A System for Large Vocabulary Sign Search

Haijing Wang [1], Alexandra Stefan [1], Sajjad Moradi [1], Vassilis Athitsos [1], Carol Neidle [2], and Farhad Kamangar [1]

[1] Computer Science and Engineering Department, University of Texas at Arlington
Arlington, Texas 76019, USA

[2] Linguistics Program, Boston University
Boston, Massachusetts 02215, USA

Abstract. A method is presented to help users look up the meaning of an unknown sign from American Sign Language (ASL). The user submits a video of the unknown sign as a query, and the system retrieves the most similar signs from a database of sign videos. The user then reviews the retrieved videos to identify the video displaying the sign of interest. Hands are detected in a semi-automatic way: the system performs some hand detection and tracking, and the user has the option to verify and correct the detected hand locations. Features are extracted based on hand motion and hand appearance. Similarity between signs is measured by combining dynamic time warping (DTW) scores, which are based on hand motion, with a simple similarity measure based on hand appearance. In user-independent experiments, with a system vocabulary of 1,113 signs, the correct sign was included in the top 10 matches for 78% of the test queries.

Key words: American Sign Language recognition, gesture recognition

1 Introduction

This paper focuses on a specific application, namely helping users look up the meaning of a sign in American Sign Language (ASL). Looking up the meaning of a sign is not a straightforward task. ASL dictionaries typically allow look-up of ASL signs based on their English translations; that is, these dictionaries are really English to ASL dictionaries, which makes it difficult to look up a sign if the user either does not know the meaning of that sign, or does not know its translation into English. There are ASL dictionaries that allow access to ASL signs based on their articulatory properties, such as handshape [1], but these interfaces either require specification of many articulatory parameters, or else they require the user to scan long lists of signs (for example that share a particular handshape). These lookup methods may fail entirely if the signer is looking for a variant that is different in a small way from the dictionary entry or errs with respect to the specification of the articulatory parameter(s). A system that helps users look up unknown signs would be useful to the millions of users and learners of sign languages around the world (estimated 0.5 to 2 million users in the US [2, 3]). The capability to look up signs would be particularly useful to students of a sign language, as useful as it is for students of a written language (such as English) to be able to look up the meaning of unknown words.

In our approach, having encountered an unknown sign, the user can simply perform the sign in front of a webcam. Then, the system compares the input sign with videos of signs stored in the system database, and presents the most similar signs (and potentially also their English translations) to the user. The user can then view the results and decide which (if any) of those results is correct.

In order to produce a system that works well enough for public use, we have opted for a not fully automatic system, which requires knowing the bounding box of the hands in each frame of both the test and the database videos. For our dataset, we have chosen videos from the public ASLLVD resource [4], where such hand locations are provided for thousands of examples of signs. In our demo system, the user specifies/verifies hand locations, in collaboration with a semi-automated hand detector. Making hand detection more or entirely automatic is a challenging task that we have left for future work.

In our dataset, we have examples from a large vocabulary of 1,113 distinct sign classes. A key constraint is that we only have two training examples for each sign. Given the small number of examples per sign, we use an exemplar-based method, as opposed to a model-based method, such as Hidden Markov Models (HMMs). We start with a baseline similarity measure based on the popular dynamic time warping (DTW) distance [5]. DTW is applied on time series of feature vectors based on hand motion. We improve this baseline similarity measure by incorporating information from hand appearance.

We evaluate our approach in user-independent experiments with a system vocabulary of 1,113 signs. The correct sign was included in the top 10 matches for 78% of the test queries. By considering more signs per query, the user can successfully look up an even larger percentage of query signs. These results are a significant improvement over results previously reported in the literature for comparable vocabulary sizes and under user-independent settings.

2 Related Work

Several methods exist for recognizing isolated gestures or signs, as well as continuous signing. The majority of existing methods are model-based, using Hidden Markov Models [6–9] or alternative approaches such as recursive partition trees [10], boosted volumetric features [11], and hidden conditional random fields [12]. Such methods typically use ten or more training examples per gesture or sign. In contrast, in our setting, we have only two training examples per sign.

Using more examples per sign typically improves accuracy (see, e.g., [13, 14]), but may not be an option, due to lack of data. For example, the Gallaudet dictionary of ASL [15] includes 3,000 signs, and the only public video dataset currently available for a vocabulary of that size is the ASLLVD resource [4], where only two examples are available for most of the signs. Cooper et al. [16] aim at automatically generating large corpora by automatically segmenting signs from close-captioned sign language videos, but the usability of such automatically built corpora as training data was not evaluated. Another promising approach for limited numbers of examples per sign is transfer learning [17], but that approach has only been evaluated in a user-dependent scenario, where the test signs are performed by a user who has also provided data for training.

Exemplar-based approaches offer an alternative when only limited examples per class are available. Motion energy images [18] are a well-known exemplar-based approach, but perform poorly in our experiments. Gorelick et al. [19] represent videos of gestures/actions using 3D shapes extracted by identifying areas of motion in each video frame. However, applying that method to our setting would require accurate silhouette extraction of the hands, which is a challenging task even if the bounding box of the hand is known, especially when hands overlap with each other or with the face, or when the background is cluttered.

Some researchers have reported results on vocabularies of thousands of signs, using input from digital gloves, e.g., [20]. On the other hand, most existing vision-based approaches have been evaluated with vocabularies of some tens of signs, e.g., [6, 10, 8]. Kadir et al. [13] report results on 164 signs, with about 85% accuracy when only two training examples per sign are used, whereas Zieren et al. [14] use a vocabulary of 232 signs, and achieve a remarkable 99.3% accuracy rate. However, in both [13] and [14], a single user signed all the training and test examples. In Zieren et al. [14], when experiments are performed in a user-independent setting, the recognition rate drops from 99.3% to 44%, a drop that highlights the difficulty of user-independent sign recognition.

In earlier work [4, 21] we have reported results on data from the public ASLLVD resource, with vocabulary sizes of 992 and 921 signs respectively, and using methods based on motion energy images in [4] and dynamic time warping (with user-aided hand detection, as in our system) in [21]. In our experiments, the method described here outperforms our earlier approaches [4, 21] by a large margin.

3 Application Overview

When a user encounters an unknown sign, the user can perform the sign in front of a webcam, or submit an existing video of that sign. Then, the system asks the user to mark the start and end frames of the actual sign in the video, and to indicate whether the sign is one-handed or two-handed (see Figure 1), and which is the dominant hand (if there is an asymmetry in the production of the sign).

At the next step, the system detects bounding boxes of hands on all frames using features based on skin color and motion. The user views the hand detection results, and can correct those results on any frame. As soon as the user makes a correction, the system propagates information from that correction to improve the detection results in the rest of the frames.

After hand detection results have been approved by the user, the system computes the similarity between the query sign and all database signs. The system ranks the 1,113 distinct signs in decreasing order of similarity to the query. There are two examples for each of 1,113 distinct signs in the database, thus producing two similarity scores with respect to the query. For the purposes of ranking the signs, the better of those two scores is kept.

Once the signs have been ranked, the system presents to the user an ordered list of the best matching signs. The user then views the results, starting from the highest-ranked sign, until encountering the video displaying the actual sign of interest. For example, if the correct match was ranked as 5th best by the system, the user needs to view the top five results in order to view the correct match.

When the user identifies the correct database sign, the user can readily view any additional information associated with that sign. Currently, our signs are labelled with very rough English glosses. These do not necessarily provide accurate information about the meaning of the signs, however, since there is no 1-1 relationship between ASL signs and English words. The longer term plan is that this interface may provide access to sophisticated multi-media ASL language resources, which would provide more extensive information about the signs being looked up. For the time being, though, the user may find the very rough translation to be of some utility.

Fig. 1. Examples of three signs from the dataset that we use. For each sign, we show the first, middle, and last frame. Top row: a one-handed sign meaning “bad”. Middle row: a one-handed sign meaning “badge”, and exhibiting only little motion. Bottom row: a two-handed sign meaning “abandon”.

3.1 Measures of Accuracy

As far as the user is concerned, the system has succeeded on a query sign if the user has indeed managed to retrieve the sign that was being sought. One possible type of failure is a situation in which the query sign is not part of the system vocabulary. For the purposes of this paper we ignore this source of failure, as it does not depend on the quality of the underlying technology, but simply on the size of the database corpus.

A second type of failure results when the system ranks the correct match too low. A user would probably not be willing to view more than the top 10 or 20 signs from the results, although this may vary across users. If a user is willing to view at most k results, then the system fails when the correct match is not among those top k results.

Consequently, given a query Q, a key measure of performance is the rank $R(Q)$ that the system assigns to the correct result for Q. Given an integer k, we define a boolean measure of success $S(Q,k)$, that is true iff $R(Q) \leq k$. The success rate $S(k)$ over a test set $\mathbb{Q}$ of queries is simply the average success rate $S(Q,k)$ over $\mathbb{Q}$. For notational convenience we also define $K(s)$ to be (loosely) an inverse function of S: given a desired success rate s, $K(s)$ is the maximum number of results per query that the user must consider to obtain that success rate, so that $K(S(k)) = k$. More formally:

$$R(Q) = \text{the rank the system assigns to the correct result for } Q \ . \tag{1}$$

$$S(Q,k) = \begin{cases} 1 & \text{if } R(Q) \leq k \ . \\ 0 & \text{otherwise .} \end{cases} \tag{2}$$

$$S(k) = \text{mean}\{S(Q,k) \mid Q \in \mathbb{Q}, \text{where } \mathbb{Q} \text{ is a test set.}\} \ . \tag{3}$$

$$K(s) = \text{the } k \text{ such that } S(k) = s \ . \tag{4}$$

Even if the user is willing to view $K(s)$ results per query, if the correct match is ranked at $R(Q) < K(s)$, then the user can stop viewing results as soon as the user encounters the correct match. Consequently, another meaningful measure of accuracy is the *average* number of results that the user needs to consider per query until encountering the correct result, for a given success rate s. We define that measure as $A(s)$:

$$A(s) = \text{mean}\{R(Q) \mid (Q \in \mathbb{Q}) \wedge (R(Q) \leq K(s))\} \ . \tag{5}$$

4 Features and Normalization

Let X be a video of a sign. We denote by $|X|$ the number of frames in the video, and by $X(t)$ the t-th frame of that video, t ranging from 1 to $|X|$. From sign X we extract the following location, orientation, and hand appearance features:

- $L_d(X,t)$ and $L_{nd}(X,t)$: The (x,y) centroid respectively of the dominant hand and non-dominant hand of the signer at frame t.

- $L_\delta(X,t)$: The relative position of the dominant hand with respect to the non-dominant hand at frame t. $L_\delta(X,t) = L_d(X,t) - L_{nd}(X,t)$.
- $O_d(X,t)$ and $O_{nd}(X,t)$: The unit vectors representing the direction of motion from $L_d(X,t-1)$ to $L_d(X,t+1)$ and from $L_{nd}(X,t-1)$ to $L_{nd}(X,t+1)$.
- $O_\delta(X,t)$: The unit vector representing the direction of motion from $L_\delta(X,t-1)$ to $L_\delta(X,t+1)$.
- $H_{d,s}(X), H_{d,e}(X), H_{nd,s}(X)$, and $H_{nd,e}(X)$: images of the dominant and the non-dominant hand at the start and end frame of the video.

Each hand appearance image H is preprocessed using the following steps:

1. We start by simply cropping the subwindow H_1 corresponding to the bounding box of the hand.
2. We detect skin in that window, using the method of Jones et al. [22].
3. We set all non-skin pixels in H_1 to 0.
4. We create H_2 to be the grayscale version of H_1.
5. We normalize H_2 to have a mean of zero and a standard deviation of 1.
6. We create H_3 as a scaled version of H_2, so that the longest side of H has length 50.
7. The final image H is a padded version of H_3, to make sure that H has an equal number of rows and columns. Additional rows or columns are added as needed, with values of zero. The padding is applied symmetrically, so that the centroid of the hand corresponds with the center of the final image H.

For notational convenience, all features referring to the non-dominant hand (i.e., $L_{nd}, L_\delta, O_{nd}, O_\delta, H_{nd,s}, H_{nd,e}$) are set to zero vectors for one-handed signs.

4.1 Coordinate System

In defining location features, the choice of coordinate system is important. To account for differences in translation and spatial scale between the query video and the matching training videos, we use a face-centric coordinate system. We use the face detector of Rowley et al. [23] to detect the face of each signer at the first frame of the sign. The coordinate system is defined so that the center of the face is at the origin, and the diagonal of the face bounding box has length 1. The same scaling factor is applied to both the x and the y direction. Features L_d, L_{nd}, L_δ are all defined in this normalized coordinate system.

4.2 Time Series Length Normalization

Different signers may sign at different speeds. Dynamic Time Warping (DTW), which we describe in Section 5, is a similarity measure that is biased against longer database matches, and this bias is more noticeable for short queries. To account for that, we normalize each sequence, so that the length of all sequences is the same (20 in our experiments). In particular, we resample the sequences of L_d, L_{nd}, and L_δ features extracted from each sign, so that each sequence has length 20. Resampling is done using linear interpolation. As shown in our experiments, this normalization significantly improves accuracy.

5 Comparing Trajectories via Dynamic Time Warping

Let X be a video of a sign. We can represent X as a time series $(X_1, \dots, X_{|X|})$, where each X_t is simply a concatenation of the features extracted at frame t:

$$X_t = (L_d(X,t), L_{nd}(X,t), L_\delta(X,t), O_d(X,t), O_{nd}(X,t), O_\delta(X,t)) \ . \tag{6}$$

As a reminder, features $L_{nd}, L_\delta, O_{nd}, O_\delta$ are set to 0 for one-handed signs.

Dynamic Time Warping (DTW) [5] is a commonly used distance measure for time series. Given two sign videos Q and X, DTW computes a warping path W establishing correspondences between frames of Q and frames of X:

$$W = ((q_1, x_1), \dots, (q_{|W|}, x_{|W|})) \ , \tag{7}$$

where $|W|$ is the length of the warping path, and pair (q_i, x_i) means that frame q_i of Q corresponds to frame x_i of X. A warping path must follow two constraints:

- **boundary constraints:** $q_1 = 1,\ x_1 = 1,\ q_{|W|} = |Q|,\ x_{|W|} = |X|$.
- **monotonicity and continuity:** $0 \leq q_{i+1} - q_i \leq 1, \quad 0 \leq x_{i+1} - x_i \leq 1$.

The cost $C(W, Q, X)$ of a warping path W is the sum of individual local costs $c(Q_{q_i}, X_{x_i})$, corresponding to matching each Q_{q_i} with the corresponding X_{x_i}:

$$C(W, Q, X) = \sum_{i=1}^{|W|} c(Q_{q_i}, X_{x_i}) \ . \tag{8}$$

As local cost c, we use a weighted linear combination of the individual Euclidean distances between the six features extracted from the two frames:

$$\begin{aligned} c(Q_{q_i}, X_{x_i}) = {} & f_1\|L_d(Q,q_i) - L_d(X,x_i)\| + f_2\|L_{nd}(Q,q_i) - L_{nd}(X,x_i)\| + \\ & f_3\|L_\delta(Q,q_i) - L_\delta(X,x_i)\| + f_4\|O_d(Q,q_i) - O_d(X,x_i)\| + \\ & f_5\|O_{nd}(Q,q_i) - O_{nd}(X,x_i)\| + f_6\|O_\delta(Q,q_i) - O_\delta(X,x_i)\| \end{aligned} \tag{9}$$

In the above equation, $\|\cdot\|$ stands for the Euclidean norm. In our experiments, weights f_j are optimized using cross-validation on the training set.

The DTW distance $D_{\mathrm{DTW}}(Q, X)$ between sign videos Q and X is defined as the cost of the lowest-cost warping path between Q and X:

$$D_{\mathrm{DTW}}(Q, X) = \min_W C(W, Q, X) \tag{10}$$

The optimal warping path and the distance $D_{\mathrm{DTW}}(Q, X)$ can be computed using dynamic programming, with a time complexity of $O(|Q||X|)$ [5].

6 Incorporating Hand Appearance

The D_{DTW} distance measure defined above depends only on the trajectories of the two hands. At the same time, the appearance of the hand is an important additional source of information about the identity of a sign. Recognizing handshape is a challenging task, especially when a hand appears in front of another skin-colored object such as the other hand or the face. Given the difficulty of this topic, we have postponed the task of implementing a sophisticated similarity measure for hand appearance for future work. Instead, in this paper we have opted for the simplest possible option, which is the Euclidean distance between hand appearance images. Despite its simplicity, this approach has led to significant improvements in accuracy, as shown in the experiments.

In particular, we define a distance $D_{\text{hand}}(Q, X)$ between two sign videos Q and X as follows:

$$D_{\text{hand}}(Q, X) = \|H_{d,s}(Q) - H_{d,s}(X)\| + \|H_{d,e}(Q) - H_{d,e}(X)\| + \|H_{nd,s}(Q) - H_{nd,s}(X)\| + \|H_{nd,e}(Q) - H_{nd,e}(X)\| \ . \quad (11)$$

As a reminder (see Section 4), each hand image has been preprocessed, by scaling/padding to a canonical size, and removing non-skin pixels.

Combining D_{DTW} and D_{hand} can be done by simply taking a weighted sum of the two distances:

$$D(Q, X) = D_{\text{DTW}}(Q, X) + f_{\text{hand}} D_{\text{hand}}(Q, X) \ . \quad (12)$$

The weight f_{hand} is chosen by searching over many possible values, so as to optimize performance on the training data.

7 Experiments

Our dataset includes 1,113 distinct sign classes. For each sign class there are three examples, each from a different user. All sign videos and annotations have been downloaded from the ASLLVD website [4]. Although the ASLLVD website includes four synchronized camera views for each sign, only a single frontal view is used for each sign in our experiments.

The dataset was divided into three groups, each group containing a single example from each of the 1,113 classes. Experiments were performed in a user-independent manner, by ensuring that each signer appeared in only a single group out of those three groups. Each group was in turn used as the test set, with the other two groups used as training. All experimental measurements are averaged over the three test groups. All weights involved in defining the overall distance measure were computed exclusively from the training data, and thus a different combination of weights was applied for each test group.

We use the measures of accuracy defined in Section 3.1, and in particular $K(s)$ and $A(s)$, which are respectively the maximum and average number of results that a user must consider for a query in order to encounter the correct result for a fraction s of all queries.

Since the user indicates at query time whether a sign is one-handed or two-handed, signs using a different number of hands than the query are not considered for that query. (For real use cases in the longer term, we will have to allow a small probability for a canonically 2-handed sign being produced with just 1 hand, and for a 1-handed sign to be produced with 2 hands.) Signs performed with the left hand as the dominant hand are replaced by mirrored versions, so that we can treat all database and query signs as right-handed. These rules have been applied in all experiments for all methods.

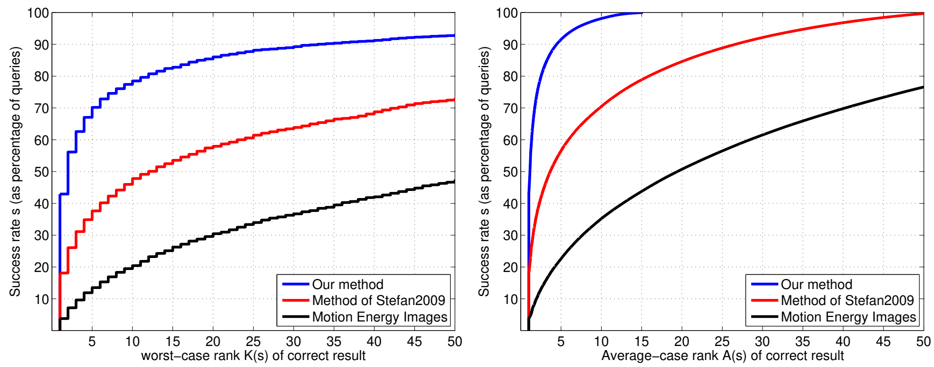

Fig. 2. Comparison of our method, the method described by Stefan et al. in [21], and the MEI-based method used in [4]. The y-axis corresponds to success rates s. The x-axis corresponds to values of $K(s)$ on the left, and to values of $A(s)$ on the right.

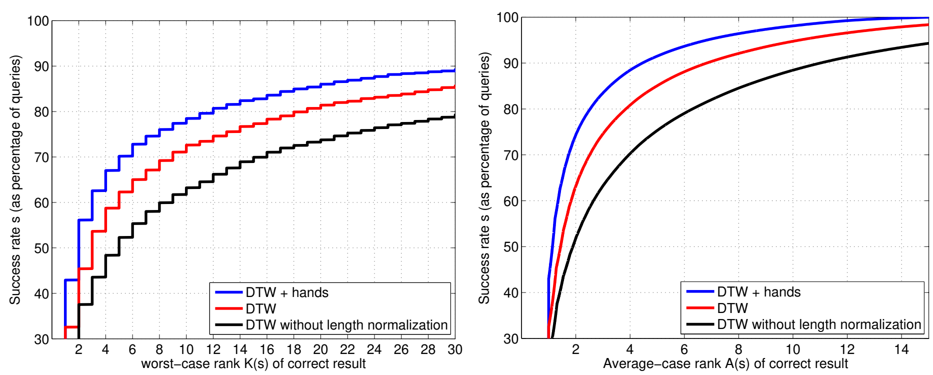

Fig. 3. Comparison of DTW without length normalization, DTW, and our full method that combines DTW scores with hand appearance similarity scores. The y-axis corresponds to success rates s. The x-axis corresponds to values of $K(s)$ on the left, and to values of $A(s)$ on the right.

7.1 Results

We compare our method to previous methods applied to data from the ASLLVD dataset, namely to the approach of using motion energy images (MEI) described in [4], and the DTW-based approach reported in [21]. With respect to the DTW method of [21], we should note that it corresponds to a stripped-down version of our method, which only uses the L_d and L_{nd} features, does not normalize the length of all time series to a fixed constant, and does not use hand appearance.

Figure 2 shows comparative results for the three methods. As measures of accuracy, we use functions $K(s)$ and $A(s)$ (defined in Section 3.1), which are, respectively, the maximum and average number of results per query that a user must consider in order to encounter the correct result for a fraction s of all queries. Our method clearly outperforms the two other methods. As an example, the percentage of queries for which the the correct sign is not included in the top 25 results is, respectively, 66% for MEI, 38.6% for Stefan et al. [21], and 11.9% for our method. Similarly, the correct sign is successfully included in the top 10 results for only 20.4% of the queries using MEI, for 47.8% of the queries using the method of Stefan et al. [21], and for 78.4% of queries for our method. For our method, for a success rate of 78.4% the user needs to consider at most 10, and on average 2.36 results per query, until encountering the correct result.

In Figure 3 we evaluate three different variations of our method: the first variation, denoted as "DTW without length normalization", does not use hand appearance and also does not use the resampling step described in Section 4.2, which normalizes all time series to length 20. The second variation, denoted as DTW, does not use hand appearance. The third variation, denoted as "DTW + hands", is the full method described in this paper, that incorporates information from both DTW and hand appearance. As see from the results, normalizing all time series to the same length significantly improves accuracy, and incorporating hand appearance leads to a noticeable additional improvement.

In terms of running time, the system takes on average about one second to compare a query video to the 2,226 database videos. Running time was measured on a PC with an Intel quad-core CPU, running at 2.4GHz, and with 3GB of memory. Our method has been implemented as a single-threaded application.

8 Discussion and Future Work

We have presented a method for helping users look up unknown signs, using similarity-based retrieval in a database containing examples of signs from a large vocabulary. In our method, feature vectors are defined based on hand motion and hand appearance. Similarity between signs is measured by combining dynamic time warping scores, which are based on hand motion, with Euclidean distances between hand appearances.

There are several research topics that we are interested in pursuing as future work, with the goal of further improving system performance and the overall user experience. While in the current system hand detection is only semi-automatic, a more (or entirely) automated hand detector will significantly enhance the user experience, and this is a topic that we intend to explore in the near term. Also, while our simple way of using hand appearance led to good results, there is clearly room for improvement in how we use hand appearance, and that is another topic that we plan to explore. Our current approach of not allowing one-handed signs to be matched with two-handed signs, and of requiring the user to specify the dominant hand for the query sign, has limitations that need to be addressed. Finally, although the proposed approach works reasonably well in our experiments, we believe that more work is needed in order to satisfactorily address the question of how to learn a good similarity measure for a large vocabulary of signs, given only one or two training examples per sign.

Acknowledgements

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. For assistance with data collection and annotations, we thank Jaimee DiMarco, Joan Nash, Chrisann Papera, Jessica Scott, Jon Suen, Ashwin Thangali, Iryna Zhuravlova, Kishan Kumar, Roochi Mishra, and Muhammad Yousaf. The research reported here has been partially funded by grants from the National Science Foundation: IIS-0705749, IIS-0812601, MRI-0923494.

References

1. Tennant, R.A., Brown, M.G.: The American Sign Language Handshape Dictionary. Gallaudet U. Press, Washington, DC (Washington, DC)
2. Lane, H., Hoffmeister, R.J., Bahan, B.: A Journey into the Deaf-World. DawnSign Press, San Diego, CA (1996)
3. Schein, J.: At home among strangers. Gallaudet U. Press, Washington, DC (1989)
4. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., Thangali, A.: The American Sign Language lexicon video dataset. In: IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB). (2008)
5. Kruskal, J.B., Liberman, M.: The symmetric time warping algorithm: From continuous to discrete. In: Time Warps. Addison-Wesley (1983)
6. Bauer, B., Hienz, H., Kraiss, K.F.: Video-based continuous sign language recognition using statistical methods. In: International Conference on Pattern Recognition. (2000) 2463–2466
7. Dreuw, P., Deselaers, T., Keysers, D., Ney, H.: Modeling image variability in appearance-based gesture recognition. In: ECCV Workshop on Statistical Methods in Multi-Image and Video Processing. (2006) 7–18
8. Starner, T., Pentland, A.: Real-time American Sign Language recognition using desk and wearable computer based video. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1371–1375
9. Vogler, C., Metaxas, D.N.: Parallel hidden markov models for american sign language recognition. In: IEEE International Conference on Computer Vision (ICCV). (1999) 116–122

10. Cui, Y., Weng, J.: Appearance-based hand sign recognition from intensity image sequences. Computer Vision and Image Understanding **78** (2000) 157–176
11. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: IEEE International Conference on Computer Vision (ICCV). Volume 1. (2005) 166–173
12. Wang, S.B., Quattoni, A., Morency, L.P., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 2. (2006) 1521–1527
13. Kadir, T., Bowden, R., Ong, E., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: British Machine Vision Conference (BMVC). Volume 2. (2004) 939–948
14. Zieren, J., Kraiss, K.F.: Robust person-independent visual sign language recognition. In: Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). Volume 1. (2005) 520–528
15. Valli, C., ed.: The Gallaudet Dictionary of American Sign Language. Gallaudet U. Press, Washington, DC (2006)
16. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009) 2568–2574
17. Farhadi, A., Forsyth, D.A., White, R.: Transfer learning in sign language. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2007)
18. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **23** (2001) 257–267
19. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 2247–2253
20. Yao, G., Yao, H., Liu, X., Jiang, F.: Real time large vocabulary continuous sign language recognition based on OP/Viterbi algorithm. In: International Conference on Pattern Recognition. Volume 3. (2006) 312–315
21. Stefan, A., Wang, H., Athitsos, V.: Towards automated large vocabulary gesture search. In: Conference on Pervasive Technologies Related to Assistive Environments (PETRA). (2008)
22. Jones, M., Rehg, J.: Statistical color models with application to skin detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (1999) I:274–280
23. Rowley, H., Baluja, S., Kanade, T.: Rotation invariant neural network-based face detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (1998) 38–44