To appear in Proceedings of IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB), June 2008.

The American Sign Language Lexicon Video Dataset

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

[1]Computer Science and Engineering Department, University of Texas at Arlington, USA

[2]Linguistics Program, Boston University, Boston, Massachusetts, USA

[3]Computer Science Department, Boston University, Boston, Massachusetts, USA

Abstract

The lack of a written representation for American Sign Language (ASL) makes it difficult to do something as commonplace as looking up an unknown word in a dictionary. The majority of printed dictionaries organize ASL signs (represented in drawings or pictures) based on their nearest English translation; so unless one already knows the meaning of a sign, dictionary look-up is not a simple proposition. In this paper we introduce the ASL Lexicon Video Dataset, a large and expanding public dataset containing video sequences of thousands of distinct ASL signs, as well as annotations of those sequences, including start/end frames and class label of every sign. This dataset is being created as part of a project to develop a computer vision system that allows users to look up the meaning of an ASL sign. At the same time, the dataset can be useful for benchmarking a variety of computer vision and machine learning methods designed for learning and/or indexing a large number of visual classes, and especially approaches for analyzing gestures and human communication.

1. Introduction

American Sign Language (ASL) is used by 500,000 to two million people in the U.S. [10, 17]. Unfortunately, many resources that are taken for granted by users of spoken languages are not available to users of ASL, given its visual nature and its lack of a standard written form. One such resource is the ability to look up the meaning of an unknown sign. When we encounter an English word that we do not understand, we can look it up in a dictionary. Unfortunately, when an ASL user encounters an unknown sign, it is anything but straightforward to find the meaning of that sign. Using a typical printed ASL/English dictionary, one can easily find out what sign corresponds to an English word, but this does not work in the other direction, to enable discovery of the meaning of an unknown sign. There are dictionaries that allow look-up based on articulatory properties of the signs. For example, the *American Sign Language Handshape Dictionary* [20] arranges signs based on the initial handshape, from among 40 basic handshapes. However, even with that dictionary, substantial effort is needed to find a specific sign from among the 1600 included.

This paper introduces the ASL Lexicon Video Dataset, a new and expanding public dataset that contains high-quality video sequences of thousands of distinct ASL signs. This dataset is being created as part of a project to develop a real-world vision-based system that allows users to look up the meaning of an ASL sign. An important aspect of this dataset is its comprehensiveness: we plan to include a set of signs similar in scale and scope to the set of lexical entries in existing English-to-ASL dictionaries [4, 19, 20, 21].

A number of approaches have been proposed for sign language recognition (see [14] for a recent review). Many approaches are not vision-based, but instead use magnetic trackers and sensor gloves, e.g., [8, 11, 16, 22, 23, 25]. Such methods achieve good recognition results on continuous Chinese Sign Language with vocabularies of about 5,000 signs [8, 23, 25]. On the other hand, vision-based methods, e.g., [2, 5, 6, 7, 9, 18, 24] use smaller vocabularies (20-300 signs) and often rely on color markers, e.g., [2, 6]. We hope that our dataset will help promote research towards developing vision-based methods that operate on markerless images and can handle a more comprehensive vocabulary.

In the ASL Lexicon Video Dataset, each sign is performed by native ASL signers. The video sequences are collected in our recording facility, using a four-camera system that simultaneously captures two frontal views, one side view, and one view zoomed in on the face of the signer. The annotations include, for each sign in a video sequence, information such as class label, type of sign (one-handed or two-handed), start/end frame for that sign, and signer ID.

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.

We believe that the video and annotations in our dataset will be a valuable resource for researchers in sign language recognition, gesture recognition, and human activity analysis. The video and annotations can be used to build models and train classifiers for a comprehensive set of ASL signs. The dataset can also be useful for benchmarking different methods and for helping showcase new and better methods for gesture and sign recognition that significantly improve on the current state of the art.

At the same time, we believe that our dataset presents challenges that are relevant to the broader areas of computer vision, machine learning, and data mining. Open research problems in these areas that are highlighted by our dataset include discriminating among thousands of visual motion classes, and achieving memory and runtime efficiency in the presence of massive amounts of video data. We believe that the dataset can serve both as motivation and as a challenging benchmarking platform for researchers investigating such problems.

In addition to describing the dataset, we also provide some baseline experimental results, where a simple computer vision technique, motion energy images [3], is applied for retrieving similar signs given a query video sequence. The experiments highlight the challenge of creating a vision-based sign lookup system, but they also provide some encouraging results, as for a small but significant fraction of test sequences the correct sign class is ranked relatively high among the set of all sign classes.

2. Scope of the Dataset

The number and type of signs included in our dataset is similar in scale and scope to the set of lexical entries in existing English-to-ASL dictionaries [4, 19, 20, 21]. 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 of American Sign Language* [21]. Our goal is to collect, for each sign, examples from four native users of ASL.

In regular ASL discourse, signers sometimes use fingerspelling, i.e., they spell out English words (typically proper nouns, technical terms, or borrowings) using the letters of the manual alphabet. With the exception of some commonly used signs composed of such letters, frequently referred to as "loan signs," fingerspelled items would not normally be included in an ASL dictionary [4, 19, 20, 21], and they will not be included in our lexicon dataset.

We also do not plan to include constructions that involve what linguists refer to as "classifiers," signs that convey information about the size or shape or other classification of an entity. In "classifier constructions," the classifier undergoes iconic movement, to illustrate the path or manner of motion, or the interaction of entities. For example, one might sign CAR followed by a classifier for VEHICLE that might then trace the path of the car, up hill, turning right, etc. Such constructions do not involve a finite repertoire of movements, and therefore are not appropriately stored in a lexicon (although the individual classifier, e.g., VEHICLE, could be looked up). The signs included in our dataset will be restricted to the remaining (most prevalent) class of signs in ASL, which we refer to as "lexical signs."

3. 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. In both the side view and two frontal views the upper body occupies a relatively large part of the visible scene. In the face view, a frontal view of the face occupies a large part of the image. All sequences are in color. 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, first frontal view, and 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.

This high-resolution frontal view may facilitate the application of existing hand pose estimation and hand tracking systems on our dataset, by displaying the hand in significantly more detail than in the 640x480 views.

The video sequences are stored in a format that employs lossless compression. C++ code for reading these video files is also available on the website. For each video sequence, a compressed QuickTime version is also available, for viewers who want to quickly browse our dataset and identify content of interest. Given the very large amount of storage required for all this data, for the time being we are not making the high-resolution frontal view available on the web, although we can still provide it to interested researchers by shipping hard drives with the data.

At this point, the video sequences collected employ a neutral backdrop (e.g., Figure 1). A simple background can facilitate tasks like semi-automated annotation of hand and face locations, and training of models for sign classification. In the future, we may include video sequences (to be used for testing recognition algorithms) with more complex backgrounds, containing clutter and moving objects, so as to simulate realistic conditions under which a computer vision-based sign lookup system would operate.

4. Annotations

One important piece of annotation is the class label for every sign. Since there is no written form for ASL signs, transcriptions frequently associate each ASL sign with some approximate English translation, called a "gloss." Some conventions have developed for these associations (although these differ to some degree), since in fact there is no one-to-one correspondence between ASL and English words, any more than would be the case for English and French, or English and Greek. In our annotations, the class label is a gloss of the sign. Since glosses are used as class labels, two signs are assigned the same gloss if and only if they correspond to the same ASL lexical item.

Each video sequence contains multiple signs. The annotation for a video sequence contains, for each sign in that sequence, the start and end frames for that sign, the gloss of the sign, whether the sign is one-handed or two-handed, and a signer ID. Signer IDs will allow researchers to set up experiments for user-dependent and user-independent sign recognition.

In the near term we plan to include annotations of the locations of the two hands and the face at each frame. This type of annotation has only been performed for relatively few video sequences at this point; we are in the process of annotating the remaining sequences that we have collected. For each hand and the face, we mark the approximate location as a bounding square.

5. Collecting Experimental Results

We plan to include on the dataset website a table of results achieved on this dataset by different methods. Researchers experimenting on the dataset are encouraged to contact us to have their results included. In order to allow direct comparisons between methods, we will encourage researchers to specify:

- The set of sequences used for training and testing.
- The set of views (out of the three camera views available) used for training and testing.
- Whether recognition experiments were performed in a user-dependent or user-independent manner.

In light of the large scale of the dataset, several statistics are relevant for evaluating performance:

- **Rank 1 accuracy:** This performance statistic is a single number: the fraction of test sequences that were correctly classified, i.e., for which the correct class was ranked as the most likely class by the system. However, given the large number of classes, we do not expect rank 1 accuracy to be as important a measure as in other benchmark datasets, which contain no more than a few tens of classes.
- **Ranking statistics:** This performance statistic consists of a number per test sequence: the rank of the correct sign class, among all sign classes.
- **Runtime:** Runtime can be measured in seconds or in CPU cycles. Naturally, to make comparisons fair, information about the platform and implementation language should also be provided.
- **Memory requirements:** The video sequences currently available are already too large (total size of hundreds of gigabytes) to be all loaded in memory simultaneously, at least for typical current computer systems. Utilizing feature and model representations that achieve memory efficiency is a key aspect of designing methods that are applicable in real-world settings.

Naturally, the dataset can also be used for evaluating algorithms that do not focus on sign recognition and retrieval per se, but focus instead on other video analysis tasks, such as hand detection and tracking, human body tracking, facial expression analysis, etc. We also are interested in including such results on our website.

6. Availability

Video sequences containing a total of about 3,800 signs (corresponding to about 3,000 unique class labels) have

been collected so far. Compressed versions of those sequences, and annotations for at least 1,200 of those signs, are publicly available from the project websites:

- `http://www.bu.edu/asllrp/lexicon/`
- `http://crystal.uta.edu/~athitsos/asl_lexicon`

The website contents are updated frequently, with hundreds of new sequences and annotations added on a monthly basis.

We plan on making uncompressed sequences available online in the near term. However, downloading the entire set of uncompressed sequences may not be a realistic option, due to the sheer volume of data (hundreds of gigabytes). We can also provide the uncompressed sequences to interested researchers by shipping hard drives with the data.

In addition to the ASL Lexicon Video Dataset, a large quantity of ASL video and annotations that we collected for previous projects is available on the web, at `http://ling.bu.edu/asllrpdata/queryPages/`. This video dataset includes 15 short narratives (2-6 minutes in length) plus hundreds of elicited utterances, for a total of about 2,000 utterances with over 1700 distinct signs (up to 200 tokens for some signs), and a total of over 11,000 sign tokens altogether.

7. Experiments

While our dataset can be used for testing a wide variety of computer vision, machine learning, and database indexing algorithms, our primary goal is to use this dataset as part of a computer vision system that allows users to look up the meaning of a sign automatically. In such a system, the user performs the sign in front of a camera (or, possibly, in a multi-camera set up), and the computer retrieves and displays the most similar signs in the lexicon dataset. Naturally, achieving good accuracy and efficient performance will be challenging tasks, as existing vision-based gesture recognition methods are typically designed for far smaller gesture vocabularies.

7.1. Baseline Method: Motion Energy Images

To kickstart the sign lookup project, and to put an initial entry in our table of experimental results, we have implemented a very simple method for sign retrieval, which is a variation of motion energy images [3]. Each test and training video sequence consists of a segmented sign. For each such sequence V, we compute a summary motion energy image $M(V)$ as follows: We denote by $V_m(i,j)$ the pixel value of video frame V_m at pixel location (i,j) (where $m \in \{1,\ldots,|V|\}$), and by $|V|$ the number of frames in a video sequence V. We define the motion energy $D_{V,m}$ at frame m of V as:

$$D_{V,m}(i,j) = \min(\{|V_{m+1}(i,j) - V_m(i,j)|, |V_m(i,j) - V_{m-1}(i,j)|\} \quad (1)$$

The motion energy image M_V for sequence V is defined as:

$$M_V(i,j) = \sum_{m=2}^{|V|-1} D_{V,m}(i,j)\ . \quad (2)$$

Each motion energy image is normalized for translation, based on the center of the face in the first frame where a face is detected in that sequence. If no face is detected, no translation normalization is performed. Face locations are obtained using a face detector developed by Rowley, et al. [15]. Also, each motion energy image is blurred, by applying 10 times successively the filter $[.25,.5,.25]^t[.25,.5,.25]$, where A^t denotes the transpose of matrix A.

Given a test video sequence Q and a training video sequence V, the similarity between Q and V is simply the normalized correlation between M_Q and M_V. Only the 640x480 frontal view is used for each sequence. The similarity between Q and a sign class C is defined as the highest normalized correlation between M_Q and any M_V such that V is a training sequence belonging to class C. This way, we assign a similarity between Q and each sign class, and we can obtain a ranking of all sign classes in decreasing order of similarity to Q. The best result is for the correct class to have rank 1. More generally, the higher (closer to 1) the rank of the correct class is for Q, the better the recognition result is considered to be for Q.

7.2. Results

We have applied this simple motion energy method, using a test set of 206 video sequences belonging to 108 distinct glosses (used as class labels) and a training set of 999 video sequences belonging to 992 distinct glosses. In other words, for almost all sign classes we had only one training example. Obviously, the worst possible ranking result for the correct class of any test sign is 992. The test sequences were signed by two signers, and the training sequences were signed by another signer, who did not sign in any of the test sequences. Thus, the experiments are user-independent.

Figure 2 and Table 1 show the results. As can be seen, motion energy images do not perform very well over the entire test set, even when using the one-handed vs. two-handed information. This performance is expected, as it is obtained using a simple baseline method. We believe that researchers developing gesture recognition methods will find it easy to improve upon these baseline results.

For about 20% of our test sequences the correct sign class is ranked in the top 2.2% of all classes, even without

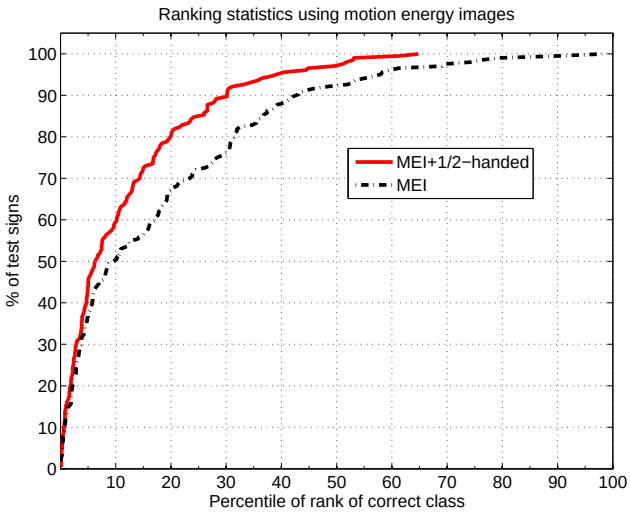

Figure 2. Ranking statistics using motion energy images, including one-handed vs. two-handed information (MEI+1/2-handed) and excluding one-handed vs. two-handed information (MEI). For every percentile R on the x axis, we show on the y axis the percentage of test signs for which the correct class was ranked in the top R percentile. For example, using only MEI, for 50% of the test examples, the correct class was ranked in the top 9.5% among all classes.

percentage of test signs	percentile of rank of correct class	
	only MEI	MEI + 1/2-handed
10	0.8	0.7
20	2.2	1.8
30	3.7	2.8
40	5.7	4.7
50	9.5	6.2
60	17.1	10.2
70	23.5	14.4
80	31.4	20.0
90	43.0	30.1
100	98.1	64.8

Table 1. Some cumulative ranking statistics obtained using motion energy images (MEI), and motion energy images plus one-handed vs. two-handed information (MEI + 1/2-handed). For every row, if P is the percentage indicated on the left column, then the middle and the right column indicate (for the respective methods) the percentile R so that for P percent of test signs (out of 206 test signs) the correct class ranked in the top R percentile among all 992 classes. For example, using only MEI, for 20% of the test examples, the correct class was ranked in the top 2.2% among all classes.

using the one-handed vs. two-handed information. For a lexicon size of about 3,000, such as in the Gallaudet dictionary [21], 2.2% of all signs would include 66 of the 3,000 signs. On the other hand, for 50% of the test signs, the correct class was ranked lower than the top 9.5% percentile (using only MEI information). A significantly higher accuracy is needed to make a purely vision-based sign lookup system ready for real-world deployment.

In addition to results that only use motion energy images, we also include results where, for each test sign, the system knows whether the sign is one-handed or two-handed, and uses that information as well. Specifying whether the sign of interest is one-handed or two-handed is a piece of information that can easily be provided by the user. Recognition performance improves when this extra information is provided to the system. At the same time, we should point out that, in actual signing, it is not uncommon for two-handed signs to be produced with a single hand, and it is also possible (although less likely) to find a one-handed sign produced with two hands. Consequently, specifying whether the query sign is one-handed or two-handed may not be as useful in practice as indicated in our experiments.

Figures 3 and 4 illustrate respectively test sequences that produce really good and really bad ranking results. Matching motion energy images is correlation-based, and thus very sensitive to the spatial alignment of the two videos. Even if faces are successfully detected and used for translation normalization, the variations in which the same sign can be performed by different signers can cause spatial misalignments. Furthermore, scale differences can be an issue, and we have not used any type of scale normalization in these experiments. Perhaps similarity measures that are more robust to spatial misalignments, such as the chamfer distance [1], can yield improved accuracy in this context.

The average processing time per query sequence was about four seconds when information about one-handed vs. two-handed signs was used by the system, and six seconds when one- vs. two-handed information was not used. This includes the time needed to compute the motion energy image of the test sequence, detect the face in at least one frame, and compute normalized correlations with the motion energy images of all training examples. The training motion energy images were precomputed and stored in main memory offline, taking up about 300MB of RAM (640x480x999 bytes). We note that this processing time is acceptable for our application scenario where a user wants to look up the meaning of a sign. Experiments were performed on a PC with a 2.40GHz Intel Core2 Quad CPU, using only one of the four cores. Our system was implemented in C++.

8. Discussion and Conclusions

We have introduced a new large-scale dataset, the ASL Lexicon Video Dataset, containing video sequences of thousands of distinct sign classes of American Sign Language. This dataset is publicly available, and expanding rapidly. We believe that this dataset will be an important resource for researchers in sign language recognition and human ac-

Figure 3. Examples of signs for which the correct class was ranked first (which is the best possible result) even without using one-handed vs. two-handed information. For each sign, we show, from left to right, the first frame, a middle frame, the last frame, and the blurred motion energy image. Top: an example of the sign DIRTY. Bottom: an example of the sign EMBARRASSED.

Figure 4. Examples of signs for which the correct class was ranked very low, even using using one-handed vs. two-handed information. For each sign, we show, from left to right, the first frame, a middle frame, the last frame, and the blurred motion energy image. Top: an example of the sign COME-ON, where no face was detected and the rank of the correct class was 299. Bottom: an example of the sign DISAPPEAR, where the rank of the correct class was 643.

tivity analysis, by providing a large amount of data that can be used for training and testing, and by providing a public benchmark dataset on which different methods can be evaluated. At the same time, some of the challenges posed by this dataset are relevant more broadly to researchers in the computer vision, machine learning, and data mining communities. Such challenges include the difficulty of discriminating among thousands of sign classes, and the need for efficiency in training and classification algorithms involving massive amounts of video data.

Currently the annotations for each video sequence include, for each sign in that sequence, the start and end frames, gloss, signer ID, and whether the sign is one-handed or two-handed. In the longer term, we plan to expand the annotations to include additional phonological and morphological information, e.g., handshape, type of hand motion, and position of the hand with respect to the body. Enhancements currently under development for SignStream®[12, 13] (an application designed for linguistic analysis of visual language data; SignStream is currently being reimplemented in Java with many new features) will facilitate more detailed annotations.

We are creating this dataset as part of a project to build a computer vision-based system that allows users to look up

the meaning of an ASL sign. We believe that an easy-to-use sign lookup tool will be a useful resource for both users and learners of ASL. Our preliminary experiments, with a simple method based on motion energy images, indicate both the promise and the challenges of the sign lookup project. We are actively working on developing better methods for sign matching. At the same time, we hope that the availability of the ASL Lexicon Video Dataset will encourage and help other researchers to study the problems of sign language recognition, gesture recognition, and human activity analysis, so as to significantly improve the current state of the art.

Acknowledgments

This work was supported in part by NSF grants IIS-0705749 and CNS-0427988.

References

[1] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *International Joint Conference on Artificial Intelligence*, pages 659–663, 1977.

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

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(3):257–267, 2001.

[4] E. Costello. *Random House American Sign Language Dictionary*. Random House, New York, 1994.

[5] Y. Cui and J. Weng. Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, 78(2):157–176, 2000.

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

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

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

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

[10] H. Lane, R. Hoffmeister, and B. Bahan. *A Journey into the Deaf-World*. DawnSign Press, San Diego, CA, 1996.

[11] J. Ma, W. Gao, J. Wu, and C. Wang. A continuous Chinese Sign Language recognition system. In *Automatic Face and Gesture Recognition*, pages 428–433, 2000.

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

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

[14] S. C. W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891, 2005.

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

[16] H. Sagawa and M. Takeuchi. A method for recognizing a sequence of sign language words represented in a Japanese Sign Language sentence. In *Automatic Face and Gesture Recognition*, pages 434–439, 2000.

[17] J. Schein. *At home among strangers*. Gallaudet U. Press, Washington, DC, 1989.

[18] T. Starner and A. Pentland. 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, 1998.

[19] M. L. A. Sternberg. *American Sign Language Dictionary Unabridged*. Collins, 1998.

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

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

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

[23] C. Wang, S. Shan, and W. Gao. An approach based on phonemes to large vocabulary Chinese Sign Language recognition. In *Automatic Face and Gesture Recognition*, pages 411–416, 2002.

[24] M. Yang and N. Ahuja. Recognizing hand gesture using motion trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 466–472, 1999.

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