Scalable ASL Sign Recognition using Model-based Machine Learning and Linguistically Annotated Corpora

Dimitris Metaxas*, Mark Dilsizian*, Carol Neidle**
*Rutgers University Computer Science Department, **Boston University Linguistics Program
*110 Frelinghuysen Rd., Piscataway, NJ 08854-8019, **621 Commonwealth Ave., Boston, MA 02215
dnm@cs.rutgers.edu, mdil@cs.rutgers.edu, carol@bu.edu

Abstract

We report on the high success rates of our new, scalable, computational approach for sign recognition from monocular video, exploiting linguistically annotated ASL datasets with multiple signers. We recognize signs using a hybrid framework combining state-of-the-art learning methods with features based on what is known about the linguistic composition of lexical signs. We model and recognize the sub-components of sign production, with attention to hand shape, orientation, location, motion trajectories, plus non-manual features, and we combine these within a CRF framework. The effect is to make the sign recognition problem robust, scalable, and feasible with relatively smaller datasets than are required for purely data-driven methods. From a 350-sign vocabulary of isolated, citation-form lexical signs from the American Sign Language Lexicon Video Dataset (ASLLVD), including both 1- and 2-handed signs, we achieve a top-1 accuracy of 93.3% and a top-5 accuracy of 97.9%. The high probability with which we can produce 5 sign candidates that contain the correct result opens the door to potential applications, as it is reasonable to provide a sign lookup functionality that offers the user 5 possible signs, in decreasing order of likelihood, with the user then asked to select the desired sign.

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

1. Introduction

Whereas many older approaches to computer-based sign recognition from video had focused on a selection of features known to be linguistically relevant to sign production, more recent research that has exploited neural nets has generally not attended to what is known about linguistic structure. The latter approaches do not work well, however, in the absence of large quantities of annotated data, quantities that exceed what is generally available for sign languages currently. Furthermore, they fail to provide insights into cases where the recognition fails.

To address the linguistic and computer vision complexities associated with automatic sign recognition, we have developed a novel hybrid approach that utilizes a set of known linguistic properties of the language to optimize the parameterization for state-of-the-art machine learning methods. These methods also rely on linguistically annotated data for citation-form signs from our American Sign Language Lexicon Video Dataset (ASLLVD) (Neidle, Thangali, and Sclaroff, 2012).[1]

Our 3-step approach differs from most other methods since it uses parameters related to upper body and hand and face configuration, coupled with linguistic constraints (as reflected in the statistics from the dataset).

1) We first use neural networks to automatically extract the 2D upper body and facial features from a signer's video sequence. These features are then used to estimate the 2D pose of the signer, and then, using dynamic programming, to fit a 3D model to estimate the related parameters. We also extract hand features using another neural net trained for handshape recognition.

2) We then introduce linguistic dependencies to adjust the probabilities of estimated start and end handshapes; these are based on precomputed co-occurrence probability priors for start/end handshape combinations. We also add a parameter related to the possible relationships between handshapes on the 2 hands in 2-handed signs.

3) The previously estimated parameters related to the upper body and handshape probabilities, modified with linguistically based information, are then used in a modified Hidden Conditional Ordinal Random Field (HCORF) for sign recognition.

This unified hybrid framework for sign recognition offers impressive sign recognition results in a fully scalable manner. Using a 350-sign vocabulary of isolated, citation-form lexical signs, we achieve a top-1 accuracy of 93.3% and a top-5 accuracy of 97.9%.

Section 2 briefly situates our current approach in the context of previous attempts at sign recognition. Section 3 presents our framework; the experiments and results are summarized in Section 4. In Section 5, we discuss possible applications of this technology.

2. Previous Achievements in Sign Recognition

In the early 2000's, isolated sign recognition from video or RGBD sensors, often using features of the signing known to be linguistically significant (e.g., Bowden et al., 2004), demonstrated some success on small vocabularies.

Signer independence poses additional challenges. Von Agris et al. (2006), using extracted image features, achieved 96.9% signer-independent recognition of 153 signs from 4 native signers of British Sign Language. Later, von Agris, Knorr, and Kraiss (2008), by combining 2D motion trajectories, facial features, and a hand model, achieved 88.3%, 84.5%, and 80.2% respectively for signer-independent recognition of vocabularies of 150, 300, and 450 signs from 25 native signers of German Sign Language. These results indicate that scalability is an issue.

Zaki and Shaheen (2011), using hand-crafted features describing handshape and orientation, place of articula-

[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). The Sign Bank examples that were recorded as isolated signs, in citation form, are taken from the ASLLVD; the Sign Bank also includes additional examples taken from continuous signing.

tion, and hand motion, report 89.9% success in recognizing 30 ASL signs from 3 signers from the RWTH-BOSTON-50 database (Zahedi et al., 2005; that database is, in fact, comprised of a subset of 50 signs taken from the ASL data we had made publicly available and which are now shared through our Data Access Interface (DAI, and the new DAI 2); see Footnote 1).

For larger vocabularies, Cooper et al. (2011) attained 71.4% top-1 accuracy on a set of 984 signs from British Sign Language, but all from a single signer. Wang et al. (2016) achieved 70.9% accuracy on 1,000 isolated signs in Chinese Sign Language across multiple signers. However, they relied on an RGBD sensor for 3D information.

More recent approaches to sign language recognition, although focused on continuous signing rather than isolated signs, have been spurred by advances in neural nets. Such purely data-driven end-to-end approaches have been based on Recurrent Neural Net (RNN) architectures (e.g., Cui, Liu, and Zhang, 2017). Koller, Zargarin, and Ney (2017) use such an architecture, incorporating HMMs and 2D motion trajectories (but without integration of linguistic knowledge) to achieve 45.1% accuracy. Their multi-signer performance (27.1%) demonstrates that such methods do not generalize easily.

It is difficult to make direct comparisons with other sign recognition results because of vast differences in the nature of the data and conditions for research reported in the literature. In general, however, as the size of the dataset increases, the accuracy of isolated sign recognition has decreased. Methods used have not proved to be scalable. Our methods achieve both high accuracy in sign recognition on sizable vocabularies and scalability.

3. Overview of our Sign Recognition Framework

Our hybrid approach uses 1) discriminative neural net based computer vision methods coupled with generative methods for hand and pose feature extraction and related parameters, 2) additional linguistically driven parameters (Sections 3.1, 3.2), with enhancement of parameters from known linguistic dependencies (Section 3.3); and 3) scalable machine learning methods for sign recognition using the extracted parameters (Section 3.4); see Figure 2.

This results in improved sign recognition compared to previous approaches, because of the reduced parameterization and the efficiency of the algorithms, which are capable of coping with limited quantities of annotated data.

3.1 Summary of Features

Using the framework just described, we estimate a comprehensive set of features, 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 e) contact.

a) Features related to **handshape** are extracted from a neural net.

b) Signs are categorized based on the **number of hands** (1 vs. 2 hands) and the **degree of similarity of the handshapes** on the 2 hands for 2-handed signs.

c) The **upper body** parameters include 3D joint locations for the shoulders, arms, and wrists; velocities; and the **distance between the hands.**

d) The features for the **face** include 66 points (visible in Figure 1) from 3D estimates for the forehead, ear, eye, nose, and mouth regions, and their velocities across frames.

e) The **contact** parameters are extracted from our 3D face and upper body movement estimation, and relate to the possibilities of the hand touching specific parts of the body, e.g., the forehead or other parts of the face, arms, upper body, or the other hand.

The initial parameter values will, in some cases, be subsequently modified based on linguistic considerations, to be discussed in Section 3.3. This comprehensive set of parameters is then used within our CRF-based machine learning framework for purposes of sign recognition.

3.2 Feature Parameter Extraction

Next we describe how these parameters are extracted.

3.2.1 Upper Body, Hands, and Arms

We model upper body pose and use the 3D joint locations as features. We use Convolutional Neural Nets (CNNs) for initial estimation of 2D pose. We then apply a nearest neighbor matching coupled with a dynamic programming approach to search for the optimal 3D pose and part confidence maps (Dilsizian et al., 2016).

Using this 3D approach, we also extract linguistically important parameters, such as 3D motion trajectories, information about the number of hands (1- vs. 2-handed) and events involving contact between the 2 hands or contact with the face or body, as shown in Figure 1.

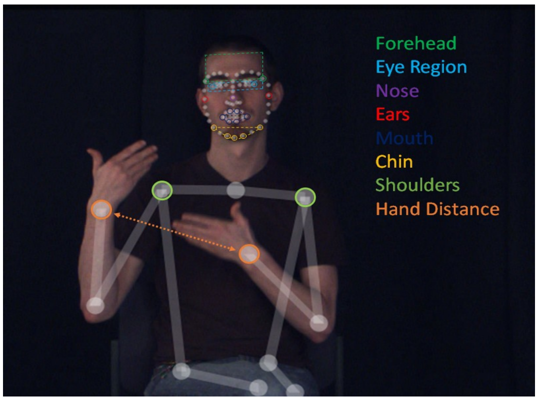

Figure 1. Locations where contact occurs

Handshape feature extraction and recognition have previously been demonstrated (Dilsizian et al., 2014; Ricco and Tomasi, 2009) with reasonable accuracy on limited datasets. More recently, CNNs have been used for robust recognition of New Zealand Sign Language handshapes from a large dataset with high variability (Koller, Ney, and Bowden, 2016). In our approach, we generate an ASL dataset for handshape recognition based on the publicly available ASLLVD corpus, which we plan to make available on the Web. We use wrist locations and forearm orientation to identify bounding boxes around the hands. We consider handshapes for which sufficient examples are available in the dataset (as is the case for 74 of the 86 handshapes). We balance the dataset by taking perturbations of shapes with fewer examples. Then we separate out 80% of all the obtained handshapes to be used as training exemplars in a CNN.

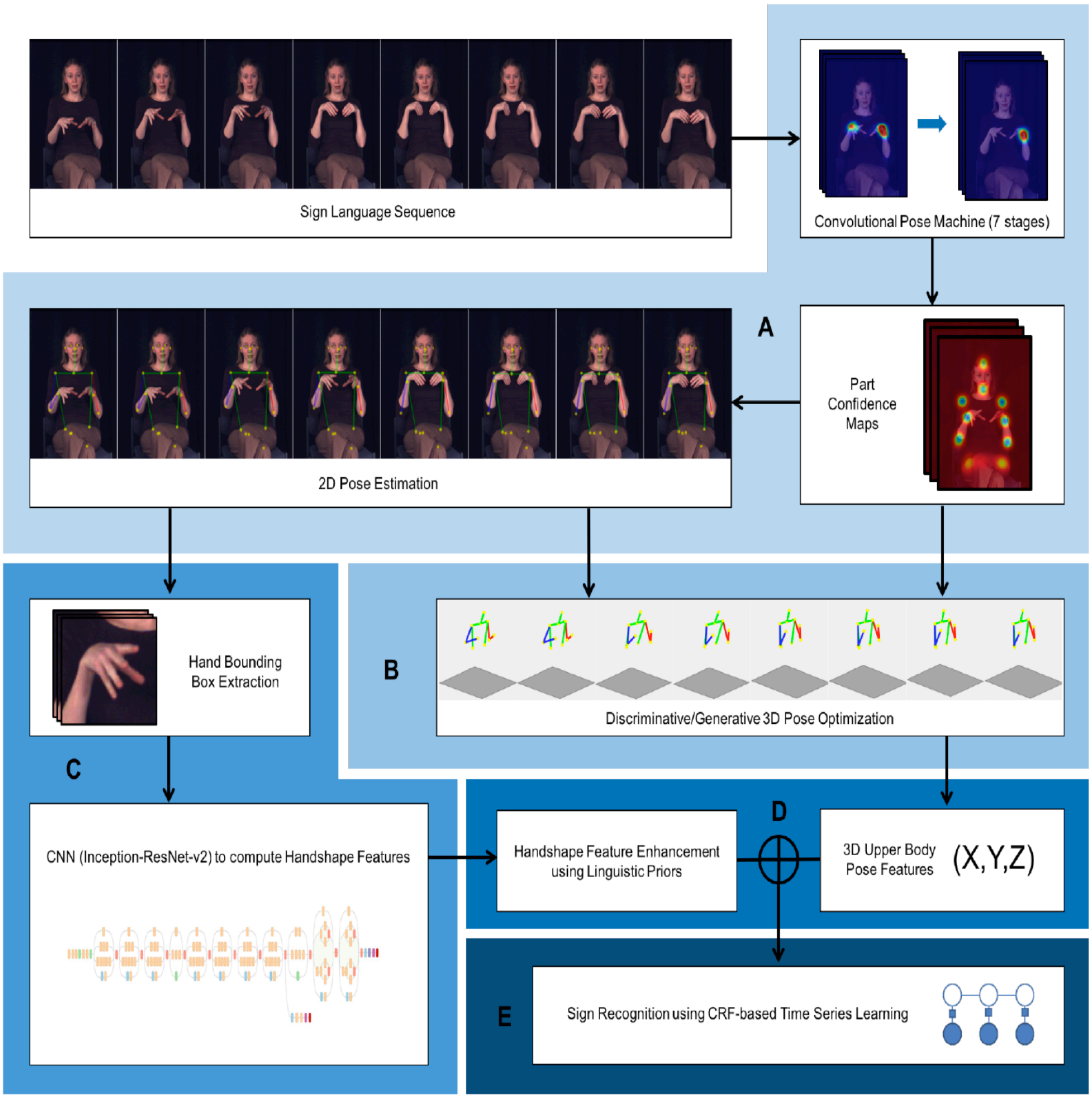
Sign Language Sequence
Convolutional Pose Machine (7 stages)
A
Part Confidence Maps
2D Pose Estimation
B
Discriminative/Generative 3D Pose Optimization
Hand Bounding Box Extraction
C
CNN (Inception-ResNet-v2) to compute Handshape Features
D
Handshape Feature Enhancement using Linguistic Priors
3D Upper Body Pose Features
(X,Y,Z)
E
Sign Recognition using CRF-based Time Series Learning

Figure 2. Overview of our sign recognition framework. (A) CNN-based 2D pose estimation, (B) novel discriminative/generative 3D pose estimation, (C) CNN-based handshape recognition, (D) linguistic enhancement and feature combination, and (E) CRF-based sign recognition

The human annotations make use of a set of discrete handshape labels. However, hand configurations exhibit variations along a continuum (i.e., they are not discrete). In addition, actual handshape configurations produced in the course of signing frequently differ from the canonical handshapes we are using in our idealization, and even humans may have difficulty in determining which is the closest canonical handshape for a given realization. To capture the varying production of the handshapes that are key to sign identification (the start and end handshapes being the most informative), we consider the entire set of output probabilities (for each handshape) of the CNN to be features for sign recognition, rather than focusing on a single handshape label with the highest probability.

In order to capture a set of output probabilities that is sufficiently descriptive, we must avoid overfitting to prevent the CNN from converging entirely to the most probable handshape labels during the course of a sign. We train Inception-ResNet-v2 (Szegedy et al., 2017) on the hand images because of its ability to capture information from both local and global appearance.

Although we use the entire set of handshape output probabilities computed by our CNN as features for sign recognition, we report handshape prediction accuracy to demonstrate the effectiveness of our approach. We achieve a top-1 accuracy of 70.1% on the testing dataset after 20 epochs of training. The top-5 accuracy reaches 92.3%. The top-1 and top-5 accuracy for the test set is shown over training epochs in Figure 3.

Thus, in the initial phase of our handshape feature extraction, we compute a vector of handshape probabilities (with a length of 74, as we are using the 74 handshapes for which we have a sufficient number of examples) for each hand in each frame during the production of signs in our sign recognition dataset.

3.2.3 Face and Head

Non-manual features have been shown to improve recognition of manual signs (von Agris, Knorr, and Kraiss, 2008; Koller, Forster, and Ney, 2015). Thus we estimate the 3D locations of 66 points on the face, as well as head movement, to include all possible informative non-manual information.

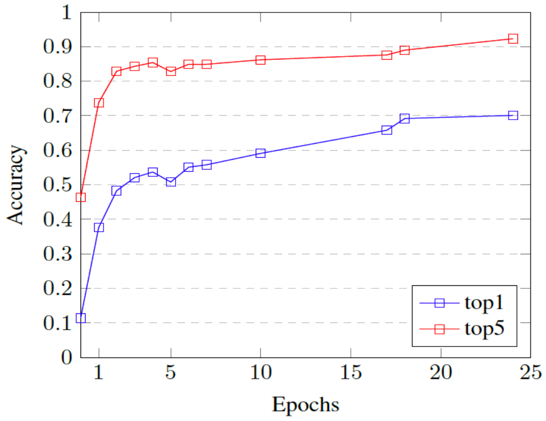

Figure 3. Top-1 and top-5 handshape recognition accuracy on test set by training epoch.

3.3 Incorporation of Linguistic Modeling for Enhancement of Parameter Estimates

The initial estimates of several of the above parameters can be refined based on known linguistic dependencies.

3.3.1 Dependencies between Start & End Handshapes

We exploit phonological constraints that hold between start and end handshapes in lexical signs to refine the handshape estimates for start and end handshapes (Thangali et al., 2011; Thangali 2013; Dilsizian et al., 2014). These dependences are reflected in the co-occurrence probabilities from our dataset.

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

We distinguish 2-handed signs that have essentially the same handshape on both hands from those that involve different handshapes, based in part on the handshape similarity parameter mentioned earlier. In the former case, we can boost handshape accuracy by combining information from the independent handshape estimates for the 2 hands. In the latter case, handshape possibilities for the non-dominant hand are significantly constrained.

3.4 Sign Recognition

We use the above extracted parameters as input to a structured Conditional Random Field (CRF) method—a modified Hidden Conditional Ordinal Random Field (HCORF) (Walecki et al., 2015)—to recognize signs. In addition, for each sequence, our modified HCORF includes an additional error term that measures the error between start/end handshape predictions and ground truth labels.

The advantages of our linguistically motivated, reduced parameter approach are demonstrated in the next section.

4. Sign Recognition Experiments and Results

4.1 Dataset

In this research we focus on lexical signs, the largest morphological class of signs. For training, we used the most comprehensive publicly accessible, linguistically annotated, video collection of isolated ASL signs, the American Sign Language Lexicon Video Dataset (ASLLVD) (Neidle, Thangali, and Sclaroff, 2012); see also Footnote 1. The ASLLVD itself includes over 8500 examples corresponding to almost 2800 monomorphemic lexical signs in citation form from 6 native signers. However, for these experiments, we selected a set of 350 signs from among those that had the greatest number of examples and signers per sign. 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). This was sufficient to train our neural nets.

4.2 Experiments

For each frame in each video sequence, we extract a feature vector of dimension 110, which includes the previously discussed features (handshape, motion trajectory, and other linguistically motivated features). This feature vector is used as input to our machine learning framework for sign recognition. We trained on our dataset, which generally contained 4-6 signers per example, using 80% of the data for training and 20% for testing. For each sign, 2 examples were randomly selected to be in the testing set, and the remaining examples were used for training. We tested on vocabularies of differing sizes (175 vs. 350 signs) to test the efficiency and scalability of our approach. We also performed a series of experiments to separate out the contributions of the different parameters.

4.3 Results

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 demonstrates the advantage of: 3D pose over 2D (green vs. amber); the addition of contact parameters (red); and the inclusion of all linguistic parameters and constraints in our framework (blue).

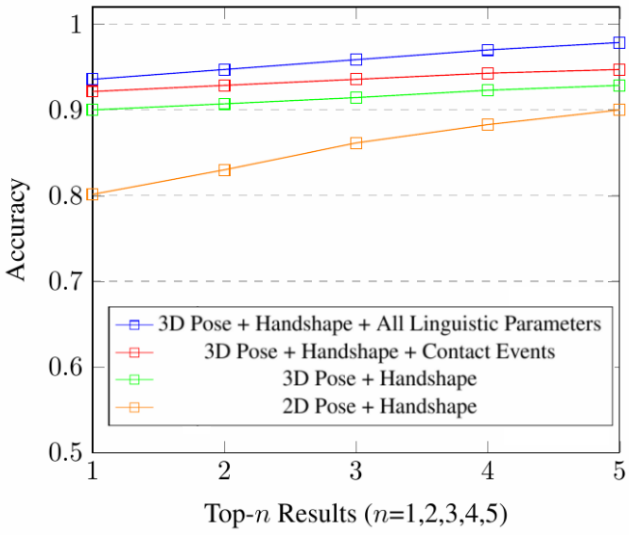

Figure 4. Contribution of Parameters to Accuracy

Comparing the results of vocabularies of 175 vs. 350 signs (Figure 5), accuracy declines by only 2.1% for top 1, and by only 1.3% for top 5 with the larger vocabulary. This provides evidence for the scalability of the approach.

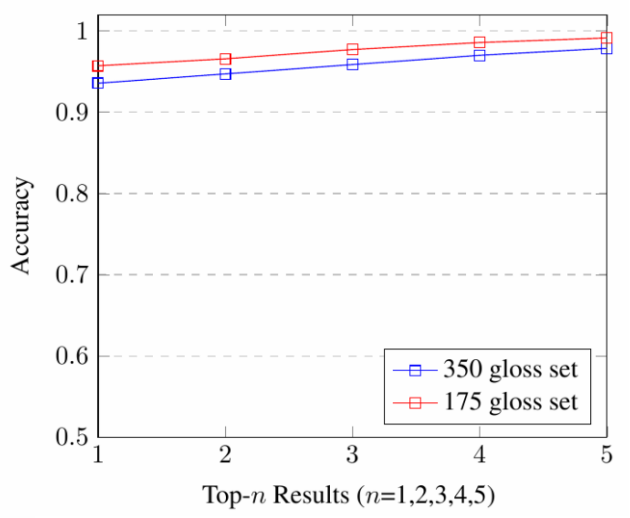

Figure 5. Comparing the Results on Vocabularies of 175 vs. 350 Signs

5. Significance for Potential Future Applications

There are many possible practical applications of technology for sign identification from video. For example, sign lookup capability would present significant benefits to Deaf communities, and to others wanting access to sign language resources such as dictionaries. Sign language dictionaries are currently often accessed by means of the written language, e.g., looking up a sign in an ASL resource by searching for a possible English translation of that sign. This has obvious drawbacks, as the user (whether Deaf or hearing) may not know the corresponding word from the spoken/written language. Available alternatives, which are in use for some sign language resources, generally involve laboriously having the user specify multiple features of the sign, such as handshape, location, movement type; this constitutes a very inefficient and unsatisfying lookup mechanism.

Our goal is to develop a lookup functionality that would enable users to search through our own electronic resources (Neidle et al., 2018), or to use our lookup interface to access other resources, through one of two input methods: either by producing the target in front of a webcam, or by identifying the start and end frames of the sign of interest from a video with continuous signing.

Although additional research will be required before such a lookup mechanism can be provided, the fact that we currently achieve about 98% success, using scalable methods, in identifying five candidate signs that include the target sign is extremely encouraging. It would be practically reasonable to offer the user 5 choices, in decreasing order of likelihood, as part of the lookup process, with the user able to view those sign videos and choose among the signs before confirming the selection and proceeding with the lookup, as sketched in Figure 6. Final design of such an interface will also involve consultation with prospective users of such tools.

6. Conclusions

We have demonstrated a general framework for recognition of isolated signs produced by multiple signers. Our framework leverages linguistic structure and dependencies, thereby enabling it to work from limited quantities of annotated data and to outperform previous methods. Our parameter extraction methods are based on state-of-the-art 3D handshape, face, and upper body parameter estimation, as well as integration of linguistic properties and constraints. The resulting modified parameter vector allows for a scalable and efficient approach to sign recognition.

In the future, we plan to expand the corpus and associated annotation sets to further improve the performance of our methods. We also intend to refine/augment the linguistically motivated features to enhance recognition accuracy, which would not be possible with purely data-driven methods. Furthermore, the methods being developed will, we hope, have beneficial practical applications, which we intend to pursue.

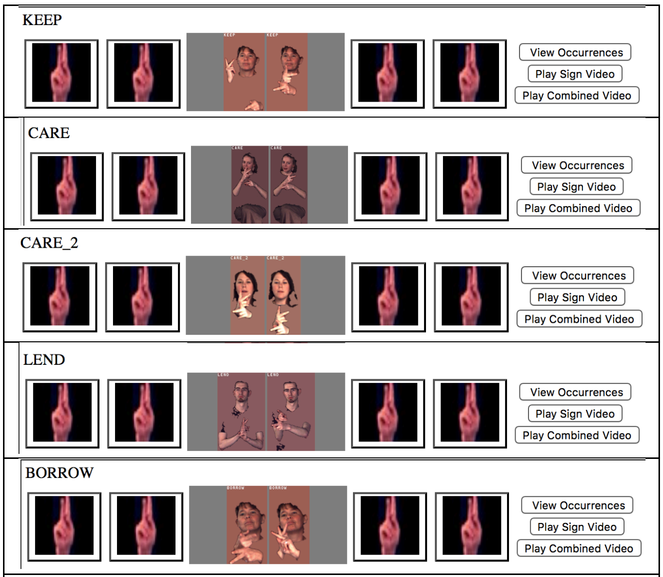

Figure 6. Example of Choices to be Offered to a User before Confirmation of a Lookup Selection – Based on the Interface for our Current ASLLRP Sign Bank

7. Acknowledgments

This research was partially funded by grants from the National Science Foundation (#1748016, 1748022, 1059218, 0705749, 0958247, 1703883, 1567289, 1555408, 1451292, 1447037). We also gratefully acknowledge contributions by many students and colleagues at Boston, Gallaudet, and Rutgers Universities and we thank our many ASL linguistic consultants. Special thanks to 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, and to Christian Vogler and Augustine Opoku for their help in developing our Web interfaces for data sharing.

8. Bibliographical References

Bowden, R., Windridge, D., Kadir, T., Zisserman, A. and Brady, M. (2004). A Linguistic Feature Vector for the Visual Interpretation of Sign Language. Proceedings of the ECCV.

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., 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.

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

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.

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.

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.

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. (2013) *Exploiting Phonological Constraints for Handshape Recognition in Sign Language Video*. Unpublished, Doctoral Dissertation, Boston University.

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.

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.

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).

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*.

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.