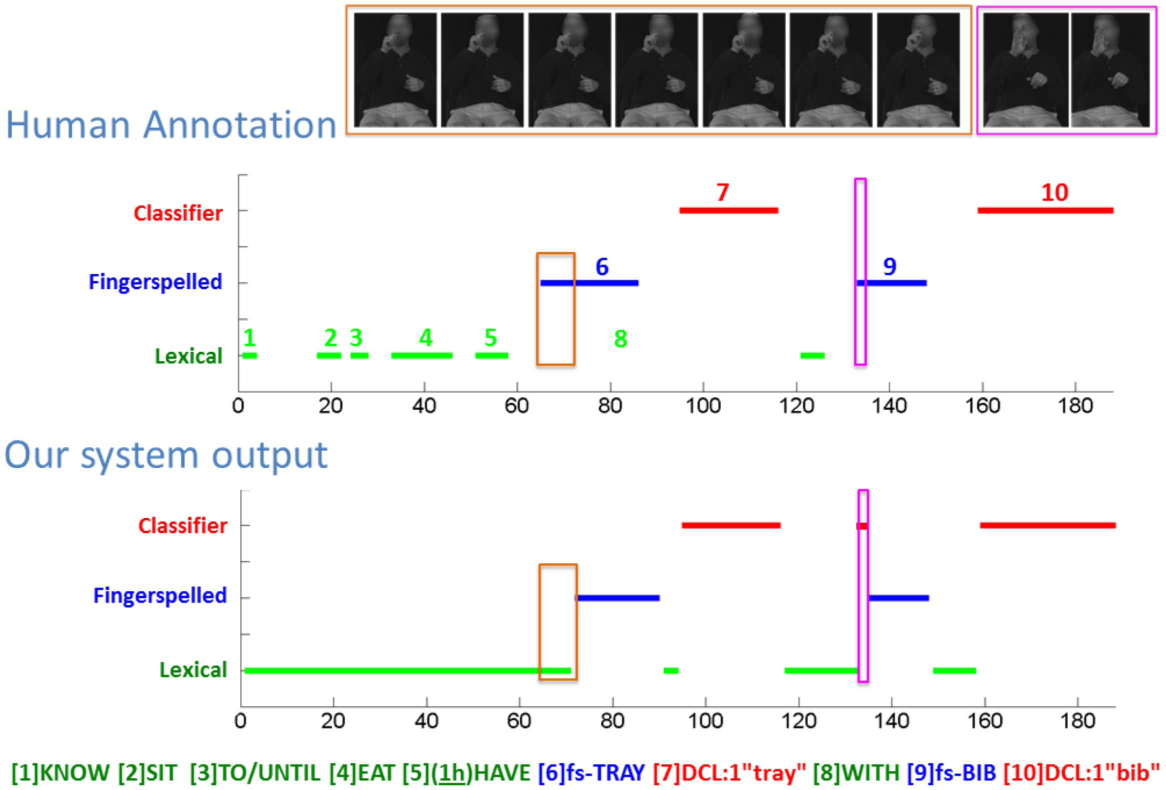

Figure 3: Some transitional motions that were labeled as fingerspelling in the human annotations and are not actually part of the immediately following fingerspelled signs [orange and magenta insets].

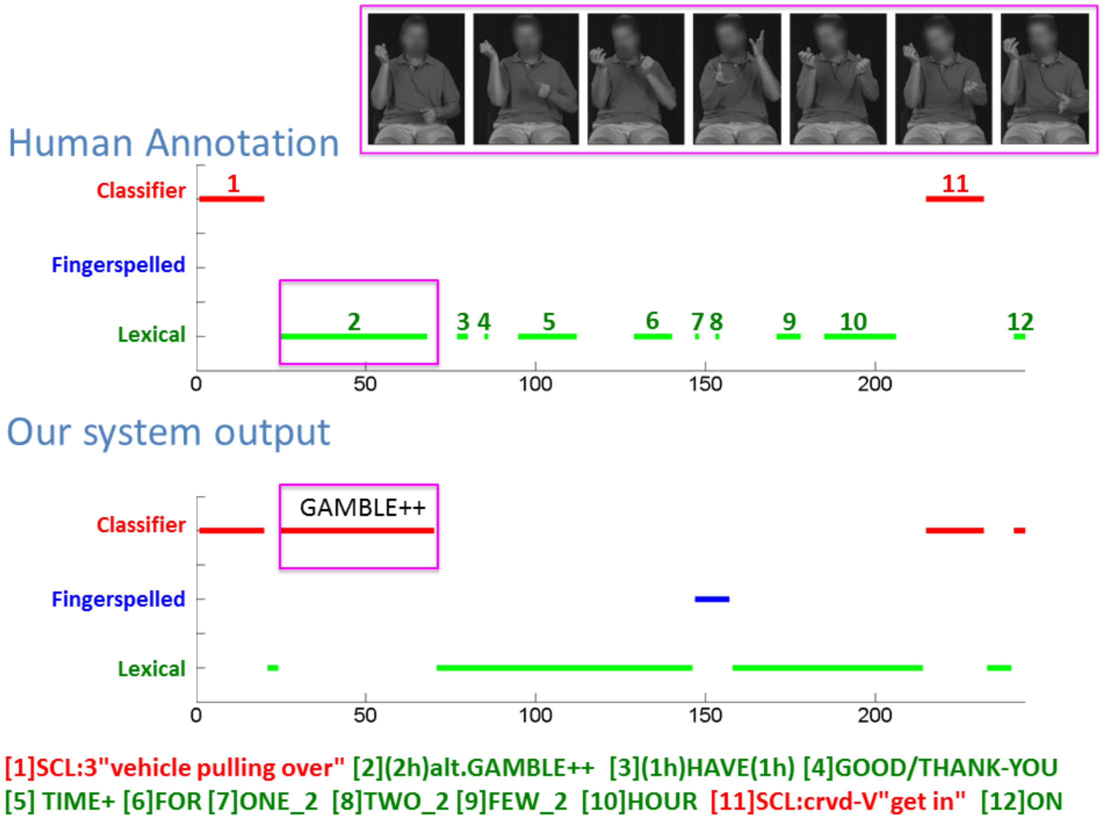

Figure 4: "GAMBLE++", a classifier-like lexical sign has been labeled as a classifier [magenta inset].

System Result Human Annotation	Lexical	Finger-spelled	Classifier	Total Matches
Lexical	29806	774	433	96.11%
Fingerspelled	1225	4120	0	77.08%
Classifier	1181	2	5406	82.05%

Table 2: Per-frame confusion matrix.

The remaining 8.73% include cases involving human inaccuracy in labeling sign boundaries (Figure 3), signs with properties consistent with more than one sign class (Figure 4) or short fingerspelled signs. In Figure 3, orange and magenta insets display frames in which the hand is getting into position for fingerspelling. That transition is included in the region that had been annotated as fingerspelling; i.e., a small human labeling error as to the precise start point of the fingerspelling is corrected by our system.

In Figure 4, "(2h)alt.GAMBLE++", which had been annotated as lexical, was identified by our system as a classifier. However, this sign is profoundly classifier-like and was only considered lexical by the annotators because it has come into frequent usage. (It involves a kind of acting out of rolling dice; the linguistic properties are not typical of lexical signs.) Thus, the result from the system is reasonable in light of the nature of this sign.

Table 2 shows the per-frame confusion matrix. Fingerspelled signs and classifiers have a bigger overlap with lexical signs than with each other. In the presentation, we discuss some of the sources of confusion seen in Table 2 (such as the confusion between wrist rotation and finger movement).

A more comprehensive analysis of the cases where the system and the human disagreed on classification is reported in the conference presentation. It should be noted that this performance is based on a clean dataset with no background clutter or movement. Therefore, the regions of interest can be trivially extracted based on local motion. In future work, we will use background subtraction methods (Cui et al., 2012) to test our system in cases with cluttered backgrounds.

8. Conclusion and Future Work

In order to tailor sign recognition strategies to structurally different types of signs, we have developed a system for segmentation of continuous signing based on the linguistic type of sign production (lexical signs, fingerspelling, classifier constructions). The system labeling matched that of human annotators for 91.27% of the frames from 500 utterances consisting of 3,085 signs. The remaining 8.73% contain cases that clearly reflect either human inaccuracy in the "ground truth" labeling of boundaries, or signs with properties consistent with more than one class of signs. The segmentation results could be improved by running different sign type HMMs simultaneously in cases where the segmentation system produces a low-confidence result.

This is a first step towards creating a complete system for real-time sign recognition that leverages the linguistic properties of the distinct sign classes. Incorporation of such linguistic information would 1) lead to improvements in the segmentation itself, and 2) be used for sign identification within the segmented regions. There is considerable research on recognition strategies for fingerspelled signs (Rioux-Maldague and Giguere, 2014; Kim et al., 2013; Pugeault and Bowden, 2011; Ricco and Tomasi, 2009, e.g.). For lexical signs, linguistic constraints on the relationships between the start and end handshapes of a given sign, and between the handshapes used on the left and right hands, have already been demonstrated to improve accuracy of handshape recognition (a crucial linguistic component) (Dilsizian et al., 2014). Work is now in progress on incorporating 3D tracking of hands, arms, and upper body to exploit the motion properties of lexical signs for sign identification (Dilsizian et al., 2016). This approach, sensitive to the fundamentally different internal structure of distinct sign types, holds great promise for recognition of large-scale vocabulary from continuous signing.

9. Acknowledgments

We thank all of the many Deaf consultants and students at Boston University who have participated in and contributed to the data collection, annotation, and analysis. We thank especially Rachel Benedict, Ben Bahan, Tory Sampson, Corbin Kuntze, Blaze Travis, Chelsea Hammond, Indya Oliver, Emma Preston, and Rebecca Lopez. We are also grateful for support in software development from Augustine Opoku and Gregory Dimitriadis (Rutgers University, Laboratory for Computer Science Research). This research was partially funded by grants from the National Science Foundation (#1065013, 0964385, 1059218, 1064965, 1355304, 1447037).

10. Bibliographical References

Assaleh, K., Shanableh, T., Fanaswala, M., Bajaj, H., and Amin, F. (2008). Vision-based system for continuous Arabic Sign Language recognition in user dependent mode. In *Proceedings of the 2008 International Symposium on Mechatronics and Its Applications*, pages 1–5.

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Thangali, A., Wang, H., and Yuan, Q. (2010). 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.

Battison, R. (1978). *Lexical Borrowing in American Sign Language*. Linstok Press.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522.

Ben-Hur, A. et al. (2012). Multiple instance learning of Calmodulin binding sites. *Bioinformatics*, 28(18):i416–i422.

Bowden, R., Windridge, D., Kadir, T., Zisserman, A., and Brady, M. (2004). A linguistic feature vector for the visual interpretation of sign language. In *Proceedings of the 2004 European Conference on Computer Vision (ECCV)*, pages 390–401. Springer.

Brentari, D. (1998). *A prosodic model of sign language phonology*. MIT Press.

Cooper, H., Ong, E.-J., Pugeault, N., and Bowden, R. (2012). Sign language recognition using sub-units. *The Journal of Machine Learning Research*, 13(1):2205–2231.

Cui, X., Huang, J., Zhang, S., and Metaxas, D. N. (2012). Background subtraction using low rank and group sparsity constraints. In *Proceedings of the 2012 European Conference on Computer Vision (ECCV)*, pages 612–625.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE.

Dilsizian, M., Yanovich, P., Wang, S., Neidle, C., and Metaxas, D. (2014). A new framework for sign language recognition based on 3D handshape identification and linguistic modeling. In *Proceedings of the 2014 International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Dilsizian, M., Tang, Z., Metaxas, D., and Neidle, C. (2016). The importance of 3D motion trajectories for computer-based sign recognition. In *In Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, LREC 2016.*

Elons, A. S., Aboul-Ela, M., and Tolba, M. F. (2013a). 3D object recognition technique using multiple 2D views for Arabic Sign Language. *Journal of Experimental & Theoretical Artificial Intelligence*, 25(1):119–137.

Elons, A. S., Abull-Ela, M., and Tolba, M. F. (2013b). A proposed PCNN features quality optimization technique for pose-invariant 3D Arabic Sign Language recognition. *Applied Soft Computing*, 13(4):1646–1660.

Emmorey, K. (2013). *Perspectives on Classifier Constructions in Sign Languages*. Taylor & Francis Group.

Fleet, D. and Weiss, Y. (2006). Optical flow estimation. In *Handbook of Mathematical Models in Computer Vision*, pages 237–257. Springer.

Gao, W., Fang, G., Zhao, D., and Chen, Y. (2004). Transition movement models for large vocabulary continuous sign language recognition. In *Proceedings of the 2004 IEEE International Conference on Automatic Face and Gesture Recognition*, pages 553–558. IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Han, J., Awad, G., and Sutherland, A. (2013). Boosted subunits: a framework for recognising sign language from videos. *IET Image Processing*, 7(1):70–80.

Kim, T., Shakhnarovich, G., and Livescu, K. (2013). Fingerspelling recognition with semi-Markov Conditional Random Fields. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1528.

Kong, W. and Ranganath, S. (2014). Towards subject independent continuous sign language recognition: a segment and merge approach. *Pattern Recognition*, 47(3):1294–1308.

Kotz, S., Kozubowski, T., and Podgorski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, exonomics, engineering, and finance*. Number 183. Springer Science & Business Media.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 2001 International Conference on Machine Learning (ICML)*, pages 282–289.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the 1999 IEEE Conference on Computer Vision*, volume 2, pages 1150–1157. IEEE.

Madeo, R. C., Peres, S. M., Lima, C. A., and Boscarioli, C. (2012). Hybrid architecture for gesture recognition: integrating fuzzy-connectionist and heuristic classifiers using fuzzy syntactical strategy. In *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Mohandes, M., Deriche, M., Johar, U., and Ilyas, S. (2012). A signer-independent Arabic Sign Language recognition system using face detection, geometric features, and a Hidden Markov Model. *Computers & Electrical Engineering*, 38(2):422–433.

Nayak, S., Duncan, K., Sarkar, S., and Loeding, B. L. (2012). Finding recurrent patterns from continuous sign language sentences for automated extraction of signs. *Journal of Machine Learning Research*, 13:2589–2615.

Neidle, C. and Vogler, C. (2012). A new Web interface to facilitate access to corpora: development of the ASLLRP data access interface (DAI). In *Proceedings of the 2012 Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC.*

Pfister, T., Charles, J., and Zisserman, A. (2013). Large-scale learning of sign language by watching TV (using co-occurrences). In *Proceedings of the 2013 British Machine Vision Conference (BMVC).*

Pfister, T., Charles, J., and Zisserman, A. (2014). Domain-adaptive discriminative one-shot learning of gestures. In *Proceedings of the 2014 European Conference on Computer Vision (ECCV)*, pages 814–829.

Pugeault, N. and Bowden, R. (2011). Spelling it out: real-time ASL fingerspelling recognition. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1114–1119. IEEE.

Ricco, S. and Tomasi, C. (2009). Fingerspelling recognition through classification of letter-to-letter transitions. In *Computer Vision–ACCV 2009*, pages 214–225. Springer.

Rioux-Maldague, L. and Giguere, P. (2014). Sign language fingerspelling classification from depth and color images using a Deep Belief Network. In *Canadian Conference on Computer and Robot Vision (CRV)*, pages 92–97. IEEE.

Sarkar, S., Loeding, B., Yang, R., Nayak, S., and Parashar, A. (2011). Segmentation-robust representations, matching, and modeling for sign language. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 13–19.

Stacy, E. W. (1962). A generalization of the Gamma distribution. *The Annals of Mathematical Statistics*, pages 1187–1192.

Thangali, A., Nash, J. P., Sclaroff, S., and Neidle, C. (2011). Exploiting phonological constraints for handshape inference in ASL video. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '11, pages 521–528, Washington, DC, USA. IEEE Computer Society.

Theodorakis, S., Pitsikalis, V., Rodomagoulakis, I., and Maragos, P. (2012). Recognition with raw canonical phonetic movement and handshape subunits on videos of continuous sign language. In *Proceedings of the International Conference on Image Processing*.

Tsechpenakis, G., Metaxas, D., and Neidle, C. (2006). Learning-based dynamic coupling of discrete and continuous trackers. *Computer Vision and Image Understanding*, 104(2):140–156.

Tsechpenakis, G., Metaxas, D., and Neidle, C. (2008). Combining discrete and continuous 3D trackers. In Bodo Rosenhahn, et al., editors, *Human Motion*, volume 36 of *Computational Imaging and Vision*, pages 133–158. Springer Netherlands.

Vogler, C. and Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of American Sign Language. *Computer Vision and Image Understanding*, 81(3):358–384.

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.

Wang, J. and Zucker, J.-D. (2000). Solving multiple-instance problem: a lazy learning approach. In *Proceedings of the 2000 International Conference on Machine Learning (ICML)*, pages 1119–1126.

Wang, S. B., Quattoni, A., Morency, L., Demirdjian, D., and Darrell, T. (2006). Hidden Conditional Random Fields for gesture recognition. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1521–1527. IEEE.

Yang, H.-D. and Lee, S.-W. (2013). Robust sign language recognition by combining manual and non-manual features based on Conditional Random Field and Support Vector Machine. *Pattern Recognition Letters*, 34(16):2051–2056.

Yang, R. and Sarkar, S. (2006). Detecting coarticulation in sign language using Conditional Random Fields. In *Proceedings of the 2006 International Conference on Pattern Recognition (ICPR)*, ICPR '06, pages 108–112, Washington, DC, USA. IEEE Computer Society.