2018

Linguistically-driven framework for computationally efficient and scalable sign recognition

Metaxas, Dimitris N.

European Language Resources Association (ELRA)

Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition

Dimitris Metaxas*, Mark Dilsizian*, Carol Neidle**

*Rutgers University, **Boston University
*Rutgers University, CBIM, Department of Computer Science, 617 Bowser Road, Piscataway, NJ 08854
**Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215
dnm@cs.rutgers.edu, mdil@cs.rutgers.edu, carol@bu.edu

Abstract

We introduce a new general framework for sign recognition from monocular video using limited quantities of annotated data. The novelty of the hybrid framework we describe here is that we exploit state-of-the art learning methods while also incorporating features based on what we know about the linguistic composition of lexical signs. In particular, we analyze hand shape, orientation, location, and motion trajectories, and then use CRFs to combine this linguistically significant information for purposes of sign recognition. Our robust modeling and recognition of these sub-components of sign production allow an efficient parameterization of the sign recognition problem as compared with purely data-driven methods. This parameterization enables a scalable and extendable time-series learning approach that advances the state of the art in sign recognition, as shown by the results reported here for recognition of isolated, citation-form, lexical signs from American Sign Language (ASL).

Keywords: Sign Recognition, Model-based Machine Learning, Computer Vision, American Sign Language (ASL).

1. Introduction

Automatic sign recognition is a difficult problem given the complexity of the linguistic structures in sign languages and the challenges in modeling 3D configurations and movements from 2D video. To address this problem, we use certain known linguistic properties of the language to structure the problem, and to inform, enhance, and correct visual recognition tasks. By combining these components into a unified optimization framework, we recognize isolated, citation-form lexical signs from American Sign Language (ASL) in a fully scalable manner.

Whereas prior vision-based approaches to sign recognition by computer had focused on detection of linguistically important components, such as handshape and motion trajectory, neural networks have recently been applied to the overall problem of end-to-end sign recognition without attending to linguistic structure. The framework described here (1) exploits recent discriminative neural net-based learning approaches, coupled with generative model-based methods, to improve the detection and analysis of the linguistically relevant components and features; (2) integrates knowledge of linguistic structures and dependencies to derive additional parameters; and (3) uses CRF learning methods to integrate these features for sign recognition. This enhances visual recognition capabilities for the critical sign components and offers a unified framework for sign recognition. This approach is successful working with limited quantities of annotated data and is scalable.

2. Previous Work

Previous computer vision research on sign recognition has generally focused on aspects of sign production known to be linguistically important, including analysis of handshapes, upper body pose, and movement trajectories.

Prior work on hand pose recognition in general includes Heap and Hogg (1996), Athitsos and Sclaroff (2001, 2003) and Tompson et al. (2014). Lu et al. (2003), Vogler and Metaxas (2004), and Isaacs and Foo (2004) focus specifically on the recognition of handshapes in sign languages. Yuntao and Weng (2000), Ding and Martinez (2007, 2009) , Ricco and Tomasi (2009), Thangali et al. (2011), Dilsizian et al. (2014), and Koller et al. (2016) constrain handshape recognition to fit those handshapes that are used linguistically in the sign language. Koller et al. (2016) is notable in the use of convolutional neural nets to achieve state-of-the art handshape recognition on a large dataset. Thangali et al. (2011) and later Dilsizian et al. (2014) leverage phonological constraints on start and end handshape co-occurrence to improve handshape recognition accuracy.

Other research has explored hand motion trajectories as an intermediate step towards sign recognition (Han, Awad, and Sutherland, 2009; Dilsizian et al., 2016; Pu et al., 2016). Ding and Martinez (2007, 2009) combine motion trajectories with face and hand configuration for sign recognition. Dilsizian et al. (2016) demonstrates the importance of 3D motion trajectories for sign recognition.

There have been some attempts to build full sign recognition frameworks for isolated signs, which have had limited success. Cooper, Holt, and Bowden (2011) combine 2D motion trajectories and handshape features to achieve 71.4% top-1 accuracy on 984 signs from a single signer. Wang et al. (2016) achieve 70.9% accuracy on 1000 isolated signs across multiple signers. Guo et al. (2016) propose an adaptive Gaussian Mixture Model HMM framework, and from vocabulary of 370 signs, they achieve a top-1 accuracy of 33.54% and a top-5 accuracy of 59.79%. However, they rely on an RGBD sensor for 3D information.

More recently there have been several purely data-driven end-to-end approaches to sign recognition from continuous signing based on Recurrent Neural Net (RNN) architectures (Cui, Liu, and Zhang, 2017; Koller, Zargaran, and Ney, 2017). However, the performance of these image-based approaches is held back by limitations in the data-

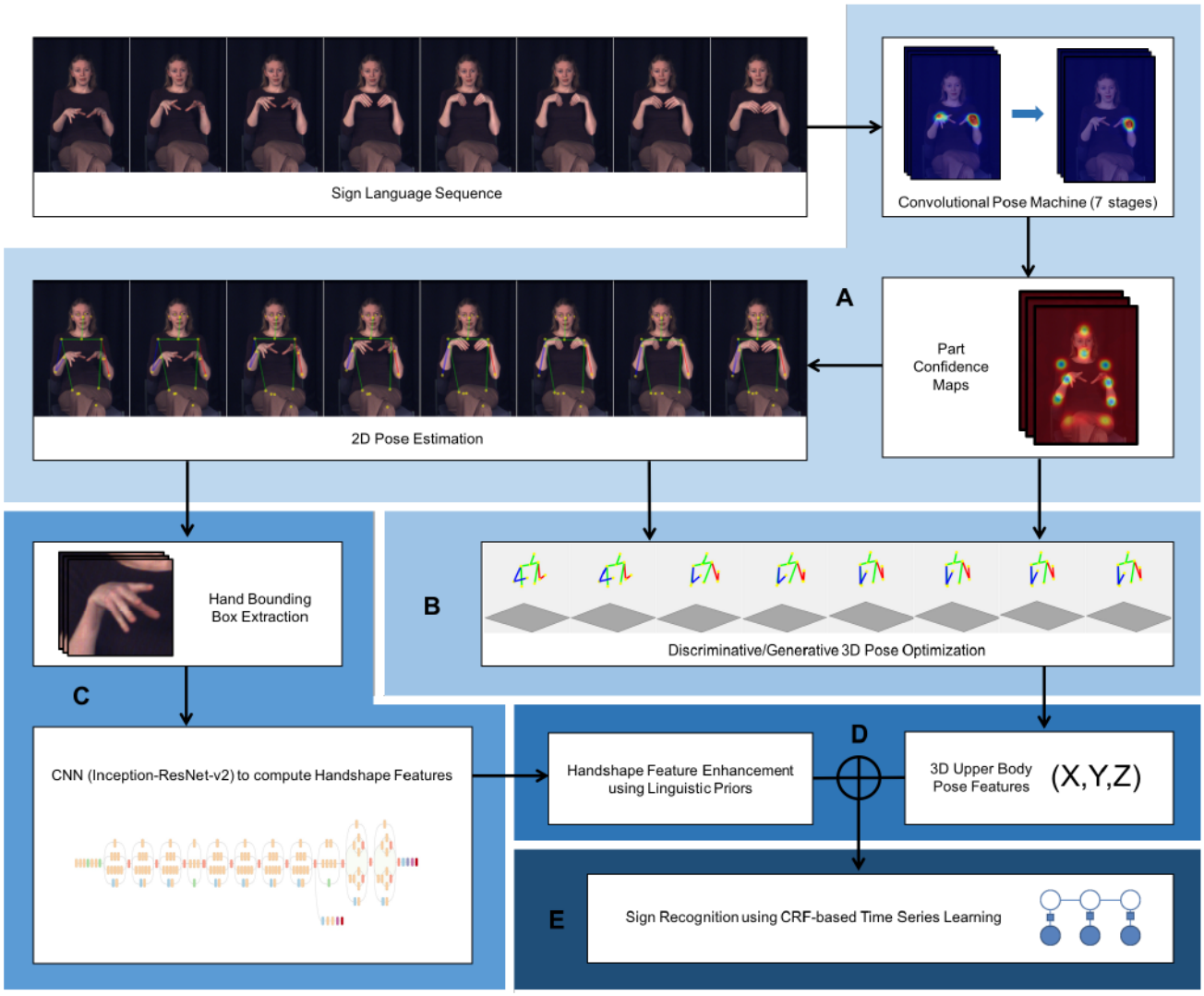

Figure 1: Framework Overview

sets and the fact that they do not integrate linguistic knowledge and perform 3D analysis. They report limited success in sign recognition, and these methods do not generalize well across multiple signers. The previous work in sign language recognition clearly demonstrates the need for a new computational approach.

3. New Framework for Sign Recognition

We propose a learning-based approach with three sub-components: 1) new discriminative learning-based computer vision methods (based on advances in deep learning) coupled with generative methods for hand and pose feature extraction and related parameters (Section 3.1); 2) additional linguistically driven parameters (Section 3.2), with enhancement of parameters from known linguistic dependencies (Section 3.3); and machine learning methods for sign recognition using the extracted parameters (Section 3.4).

This gives rise to a reduced parameterization and a significantly more efficient algorithm capable of coping with limited quantities of annotated data. This results in improved sign recognition compared to previous approaches. See Figure 1 for an overview of our framework.

.

3.1 Summary of Features Used for Sign Recognition

Using the methods to be described below, we estimated a comprehensive set of features (a total of 110), with regard to: **a)** handshapes, **b)** number of hands, **c)** 3D upper body locations, movements of the hands and arms, and distance between the hands, **d)** facial features and head movements (which have been shown to improve manual sign recognition (von Agris, Knorr, and Kraiss, 2008; Koller, Forster, and Ney, 2015)), and **e)** contact. The features for the face include 66 points (visible in Figure 2) from 3D estimates for the forehead, ear, eye, nose, and mouth regions, and their velocities across frames. The contact features are extracted from our 3D face and upper body movement estimation, and relate to the possibilities of the hand touching specific parts of the head or body. The parameter extraction is described in the next section.

3.2 Coupling of Discriminative and Generative Methods for Feature Extraction

In order to build a robust and scalable framework for sign recognition, we model individual components of the sign recognition problem. In this section, we present our methodology for upper body trajectory and handshape estimation, and related feature extraction.

3.2.1 Upper Body and Hand/Arm Movement Trajectories

Previous work has shown that tracking upper body pose, especially in 3D, is critical to sign recognition (Fillbrandt, Akyol, and Kraiss, 2003; Vogler and Metaxas, 2004; Zafrulla et al., 2011; Dilsizian et al., 2014). In our framework, we model upper body pose and use the 3D joint locations as features.

To develop an accurate 3D pose estimation suitable for ASL, we integrate state-of-the-art neural net-based 2D and 3D pose estimation (fine-tuned on ASL upper body videos) and a generative, deformable model-based fitting approach to further refine the 3D pose (Dilsizian, 2016). We start with a Convolutional Pose Machine (Wei et al., 2016) trained on a combination of the MPII human pose dataset (Andriluka et al., 2014) and the Kinect-based dataset for upper body pose in Dilsizian et al. (2016) to better match the 2D pose projection. Next, we use nearest neighbor matching with a 3D Pose library that includes the Human 3.6M dataset (Ionescu et al., 2014) and is also combined with the Kinect-based dataset from Dilsizian et al. (2016).

We formulate generative human pose recognition as a search problem in 3D (Euclidean) space (Dilsizian, 2016); solving this problem entails finding optimal pose parameters of a human model whose learned part appearance representation has the best matching score based on the image. Confidence maps returned by the Convolutional Pose Machine are used as the cost surface and inversely projected in 3D space along the tensor normal to the camera. The neural-net based 3D prediction is then used as an initialization to our generative approach to search for 3D candidate locations for each part.

In our novel 3D generative approach (Dilsizian, 2016), the global 3D upper body pose is modeled by an I-node relational graph representing I human skeleton parts and joints. Each node represents a part center, and edges denote skeletal links. For image X, the function

$$\varphi^{\lambda} : \boldsymbol{X} \times \boldsymbol{L} \rightarrow R^{d}$$

extracts features for C candidate 2D image projections

$$\boldsymbol{L} = l_1, l_2, \ldots, l_C$$

of each 3D candidate location

$$\mathbf{Y} = \mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_C$$

at the scale space associated with the depth parameter

$$\boldsymbol{\lambda} = \lambda_1, \lambda_2, \ldots, \lambda_C.$$

In order to conduct a local search for the j^{th} part across 3D candidate locations $\boldsymbol{Y}$j , we compute the corresponding 2D projections $\boldsymbol{L}_j = f_c(\boldsymbol{Y}_j)$ and the quantized depth parameters $\boldsymbol{\lambda}_j$ of the part. The local matching score makes use of the learned part templates $t^{\lambda j, mj}$ and the local mixture parameter $\boldsymbol{b}_i^{mj}$; it is computed as:

$$\psi_j(\boldsymbol{Y}_j, \boldsymbol{m}_j) = \boldsymbol{t}_j^{\boldsymbol{\lambda}_j, \boldsymbol{m}_j} * \phi_j^{\boldsymbol{\lambda}_j}(\boldsymbol{X}, f_c(\boldsymbol{Y}_j)) + \boldsymbol{b}_j^{\boldsymbol{m}_j} \quad (1)$$

These local scores are optimized efficiently through the use of a dynamic programming approach similar to those of Felzenszwalb and Huttenlocher (2005) and Yang and Ramanan (2011). The optimization includes the passing of a message $\boldsymbol{S}$ from child to parent nodes. The optimal parameters of a part i over the candidate locations, mixtures, and scales are obtained by optimizing the local score and the sum over each message S passed from each child node j:

$$\boldsymbol{\Psi}(\boldsymbol{Y}_i, \boldsymbol{m}_i) = \boldsymbol{\psi}_i(\boldsymbol{Y}_i, \boldsymbol{m}_i) + \sum_{j \in \mathrm{child}(i)} \boldsymbol{S}_j(\boldsymbol{Y}_i, \boldsymbol{m}_i) \quad (2)$$

Message $\boldsymbol{Sj}$ makes use of the learned pairwise mixture parameter $\boldsymbol{b}_{ij}$, as well as the pairwise distance term $\boldsymbol{w}$, which scores the relative location of part j with respect to its parent i.

$$\boldsymbol{S}_j(\boldsymbol{Y}_i, \boldsymbol{m}_i) = \max_{\boldsymbol{y}_j \in \boldsymbol{Y}_j, m_j \in \boldsymbol{m}_j} \left[\boldsymbol{\Psi}(\boldsymbol{y}_j, m_j) + \boldsymbol{w}(\boldsymbol{Y}_i, \boldsymbol{y}_j) + \boldsymbol{b}_{ij}^{\boldsymbol{m}_i, m_j} \right] \quad (3)$$

Equations 2 and 3 recursively compute the score of a part at each location and scale using the learned part templates. The global optimization is computed from the head node, which has index 1. We find the joint configuration with the maximum score:

$$\hat{\boldsymbol{y}}_1, \hat{m}_1 = \underset{\boldsymbol{y}_1 \in \boldsymbol{Y}_1, m_1 \in \boldsymbol{m}_1}{\arg\max} \boldsymbol{\Psi}(\boldsymbol{y}_1, m_1) \quad (4)$$

Starting with $\hat{\boldsymbol{y}}_1$ and $\hat{m}_1$, we backtrack through location and mixture indices to find all the part locations $\hat{\boldsymbol{Y}}$ of the optimal configuration.

Examples of 3D upper body pose reconstructions projected onto the image and from an alternative view can be seen in Figure 2.

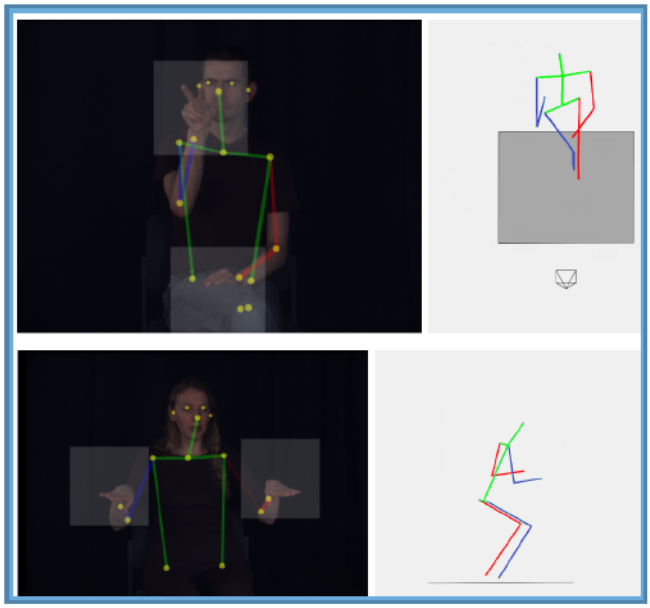

Figure 2: 3D upper body pose reconstruction and hand bounding boxes during an example of the signs glossed as LOOK (top row) and SHELF/FLOOR (bottom row). [The pictures depict the full body reconstruction that our system generates, although we only use the upper body for this research.]

Because our framework is capable of integrating a wide variety of disparate features, we can further mine the 3D trajectories for additional discriminative information. Another relevant set of features is extracted from linguistically recognized events, namely: we identify 1- vs. 2-handed signs, categorize hand touching events, and analyze motion trajectory. These additional features are particularly discriminative when combined with features relating to handshape appearance.

3.2.2 Handshape Parameters

Our parameterization also includes feature extraction from hand images specifically. We focus on the handshapes at the start and end of each sign, because those are the most linguistically informative handshapes.

We extract features derived from a neural net trained for handshape recognition. Additional features are then derived based on the relationship between handshapes on the dominant and non-dominant hands, as well as at the start and end of the sign (factoring in linguistic dependencies derived from frequencies of co-occurrence in our dataset).

In order to avoid overfitting and capture both the local and global appearance of the hand, we train Inception-ResNet-v2 (Szegedy et al., 2017) on hand images extracted from our upper body pose prediction. The handshape CNN returns top-1 accuracy of 70.1%; top-5 accuracy reaches 92.3%. However, we use the entire set of handshape probabilities from the output of the neural net as features for sign recognition.

3.3 Linguistic Structure

Once motion trajectories and handshape parameters have been extracted, we enhance features from Section 3.1 by focusing on properties known to be linguistically important and by leveraging known linguistic dependencies.

3.3.1 Upper Body Parameters Related to Contact

Some signs involve linguistically significant contact between the two hands, or hands contacting some specific part of the head, face, or body. We include parameters extracted from recognizing when such contact occurs and classifying the types of contact. For present purposes, the face and body were divided into different regions, based on the linguistically significant distinctions.

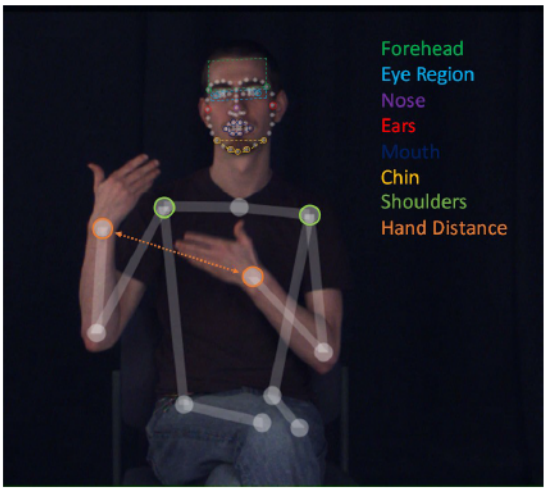

Figure 3: Visualization of Locations where Contact Occurs

Contact events are considered relative to 3D keypoints identified on the face and body from our pose estimation. A threshold is used for each touch event that is determined empirically based on what leads to improvement to sign recognition accuracy. Locations for contact events are visualized in Figure 3.

3.3.2 Number of Hands

Whereas some signs are produced using only one hand, others are normally produced using two hands (and in such cases, there are some dependencies with regard to what happens on the two hands, as discussed in Section 3.4). Thus we introduce an additional parameter related to this distinction. Our dataset of motion trajectories from citation-form ASL examples is used to train an HMM to predict 1- vs. 2- handed sequences from the dataset.

In summary, flags for signs involving contact between hands, face, and body, and flags for 1- vs. 2-handed signs function as additional parameters in our sign recognition feature vector.

3.4 Linguistic Dependencies

3.4.1 Dependencies between Start & End Handshapes

Following Thangali et al. (2011), Thangali (2013), and Dilsizian et al. (2014), we enhance handshape recognition by leveraging phonological constraints that hold between start and end handshapes in lexical signs, as reflected in the co-occurrence probabilities from our data set. After extracting, for each sign, the above per-frame start/end handshape parameters, we adjust those parameters based on the computation of the probabilities of co-occurrence of specific start and end handshapes. These co-occurrence statistics are multiplied by the joint probabilities of one window from the beginning of the sequence, which contains the start handshape, and one from the end, which contains the end handshape. The new start/end handshape priors are then used to adjust and improve handshape parameters for each frame of the sequence that falls in a start or end window.

3.4.2 Dependencies between Dominant & Non-dominant Handshapes in 2-handed Signs

In addition to enhancing handshape parameters through use of start/end handshape co-occurrence statistics, we extract another relevant parameter from comparing left and right handshapes. Many 2-handed lexical signs are produced with the same handshape on both hands. Furthermore, when the handshapes differ, the options for the non-dominant hand are severely reduced to a small set of unmarked handshapes. By computing the probability that the two hands are producing the same shape, the learning algorithm can benefit from this additional parameter with known discriminative value.

To compare the shapes, we compute a simple Mahalanobis distance between the dominant and non-dominant hand probability sets for each frame. This distance, which captures the similarity between the two handshapes, is added to our sign recognition feature vector.

3.5 Summary of Feature Vector Extraction

Based on our previously described hybrid approach, we assemble the features from the upper body and the handshapes into a feature vector to be used for sign recognition. This final feature vector consists, for each frame of a sign, of the features outlined in 3.1.

3.6 Sign Recognition

Our current sign recognition approach combines detection of upper body pose and hand configurations, the latter leveraging statistical properties resulting from linguistic constraints on sign formation, as just discussed. In doing this, we face several challenges. First, the signal for handshape recognition is noisy. Although for the training samples, we can rely on human annotations of start and end frames of a given sign (and these are the handshapes that are most important for sign identification), for the testing samples, there needs to be estimation of the start and end frames containing the handshapes to be taken as representative for the given sign. Second, the motion trajectories for a sign vary spatially and temporally from one instance to another of a given sign produced by same or different signers.

To process the motion trajectories, we normalize all upper body locations to the sternum location. In order to capture dependencies between our various features and to explicitly model the structure of the language, we employ a structured CRF-based method. We employ Hidden Conditional Ordinal Random Fields (HCORF), which explicitly model sequence dynamics as the dynamics of ordinal categories (Walecki et al., 2015); in our case, the ordinal categories are start and end handshape labels. We modify the HCORF objective function to include an additional error term that compares handshape predictions to ground truth labels for the two ordinal states (start/end).

Given normalized 3D body part locations (including the face and the head) and handshape features for each frame of a sequence, our resulting optimization minimizes the error of sign recognition while locally minimizing the error of start/end handshape prediction. As demonstrated in our Experiments section, this ordinal, structured approach is flexible and robust enough to overcome various types of failures in the different components of our framework.

In summary, our approach to sign recognition takes advantage of the fact that ASL has structure, and we achieve a significant reduction in the parameters used, which results in more efficient and robust ASL learning, as demonstrated in the next section.

4. Experiments and Results

4.1 Dataset

This research exploits the publicly accessible American Sign Language Lexicon Video Dataset (ASLLVD) (Neidle, Thangali, and Sclaroff, 2012).[1] This includes over 8500 examples corresponding to almost 2800 monomorphemic lexical signs in citation form from 6 native ASL signers. Although the entire ASLLVD dataset contains between 1 and 6 signers for all signs, we chose to use a subset of 350 signs, from among those with the highest numbers of signers and examples. On average, there were 4.7 signers and 6.9 total examples per sign for this set of 350 signs (a total of about 2400 examples). For each sign, 2 examples were randomly selected to be in the testing set, and the remaining examples were used for training.

4.2 Experiments

For each frame in each sequence, we extract a feature vector of dimension 110, which, as explained in Section 3.4, includes features for handshape, motion trajectory, and other linguistically motivated features discussed above. Then this feature vector is used as input to our modified HCORD-based framework for sign recognition. We trained on our data from 6 signers, using about 80% of the data for training and 20% for testing. We tested on vocabularies of differing sizes (175 vs. 350 signs) as a first step in demonstrating the efficiency and scalability of our approach. The set of 175 was chosen randomly from the signs in the larger set of 350.

We also performed a series of experiments to separate out the contributions of the different parameters, including those based on linguistically motivated features. This linguistic parameterization is especially useful in the current context of sign recognition research, where large amounts of data with ground truth are not available.

4.3 Results

4.3.1 Recognition Accuracy

As shown in Figure 4, from a vocabulary of 350 signs (including both 1- and 2-handed signs), using all of our parameters, we achieve a top-1 accuracy of 93.3% and a top-5 accuracy of 97.9%.

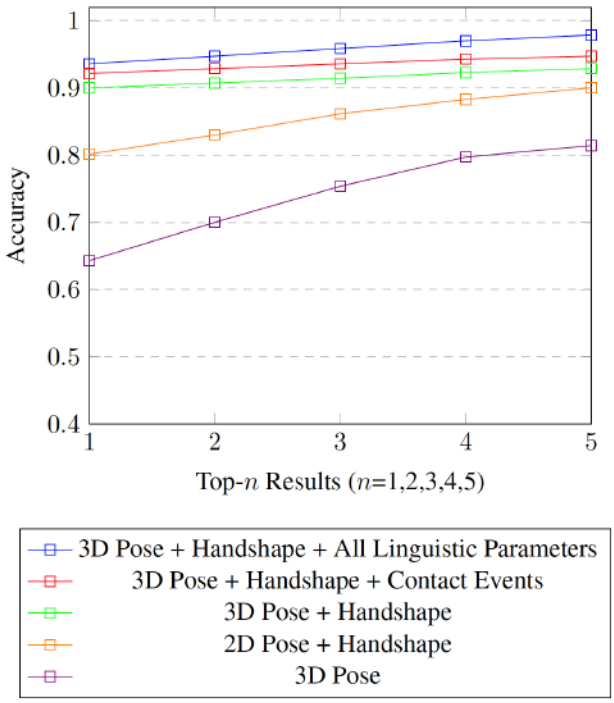

Figure 4: ASL sign recognition accuracy for top-n predictions over a 350 sign vocabulary.

[1] See http://www.bu.edu/av/asllrp/dai-asllvd.html. This dataset is also available at http://secrets.rutgers.edu/dai/queryPages/search/search.php and forms the basis for our new Web-accessible ASLLRP Sign Bank, accessible at http://dai.cs.rutgers.edu/dai/s/signbank (Neidle et al., 2018).

Figure 4 demonstrates the importance of integrating the different features computed by our framework. Sign recognition based solely on 3D upper body trajectories (violet) achieves reasonable accuracy. Although the handshape features by themselves are not sufficient to recognize signs, when combined with 3D trajectories, a significant boost in recognition accuracy is achieved (green). The addition of the contact events (shown in red) and linguistic information (blue) also improves recognition accuracy, particularly when considering top-5 accuracy, where it penalizes low probability observations that would result in impossible or improbable start/end combinations.

In addition, the importance of 3D trajectories over 2D is demonstrated here. The use of 3D motion trajectories (green) results in a significant boost in performance over 2D (yellow). This shows that there is significant discriminatory information built into the depth component of the motion trajectories. This important dimension is captured through the explicitly modeling of upper body 3D pose, and would not be captured by an end-to-end sign recognition CNN that implicitly encodes 2D motion patterns.

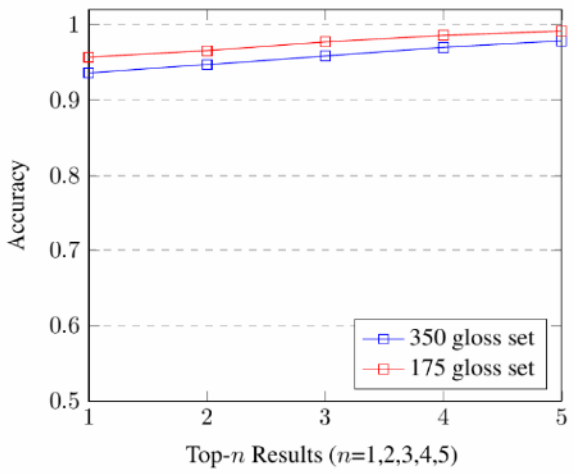

Figure 5: Comparing the Results on Vocabularies of 175 vs. 350 Signs

When increasing the vocabulary size from 175 to 350 signs (Figure 5), accuracy declines by only 2.1% for top-1, and by only 1.3% for top-5. This provides evidence for the scalability of the approach.

4.3.2 Error Analysis

An analysis of the errors in sign recognition revealed that some of the confusion involved signs with strong similarities in handshapes or movement trajectories. For example, there was one case where CAN was confused with COLD. Both signs involve closed fists in front of the body and have a similar movement pattern, but they differ critically in orientation (with the palm/fist facing downward for CAN but sideward, facing the center of the body (and shaking a bit from side to side), for COLD) as seen in Figure 6. Other recognition errors similarly involved orientation of the hands or confused locations relative to the body, or handshapes or movement patterns in some cases.

In any case, the success rate already achieved through the reduction in the number of parameters offers promise for scalability of these methods to large databases. We will investigate modifying the current parameterizations to better capture certain types of distinctions that are linguistically significant but that were not reliably exploited in the current set of results. For example, we expect that further improvements can be achieved by incorporating additional information about linguistic dependencies related to movement patterns of the two hands.

4.4 Discussion

There is a limited number of previous studies on isolated sign recognition available for comparison. In addition, many of the reports in the literature are for different sign languages (e.g., (von Agris et al., 2006; von Agris, Knorr, and Kraiss, 2008; Cooper, Holt, and Bowden, 2011; Wang et al., 2016)). Furthermore, research focused on ASL generally uses datasets we don't have (which in many cases contain small numbers of signs (e.g., (Zahedi et al., 2005; Zaki and Shaheen, 2011)), and/or the authors do not provide enough details or code to enable direct comparisons).

One relevant comparison is Guo et al. (2016), which uses a dataset of a size comparable to ours (370 signs) and a similar number of signers (5). Despite using RGBD data and a number of examples for training about 4 times greater than in our experiments, their adaptive GMM-HMM method results in recognition accuracy of only 33.54% for top-1, 59.79% for top-5, and 69.41% for top-10.

Conly (2016) is based on the same ASLLVD data we use, but he supplements that with additional data that he

Figure 6: CAN on the left (hands move downward); COLD on the right (hands move side to side)

collected that include depth information; and, despite that, and even taking into account the fact that he is working with a larger set of signs, he still gets quite a bit lower recognition accuracy. From a vocabulary of just over 1,100 signs, Conly reports the correct sign match 14.7% (top-1) of the time, and 36.0% for top-5. Although we use a smaller set of signs, 350 total, we get the correct match 93.3% (top-1) and 97.6% (top-5) of the time.

Thus, our recognition accuracy compares favorably with the two approaches just mentioned, and furthermore has the advantage of being scalable.

5. Conclusions

We have established a framework for sign recognition that relies on combining 3D modeling of start and end handshapes as distributions based on a few initial and final frames (enhanced by statistical information about their linguistic dependencies) and of 3D movement patterns of the hands, arms, and upper body during sign production. In particular, we have developed a statistical approach that combines distributions of the initial and final hand shapes and pose coupled with the spatiotemporal patterning of the arm and upper torso. Using this approach, we achieve high accuracy for recognition of ASL signs. This statistical parameterization of the linguistically important components of lexical signs makes it possible to employ learning methods that can take advantage of large amounts of data relevan t to each parameter, without requiring large numbers of examples of each individual sign in the vocabulary. As a result of the use of a reduced parameter representation, this method will also scale to larger sign vocabularies. To improve our sign recognition results in the future, we intend to expand the proposed parameterization to incorporate additional linguistic information about location, orientation, and movement patterns that are relevant to discrimination of signs. The ability to incorporate such improvements represents yet another advantage over purely data-driven approaches.

6. Acknowledgments

The research reported here has been supported in part by grants from the National Science Foundation (#1748016, 1748022, 1059218, 0705749, 0958247, 1703883, 1567289, 1555408, 1451292, 1447037). We are also grateful for contributions from many students and colleagues at Boston, Gallaudet, and Rutgers Universities. Thanks also to our many ASL linguistic consultants. We wish also to thank Stan Sclaroff, Ashwin Thangali, Joan Nash, and Vassilis Athitsos for their participation in data collection efforts for the ASLLVD, the dataset that served as the basis for the research reported here.

7. Bibliographical References

Andriluka, M., Pishchulin, L., Gehler, P., and B., S. (2014). 2D Human Pose Estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*.

Athitsos, V. and Sclaroff, S. (2001). 3D hand pose estimation by finding appearance-based matches in a large database of training views. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2001*.

Athitsos, V. and Sclaroff, S. (2003). Estimating 3D hand pose from a cluttered image. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2003,* Wisconsin.

Conly, C. (2016) Improving Accuracy in Large Vocabulary Sign Search Systems. Unpublished Doctoral Dissertation, University of Texas at Arlington.

Cooper, H., Holt, B., and Bowden, R. (2011) Sign Language Recognition. In Moeslund, T. B., Hilton, A., Krüger, V. and Sigal, L., (eds.) *Visual Analysis of Humans: Looking at People*: Springer. pp. 539–562.

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. *Proceedings of the CVPR 2017,* Honolulu, Hawaii.

Dilsizian, M., Yanovich, P., Wang, S., Neidle, C., and Metaxas, D. (2014). New Framework for Sign Recognition based on 3D Handshape Identification and Linguistic Modeling. *Proceedings of the LREC 2014, Reykjavik, Iceland. May 2014.*

Dilsizian, M. (2016) Hybrid Discriminative-Generative Methods for Human Pose Reconstruction from Monocular Imagery. Unpublished PhD Dissertation, Rutgers University.

Dilsizian, M., Tang, Z., Metaxas, D., Huenerfauth, M., and Neidle, C. (2016). The Importance of 3D Motion Trajectories for Computer-based Sign Recognition. *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. LREC 2016, Portorož, Slovenia. May 2016.*

Ding, L. and Martinez, A.M. (2007) Recovering the linguistic components of the manual signs in American Sign Language. *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, September 5-7, 2007.*, pp. 447-452.

Ding, L. and Martinez, A.M. (2009) Modelling and Recognition of the Linguistic Components in American Sign Language. *Image and Vision Computing, 27*(12), pp. 1826-44.

Felzenszwalb, P.F. and Huttenlocher, D.P. (2005) Pictorial Structures for Object Recognition. *International Journal of Computer Vision, 61*(1), pp. 55–79.

Fillbrandt, H., Akyol, S., and Kraiss, K.-F. (2003) Extraction of 3D Hand Shape and Posture from Image Sequences for Sign Language Recognition. *AMFG, Vol. 3*.

Guo, D., Zhou, W., Wang, M., and Li, H. (2016). Sign Language Recognition Based on Adaptive HMMs with Data Augmentation. *Proceedings of the IEEE Conference on Image Processing*.

Han, J., Awad, G., and Sutherland, A. (2009) Modelling and segmenting subunits for sign language

recognition based on hand motion analysis. *Pattern Recognition Letters, 30*(6), pp. 623-633.

Heap, T. and Hogg, D. (1996). Towards 3D Hand Tracking Using a Deformable Model. *Proceedings of the International Conference on Automatic Face and Gesture Recognition*.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2014) Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(7), pp. 1325-1339.

Isaacs, J. and Foo, S. (2004) System Theory. Proceedings of the Thirty-Sixth Southeastern Symposium on IEEE, 2004. *Hand pose estimation for American sign language recognition*.

Koller, D., Ney, H., and Bowden, R. (2016). Deep Hand: How to Train a CNN on 1 Million Hand Images when your Data is Continuous and Weakly Labelled. *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Koller, O., Forster, J., and Ney, H. (2015) Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding, 141*, pp. 108-125.

Koller, O., Zargaran, S., and Ney, H. (2017). Re-Sign: Re-Aligned End-To-End Sequence Modelling With Deep Recurrent CNN-HMMs. *Proceedings of the CVPR 2017,* Honolulu, Hawaii.

Lu, S., Metaxas, D., Samaras, D., and Oliensis, J. (2003). Using Multiple Cues for Hand Tracking and Model Refinement. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2003*.

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

Neidle, C., Opoku, A., Dimitriadis, G., and Metaxas, D. (2018). NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. LREC 2018, Miyagawa, Japan. May 2018. .*

Pu, J., Zhou, W., Zhang, J., and Li, H. (2016). Sign Language Recognition based on Trajectory Modeling with HMMs. *Proceedings of the International Conference on Multimedia Modeling*.

Ricco, S. and Tomasi, C. (2009) Fingerspelling Recognition through Classification of Letter-to-Letter Transitions. *9th Asian Conference on Computer Vision, Xi'an, China*.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A.A. (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*, pp. 4278–4284.

Thangali, A., Nash, J.P., Sclaroff, S., and Neidle, C. (2011). Exploiting Phonological Constraints for Handshape Inference in ASL Video. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2011*.

Tompson, J., Stein, M., LeCun, Y., and Perlin, K. (2014) Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG), 33*(5), pp. 169.

Vogler, C. and Metaxas, D. (2004) Handshapes and movements: Multiple-channel ASL recognition. *Springer Lecture Notes in Artificial Intelligence (Proceedings of the Gesture Workshop '03, Genova, Italy), 2915*, pp. 247-258.

von Agris, U., Schneider, D., Zieren, J., and Kraiss, K.-F. (2006). Rapid signer adaptation for isolated sign language recognition. *Proceedings of the Workshop on Vision for Human Computer Interaction (V4HCI)*.

von Agris, U., Knorr, M., and Kraiss, K.F. (2008). The significance of facial features for automatic sign language recognition. *Proceedings of the International Conference on Automatic Face & Gesture Recognition*.

Walecki, R., Rudovic, O., Pavlovic, V., and Pantic, M. (2015). Variable-state latent conditional random fields for facial expression recognition and action unit detection. *Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, IEEE*.

Wang, H., Chai, X., Hong, X., Zhao, G., and Chen, X. (2016) Isolated Sign Language Recognition with Grassmann Covariance Matrices. *ACM Transactions on Accessible Computing (TACCESS), 8*(14).

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yang, Y. and Ramanan, D. (2011). Articulated Pose Estimation with Flexible Mixtures-of-parts. *Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE, 2011*.

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

Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., and Presti, P. (2011). American Sign Language Recognition with the Kinect. *Proceedings of the 13th international conference on multimodal interfaces. ACM, 2011*.

Zahedi, M., Keysers, D., Deselaers, T., and Ney, H. (2005) Combination of Tangent Distance and an Image Distortion Model for Appearance-Based Sign Language Recognition In *Pattern Recognition*, Berlin / Heidelberg: Springer. pp. 401-408.

Zaki, M.M. and Shaheen, S.I. (2011) Sign Language Recognition using a Combination of New Vision Based Features. *Pattern Recognition Letters, 32*(4), pp. 572-77.