Unsupervised Modeling of Signs Embedded in Continuous Sentences

Sunita Nayak Sudeep Sarkar
Department of Computer Science & Engineering
University of South Florida
Tampa, FL 33620, USA
{snayak,sarkar}@csee.usf.edu

Barbara Loeding
Department of Special Education
University of South Florida, Lakeland Campus
Lakeland, FL 33803, USA
bloeding@lklnd.usf.edu

Abstract

The common practice in sign language recognition is to first construct individual sign models, in terms of discrete state transitions, mostly represented using Hidden Markov Models, from manually isolated sign samples and then to use it to recognize signs in continuous sentences. In this paper we (i) propose a continuous state space model, where the states are based on purely image-based features, without the use of special gloves, and (ii) present an unsupervised approach to both extract and learn models for continuous basic units of signs, which we term as signemes, from continuous sentences. Given a set of sentences with a common sign, we can automatically learn the model for part of the sign, or signeme, that is least affected by coarticulation effects. While there are coarticulation effects in speech recognition, these effects are even stronger in sign language.

The model itself is in term of traces in a space of Relational Distributions. Each point in this space represents a Relational Distribution, capturing the spatial relationships between low-level features, such as edge points. We perform speed normalization and then incrementally extract the common sign between sentences, or signemes, with a dynamic programming framework at the core to compute warped distance between two subsentences. We test our idea using the publicly available Boston SignStream Dataset by building signeme models of 18 signs. We test the quality of the models by considering how well we can localize the sign in a new sentence. We also present preliminary results for the ability to generalize across signers.

1. Introduction

While speech recognition has made rapid advances, sign language recognition is lagging behind. With gradual shift to speech based I/O devices, there is great danger that persons who rely solely on sign languages for communication will be deprived access to state-of-the-art technology unless there are significant advances in automated recognition of sign languages. Sign language is quite different from gestures. American sign language, with its own grammar, is also different from Signed English. Like in speech, the ability to recognize isolated sign does not guarantee the recognition of that sign in continuous sentences. When a signer signs a sentence, each sign is affected by the adjacent signs – this is the coarticulation effect. The beginning and end of a sign depends strongly on its context in the sentence. This coarticualtion effect makes the extraction, modeling, and recognition of signs from continuous sentences more difficult when compared to just plain gestures.

Previous related works have been mostly in the recognition of static gestures, dynamic gestures, and isolated ASL signs. Starner and Pentland [8] were the first to seriously consider *continuous* sign recognition. Using Hidden Markov Model (HMM) based representations, they achieved near perfect recognition with sentences with fixed structure, i.e. containing personal pronoun, verb, noun, adjective, personal pronoun in that order. Vogler and Metaxas [9] have been instrumental in significantly pushing the state-of-the-art in automated ASL recognition using HMMs. In terms of the basic HMM formalism, they have explored many variations, such as context dependent HMMs, HMMs coupled with partially segmented sign streams, and parallel HMMs. The use of HMM is also seen in other sign language recognizers. Bauer and Krass [1] do automatic German continuous Sign Language recognition based on word level subunits, called *fenones*, instead of whole sign models. Fang *et.al.* [3] also proposed a temporal clustering algorithm to extract subunits at word level from Chinese Sign Language, which they use to build HMMs. We also seek to find subunits but at sentence level. Also instead of using HMMs that are good for discretized state space representations, we use a continuous state space and a time-normalized representation.

Most of the works in continuous sign language recognition have avoided the basic problem of segmentation and tracking of hands by using wearable devices, such as colored gloves, or magnetic markers, to directly get the location features. For example Vogler and Metaxas [9] have used 3D magnetic tracking system; Starner and Pentland [8] have used colored gloves while Ma *et al.* [5, 10] have used

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

Cybergloves. Bauer and Krass [1] use colored gloves for feature extraction. Fang *et al.* [3] used cybergloves and 3D position trackers. Only recently there has been effort to extract information and to track directly from color images, without the use of special devices [12], but it has only been used for *isolated* sign recognition. In this paper, we use only the plain color images, without using any augmenting wearable devices, for sentence level recognition of continuous signs with unconstrained sentence structure. This has not been done earlier. Because of the inherent difficult nature of the problem, one should not expect the high level of performance reported in other work.

Unlike the majority of previous work, our representation does not require tracking of hands. We work with edge pixel that are within skin colored patches as low-level features. We emphasize the relationships among these edge pixels by building relational distributions that capture the statistics of the pairwise relationship between them. These relational distributions are efficiently represented as points in the Space of Relational Distributions (SoRD). Any given sentence is a sequence of points in this space, from which a linearly interpolated track is derived. Given a set of sentences with one common sign, the part of the track that is common to all the sentences is arrived at by dynamic programming.

2. Relational Distributions

Grounded on the observation that the *organization* or *structure* or *relationships* among low-level primitives are more important than the primitives themselves, we focus on the statistical distribution of the relational attributes observed in the image, which we refer to as *relational distributions*. Such statistical representation also removes the need for primitive level correspondence or tracking across frames. These kinds of representations have been successfully used for modeling periodic human motion [13]. Here, we use it to extract statistical models automatically for signs from continuous ASL sentences. Of course, non-relational, primitive level statistical distributions, such as orientation histograms are fairly commonly used and have been used for gesture recognition [4]. The novelty of relational distributions lies in that it offers a strategy for incorporating dynamic aspects.

Let $F = \{f_1, ..., f_N\}$ represent the set of N primitives in an image. For us these are Canny edge pixels of the image. Let F_k represent a random k-tuple of primitives, and the relationship among k-tuple primitives be denoted by R_k. Let the relationships R_k be characterized by a set of M attributes $A_k = \{A_{k1}, ..., A_{kM}\}$. For ASL, we use the distance of the two edge pixels in the vertical and horizontal direction (dx, dy) as the attributes. We normalize and represent the distance between the pixels in an image size of 32 x 32 to reduce the size for further processing. The shape of the pattern can be represented by joint probability functions: $P(\mathbf{A_k} = \mathbf{a_k})$, also denoted by $P(a_{k1}, ..., a_{kM})$ or $P(\mathbf{a_k})$, where a_{ki} is the (discretized in practice) value taken by the relational attribute A_{ki}. We term these probabilities as the *Relational distributions*.

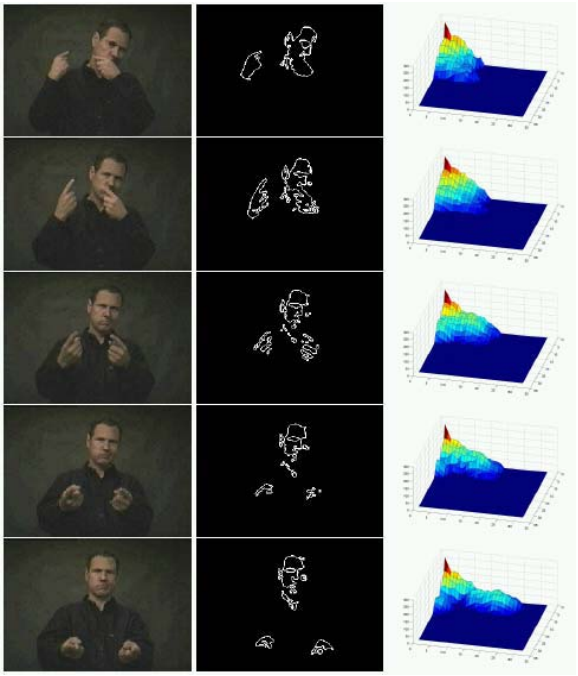

Figure 1: Variations in relational distributions with motion. The left column shows the image frames in the sign 'UP-TO-NOW'. The middle column shows the edge pixels, and the right column shows the relational distributions

One interpretation of these distributions is:

> Given an image, if you randomly pick k-tuples of primitives, what is the probability that it will exhibit the relational attribute $\mathbf{a_k}$? What is P($\mathbf{A_k} = \mathbf{a_k}$)?

Given that these relational distributions exhibit complicated shapes that do not readily afford modeling using a combination of simple shaped distribution, we adopt non-parametric histogram based representation. However, to reduce the size that is associated with a histogram based representation, we use the Space of Relational Distributions (SoRD).

As the hands of the signer move, the relational distribution changes. Motion of hands introduces non-stationarity in the relational distributions. Figure 1 shows example of the 2-ary relational distributions for the sign 'UP-TO-NOW'. In the relational distribution's plot, the vertical axis represents the joint probability and the two horizontal axes represent the attributes. Notice the change in the distributions as the hands come down. The change in one attribute dimension (vertical distance between edge pixels) in the

plots can be seen clearly as the hands come down, while there is comparatively less change in the other attribute dimension.

Let $P(\mathbf{a_k}, t)$ represent the relational distribution at time t. Let

$$\sqrt{P(\mathbf{a_k}, t)} = \sum_{i=1}^{n} c_i(t)\Phi_i(a_k) + \mu(\mathbf{a_k}) + \eta(\mathbf{a_k}) \qquad (1)$$

describe the *square root* of each relational distribution as a linear combination of orthogonal basis functions, where $\Phi_i(\mathbf{a_k})$'s are orthonormal functions, the function $\mu(\mathbf{a_k})$ is a mean function defined over the attribute space, and $\eta(\mathbf{a_k})$is a function capturing small random noise variations with zero mean and small variance. We refer to this space as the Space of Relational Distributions (SoRD).

We use the square root function so that we arrive at a space where Euclidean distances are related to the Bhattacharya distance between the relational distributions, which is an appropriate distance measure for probability distributions. Its proof can be found in [13]. This lets us work with Euclidean distances in this space. Given a set of relational distributions, $\{P(\mathbf{a_k}, t_i) \mid i = 1, ..., T\}$, the SoRD can be arrived at by principal component analysis (PCA). In practice, we can consider the subspace spanned by a few ($N \ll n$) dominant vectors associated with the large eigenvalues. For us, most of the variation is captured by the eigen vectors associated with the top 12 (largest) eigen values. Thus, a relational distribution can be represented using these N coordinates ($c_i(t)s$), which is more compact representation than a normalized histogram based representation. The ASL sentences form sequences of points in this Space of Relational Distributions.

3. Speed-Normalized Representation

Speed normalization is an important issue in gesture recognition [2]. It is important to be able to compare two signs executed at different speeds. We proceed as follows. Let the initial SoRD coordinates of a sentence be denoted by $\vec{A_i}, 1 \leq i \leq n$, where n is the number of image frames in the whole sentence. We linearly interpolate a set of points $\vec{q_1}, \vec{q_1}, \vec{q_2}, \vec{q_3}...\vec{q_k}$ between $\vec{A_1}$ and $\vec{A_2}$ using $\vec{q_j} = \vec{A_1} + j\vec{v}$, where $\vec{v}$ is an increment vector given by $\vec{v} = \frac{\vec{A_2} - \vec{A_1}}{|\vec{A_2} - \vec{A_1}|} d$ and $j = 0, 1, 2, 3..k$, with k satisfying $\mid \vec{q_k} - \vec{A_1} \mid > \mid \vec{A_2} - \vec{A_1} \mid$, and $\mid \vec{q_{k-1}} - \vec{A_1} \mid < \mid \vec{A_2} - \vec{A_1} \mid$. Let d denote a fixed distance between the interpolated points. We drop $\vec{q_k}$ from the series of interpolated points, if $\mid \vec{q_k} - \vec{A_2} \mid > \frac{d}{2}$. For a successive interpolation stage, we choose the last point in the previously obtained series as the starting point. Finally we get a series of equidistant points representation for the whole sentence in which the distance between any two consecutive points is given by d. Given that the relational distributions capture the shape of the underlying distribution of points, two consecutive points on the interpolated curve represent roughly similar amount of change in shape. Figure 2 shows the interpolated points for a sentence. It also depicts the frames for the sign 'UP-TO-NOW' and their corresponding points in the sentence curve.

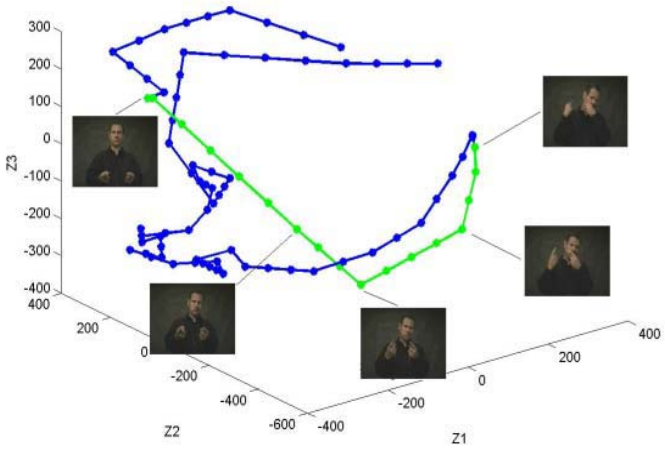

Figure 2: Representation of the sentences as a series of points. We show only the first three dimensions (Z1, Z2 and Z3 are the first, second and third SoRD coefficients respectively). Frames of the sign 'UP-TO-NOW'(see clockwise) in the sentence 'JOHN STUDY PHYSICS UP-TO-NOW' are shown against their corresponding points

A linear-interpolated representation helps in explicitly representing the motion between any two given image frames. This aspect is quite important. Firstly, if the signer signs the same sign at a fast speed in some instance than the other, it is quite possible that some motion frames are missed while capturing data for the faster instance. This would directly affect the sign extraction and detection results from the sentence. One existing and widely used approach to deal with such situations is using dynamic time warping to match the sequences. Wu *et al.* [11] experimented various improvements of continuous dynamic programming on human gesture recognition. Their results show that all these improvements as well as the conventional method works best when the speed ratio of the input and output gestures is one. This is generally not the case, especially in signed languages, where a signer signs the same words (signs) with varying speeds in different contexts/sentences. Secondly, even a mild difference in the data capture frame rates can significantly affect the existing methods of using sole dynamic programming. Thirdly, the speed-normalized modeling approach makes it possible to work with compressed data, where frames might be

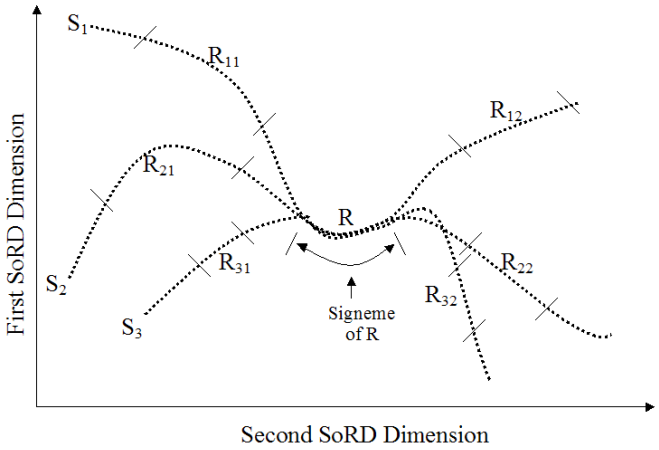

Figure 3: Concept of signemes. First vs. second SoRD dimensions of sentences S_1 with signs R_{11}, R, R_{12} in order, S_2 with signs R_{21}, R, R_{22} and S_3 with signs R_{31}, R, R_{32}. The common sign is R. The portion of R that is most similar across sentences is the signeme representative of R.

dropped. We illustrate this aspect by showing results on compressed datasets in our experiments.

Our approach finds a representation that has more number of points between consecutive image frames where there is large motion than where there is less motion. It merges the points corresponding to consecutive frames that have very small motion between them or almost no shape change, when the distance between the SoRD points between the frames is less than $d/2$. This feature of the representation helps to move towards a speed-invariant representation. As the value of d is decreased, we get higher speed invariance. Ideally, as d tends to zero the approach becomes completely speed invariant. But by decreasing d, the number of points generated for further processing for extraction/matching increases.

4. Signeme: Definition

ASL sentences are series of signs with coarticulation effects in between the signs. Figure 3 shows the traces of the first vs. second SoRD dimensions of three sentences S_1, S_2 and S_3 with only one common sign, R, among them. Although the sign R, is common among the sentences, only a portion of the sign is similar across sentences. We define this common part to be the signeme representative of R. Mathematically, signeme is the longest subsequence in the sentences that is closest. Signeme represents the portion of the sign that would be present in any sentence with the sign, irrespective of the adjacent signs in the sentence.

Formally, signemes are specified as follows. Let S_1, S_2, ... S_N be the curves representing N sentences with one common sign. Let $\alpha_{ij}(r)$ be a function that maps points on sentence S_j to points on sentence S_i. Then

$$d^{\alpha}_{ijkr_0} = \sum_{r=r_0}^{r_0+k} (S_i(r) - S_j(\alpha_{ij}(r)))^2 \qquad (2)$$

represents the Euclidean distance between sub-segments of length k, starting at r_0 on S_i, with the warping function α_{ij}. The most similar common part between two sentences is defined by the minimum of this distance over various choices of k, r_0 and α_{ij}.

$$(k^m, r_0^m, \alpha_{ij}^m) = \arg \min_{k,r_0,\alpha_{ij}} d^{\alpha}_{ijkr_0} \qquad (3)$$

Since a trivial solution of this minimization is obtained for $k = 0$, we constrain the domain of possible choices of k, as we shall see in the next section. The signeme is then specified by the common segments between pairs of sentences, defined by the following minimization

$$(k^m, r_0^m, \alpha_{1j}^m) = \arg \min_{k,r_0,\alpha_{1j}} d^{\alpha}_{1jkr_0} \qquad (4)$$

where S_1 is used as the anchoring sentence to find the common segments from all the other sentences. The signeme is then the following set of sentence fragments. $\{\{S_1(r_0^m), ...S_1(r_0^m+k^m)\}, \{S_2(\alpha_{12}(r_0^m)), ...S_2(\alpha_{12}(r_0^m+k^m))\}, ...\{S_N(\alpha_{1N}(r_0^m)), ...S_N(\alpha_{1N}(r_0^m+k^m))\}\}$

5. Signeme: Extraction

We start with two sentences, S_1 with U points specified by $S_1(1)$,$S_1(2)...S_1(U)$, and sentence S_2 with V points $S_2(1), S_2(2)...S_2(V)$. We first consider the segments $\{S_1(1), S_1(2), S_1(3)...S_1(k)\}$ and $\{S_2(1), S_2(2), S_2(3)...S_2(k)\}$, each with k points and find a matching score between them using dynamic programming that satisfies endpoint, monotonicity and continuity constraints. The dynamic programming score table, D, can be built using the following recursion:

$$D(p,q) = \min \begin{cases} \frac{D(p-1,q)+e(p,q)}{L(p-1,q)+1}, \\ \frac{D(p-1,q-1)+e(p,q)}{L(p-1,q-1)+1}, \\ \frac{D(p,q-1)+e(p,q)}{L(p,q-1)+1} \end{cases} \qquad (5)$$

where $D(p,q)$ represents the matching score between the sub-segments $\{S_1(1), S_1(2)...S_1(p)\}$ and $\{S_2(1), S_2(2)...S_2(q)\}$, $e(p,q)$ is the Euclidean distance between $S_1(p)$ and $S_2(q)$ and $L(p,q)$ represents the length of the optimal warping path from $(1,1)$ to (p,q). L is incremented by 1 each time the warping path adds a new element to itself. $D(k,k)$ gives the matching score between the segments $\{S_1(1), S_1(2)...S_1(k)\}$ and $\{S_2(1), S_2(2)...S_2(k)\}$. In a similar manner, we find the matching score of the segment $\{S_1(1), S_1(2)...S_1(k)\}$

with each of the segments $\{S_2(2), S_2(3)...S_2(k+1)\}$, $\{S_2(3), S_2(4)...S_2(k+2)\}$ and so on. Then we shift the segment of S_1 by 1, and compare it with all segments of length k from S_2 and get a matching score for each pair of segments compared. We repeat the process till we get matching scores between all the segments of length k from S_1 and S_2. We consider the pair of segments that has the highest similarity (i.e. minimum score) as representing the common sign in the sentences S_1 and S_2. Then, we randomly pick one of the two segments and match it to all segments of length k in other training sentences to extract the sign from them. A mean model for the sign is formed from the segments extracted as the sign from all the training sentences. We call the mean model as a *signeme*. In this extraction process, the first two sentences chosen play a critical role. The length parameter, k, is chosen based on the quantity of motion involved. For example, for signs that have very less movement, we choose smaller lengths than what is chosen for signs involving more motion.

6. Results & Discussion

At Boston University, Neidle *et al.* have annotated an ASL dataset [7] using SignStream [6], which is a system for linguistic annotation, storage, and retrieval of ASL and other forms of gestural communication. This dataset has no wearable aids. The compressed version of 117 ASL sentences is publicly available. We used this *compressed* data for our experiments. The images are 240 by 180 in size and has plain black background and clothing.

We modeled 18 signs from the set of 117 sentences. The signs were BOOK, UPTONOW, YESTERDAY, WHO, WHAT, SEE, CAR, BUY, CORN, GO, CAN, TEACHER, HOUSE, (distant)FUTURE, (near)FUTURE, TOMORROW, NOT and FINISH. We chose only those signs that had atleast two sentences with the target sign as the only common sign. Each sign had at least 5 different sentences in the data set. The number of sentences used for extracting the signemes varied from 4 to 18 sentences, the average being 9 sentences for each sign.

We used the learned signeme models to localize the signs in new test sentences. Given the small number of signs involved reporting detection rates does not make statistical sense. We tested with 15 test sentences and their length varied from 3 to 15 signs. The test sentences were not used during training. The set of points representing the signeme are matched with the segments of the SoRD points from the test sentences to find the segment with the minimum matching score, which would represent the sign in the test sentence. The dynamic programming approach used for the extraction of signemes is used for localizing signs as well. The SoRD points of the signeme retrieved from the test sentence are mapped back to their nearest frames and compared

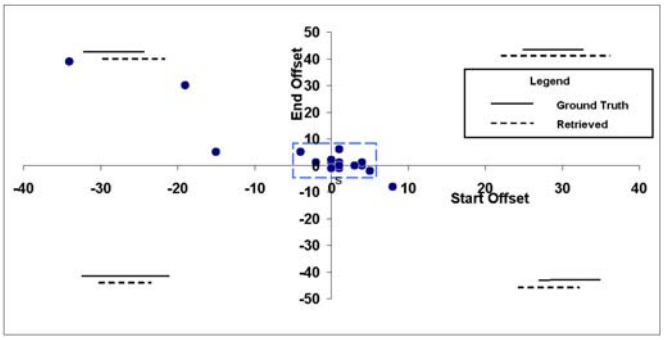

Figure 4: The solid and the dashed lines in the four quadrants show the nature of overlap between the retrieved and ground truth signs represented by that quadrant. The rectangle surrounding the origin encompasses all the points that correspond to test signs whose retrieved signeme had common frames with the ground truth sign.

with the ground truth frame series representing the sign in the sentence.

We characterize localization performance as follows. Let a_1 and b_1 denote the start and end frame numbers of the underlying ground truth sign, as defined in the Boston SignStream Dataset, in the test sentence, and a_2 and b_2 denote the start and end frame numbers of the subsequence retrieved as signeme in the sentence. We define Start Offset, ΔS, and End Offset, ΔE, as $\Delta S = a_1 - a_2$ and $\Delta E = b_2 - b_1$. The plot of the Start Offset vs. End Offset is shown in Figure 4. Ideally, both the offsets should be zero. The points for different signs are scattered in the four quadrants depending on nature of overlap between the ground truth sign and the retrieved signeme. First quadrant depicts the case when no frames are missed, but extra frames are added in both start and end, in the retrieved signeme. Second quadrant represents case when some frames at the start are missed, and some are added towards the end. The third quadrant represents the case when frames are missed at both the start and end. Finally the fourth quadrant shows the cases where extra frames are present towards the start, and some frames are missed towards the end, in the signeme. Each point in the plot corresponds to a separate test sign. Its distance from the origin indicates the localizing quality of the signeme in its test sentence. The nearer it is, the better is the localization.

From Figure 4 we see that the localization is good except for some outliers. In particular we had some cases where the signemes could not be extracted accurately, the reason being the high difference in the ways the signs were signed in the training sentences. Another reason for some of the outliers was the presence of quite less motion involved in the sign. The results are expected to improve with the use of uncompressed data. An example of best performance

is shown in Figure 5, where the ground truth sign for the sign 'BUY' and the signeme localized in the test sentence 'JOHN BUY WHAT?' matched completely. Figure 7 and Figure 6 depict cases with positive end offsets of one and two frames respectively, while Figure 8 and Figure 9 show cases with negative end offset and positive start offset of one frame each respectively.

We also did some preliminary experiments for the case of *across signers*. We tested the model for the sign 'CAN' learned from the data of the Boston dataset used above, by localizing the sign in a test sentence signed by a different signer. Figure 10 shows the result. We see that the model has the potential of generalizing across signers. However more experiments needs to be done.

7. Conclusion

We presented a novel approach for automatic extraction of sign models from continuous American Sign Language sentences. Automatic extraction of sign units from a continuous sentences is new. We modeled the configuration of hands in each frame using a relational distribution, which is a statistical representation of the observed relationships between the low-level features in each image frame. These relational distributions are further represented in a compact space as points. Any ASL sentence is a curve in this space. We work with a time-normalized, curve, based representation of signs and sentences. To model signs we proposed the concept of signemes which is robust with respect to coarticulation effects. Given a set of sentences with one common sign, whose position is unknown, we automatically extract the signeme model. We demonstrated that the effectiveness of these learned model in locating signs in test sentences. We also presented results that indicated its ability to model across signers, though more work is needed.

Acknowledgment

This work was supported by the US National Science Foundation ITR Grant No. IIS 0312993.

References

[1] M. Bauer and K.F. Kraiss, "Towards an Automatic Sign Language Recognition System Using Subunits", *International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pp. 64-75, 2001.

[2] A.F. Bobick and A.D. Wilson, "A State-Based Approach to the Representation and Recognition of Gesture", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 1325-1337, 1997.

[3] G. Fang, X. Gao, W. Gao, Y. Chen, "A Novel Approach to Automatically Extracting Basic Units from Chinese Sign Language", *International Conference on Pattern Recognition*, Vol. 4, pp. 454-457, 2004.

[4] W. Freeman and M. Roth, "Orientation histograms for hand and gesture recognition", *International Workshop on Face and Gesture Recognition*, pp. 296-301, 1995.

[5] J. Ma, W. Gao, C. Wang and J. Wu, "A continuous Chinese sign language recognition system", *International Conference on Automatic Face and Gesture Recognition*, pp. 428-433, 2000.

[6] C. Neidle, and D. MacLaughlin, "SignStream: A Tool for Linguistic Research on Signed Languages", *Sign Language & Linguistics*, vol 1, pp. 111-114, 1998.

[7] ASLLRP, SignStream Sample Database 2.0 American Sign Language Linguistic Research Project, Boston University, 2000.

[8] T. Starner and A. Pentland, "Real-time American Sign Language recognition from video using hidden Markov models", *Symposium on Computer Vision*, pp. 265-270, 1995.

[9] C. Vogler and D. Metaxas,"A framework of recognizing the simultaneous aspects of American Sign Language", *Computer Vision and Image Understanding*, Vol. 81, pp. 358-384, 2001.

[10] C. Wang, W. Gao and S. Shan, "An approach based on phonemes to large vocabulary Chinese sign language recognition", *International Conference on Automatic Face and Gesture Recognition*, pp. 393-398, 2002.

[11] H. Wu, R. Kido, T. Shioyama, "Improvement of continuous dynamic programming for human gesture recognition". *International Conference on Pattern Recognition*, vol.2, pp. 945 - 948, Sept. 2000.

[12] M. H. Yang, N. Ahuja and M. Tabb, "Extraction of 2d motion trajectories and its application to hand gesture recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp. 168-185, 2002.

[13] I. R. Vega and S. Sarkar, "Statistical motion model based on the change of feature relationships: human gait-based recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25 , pp. 1323-1328, Oct. 2003.

Figure 5: Signeme for the sign 'BUY' in the test sentence 'JOHN BUY WHAT?', Ground truth sign frames are marked with the RED(dark gray in black and white) line, and the resulting localized signeme frames are marked with GREEN(light gray in black and white) line (Similar color notations are used for the following figures as well). Only left half of some frames of the sentence are being shown here and some of the sentence frames irrelevant to the case are not shown due to the space constraint.

Figure 6: 'GO' in test sentence 'JOHN CAN GO, CAN HIM'

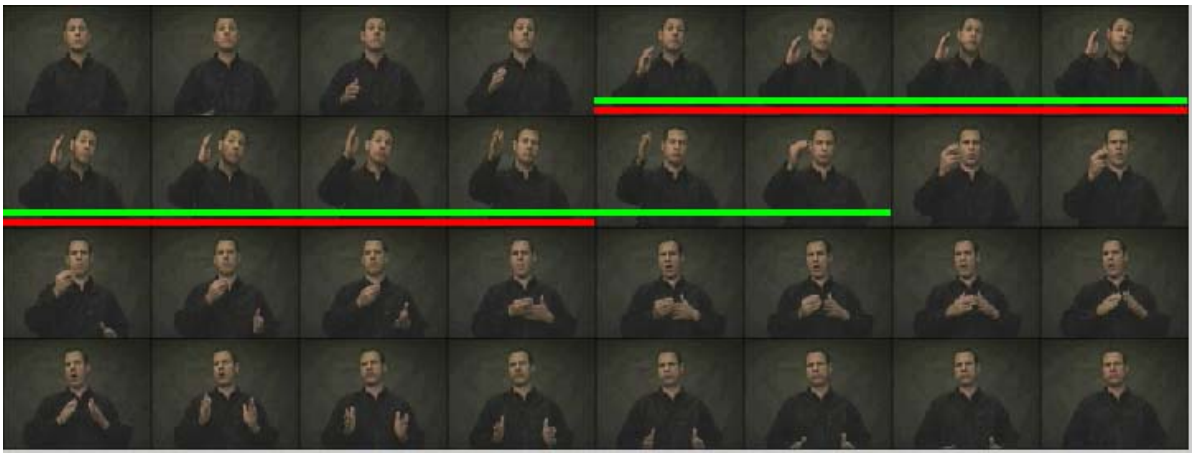

Figure 7: 'FUTURE' in test sentence 'FUTURE JOHN BUY HOUSE'

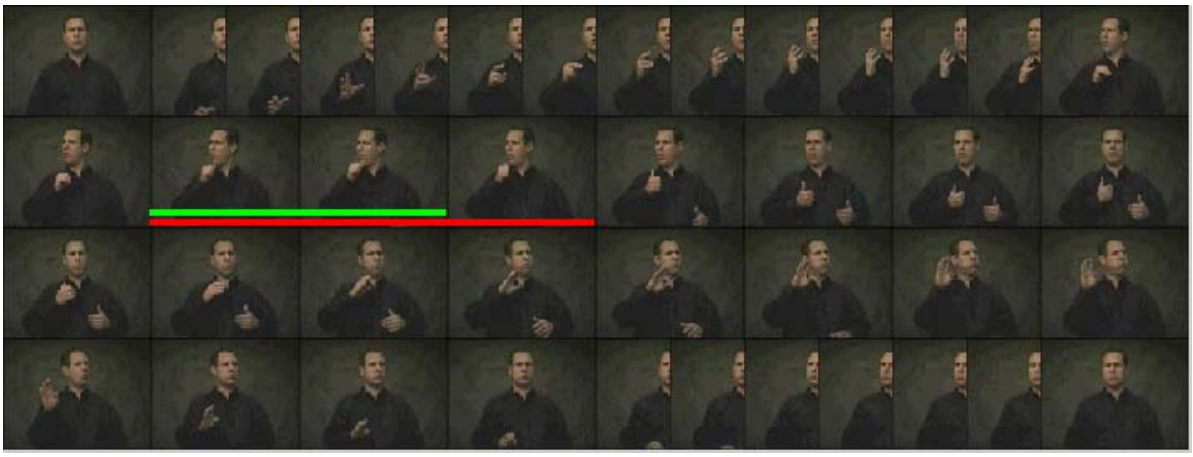

Figure 8: 'NOT' in test sentence 'JOHN PAST NOT LIVE BOSTON'

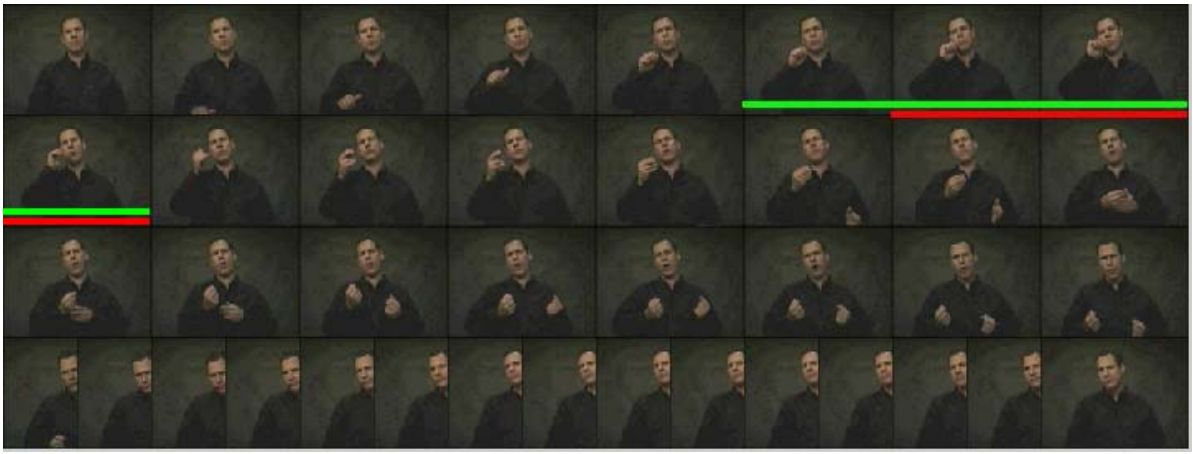

Figure 9: 'TOMORROW' in test sentence 'TOMORROW JOHN BUY CAR'

Figure 10: Across signer test with signeme for 'CAN' in test sentence 'MY SUITCASE MOVE CAN I'. The signeme for the sign 'CAN' formed by training sentences signed by the previous signer, now being tested on a new signer.