Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms

Vassilis Athitsos [1], Carol Neidle [2], Stan Sclaroff [3], Joan Nash [2], Alexandra Stefan [1], Ashwin Thangali [3], Haijing Wang [1], and Quan Yuan [3]

[1] Computer Science and Engineering Department, University of Texas at Arlington, Arlington, TX 76019, USA
[2] Linguistics Program, Boston University, Boston, MA 02215, USA
[3] Computer Science Department, Boston University, Boston, MA 02215, USA

Abstract

Looking up the meaning of an unknown sign is not nearly so straightforward as looking up a word from a written language in a dictionary. This paper describes progress in an ongoing project to build a system that helps users look up the meaning of ASL signs. An important part of the project is building a video database with examples of a large number of signs. So far we have recorded video examples for almost all of the 3,000 signs contained in the Gallaudet dictionary (and some others not listed there). Locations of hands and the face have been manually annotated for a large number of videos. Using this data, we have built an application that lets the user submit a video of a sign as a query, and presents to the user the most similar signs from the system database. System performance has been evaluated in user-independent experiments with a system vocabulary of 921 signs. For 67% of the test signs, the correct sign is included in the 20 most similar signs retrieved by the system.

1. Introduction

Looking up the meaning of an unknown sign is not nearly so straightforward as looking up a word from a written language in a dictionary. This paper describes progress in an ongoing project to build a system that helps users look up the meaning of ASL signs. Our efforts in this project include construction of a large annotated video dataset, as well as system implementation.

Our dataset contains video examples for almost all of the 3,000 signs contained in the Gallaudet dictionary (and some others not listed there). Each video sequence is captured simultaneously from four different cameras, providing two frontal views, a side view, and a view zoomed in on the signer's face. Our video dataset is available on the Web.

In the current system, the user submits a video of the unknown sign to look up its meaning. The system evaluates the similarity between the query video and every sign video in the database, using the Dynamic Time Warping (DTW) distance. System performance has been evaluated in user-independent experiments with a system vocabulary of 921 signs. In our experiments we only use a single frontal view for both test and training examples. For 67% of the test signs, the correct sign is included in the 20 most similar signs retrieved by the system. More detailed results are presented in the experiments section.

Our approach is differentiated from prior approaches to sign language recognition by the fact that it is both vision-based and user-independent, while also employing a large vocabulary (921 signs). Many approaches are not vision-based, but instead use input from magnetic trackers and sensor gloves, e.g., (Gao et al., 2004; Vogler and Metaxas, 2003; Yao et al., 2006). Such methods have achieved good results on continuous Chinese Sign Language with vocabularies of about 5,000 signs (Gao et al., 2004; Yao et al., 2006).

On the other hand, computer vision-based methods typically have been evaluated on smaller vocabularies (20-250 signs) (Bauer and Kraiss, 2001; Deng and Tsui, 2002; Dreuw and Ney, 2008; Fujimura and Liu, 2006; Kadir et al., 2004; Starner and Pentland, 1998; Zieren and Kraiss, 2005). While high recognition accuracy (85% to 99.3%) has been reported on vocabulary sizes of 164 signs (Kadir et al., 2004) and 232 signs (Zieren and Kraiss, 2005), those results are on user-dependent experiments, where the system is tested on users that have also provided the training data. In contrast, in our experiments the test signs are produced by users who do not appear in the training data, and the size of the vocabulary (921 distinct sign classes) is significantly larger than the vocabulary sizes that existing vision-based methods have been evaluated on.

2. Dataset: Videos and Annotations

In this section we describe the American Sign Language Lexicon Video Dataset (ASLLVD), which we have been building as part of this project. In particular, we update the information given in (Athitsos et al., 2008), to include the additional videos and annotations that we have added to this dataset in the last two years.

Our goal is to include video examples from a vocabulary that is similar in scale and scope to the set of lexical entries in existing ASL-to-English dictionaries, e.g., (Tennant and Brown, 1998; Valli, 2006). In the system vocabulary, we do not include name signs or fingerspelled signs, with the exception of some very commonly used ones (that are typically included in ASL dictionaries). We do not include classifier constructions, in which a classifier undergoes iconic movement, to illustrate the path or manner of motion, or the interaction of entities. The signs included in our dataset are restricted to the remaining (most prevalent) class of signs in ASL, which we refer to as "lexical signs."

At this point, we already have at least one video example per sign from a native signer, for almost all of the 3,000 signs contained in the Gallaudet dictionary (Valli, 2006). For a second signer we have collected 1630 signs, for a third signer we have collected 1490 signs, and for two additional signers we have collected about 400 signs. We would

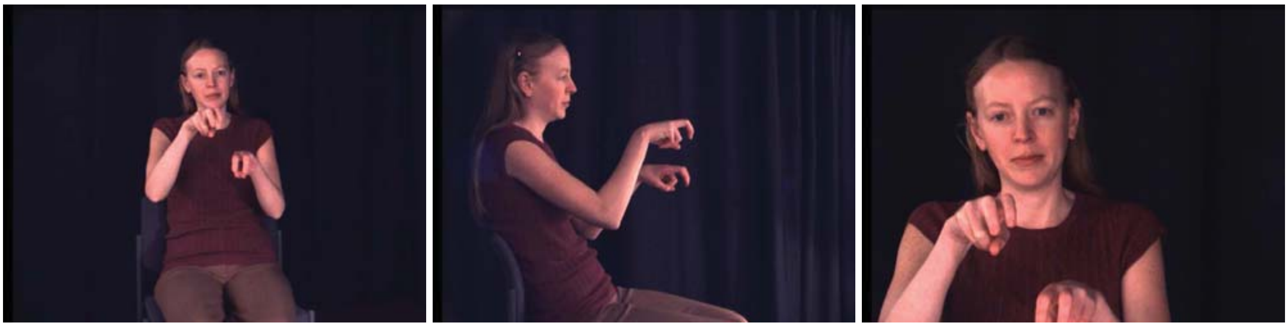

Figure 1: One of the frontal views (left), the side view (middle), and the face view (right), for a frame of a video sequence in the ASL Lexicon Video Dataset. The frame is from a production of the sign "merry-go-round."

eventually like to have at least three examples per sign for all signs in the system vocabulary.

2.1. Video Characteristics

The video sequences for this dataset are captured simultaneously from four different cameras, providing a side view, two frontal views, and a view zoomed in on the face of the signer. Figure 1 shows one of the frontal views, the side view, and the face view, for a frame of a video sequence in our dataset.

For the side view, the first frontal view, and the face view, video is captured at 60 frames per second, non-interlaced, at a resolution of 640x480 pixels per frame. For the second frontal view, video is captured at 30 frames per second, non-interlaced, at a resolution of 1600x1200 pixels per frame. All videos are available on the dataset websites, in formats employing both lossless compression (for higher video quality) and lossy compression (for faster downloading/browsing).

2.2. Annotations

The annotation for a video sequence contains, for each sign in that sequence, the start and end frames for that sign, a conventional English-based gloss of the sign, classification as one-handed or two-handed, and a signer ID. We also include manual annotations of the locations of the two hands and the face for a large number of signs. For hands, we mark at each frame the bounding box of the dominant hand, as well as the bounding box of the non-dominant hand for two-handed signs. For faces, we mark the bounding box of the face location at the first frame of each sign. Hand and face locations have been annotated for about 1500 sign examples from one signer, 1300 examples from a second signer, and 650 examples from a third signer.

The Gallaudet dictionary (Valli, 2006) includes a DVD containing a video example of every sign included in that dictionary. As those videos provide a valuable extra example per sign for almost all signs appearing in our dataset, we have annotated hand and face locations for about 1800 of the 3000 signs in that dictionary, and we intend to annotate the remaining signs in the next few months.

2.3. Availability

The ASLLVD dataset, including videos and annotations, is available for downloading on the project websites, located at the following two URLs:

- http://csr.bu.edu/asl_lexicon
- http://vlm1.uta.edu/~athitsos/asl_lexicon

In addition to the ASL Lexicon Video Dataset, a large quantity of ASL video and annotations that we have collected for previous projects is also available in various formats (on the Web from http://www.bu.edu/asllrp/ and on CD-ROM; see also (Dreuw et al., 2008)). This video dataset includes 15 short narratives (2-6 minutes in length) plus hundreds of elicited sentences, for a total of about 2,000 utterances with over 1,700 distinct signs and a total of over 11,000 sign tokens altogether. These data have been annotated linguistically, using SignStream$^{\text{TM}}$(Neidle, 2002; Neidle et al., 2001) (currently being reimplemented in Java with many new features). Annotations include information about the start and end point of each sign, part of speech, and linguistically significant facial expressions and head movements. The annotation conventions are documented (Neidle, 2002/2007) and the annotations are also available in XML format.

3. System Implementation

Signs are differentiated from one another by hand shape, orientation, location in the signing space relative to the body, and movement. In this paper we only use hand motion to discriminate between signs, leaving incorporation of hand appearance and body pose information as future work. Furthermore, we make the simplifying assumption that the system knows the location of the hands in all videos. The location of hands in all database sequences is manually annotated. Hand detection in the query sequence is performed in a semi-automatic way, where the system identifies hand locations using skin and motion information (Martin et al., 1998), and the user reviews and corrects the results.

Each sign video X is represented as a time series $(X_1, \dots, X_{|X|})$, where $|X|$ is the number of frames in the video. Each X_t, corresponding to frame t of the video, is a 2D vector storing the (x, y) position of the centroid of the dominant hand, for one-handed signs, or a 4D vector storing the centroids of both hands, for two-handed signs. For the purpose of measuring distance between the time-series representations of signs, we use the dynamic time warping (DTW) distance measure (Kruskal and Liberman, 1983).

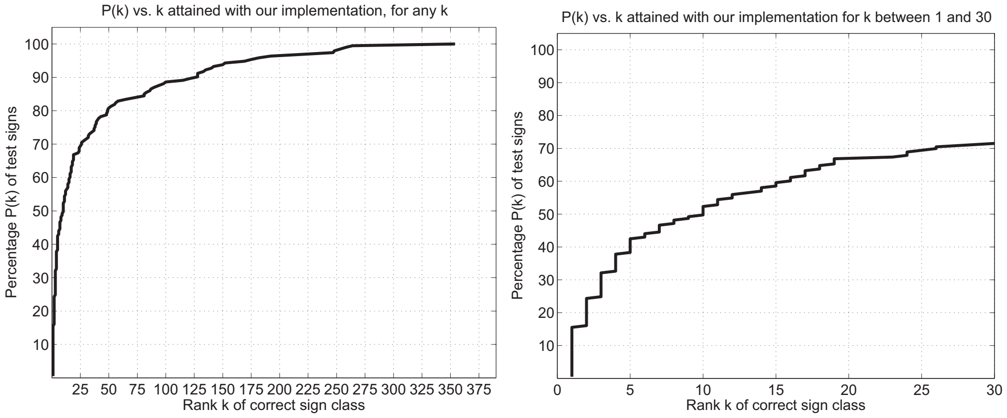

Figure 2: A plot of $P(k)$ vs. k illustrating the accuracy of our implementation. The x-axis corresponds to values of k. For each such value of k, we show the percentage of test signs $P(k)$ for which the correct sign class was ranked in the top k classes among all 921 classes. The plot on the right zooms in on a range of k from 1 to 30.

In particular, let Q be a test video and X be a training video. A warping path $W = ((w_{1,1}, w_{1,2}), \ldots, (w_{|W|,1}, w_{|W|,2}))$ defines an alignment between Q and X. The i-th element of W is a pair $(w_{i,1}, w_{i,2})$ that specifies a correspondence between frame $Q_{w_{i,1}}$ of Q and frame $X_{w_{i,2}}$ of X. Warping path W must satisfy the following constraints:

- **Boundary conditions:** $w_{1,1} = w_{1,2} = 1, w_{|W|,1} = |Q|$ and $w_{|W|,2} = |X|$.
- **Monotonicity:** $w_{i+1,1} - w_{i,1} \geq 0, w_{i+1,2} - w_{i,2} \geq 0$.
- **Continuity:** $w_{i+1,1} - w_{i,1} \leq 1, w_{i+1,2} - w_{i,2} \leq 1$.

For one-handed signs, the cost $C(W)$ of the warping path W is the sum of the Euclidean distances between dominant hand centroids of corresponding frames $Q_{w_{i,1}}$ and $X_{w_{i,2}}$. For two-handed signs, we include in the cost $C(W)$ the sum of the Euclidean distances between non-dominant hands in corresponding frames. The DTW distance between Q and X is the cost of the lowest-cost warping path between Q and X, and is computed using dynamic programming (Kruskal and Liberman, 1983), with time complexity $O(|Q||X|)$.

To address differences in translation between sign examples, we normalize all hand centroid positions based on the location of the face. The face location in database videos is manually annotated, whereas for test videos we use the face detector developed by (Rowley et al., 1998). To address differences in scale, for each training example we generate 121 scaled copies. Each scaled copy is produced by choosing two scaling parameters S_x and S_y, that determine respectively how to scale along the x axis and the y axis. Each S_x and S_y can take 11 different values spaced uniformly between 0.9 and 1.1. We should note that each of these multiple copies is not a new sign video, but simply a new time series, and thus the storage space required for these multiple copies is not significant.

4. Experiments

The test set used in our experiments consists of 193 sign videos, with all signs performed by two native ASL signers. The training set contains 933 sign videos, corresponding to 921 unique sign classes, and performed by a native ASL signer different from the signers appearing in the test videos. When submitting a test sign, the user specifies whether that sign is one-handed or two-handed. The system uses that information to automatically eliminate from the results signs performed with a different number of hands (it should be noted, however, that, especially for certain signs, there can be some variability in the number of hands used.). Although the ASLLVD dataset includes four synchronized camera views for each sign video, we only use the single 640x480 frontal view of each sign example in our experiments.

The results that we have obtained are shown in Figure 2. The measure of accuracy is a function $P(k)$ that measures the percentage of test signs for which the correct sign class was ranked in the top k out of the 921 classes. For example, in our results, $P(20) = 66.8\%$, meaning that for 66.8% of the 193 test signs, the correct sign class was ranked in the top 20 results retrieved by the system. In Figure 2, we include a plot focusing on a range of k from 1 to 30, as we believe few users would have the patience to browse through more than 30 results in order to find a video of the sign they are looking for. In Figure 3 we show an example of a query for which the correct match was ranked very low (rank 233), because of differences in the hand position between the query video and the matching database video.

On an Intel Xeon quad-core E5405 processor, running at 2.0GHz, and using only a single core, it takes on average 10 seconds to compute DTW distances and find the best matching results for a single test sign.

5. Discussion

In this paper we have provided an up-to-date description of the ASL Lexicon Video Dataset, a publicly available corpus that contains high-quality video sequences of thousands of distinct sign classes of American Sign Language, as well as manually annotated hand and face locations for a large number of those examples. We have also described an implementation of a system that allows users to look up the

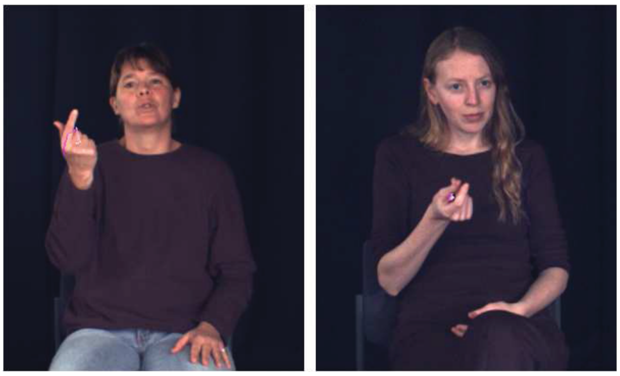

Figure 3: Example of a query sign for which the correct class ("dog") was ranked very low (rank 233). This sign exhibits small hand motion. A representative frame is shown for the query video (left) and for the correct database match (right). We note that the position of the hand is significantly different between the query and the database match.

meaning of an ASL sign, with a simple method based on hand centroids and dynamic time warping.

Using our simple implementation, the correct class is ranked in the top 20 classes, out of 921 sign classes, for 67% of the test signs. This is an encouraging result, given that we are not yet using any information from handshape, hand orientation, or body pose. At the same time, our current implementation does not work very well for a significant fraction of test signs. For example, for 19% of the test signs the correct class is not included in the top 50. We hope that including additional information, from features related to hand and body pose, will lead to significantly better results, and that is a topic that we are currently investigating.

Acknowledgments

We gratefully acknowledge the following native signers, who have served as models for the project: Naomi Berlove, Elizabeth Cassidy, Lana Cook,Tyler Richard, and Dana Schlang. Annotations were carried out with the assistance of Jaimee DiMarco, Chrisann Papera, Jessica Scott, Jon Suen, and Iryna Zhuravlova. The research reported here has been partially funded by grants from the National Science Foundation: HCC-0705749, IIS-0812601, MRI-0923494.

6. References

V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. 2008. The American Sign Language lexicon video dataset. In *IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB)*.

B. Bauer and K.F. Kraiss. 2001. Towards an automatic sign language recognition system using subunits. In *Gesture Workshop*, pages 64–75.

J. Deng and H.-T. Tsui. 2002. A PCA/MDA scheme for hand posture recognition. In *Automatic Face and Gesture Recognition*, pages 294–299.

P. Dreuw and H. Ney. 2008. Visual modeling and feature adaptation in sign language recognition. In *ITG Conference on Speech Communication*.

P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *International Conference on Language Resources and Evaluation*.

K. Fujimura and X. Liu. 2006. Sign recognition using depth image streams. In *Automatic Face and Gesture Recognition*, pages 381–386.

W. Gao, G. Fang, D. Zhao, and Y. Chen. 2004. Transition movement models for large vocabulary continuous sign language recognition. In *Automatic Face and Gesture Recognition*, pages 553–558.

T. Kadir, R. Bowden, E. Ong, and A. Zisserman. 2004. Minimal training, large lexicon, unconstrained sign language recognition. In *British Machine Vision Conference (BMVC)*, volume 2, pages 939–948.

J. B. Kruskal and M. Liberman. 1983. The symmetric time warping algorithm: From continuous to discrete. In *Time Warps*. Addison-Wesley.

J. Martin, V. Devin, and J.L. Crowley. 1998. Active hand tracking. In *Automatic Face and Gesture Recognition*, pages 573–578.

C. Neidle, S. Sclaroff, and V. Athitsos. 2001. SignStream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311–320.

C. Neidle. 2002. SignStream: A database tool for research on visual-gestural language. *Journal of Sign Language and Linguistics*, 4(1/2):203–214.

C. Neidle. 2002/2007. SignStream annotation: Conventions used for the American Sign Language Linguistic Research Project. Technical report, American Sign Language Linguistic Research Project Nos. 11 and 13 (Addendum), Boston University. Also available at http://www.bu.edu/asllrp/reports.html.

H.A. Rowley, S. Baluja, and T. Kanade. 1998. Rotation invariant neural network-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 38–44.

T. Starner and A. Pentland. 1998. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.

R. A. Tennant and M. G. Brown. 1998. *The American Sign Language Handshape Dictionary*. Gallaudet U. Press, Washington, DC.

C. Valli, editor. 2006. *The Gallaudet Dictionary of American Sign Language*. Gallaudet U. Press, Washington, DC.

C. Vogler and D. N. Metaxas. 2003. Handshapes and movements: Multiple-channel American Sign Language recognition. In *Gesture Workshop*, pages 247–258.

G. Yao, H. Yao, X. Liu, and F. Jiang. 2006. Real time large vocabulary continuous sign language recognition based on OP/Viterbi algorithm. In *International Conference on Pattern Recognition*, volume 3, pages 312–315.

J. Zieren and K.-F. Kraiss. 2005. Robust person-independent visual sign language recognition. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, volume 1, pages 520–528.