Journal of Machine Learning Research 13 (2012) 2589-2615 Submitted 11/11; Revised 5/12; Published 9/12

Finding Recurrent Patterns from Continuous Sign Language Sentences for Automated Extraction of Signs

Sunita Nayak SNAYAK@TAAZ.COM
Taaz Inc.
4250 Executive Square, Suite 420
La Jolla, CA 92037 USA

Kester Duncan KKDUNCAN@CSE.USF.EDU
Sudeep Sarkar SARKAR@CSE.USF.EDU
Department of Computer Science & Engineering
University of South Florida
Tampa, FL 33620, USA

Barbara Loeding BARBARA@USF.EDU
Department of Special Education
University of South Florida
Lakeland, FL 33803, USA

Editor: Isabelle Guyon

Abstract

We present a probabilistic framework to automatically learn models of recurring signs from multiple sign language video sequences containing the vocabulary of interest. We extract the parts of the signs that are present in most occurrences of the sign in context and are robust to the variations produced by adjacent signs. Each sentence video is first transformed into a multidimensional time series representation, capturing the motion and shape aspects of the sign. Skin color blobs are extracted from frames of color video sequences, and a probabilistic relational distribution is formed for each frame using the contour and edge pixels from the skin blobs. Each sentence is represented as a trajectory in a low dimensional space called the space of relational distributions. Given these time series trajectories, we extract signemes from multiple sentences concurrently using iterated conditional modes (ICM). We show results by learning single signs from a collection of sentences with one common pervading sign, multiple signs from a collection of sentences with more than one common sign, and single signs from a mixed collection of sentences. The extracted signemes demonstrate that our approach is robust to some extent to the variations produced within a sign due to different contexts. We also show results whereby these learned sign models are used for spotting signs in test sequences.

Keywords: pattern extraction, sign language recognition, signeme extraction, sign modeling, iterated conditional modes

1. Introduction

Sign language research in the computer vision community has primarily focused on improving recognition rates of signs either by improving the motion representation and similarity measures (Yang et al., 2002; Al-Jarrah and Halawani, 2001; Athitsos et al., 2004; Cui and Weng, 2000; Wang et al., 2007; Bauer and Hienz, 2000) or by adding linguistic clues during the recognition process

(Bowden et al., 2004; Derpanis et al., 2004). Ong and Ranganath (2005) presented a review of the automated sign language research and also highlighted one important issue in continuous sign language recognition. While signing a sentence, there exists transitions of the hands between two consecutive signs that do not belong to either sign. This is called movement epenthesis (Liddell and Johnson, 1989). This needs to be dealt with first before dealing with any other phonological issues in sign language (Ong and Ranganath, 2005). Most of the existing work in sign language assumes that the training signs are already available and often signs used in the training set are the isolated signs with the boundaries chopped off, or manually selected frames from continuous sentences. The ability to recognize isolated signs does not guarantee the recognition of signs in continuous sentences. Unlike isolated signs, a sign in a continuous sentence is strongly affected by its context in the sentence. Figure 1 shows two sentences 'I BUY TICKET WHERE?' and 'YOU CAN BUY THIS FOR HER' with a common sign 'BUY' between them. The frames representing the sign 'BUY' and the neighboring signs are marked. The unmarked frames between the signs indicate the frames corresponding to movement epenthesis. It can be observed that the same sign 'BUY' is preceded and succeeded by movement epenthesis that depends on the end and start of the preceding and succeeding sign respectively. The movement epenthesis also affects how the sign is signed. This effect makes the automated extraction, modeling and recognition of signs from continuous sentences more difficult when compared to just plain gestures, isolated signs, or finger spelling.

In this paper, we address the problem of automatically extracting the part of a sign that is most common in all occurrences of the sign, and hence expected to be robust with respect to the variation of adjacent signs. These common parts can be used for spotting or recognition of signs in continuous sign language sentences. They can also be used by sign language experts for teaching or studying variations between instances of signs in continuous sign language sentences, or in automated sign language tutoring systems. Furthermore, they can be used even in the process of translating sign language videos directly to spoken words.

In a related work inspired by the success of the use of phonemes in speech recognition, the authors sought to extract common parts in different instances of a sign and thus arrive at a phoneme-analogue for signs (Bauer and Kraiss, 2002). But unlike speech, sign language does not have a completely defined set of phonemes. Hence, we consider extracting commonalities at the sentence and sub-sentence level.

A different but a closely related problem is the extraction of common subsequences, also called motifs, from very long multiple gene sequences in biology (Bailey and Elkan, 1995; Lawrence et al., 1993; Pevzner and Sze, 2000; Rigoutsos and Floratos, 1998). Lawrence et al. (1993) used a Gibbs sampling approach based on discrete matches or mismatches of subsequences that were strings of symbols of gene sequences. Bailey and Elkan (1995) used expectation maximization to find common subsequences in univariate biopolymer sequences. In biology, researchers deal with univariate discrete sequences, and hence their algorithms are not always directly applicable to other multivariate continuous domains in time series like speech or sign language. Some researchers tried to symbolize a continuous time series into discrete sequences and used existing algorithms from bioinformatics. For example, Chiu et al. (2003) symbolized the time series into a sequence of symbols using local approximations and used random projections to extract common subsequences in noisy data. Tanaka et al. (2005) extended their work by performing principal component analysis on the multivariate time series data and projected them onto a single dimension and symbolized the data into discrete sequences. However, it is not always possible to get all the important information in

(a) Continuous Sentence 'I BUY TICKET WHERE?'

(b) Continuous Sentence 'YOU CAN BUY THIS FOR HER'

Figure 1: Movement epenthesis in sign language sentences. Frames corresponding to the common sign 'BUY' are marked in red. Signs adjacent to BUY are marked in magenta. Frames between marked frames represent movement epenthesis that is, the transition between signs. Note that the sign itself is also affected by having different signs preceding or following it.

the first principal component alone. Further extending his work, Duchne et al. (2007) find recurrent patterns from multivariate discrete data using time series random projections.

Due to the inherent continuous nature of many time series data like gesture and speech, new methods were developed that do not require approximating the data to a sequence of discrete symbols. Denton (2005) used a continuous random-walk noise model to cluster similar substrings. Nayak et al. (2005) and Minnen et al. (2007) use continuous multivariate sequences and dynamic time warping to find distances between the substrings. Oates (2002); Nayak et al. (2005) and Nayak et al. (2009a) are among the few works in finding recurrent patterns that address non-uniform sampling of time series. The recurrent pattern extraction approach proposed in this paper is based

on multivariate continuous time series, uses dynamic time warping to find distances between substrings, and handles length variations of common patterns.

Following the success of Hidden Markov Models (HMMs) in speech recognition, they were used by sign language researchers (Vogler and Metaxas, 1999; Starner and Pentland, 1997; Bowden et al., 2004; Bauer and Hienz, 2000; Starner et al., 1998) for representing and recognizing signs. However, HMMs require a large number of training data and unlike speech, data from native signers is not as easily available as speech data. Hence, non-HMM-based approaches have been used (Farhadi et al., 2007; Nayak et al., 2009a; Yang et al., 2010; Buehler et al., 2009; Nayak et al., 2009b; Oszust and Wysocki, 2010; Han et al., 2009). In this paper, we use a continuous trajectory representation of signs in a multidimensional space and use dynamic time warping to match subsequences. The relative configuration of the two hands and face in each frame is represented by a relational distribution (Vega and Sarkar, 2003; Nayak et al., 2005), which in itself is a probability density function. The motion dynamics of the signer is captured as changes in the relational distributions. It also allows us to interpolate motion, if required, for data sets with lower frame capture rates. It should also be noted that, unlike many of the previous works in sign language that perform tracking of the hands using 3D magnetic trackers or color gloves (Fang et al., 2004; Vogler and Metaxas, 2001; Wang et al., 2002; Ma et al., 2000; Cooper and Bowden, 2009), our representation does not require tracking and relies on skin segmentation.

We present a Bayesian framework to extract the common subsequences or signemes from all the given sentences simultaneously. Figure 2 depicts the overview of our approach. With this framework, we can extract the first most common sign, the second most common sign, the third most common sign and so on. We represent each sentence as a trajectory in a multi-dimensional space that implicitly captures the shape and motion in the video. Skin color blobs are extracted from frames of color video, and a relational distribution is formed for each frame using the edge pixels in the skin blobs. Each sentence is then represented as a trajectory in a low dimensional space called the space of relational distributions, which is arrived at by performing principal component analysis (PCA) on the relational distributions. There are other alternatives to PCA that are possible and discussed in Nayak et al. (2009b). The other choices do not change the nature of the signeme finding approach, they only affect the quality of the features. The starting locations $(a_1, ...a_n)$ and widths $(w_1, ...w_n)$ of the candidate signemes in all the n sentences are together represented by a parameter vector. The starting locations are initialized with random starting locations, based on uniform random sampling from each sentence, and the initial width values are randomly selected from a given range of values. The parameter vector is updated sequentially by sampling the starting point and width of the possible signeme in each sentence from a joint conditional distribution that is based on the locations and widths of the target possible signeme in all other sentences. The process is iterated till the parameter values converge to a stable solution. Monte Carlo approaches like Gibbs sampling (Robert and Casella, 2004; Gilks et al., 1998; Casella and George, 1992), which is a special case of the Metropolis-Hastings algorithm (Chib and Greenberg, 1995) can be used for global optimization while updating the parameter vector by performing importance sampling on the conditional probability distribution. However, this has a high burn-in period.

In this paper, we adopt a greedy approach based on the use of iterated conditional modes (ICM) (Besag, 1986). ICM converges much faster than a Gibbs sampler, but is known to be largely dependent on the initialization. We overcome this limitation by performing ICM a number of times equal to the average length of the n sentences, with different initializations. The most frequently occurring solution from all the ICM runs is considered as the final solution.

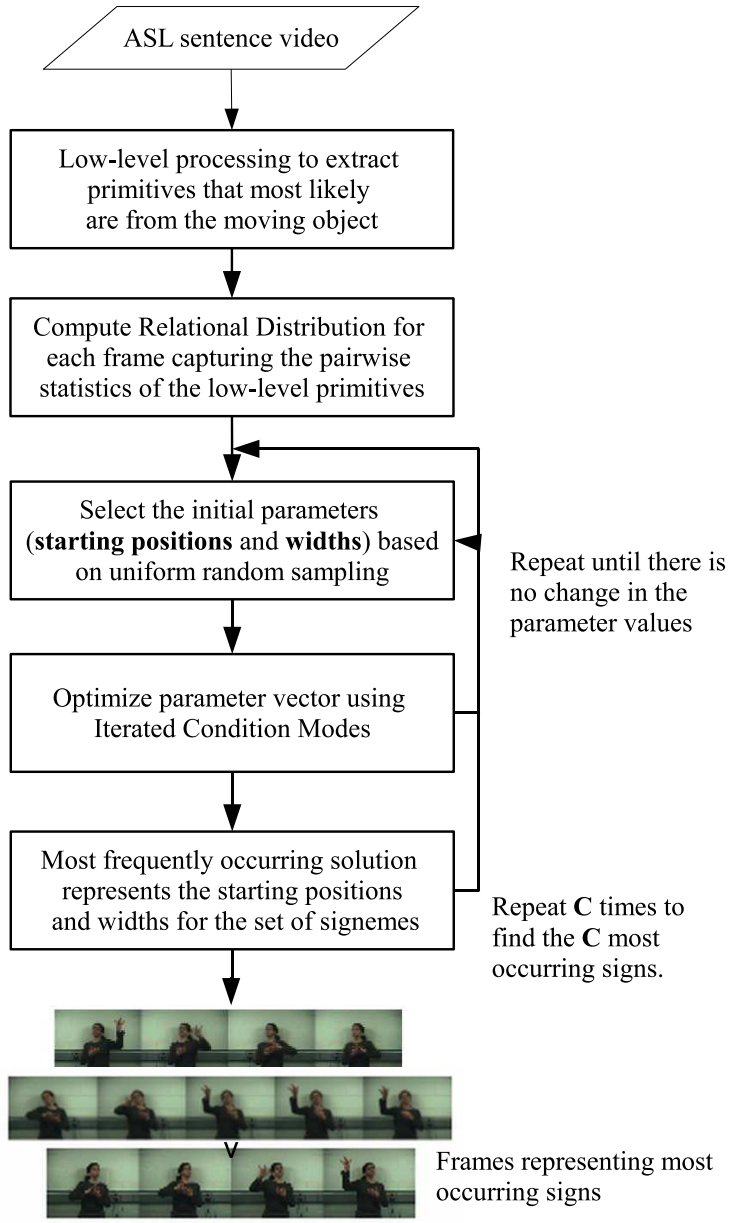


Figure 2: Overview of our approach. Each of the n sentences is represented as a sequence in the Space of Relational Distributions, and common patterns are extracted using iterated conditional modes (ICM). The parameter set $\{a_1, w_1, ...a_n, w_n\}$ is initialized using uniform random sampling and the conditional density corresponding to each sentence is updated in a sequential manner.

The work in this paper builds on the work of Nayak et al. (2009a) and is different in multiple respects. We propose a system that is generalized to extract more than one common sign from a collection of sentences (first most common sign, second most common sign and so on), whereas

in the previous work, only single signs were extracted. We also extract single signs from a mixed collection of sentences where there are more than one common sign in context. In addition to this, we present a more in-depth exposition of the underlying theory.

The contributions of this paper can be summarized as follows: (i) we present an unsupervised approach to automatically extract parts of signs that are robust to the variation of adjacent signs simultaneously from multiple sign language sentences, (ii) our approach does not consider all possible parameter combinations, instead samples each of them in a sequential manner until convergence, which saves a lot of computation, (iii) we show results on extracting signs from plain color videos of continuous sign language sentences without using any color gloves or magnetic trackers, and (iv) we show results whereby the learned signs are used for spotting signs in test sequences.

We organize the paper as follows. Section 2 presents a short review of relational distributions. In Section 3, we present the definition of signeme and then formulate the problem of finding signemes from a given set of sequences in a probabilistic framework. We describe how we solve it using iterated conditional modes. It is then followed by a description of our experiments and results in Section 4. Finally, Section 5 concludes the paper and discusses possible future work.

2. Relational Distributions

We use relational distributions to capture the global and relative configuration of the hands and the face in an image. Motion is then captured as the changes in the relational distributions. They were originally introduced by Vega and Sarkar (2003) for human gait recognition. They have also been used before for representing sign language sentences without the use of color gloves or magnetic trackers (Nayak et al., 2005, 2009b). We briefly review them here in this section.

How do we capture the global configuration of the object? We start with low-level primitives that are most likely to come from the articulated object. The exact nature of the low-level primitives can vary. Some common choices include edges, salient points, Gabor filter outputs and so on. We use edges in this work. We start from some level of segmentation of the object from the scene. These processes are fairly standard and have been used widely in gesture and sign recognition. They may involve color-based segmentation, skin-color segmentation, or background subtraction. In this work, we perform skin-color segmentation using histogram-based Bayesian classification (Phung et al., 2005). We use the contours of the skin blobs and Canny edges within the blobs as our low-level image primitives. The global configuration is captured by considering the relationships between these primitives.

We use the distance between two primitives in the vertical and horizontal directions (dx, dy) as relational attributes. Let vector $\mathbf{u} = \{dx, dy\}$ represent the vector of relational attributes. The joint probability function $P(\mathbf{u})$ then describes the distribution of primitives within an image and captures the shape of the pattern in the image. This probability is called a *relational distribution*. It captures the global configuration of the low-level primitives. Figure 3(c) illustrates how motion is captured using relational distributions. It shows the top view of the distributions. The region near to center represents points closer to each other, for example, the edge points within the face or within the hand, while farther from center represents the farther away points, for example, the relationship between edge points of a hand and the face. Notice the change in the relational distribution as the signer moves one of her hands. To be able to discriminate symmetrically opposite motion, we maintain the signs (or directions) of the horizontal and vertical distances between the edge pixels in each ordered pair. This leads to representing the probability distribution in a four quadrant system.

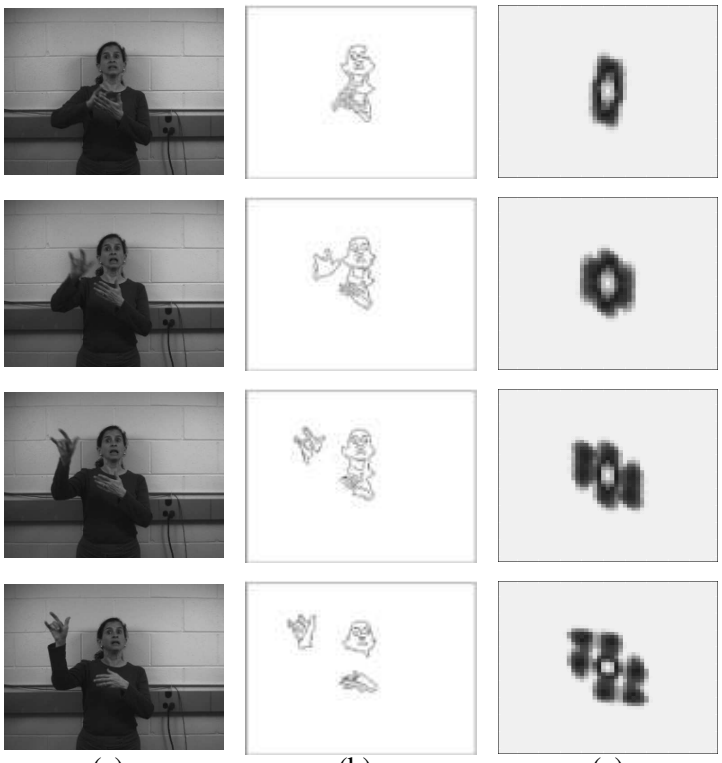


Figure 3: Variations in relational distributions with motion. (a) Motion sequence. (b) Edge pixels from the skin color blobs. (c) Relational distributions constructed from the low level features (edge pixels) of the images in the motion sequence. The horizontal axis of the relational distribution represents the horizontal distance between the edge pixels and its vertical axis represents the vertical distance between edge pixels.

Given that these relational distributions exhibit complicated shapes that are difficult to be modeled readily using a combination of simple shaped distributions such as Gaussian mixtures, we adopt non-parametric histogram-based representation. For better discrimination of the probabilities, we do not add counts to the center of the histogram which represents the distance of the edge pixels from itself or very close adjacent pixels. Each bin then counts the pairs of edge pixels between which the horizontal and vertical distances each lie in some fixed range that depends on the location of the bin in the histogram.

In our experiments, we found that an empirically-determined fixed histogram size of 51×51 was sufficient. The above range is then defined using linear mapping between the image size and the histogram size, for example, image size along the horizontal direction corresponds to half the histogram size in the horizontal direction. One could use histogram bin size optimization techniques for optimizing the histograms, but we do not address them in this paper. We then reduce the dimensionality of the relational distributions by performing PCA on the set of relational distributions from all the input sentences and retain the number of dimensions required to keep a certain percentage

of energy, typically 95%. The new subspace arrived at is called the space of relational distributions (SoRD). Each video sequence is thus represented as a sequence of points in the SoRD space.

Note that the choice of the relational distribution is not a central requirement for the signeme learning process discussed in this paper. We use relational distributions to enable us to work with pure video data, without the use of markers or colored gloves. If magnetic markers or colored gloves are available then one could use their attributes to construct a different feature space and consider trajectories in them. One advantage of our representation is that the face and head locations are implicitly taken into account in addition to the hands. In short, the first step of the process is to construct a time series representation in an appropriate feature space.

3. Problem Formulation

Sign language sentences are series of signs. Figure 4 illustrates the traces of the first vs. second dimension in the feature space, of three sentences S_1, S_2 and S_3 with only one common sign, R, among them. The signeme represents the portion of the sign that is most similar across the sentences. Table 3 defines the notations that will be used in this paper. We formulate the signeme extraction

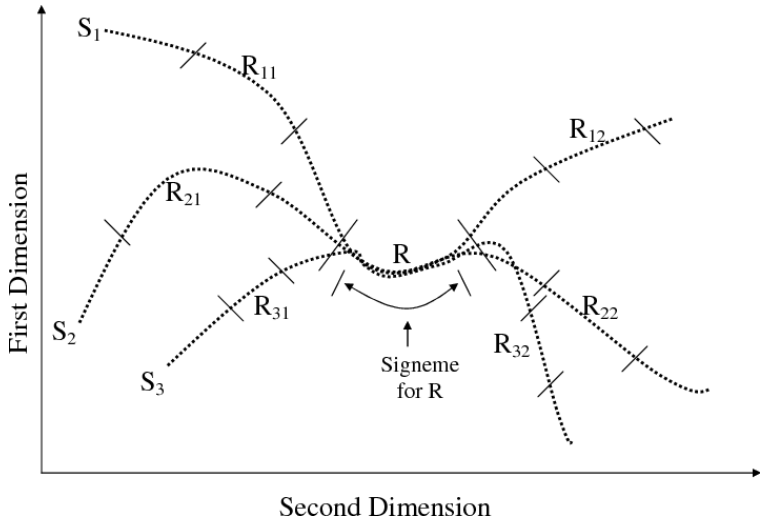


Figure 4: Concept of signemes. First vs. second dimensions of sentences S_1 with signs R_{11}, R, R_{12} in order, S_2 with signs R_{21}, R, R_{22} and S_3 with signs R_{31}, R, R_{32}. The common sign is R. The portion of R that is most similar across sentences is the signeme representative of R.

problem as finding the most recurring patterns among a set of n sentences $\{\vec{S}_1, \cdots, \vec{S}_n\}$, that have at least one common sign present in all the sentences. The commonality concept underlying the definition of a signeme can be cast in terms of distances. Let $\vec{s}_{a_i}^{w_i}$ represent a substring from the sequence $\vec{S}_i$ consisting of the points with indices $a_i, \cdots a_i + w_i - 1$, and $d(\vec{x}, \vec{y})$ denote the distance between two substrings $\vec{x}$ and $\vec{y}$ based on dynamic time warping. We define the set of signemes to be the set of substrings denoted by $\{\vec{s}_{a_1}^{w_1}, \cdots, \vec{s}_{a_n}^{w_n}\}$ that is most similar among all possible substrings from the given set of sentences. In the generalized case where C most common signs are sought, the

$\{\vec{S}_1, \cdots, \vec{S}_n\}$	Set of n sentences with at least one common sign present in all the sentences. The index within a sentence could represent time or arc length in configuration shape space
L_i	Length of sentence S_i
$\vec{s}_{a_j}^{w_j}$	Subsequence of sentence S_j starting from index a_j to $a_j + w_j - 1$. We may sometimes use $\vec{s}_{j,a}^{w}$ to make explicit the j-th index if it is not represented along with any other superscript or subscript of this term.
A, B	Possible choices of width for signemes of a sign include all integers from A to B. The values of A and B are decided based on the dynamics involved in the sign.
θ	Set of parameters $\{a_1, w_1, \cdots, a_n, w_n\}$ defining a set of substrings of the given sentences
$\theta_{(a_i)}$	Set of all parameters *excluding* the parameter a_i. We have similar interpretations for $\theta_{(w_i)}$ or $\theta_{(i)}$.
$d(\vec{x}, \vec{y})$	Distance between the subsequences $\vec{x}$ and $\vec{y}$ based on a mapping found using dynamic time warping (DTW). This distance has to be calculated carefully so that it is not biased towards finding short subsequences only.

Table 1: Notations

set of signemes are defined as $\{\vec{s}_{a_{11}}^{w_{11}}, \vec{s}_{a_{12}}^{w_{11}}, \cdots, \vec{s}_{a_{nC}}^{w_{nC}}\}$. In theory, C can extend to the number of words in the shortest sentence.

Let $\theta = \{a_1, w_1, \cdots a_n, w_n\}$ denote the parameter set representing a set of substrings, at least one from each of the n sentences, and θ_m denote the parameter set representing the target set of signemes in the n sentences. We find θ_m using the probabilistic framework of Equation 1.

$$\theta_m = \arg\max_{\theta} p(\theta) \tag{1}$$

Note that $p(\theta)$ is a probability over the space of all possible substrings. We define this probability to be a function of the inter-substring distances in Equation 2:

$$p(\theta) = \frac{g(\theta)}{\sum_\theta g(\theta)}. \tag{2}$$

The term $g(\theta)$ is defined in Equation 3 as follows:

$$g(\theta) = \exp\left(-\beta \sum_{i=1}^{n} \sum_{j=1}^{n} d(\vec{s}_{a_i}^{w_i}, \vec{s}_{a_j}^{w_j})\right) \tag{3}$$

with β being a positive constant.

Note that $g(\theta)$ varies inversely with the summation of the pair-wise distances of all the subsequences given by θ. Also note that $p(\theta)$ is hard to compute or even sample from because it is computationally expensive to compute the denominator in Equation 2, as it involves the summation over all possible parameter combinations. β acts as a scale parameter, which controls the slopes of the peaks in the probability space. It can also be looked upon as the smoothing parameter. If probability sampling algorithms like Gibbs sampling (Casella and George, 1992) are used in later steps, then the rate of convergence would be determined by this parameter.

Let θ_i represent the parameters from the i^{th} sentence, that is, $\{a_i, w_i\}$ and $\theta_{(i)}$ represent the rest of the parameters, $\{a_1, w_1 \cdots a_{i-1}, w_{i-1}, a_{i+1}, w_{i+1} \cdots a_n, w_n\}$. To make sampling easier, we construct a *conditional* density function of the parameters from each sentence, that is, θ_i, *given* the values of

the rest of the parameters, that is, $\theta_{(i)}$. In other words, we construct a probability density function of the possible starting points and widths in each sentence, given the estimated starting points and widths of the common pattern in all other sentences, that is, $f(\theta_i|\theta_{(i)})$. Of course, this conditional density function has to be *derived* from the joint density function specified in Eq. 2. This is outlined in Equation 4 as follows:

$$f(\theta_i|\theta_{(i)}) \quad = \quad \frac{p(\theta)}{p(\theta_{(i)})} = \frac{p(\theta)}{\sum_{\theta_i} p(\theta)} = \frac{g(\theta)}{\sum_{\theta_i} g(\theta)}. \tag{4}$$

Since the normalization to arrive at this conditional density function involves summation over one parameter, it is now easier to compute and sample from. The specific form for this conditional density function using the dynamic time warping (DTW) distances as described in Equation 5 is

$$f(\theta_i|\theta_{(i)}) = \frac{\exp\left(-\beta \sum_{k=1}^{n} d(\vec{s}_{a_i}^{w_i}, \vec{s}_{a_k}^{w_k}))\right)}{\sum_{\theta_i} \exp\left(-\beta \sum_{k=1}^{n} d(\vec{s}_{a_i}^{w_i}, \vec{s}_{a_k}^{w_k}))\right)}. \tag{5}$$

Note that the distance terms that do not involve a_i and w_i, that is, do not involve the i-th sentence appear both in the numerator and the denominator and so cancel out. For notational convenience, this is sometimes represented using conditional g functions described below in Equation 6 as:

$$f(\theta_i|\theta_{(i)}) = \frac{g(\theta_i|\theta_{(i)})}{\sum_{\theta_i} g(\theta_i|\theta_{(i)})}, \tag{6}$$

where $g(\theta_i|\theta_{(i)}) = \exp\left(-\beta \sum_{k=1}^{n} d(\vec{s}_{a_i}^{w_i}, \vec{s}_{a_k}^{w_k}))\right)$.

3.1 Distance Measure

The distance function d in the above equations needs to be chosen carefully such that it is not biased towards the shorter subsequences. Here, we briefly describe how we compute the distance between two substrings using dynamic time warping. Let l_1 and l_2 represent the length of the two substrings and $e(i, j)$ represent the Euclidean distance between the i^{th} data point from the first substring and the j^{th} data point from the second substring. Let D represent the score matrix of size $(l_1+1) \times (l_2+1)$. The 0^{th} row and 0^{th} column of D are initialized to infinity, except $D(0,0)$, which is initialized to 0. The rest of the score matrix, D, is completed using the following recursion of Equation 7:

$$D(i, j) = e(i, j) + \min\{D(i-1, j), D(i-1, j-1), D(i, j-1)\}, \tag{7}$$

where $1 \leq i \leq l_1$ and $1 \leq j \leq l_2$. The optimal warp path is then traced back from $D(l_1, l_2)$ to $D(0,0)$. The distance measure between the two substrings is then given by $D(l_1, l_2)$ normalized by the length of the optimal warping path.

3.2 Parameter Estimation

In order to extract the common signs from a given set of sign language sentences, we need to compute θ_i for each of the sentences sequentially. Gibbs sampling (Casella and George, 1992) is a Markov Chain Monte Carlo approach (Gilks et al., 1998) that allows us to sample the conditional probability density $f(\theta_i|\theta_{(i)})$ for all the sequences sequentially and then iterate the whole process

until convergence. Gibbs sampling results in a global optimum, but its convergence is very slow. The burn-in period is typically thousands of iterations. Therefore, we perform the optimization using iterated conditional modes (ICM), first proposed by Besag (1986). ICM has much faster convergence, but it is also known to be heavily dependent on the initialization. We address this limitation by running the optimization multiple times with different initializations and choosing the most frequently occurring solution as the final solution.

Algorithm 1: Iterated Conditional Modes($\{a_1^0, w_1^0, \cdots, a_n^0, w_n^0\}$)

comment: Choose $\{a_1, w_1, \cdots, a_n, w_n\}$ that maximizes the distribution $p(a_1, w_1, \cdots, a_n, w_n)$

comment: Initialization:

$\theta_0 \leftarrow \{a_1^0, w_1^0, \cdots, a_n^0, w_n^0\}$

repeat

- **for** $i \leftarrow 0$ **to** n
  - **do**
    - **comment:** Jointly sample a_i, w_i. L_i is the length of sequence S_i
    - **for** $w_i \leftarrow A$ **to** B
      - **do**
        - **for** $a_i \leftarrow 0$ **to** $L_i - w_i + 1$
          - **do** $g(a_i, w_i|\theta_{(a_i,w_i)}) \leftarrow \exp\left(-\beta \sum_{k=1}^{n} d(\vec{s}_{a_i}^{w_i}, \vec{s}_{a_k}^{w_k})\right)$
    - **comment:** Normalize
    - **for** $w_i \leftarrow A$ **to** B
      - **do**
        - **for** $a_i \leftarrow 0$ **to** $L_i - w_i + 1$
          - **do** $f(a_i, w_i|\theta_{(a_i,w_i)}) \leftarrow \frac{g(a_i,w_i|\theta_{(a_i,w_i)})}{\sum_{a_i,w_i} g(a_i,w_i|\theta_{(a_i,w_i)})}$
    - $a_i, w_i \leftarrow$ ARG MAX $(f(a_i, w_i|\theta_{(a_i,w_i)}))$

until CHANGE IN PARAMETERS$(\{a_1, w_1, \cdots, a_n, w_n\}) == 0$

Algorithm 1 outlines the process of ICM to extract the common patterns or signemes from a set of sentences with a given initial parameter vector. We aim to select the set of parameters that maximizes the probability $p(\theta)$ or $p(a_1, w_1, \cdots, a_n, w_n)$. We do that by estimating each of the parameters $a_1, w_1, \cdots a_n, w_n$ in a sequential manner. Since we expect the starting location and width of a subsequence representing the common sign to be strongly correlated, we estimate a_i and w_i jointly. First we compute $g(\theta_i|\theta_{(i)})$ that is, $g(a_i, w_i|\theta_{(a_i,w_i)})$ from which we compute the conditional density functions $f(\theta_i|\theta_{(i)})$ that is, $f(a_i, w_i|\theta_{(a_i,w_i)})$. Note that it involves a summation over a_i and w_i only, which involves much less computation than that required for computing $p(\theta)$ which involves a summation over $a_1, w_1, \cdots a_n, w_n$. The values for a_i and w_i are updated with those that maximize the conditional density $f(\theta_i|\theta_{(i)})$. The process is carried out sequentially for $i = 1$ to n, and then repeated iteratively till the values of the parameter vector $\{a_1, w_1, a_2, w_2, \cdots a_n, w_n\}$ do not change any more.

Figure 5 depicts the sampling process for a single iteration, r. Note the conditional and sequential nature of sampling from various sentences within the single iteration. In Figure 6, we show an example of how the conditional probability $f(\theta_{a_i,w_i}|\theta_{(a_i,w_i)})$ changes for the first seven sentences from a given set of fourteen video sentences containing a common sign 'DEPART'. The vertical axis

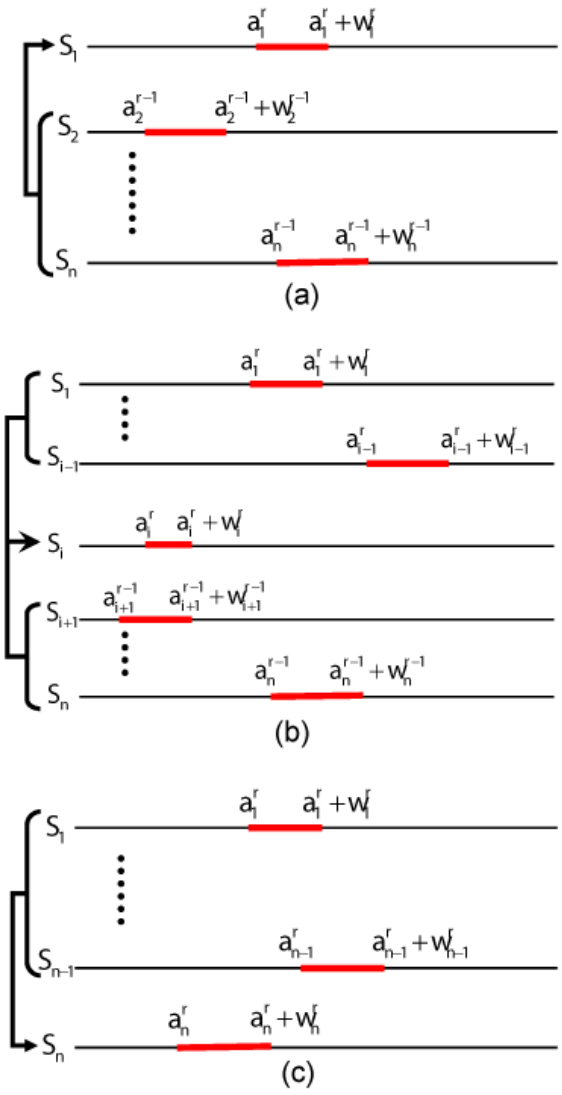


Figure 5: Sequential update of the parameter values using ICM. (a), (b) and (c) respectively show the parameter updates in the first sentence, the i^{th} and the n^{th} sentences. In the r^{th} iteration, the parameters of the common sign in i^{th} sentence is computed based on the parameter values of the previous $(i-1)$ sentences obtained in the same iteration, and those of the $(i+1)^{th}$ to n^{th} sentences obtained in the previous, that is, the $(r-1)^{th}$ iteration.

in the probabilities represents the starting locations and the horizontal axis represents the possible widths. The brighter regions represent a higher probability value. Note that the probabilities are spread out in the first iteration for each sentence and it slowly converges to a fixed starting location for each of them. They remain more spread out across the horizontal (width) axis because we vary

the width only in a small range of A to B for each sign, that is decided based on the amount of motion present in the sign.

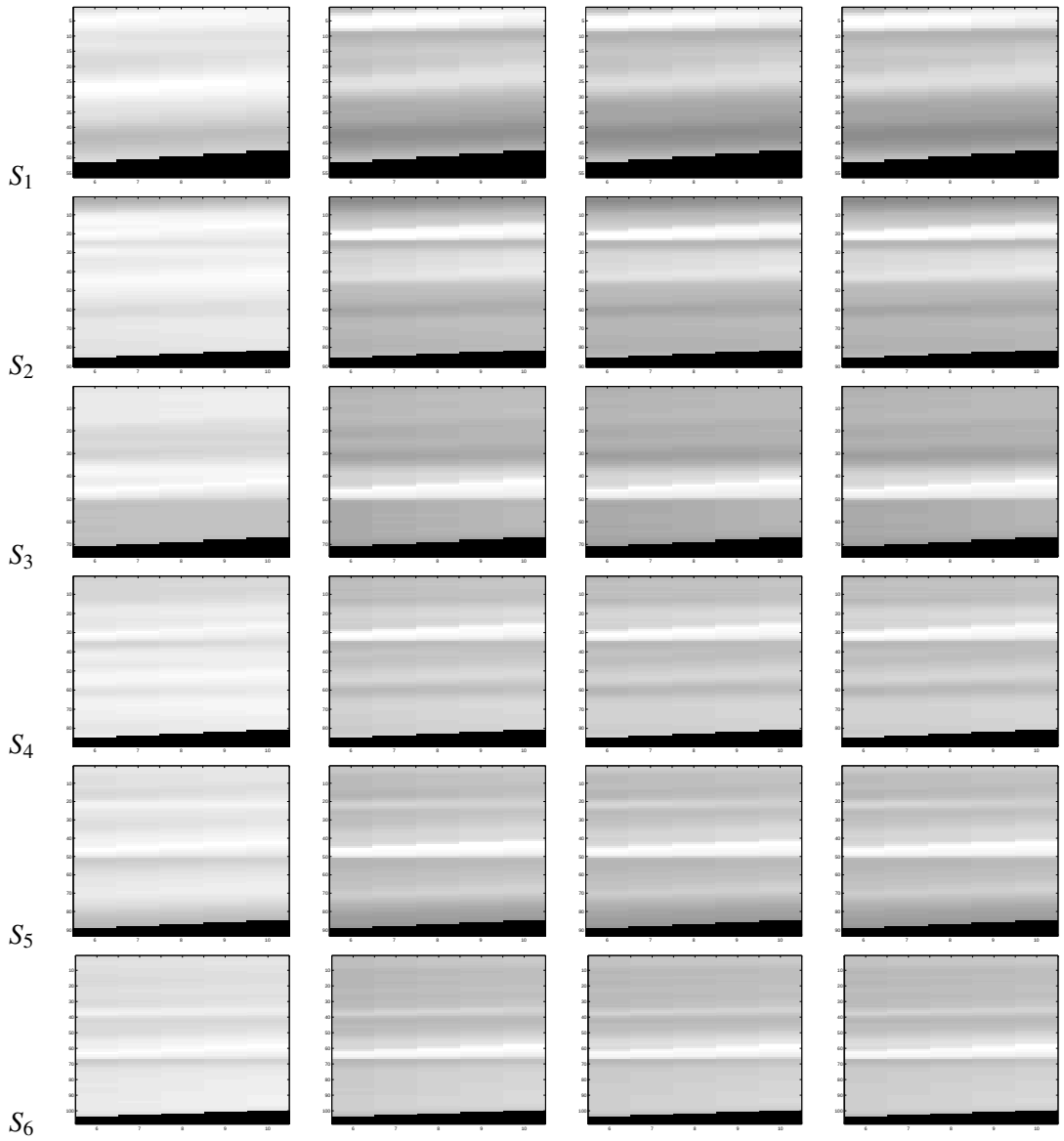


Figure 6: Convergence of the conditional probability density $f(\theta_i|\theta_{(i)})$ for sentences $S_1 ... S_6$ from a given set of sentences $S_1 ... S_{14}$. The brighter regions represent a higher probability value. The vertical axis in the probabilities represents the starting locations and the horizontal axis represents the possible widths. Note that the probabilities are spread out in the first iteration and it slowly converges to a particular starting location. They are still spread across the horizontal (width) axis because we vary the width only in a small range that is decided based on the amount of motion present in the sign.

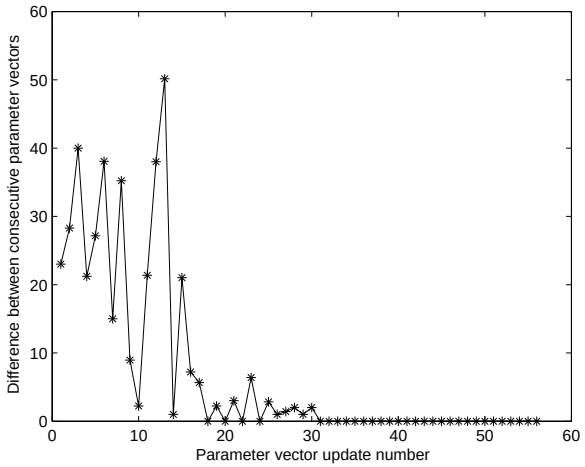


Figure 7: Convergence of values of the parameter set. The above plot shows the norm of the difference between two consecutive parameter vectors representing the set of starting points and widths of the common subsequence in the given set of sequences. It shows the typical convergence with a given initialization vector. ICM is repeated with multiple initializations and the most frequently occurring solution is considered as the final solution.

Figure 7 plots the typical convergence of the parameter values in a single ICM run. It plots the norm of difference between consecutive parameter vectors versus the parameter vector update count, which is incremented each time a parameter is sampled or selected from the probability distribution $f(\theta_i|\theta_{(i)})$. It shows that ICM converges in less than $56/14 = 4$ iterations. This, in turn, also indicates the local nature of the optimization achieved with ICM. The initialization is very important in this case. In the next subsection, we describe how we address this problem.

3.3 Sampling Starting Points For ICM

In order to address the local convergence nature of ICM, we adopt a uniform random sampling-based approach. We start by randomly assigning values to the parameter vector θ. The width w_i^0 is obtained by sampling a width value based on uniform random distribution from the set of all possible widths in a given range $[A,B]$. The value for a_i^0 is obtained by sampling a starting point based on uniform random distribution from the set of all possible starting points in the i^{th} sequence, that is, from the set $\{1 \cdots (L_i - w_i^0 + 1)\}$.

Different initial parameter vectors are obtained by independently sampling the sentences multiple times. ICM is run using each initial parameter vector generated and the most common solution is considered as the final solution. The uniform sampling of the frames in the sentences for selecting the starting locations ensures the whole parameter space is covered uniformly. The number of times we sample the initial parameter vector and run the ICM algorithm decides how densely we cover the whole parameter space. We run it the number of times equal to the average number of frames in each sentence from the given set of sentences for extracting the sign. One could choose to run a multiple of the average number of times as well, but we found the average number to be

sufficient to show the stability of the solution in our experiments. Algorithm 2 presents the process as a pseudocode.

Algorithm 2: Extract Signemes($L_1, \cdots L_n, A, B$)

comment: Generate multiple initialization vectors and call ICM with each of them.

$N = \text{MEAN}(L_1, L_2, \cdots, L_n)$
for $j \leftarrow 1$ **to** N
do $\begin{cases} \textbf{for } i \leftarrow 1 \textbf{ to } n \\ \quad \textbf{do } \begin{cases} w_i^0 = \text{UNIFORM}(A \cdots B) \\ a_i^0 = \text{UNIFORM}(1 \cdots L_i - w_i^0 + 1) \end{cases} \\ \{a_1^j, w_1^j, \cdots, a_n^j, w_n^j\} = \text{ITERATED CONDITIONAL MODES}(a_1^0, w_1^0, \\ \cdots, a_n^0, w_n^0) \end{cases}$
for $i \leftarrow 1$ **to** n
do $\begin{cases} \textbf{comment:} \text{ Assign most frequently occurring value as the final value for each parameter.} \\ w_i = \text{MODE}(w_i^j) \\ a_i = \text{MODE}(a_i^j) \end{cases}$

For extracting the sign 'DEPART' from 14 sentences, we had 89 frames per sentence on an average. Hence we ran 89 different ICM runs for extracting the common subsequence representing 'DEPART'. Figure 8 shows the plots of histograms of start and end location of the sign in each of the 14 sentences from the 89 runs. It should be noted that in most of the sentences, more than 50% of the total number of runs result in the same solution.

4. Experiments And Results

In this section, we present visual and quantitative results of our approach for extracting signemes from video sequences representing sentences from American Sign Language. We first describe the data set used then present the results of the automatic common pattern extraction.

4.1 Data Set

Our data set consists of 155 American Sign Language (ASL) video sequences organized into 12 groups (collections) based on the vocabulary (word that pervades the sentences of the group). For instance, the 'DEPART' group is comprised of all the sentences containing the word 'DEPART', the 'PASSPORT' group is comprised of all the sentences containing the word 'PASSPORT' and so on. The breakdown of these 'pure' groups and the number of sentences (sequences) in each are as follows.

- DEPART - 14 sentences
- BAGGAGE - 14 sentences
- CANT - 14 sentences

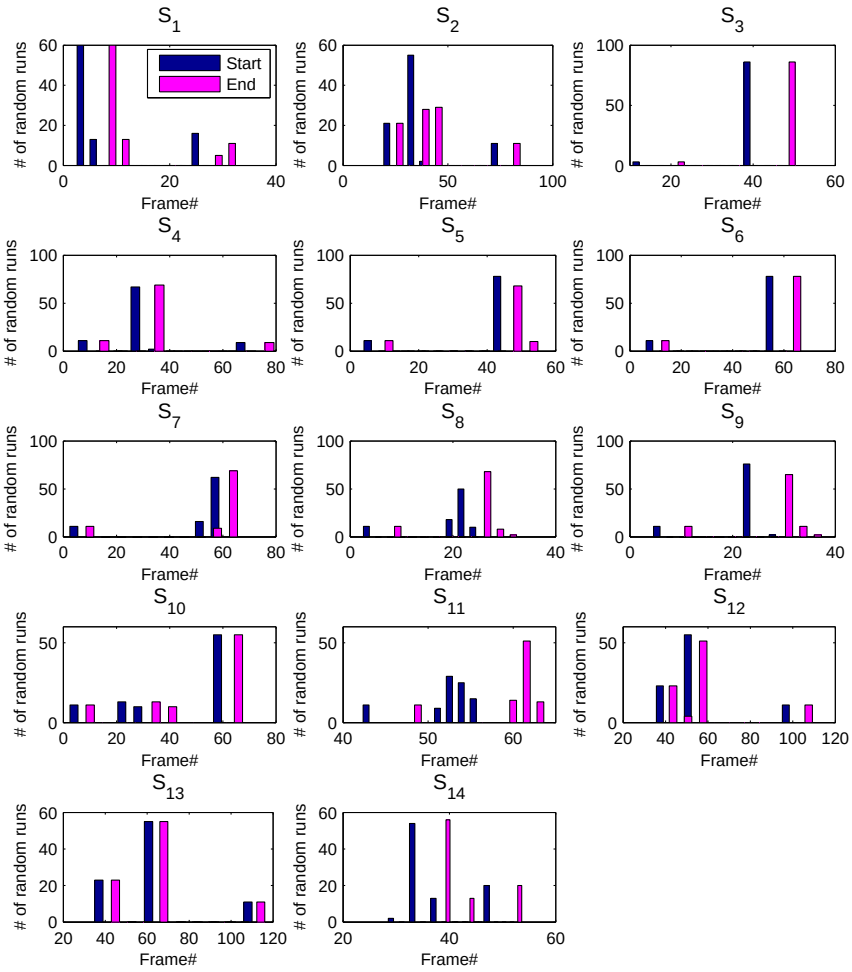


Figure 8: Histograms showing the start and end locations of signs extracted from 14 different sentences using multiple ICM runs. The initial parameter vector for each ICM run was chosen independently using uniform random sampling. As it can be seen the start and end points found by most of the runs converge to the same solution (denoted by single high bars in most of sentences). The legend shown in the plot for the first sentence, S_1, holds for other sentences as well.

- BUY - 11 sentences
- SECURITY - 16 sentences
- HAVE - 6 sentences
- MOVE - 11 sentences
- TIME - 14 sentences

- FUTURE - 12 sentences
- TABLE - 13 sentences
- PASSPORT - 14 sentences
- TICKET - 16 sentences

This data set was used to extract 12 common subsequences when we searched for the first most common sign, and 24 common subsequences when we searched for the second most common sign. We also organized the video sequences into 10 groups by combining two 'pure' groups of sentences as described above. This was used to investigate the power of our framework for selecting the common sequences in a 'mixed' collection. The breakdown of these 'mixed' groups and the number of sentences in each are as follows:

- DEPART (14 sentences) + BAGGAGE (14 sentences)
- CANT (14 sentences) + BUY (11 sentences)
- TIME (14 sentences) + TABLE (13 sentences)
- PASSPORT (14 sentences) + TICKET (16 sentences)
- SECURITY (16 sentences) + FUTURE (12 sentences)
- MOVE (11 sentences) + HAVE (6 sentences)
- BUY (11 sentences) + TABLE (13 sentences)
- DEPART (14 sentences) + FUTURE (12 sentences)
- BAGGAGE (14 sentences) + TICKET (16 sentences)
- SECURITY (16 sentences) + PASSPORT (14 sentences)

All of the signs were performed by the same signer with plain clothing and background. The video sequences were captured at 25 frames per second with a frame resolution of 490×370.

4.2 Common Pattern Extraction Results

In this section, we present the results of our method for extracting common patterns from sign language sentences. We first present results for extracting the single most common sign and multiple common signs from the 'pure' sentence groups, followed by results for the most common patterns from the 'mixed' groups.

4.2.1 EXTRACTING THE MOST COMMON PATTERN

We perform extraction of the most common patterns from the 'pure' sentence groups. We possess a priori knowledge of the most common word due to the organization of the sentence groups. However, our goal is to extract the most common sequences automatically. As an example, Figure 9 depicts the result of extraction of the sign 'DEPART' from 14 video sequences. It plots the SoRD first dimension coefficients of the frames vs. the frame number for each sentence. The highlighted

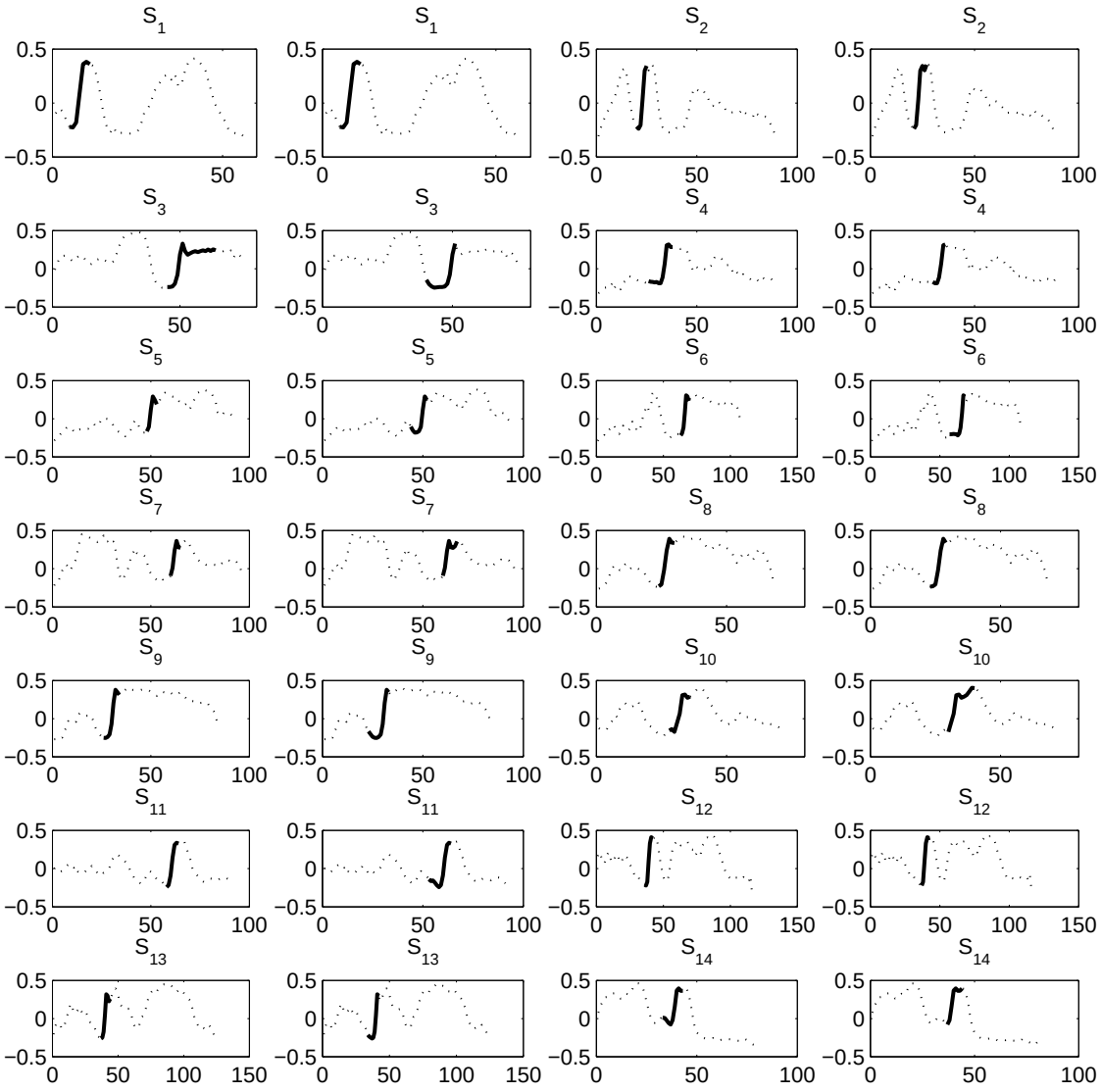
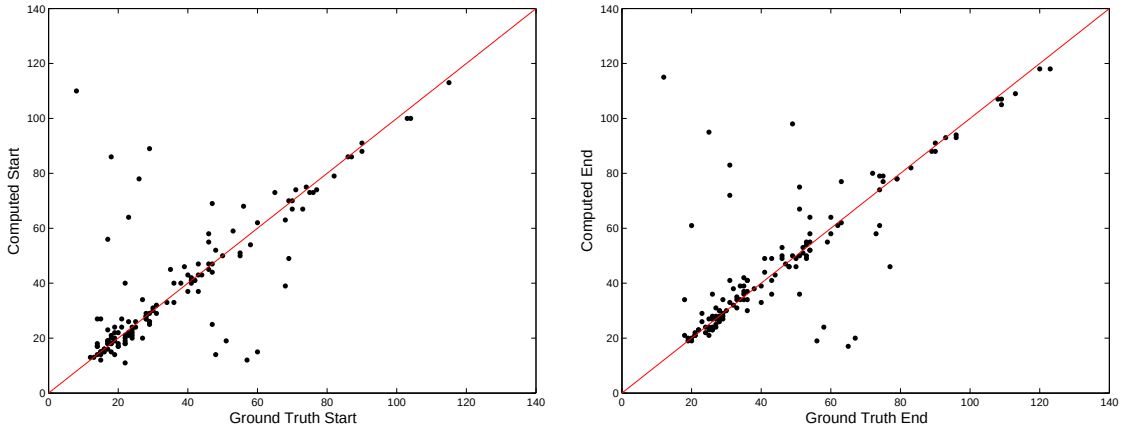


Figure 9: The first dimension of the video sequences containing a common sign 'DEPART'. The sequences are indicated by the dotted curves and the solid lines on each of them indicate the common pattern or signeme. The odd columns represent the ground truth and the even columns show the results.

portions represent the signeme. The odd columns show the ground truth and the even columns show the corresponding results. As can be seen, the extracted patterns and the corresponding ground truth patterns are quite similar, except for a few frames at the beginning and end of the some of the patterns. Note that since we deal with continuous video sequences, a difference of one or two frames between the ground truth and the extracted pattern is not considered a problem.



(a) Video Start Point Estimation

(b) Video End Point Estimation

Figure 10: Extraction of the most common patterns or signemes from the 'pure' sentence groups. The closer the points are to the diagonal, the closer the result is to the ground truth.

Figure 10(a) shows the scatter plot of the ground truth start positions vs. the estimated start positions of the pattern extracted from each of the 155 sentences in the video data set. Figure 10(b) shows the corresponding scatter plot for the end position of the patterns in the sentences. As can be seen most of the points in the scatter plots lie along the diagonal. This indicates that very few of the extracted patterns are wrong. Incorrect results correspond to the points positioned far from the diagonal. Figures 11 and 12 show one instance of the signeme extracted from group of sentences.

4.2.2 EXTRACTING MULTIPLE COMMON SIGNS

In this section we present some visual results for the extraction of the two most common signs from the 'pure' groups of sentences. We focused on extracting only two signs because the shortest ASL sentence contained two signs. Figure 13 shows the results for the two most common signs extracted from the sentence 'BAGGAGE THERE NOT MINE THERE'. The extracted subsequences correspond to the ASL words 'BAGGAGE' and 'MINE'. Consequently, the word 'BAGGAGE' appears in all the 14 sentences of the group, whereas the word 'MINE' (or 'MY') shows up in 11 sentences coinciding with what was expected. Similarly, Figure 14 shows the results for the two most common signs extracted from the sentence 'MY PASSPORT THERE STILL GOOD THERE'. The extracted subsequences correspond to the ASL words 'MY' and 'PASSPORT'. The word 'MY' appears in all the 11 sentences of the group, whereas the word 'PASSPORT' appears in all 14 sentences. These results are encouraging.

4.2.3 EXTRACTING THE MOST COMMON PATTERNS FROM MIXED SENTENCES

We perform extraction of the most common patterns from the collection of 'mixed' sentences as outlined in Section 4.1. Figure 15(a) shows the scatter plot of the ground truth start positions vs. the estimated start positions of the pattern extracted from each of the sentences. Similarly, Figure 15(b)

(a) BUY

(b) CANT

(c) DEPART

(d) FUTURE

(e) MOVE

Figure 11: Signemes extracted from sentences

shows the corresponding scatter plot for the end position of the patterns in the sentences. As can be seen, the points are more scattered as compared to the results shown in Figure 10 where the sentences used were known to contain common words. However, this result is still encouraging. A large proportion of the extracted patterns are incorrect, but there are many relatively near the diagonal. This result demonstrates the robustness of our algorithm for finding similarities in the presence of great dissimilarity. We believe that the incorrect patterns extracted are due to the differences in the frame width ranges for the mixed sentence sets. For example, sentences containing the word 'MOVE' were combined with sentences containing the word 'HAVE'. The frame width range for the sign 'HAVE' is between 4 and 6 frames with 4 being the minimum width and 6 being the maximum width. On the other hand, the frame width range for the sign 'MOVE' is between 19 and 27 frames. Combining these width ranges could be done using an average of the two or by selecting the minimum and maximum values between the two. However, these methods produced similar results. The correct combination of these range widths is a priority for future work.

(f) PASSPORT

(g) SECURITY

(h) TICKET

(i) TIME

(j) TABLE

Figure 12: Signemes extracted from sentences

4.3 Sign Localization

We used the extracted signemes to localize or spot signs in test sentences. The same process that is used for training sign models is used for sign localization. However, rather than randomly assigning initial parameter values, we use the parameters learned. We tested with 12 test sentences from the 'pure' group specified in Section 4.1 and their lengths varied from 4 to 12 signs. These test sentences were not used during training. The set of points representing the signeme were 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 SoRD points of the signeme retrieved from the test sentence are mapped to their nearest frames and compared with the ground truth frame series representing the sign in the sentence. Localization performance is characterized as follows. Let a_1 and b_1 denote the start and end frame numbers of the underlying ground truth sign in the test sentence, and a_2 and b_2 denote the start and end frame numbers of the subsequence retrieved as the

(a) Frames corresponding to the word 'BAGGAGE'

(b) Frames corresponding to the word 'MINE'

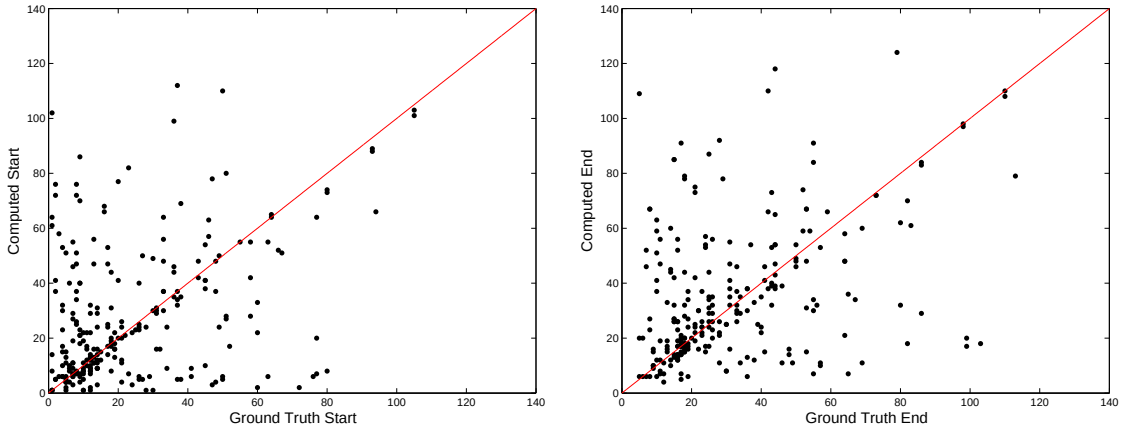
Figure 13: Extraction of the two most common patterns or signemes from the sentence 'BAGGAGE THERE NOT MINE THERE'.

(a) Frames corresponding to the word 'MY'

(b) Frames corresponding to the word 'PASSPORT'

Figure 14: Extraction of the two most common patterns or signemes from the sentence 'MY PASSPORT THERE STILL GOOD THERE'.

signeme for the test sentence. We calculate the precision and recall values of each test sentence as $\frac{m}{a_2-a_1+1}$ and $\frac{m}{b_2-a_2+1}$ respectively where m is the number of overlapping frames. Table 2 displays the results acquired. The 'Baggage', 'Cant', 'Have', and 'Table' test sequences were failure cases where there was no overlap between the extracted model frames and the localization frames (see Figure 16). Notice that the localization results heavily depend on the extracted signeme models. For a visual representation of this information, we define the Start Offset, ΔS, and End Offset, ΔE, as $\Delta S = a_1 - a_2$ and $\Delta E = b_1 - b_2$. The plot of the Start Offset vs. the End Offset is shown in Figure 16. Ideally, both the offsets should be zero. The points for different signs are scattered in the four quadrants depending on the nature of the overlap between the ground truth sign and the retrieved



(a) Video Start Point Estimation (b) Video End Point Estimation

Figure 15: Extraction of the most common patterns or signemes from the 'mixed' sentence groups. The closer the points are to the diagonal, the closer the result is to the ground truth.

Test Group	Precision	Recall
Buy	1.0	0.70
Depart	1.0	0.64
Future	0.71	0.756
Move	1.0	0.60
Passport	1.0	0.47
Security	0.57	0.67
Ticket	1.0	0.58
Time	0.63	1.0

Table 2: Localization Performance

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 closer it is to the origin, the better the quality.

5. Conclusion And Future Work

We presented a novel algorithm to extract signemes, that is, the common pattern representing a sign, from multiple long video sequences of American Sign Language (ASL). A signeme is a part of the sign that is robust to the variations of the adjacent signs and the associated movement epenthesis. We first represent each sequence as a series of points in a low dimensional space of relational distributions, and then use a probabilistic framework to locate the signemes in each sequence concurrently. We use iterative conditional modes (ICM) to sample the parameters, that is, the starting location and width of the signemes in each sentence in a sequential manner. We show results on ASL video sequences that do not involve using any magnetic trackers or gloves for extracting the

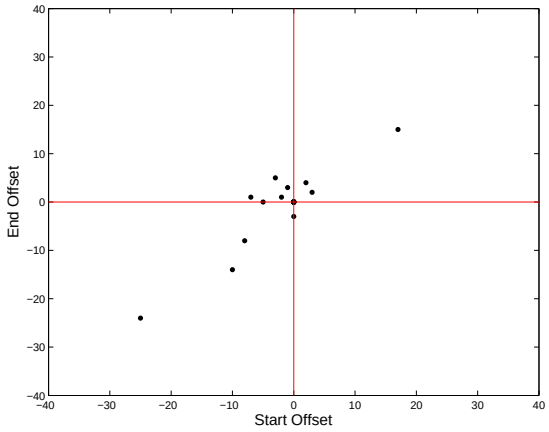


Figure 16: Start Offset vs. End Offset of Localized Signs

most common signs. The extracted signemes demonstrate that our approach is robust to some extent to the variations produced within a sign due to different contexts.

The approach in this paper can be used to speed up training set generation for ASL algorithms by drastically reducing the manual aspect of the process. Rather than manually demarcating signs in continuous sentences, which for our work took an expert approximately 5 minutes, we would just need instances of sentences containing the sign whose model is sought and based on our experiments this can be generated in approximately 2 minutes. Another contribution of this work is an empirically derived robust representation of the sign that is stable with respect to the variations due to neighboring signs and sentence context. These stable representations could be useful for detection of signs and gestures in extended gesture sequences.

There are some ways we can advance the work in this paper. One issue is the precision of the features used for representing the video sequences. Relational distributions when used as fixed size histograms perform well for discriminating global motion. However, optimizing the bin size of the histograms to the required precision might improve the accuracy. Additionally, we plan to extend our work to address the challenge of handling the large variations encountered when automatically recognizing signemes across different signers. Also, the algorithm is dependent to a large extent on the distance measure and conventional dynamic time warping cannot deal with the amplitude variations in the signs, which are very common across signers. We plan to work on a variation of dynamic time warping that is robust to amplitude differences between various instances of signs.

Acknowledgments

This work was supported in part by funds from University of South Florida's College of Engineering Interdisciplinary Scholarship Program and the National Science Foundation under ITR grant IIS 0312993.

References

O. Al-Jarrah and A. Halawani. Recognition of gestures in Arabic Sign Language using neuro-fuzzy systems. *Artificial Intelligence*, 133:117–138, 2001.

V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: A method for efficient approximate similarity rankings. *IEEE Conference On Computer Vision And Pattern Recognition*, pages 268–275, 2004.

T Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.

B. Bauer and H. Hienz. Relevant features for video-based continuous sign language recognition. In *Automatic Face And Gesture Recognition*, pages 440–445, 2000.

B. Bauer and K. F. Kraiss. Video-based sign recognition using self-organizing subunits. In *International Conference On Pattern Recognition*, volume 2, pages 434–437, 2002.

J. Besag. On the statistical analysis of dirty pictures. *Journal Of The Royal Statistical Society*, pages 259–302, 1986.

R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *European Conference On Computer Vision*, volume 1, pages 390–401, 2004.

P. Buehler, A. Zisserman, and M. Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 2961–2968, June 2009.

G. Casella and E.I. George. Explaining the Gibbs sampler. *The American Statistician*, 46:167–174, 1992.

S. Chib and E. Greenberg. Understanding the Metropolis–Hastings algorithm. *The American Statistician*, 49:327–335, 1995.

B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. *ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, pages 493–498, 2003.

H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 2568–2574, June 2009.

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

A. Denton. Kernel-density-based clustering of time series subsequences using a continuous random-walk noise model. *International Conference On Data Mining*, 2005.

K.G. Derpanis, R.R. Wildes, and J.K. Tsotsos. Hand gesture recognition within a linguistics-based framework. *European Conference On Computer Vision*, pages 282–296, 2004.

F. Duchne, C. Garbay, and V. Rialle. Learning recurrent behaviors from heterogeneous multivariate time-series. *Artificial Intelligence In Medicine*, (1):25–47, 2007.

G. Fang, X. Gao, W. Gao, and Y. Chen. A novel approach to automatically extracting basic units from Chinese Sign Language. *International Conference On Pattern Recognition*, 4:454–457, 2004.

A. Farhadi, D.A. Forsyth, and R. White. Transfer learning in sign language. In *Computer Vision And Pattern Recognition*, pages 1–8, 2007.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman and Hall, 1998.

J. Han, G. Awad, and A. Sutherland. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognition Letters*, 30(6):623–633, 2009.

C.E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

S.K. Liddell and R.E. Johnson. American Sign Language: The phonological base. *Sign Language Studies*, pages 195–277, 1989.

J. Ma, W. Gao, C. Wang, and J. Wu. A continuous Chinese Sign Language recognition system. *International Conference On Automatic Face And Gesture Recognition*, pages 428–433, 2000.

D. Minnen, C.L. Isbell, I. Essa, and T. Starner. Discovering multivariate motifs using subsequence density estimation and greedy mixture learning. *Conference On Artificial Intelligence*, 2007.

S. Nayak, S. Sarkar, and B. Loeding. Unsupervised modeling of signs embedded in continuous sentences. *IEEE Workshop On Vision For Human-Computer Interaction*, 2005.

S. Nayak, S. Sarkar, and B. Loeding. Automated extraction of signs from continuous sign language sentences using iterated conditional modes. In *IEEE Conference On Computer Vision And Pattern Recognition*, pages 2583–2590, June 2009a.

S. Nayak, S. Sarkar, and B. Loeding. Distribution-based dimensionality reduction applied to articulated motion recognition. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 31 (5):795–810, May 2009b.

T. Oates. PERUSE: An unsupervised algorithm for finding recurring patterns in time series. *International Conference On Data Mining*, pages 330–337, 2002.

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*, 27:873–891, June 2005.

M. Oszust and M. Wysocki. Determining subunits for sign language recognition by evolutionary cluster-based segmentation of time series. In *Artifical Intelligence And Soft Computing*, volume 6114 of *Lecture Notes In Computer Science*, pages 189–196. Springer Berlin / Heidelberg, 2010.

P. A. Pevzner and S. H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. *International Conference On Intelligent Systems For Molecular Biology*, pages 269–278, 2000.

S.L. Phung, A. Bouzerdoum, and D. Chai. Skin segmentation using color pixel classification: Analysis and comparison. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 27: 148–154, January 2005.

I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: The Teiresias algorithm. *Bioinformatics*, 14:55–67, 1998.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2004.

T. Starner and A. Pentland. Real-time American Sign Language recognition from video using Hidden Markov Models. *Computational Imaging And Vision*, 9:227–244, 1997.

T. Starner, J. Weaver, 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.

Yoshiki Tanaka, Kazuhisa Iwamoto, and Kuniaki Uehara. Discovery of time-series motif from multidimensional data based on MDL principle. *Machine Learning*, 58(2-3):269–300, 2005.

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*, 25:1323–1328, October 2003.

C. Vogler and D. Metaxas. Parallel Hidden Markov Models for American Sign Language Recognition. *International Conference On Computer Vision*, 1:116–122, 1999.

C. Vogler and D. Metaxas. A framework of recognizing the simultaneous aspects of American Sign Language. *Computer Vision And Image Understanding*, 81:358–384, 2001.

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*, pages 393–398, 2002.

Q. Wang, X. Chen, L.G. Zhang, C. Wang, and W Gao. Viewpoint invariant sign language recognition. In *Computer Vision And Image Understanding*, volume 108, pages 87–97, 2007.

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*, 24: 1061–1074, 2002.

R. Yang, S. Sarkar, and B. Loeding. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 32(3):462–477, March 2010.