Automated Extraction of Signs from Continuous Sign Language Sentences using Iterated Conditional Modes

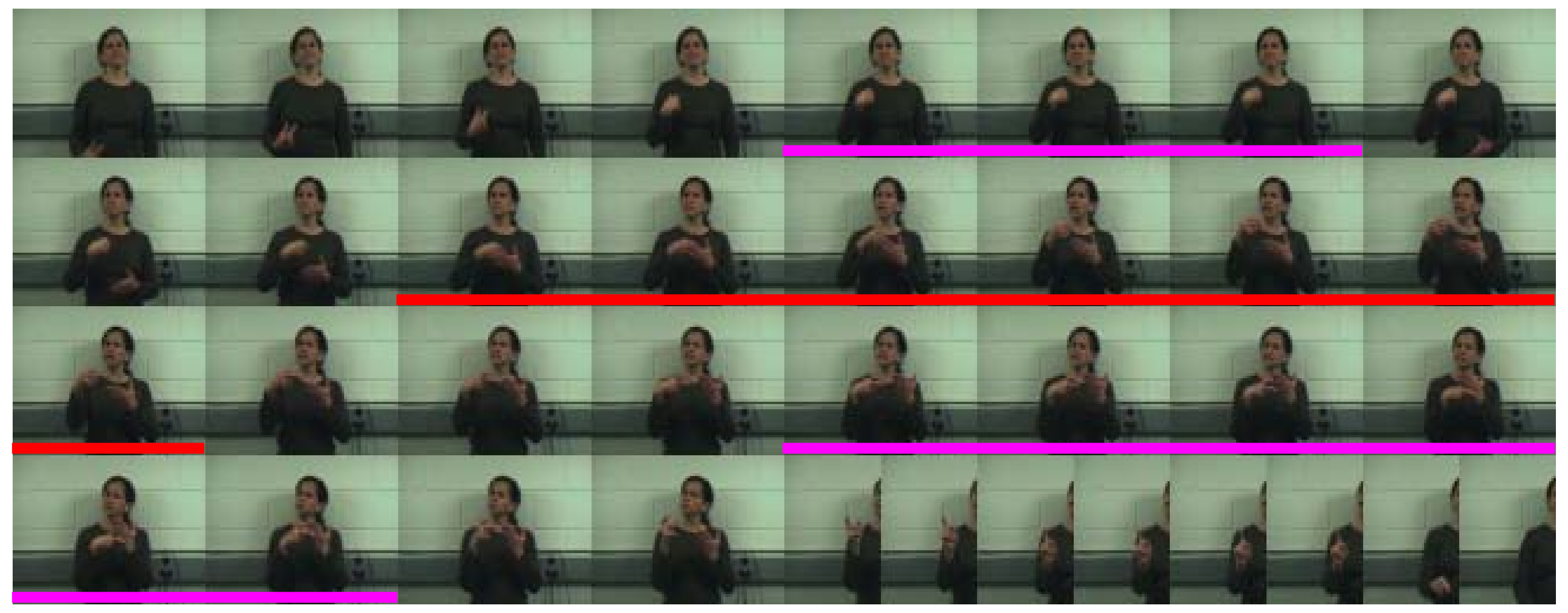
Sunita Nayak[1], Sudeep Sarkar[2] and Barbara Loeding[3]

[1]Photometria Inc., San Diego, CA 92122, USA, [2]Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA, [3]Dept. of Special Education, University of South Florida, Lakeland, FL 33803, USA

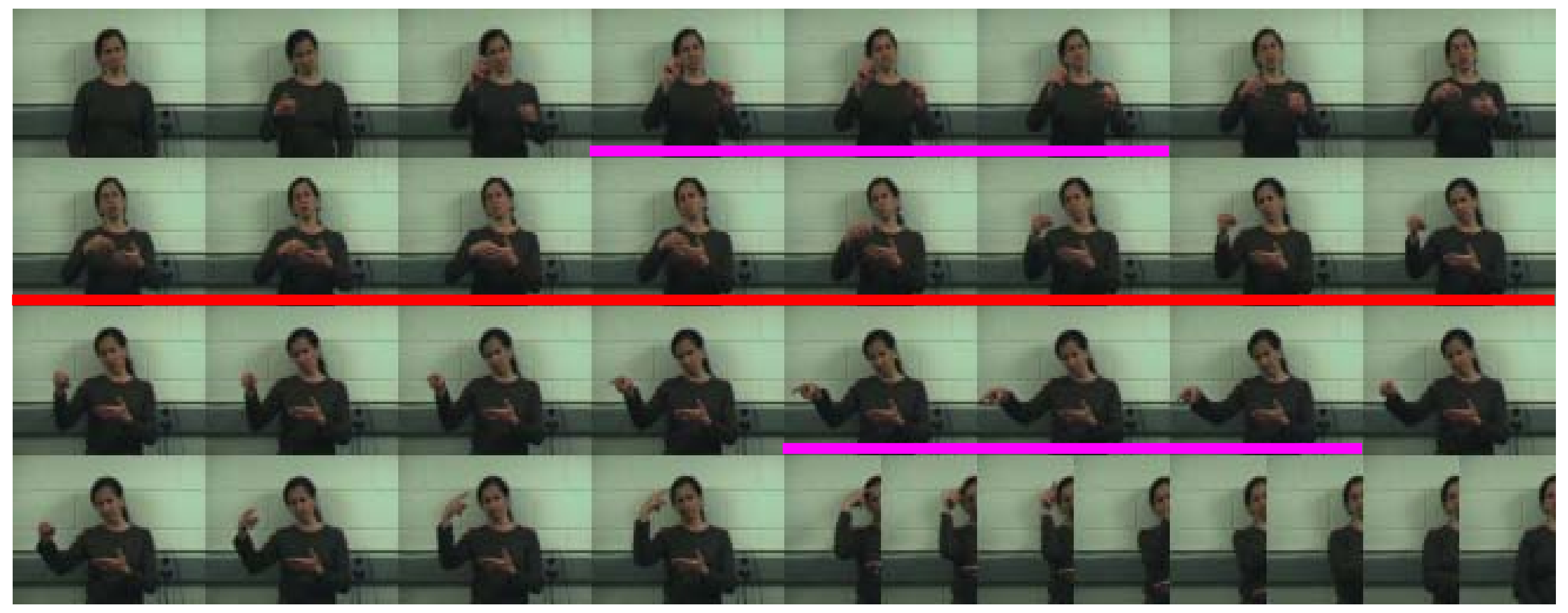
Problem Statement

Can you learn a sign model given multiple sentences containing that sign? The model should be robust to movement epenthesis.

In the following two sentences, the target word to learn is *BUY*. The ground truth frames representing the sign 'BUY' are marked in red, and neighboring signs are marked in magenta. The frames in between indicate movement epenthesis i.e. the transition between signs.

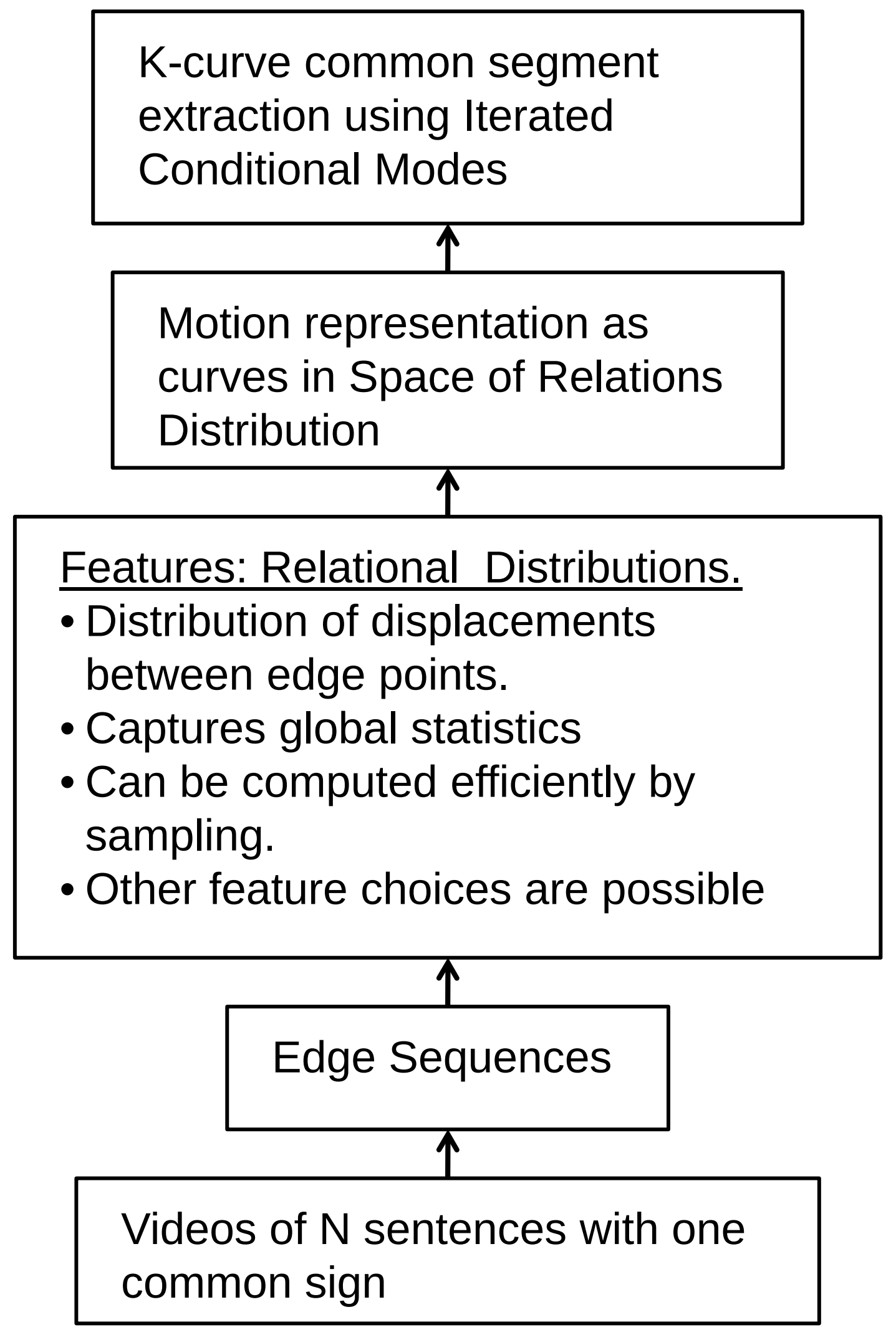


fs- I BUY TICKET WHERE



fs- YOU CAN BUY THIS FOR HER

Algorithm Overview



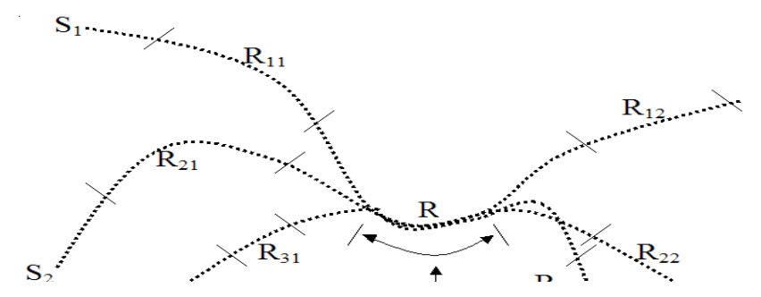
K-curve common segment extraction using Iterated Conditional Modes

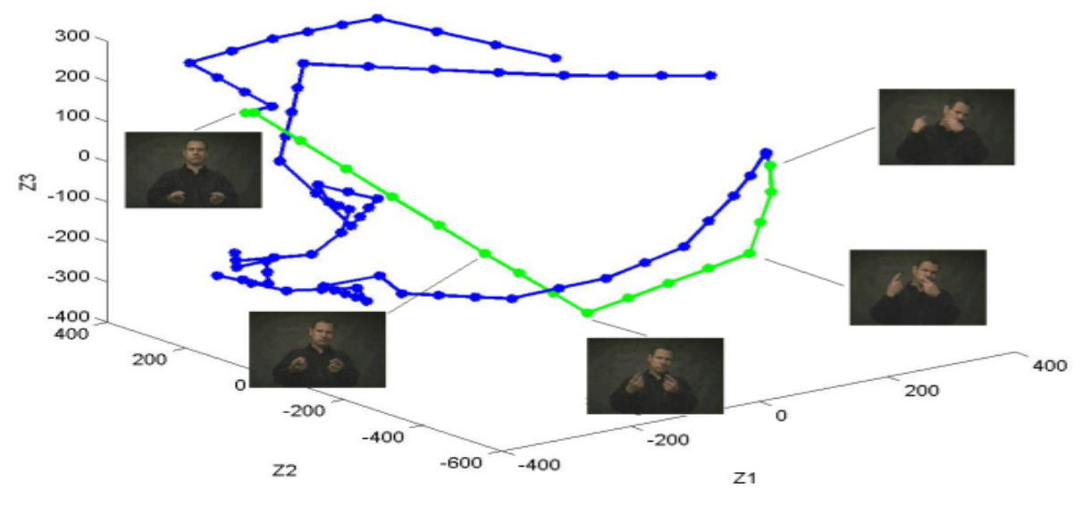
Motion representation as curves in Space of Relations Distribution

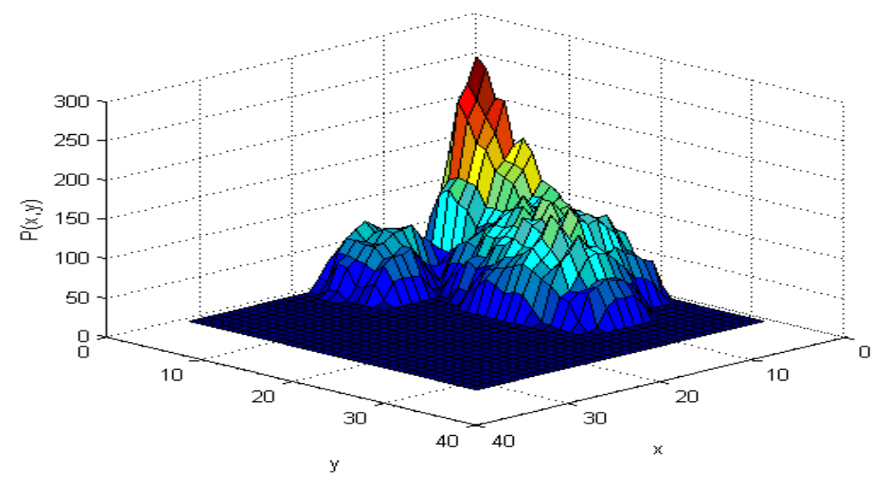
Features: Relational Distributions.
- Distribution of displacements between edge points.
- Captures global statistics
- Can be computed efficiently by sampling.
- Other feature choices are possible

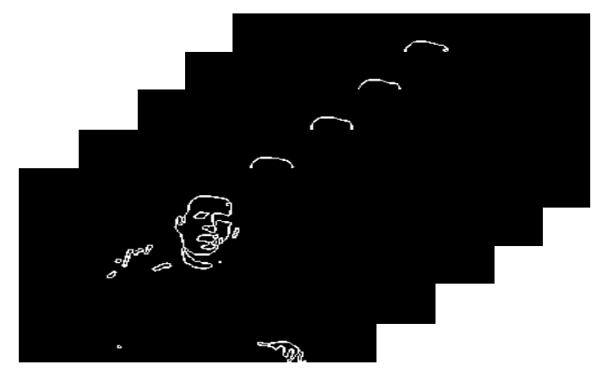
Edge Sequences

Videos of N sentences with one common sign





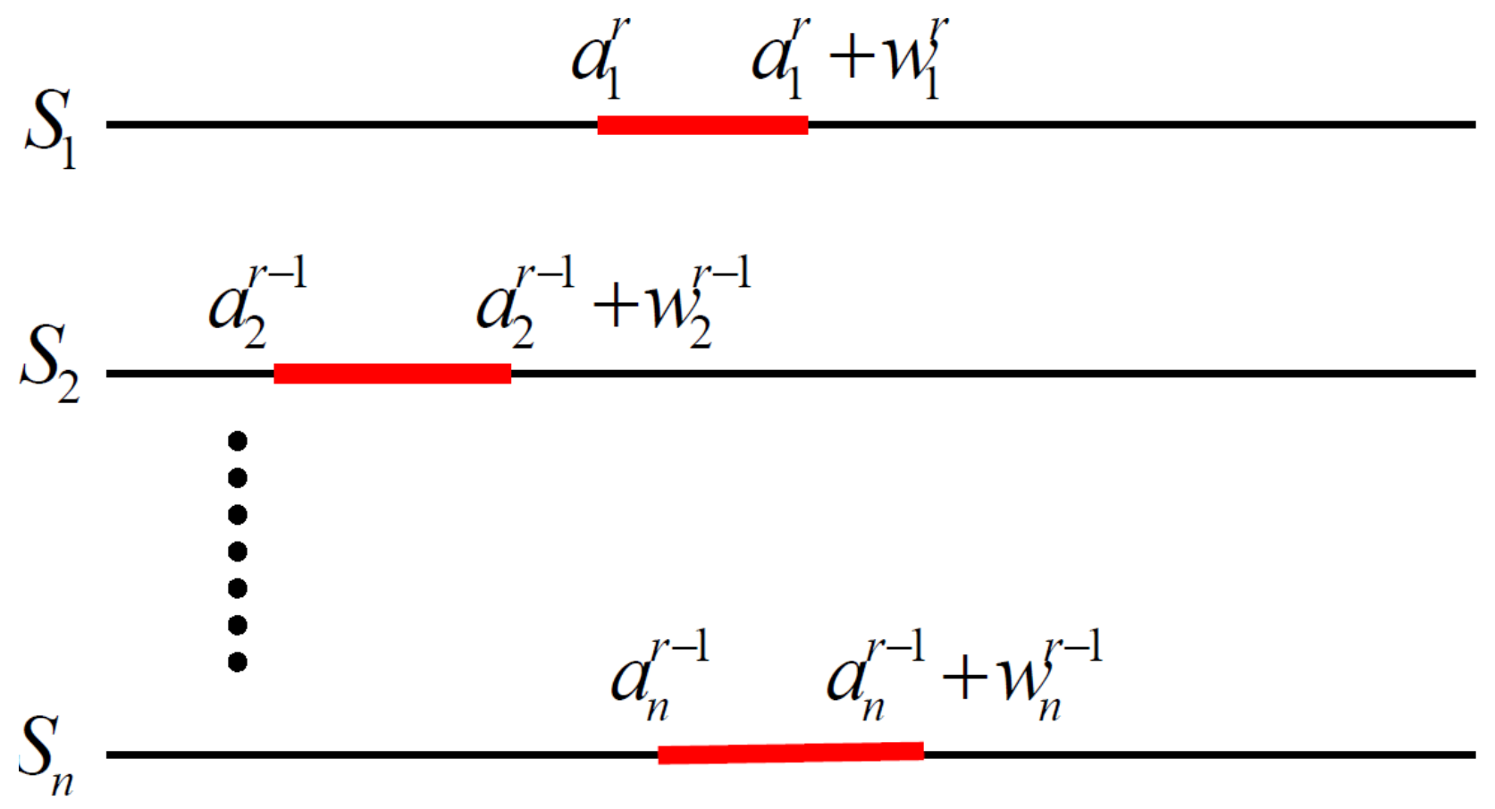




Problem Formulation

Maximize the joint probability over the substring parameters (θ) → Set of parameters (θ) defining a set of substrings (red segments) of the set of input sentences

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



Joint probability over the parameters

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

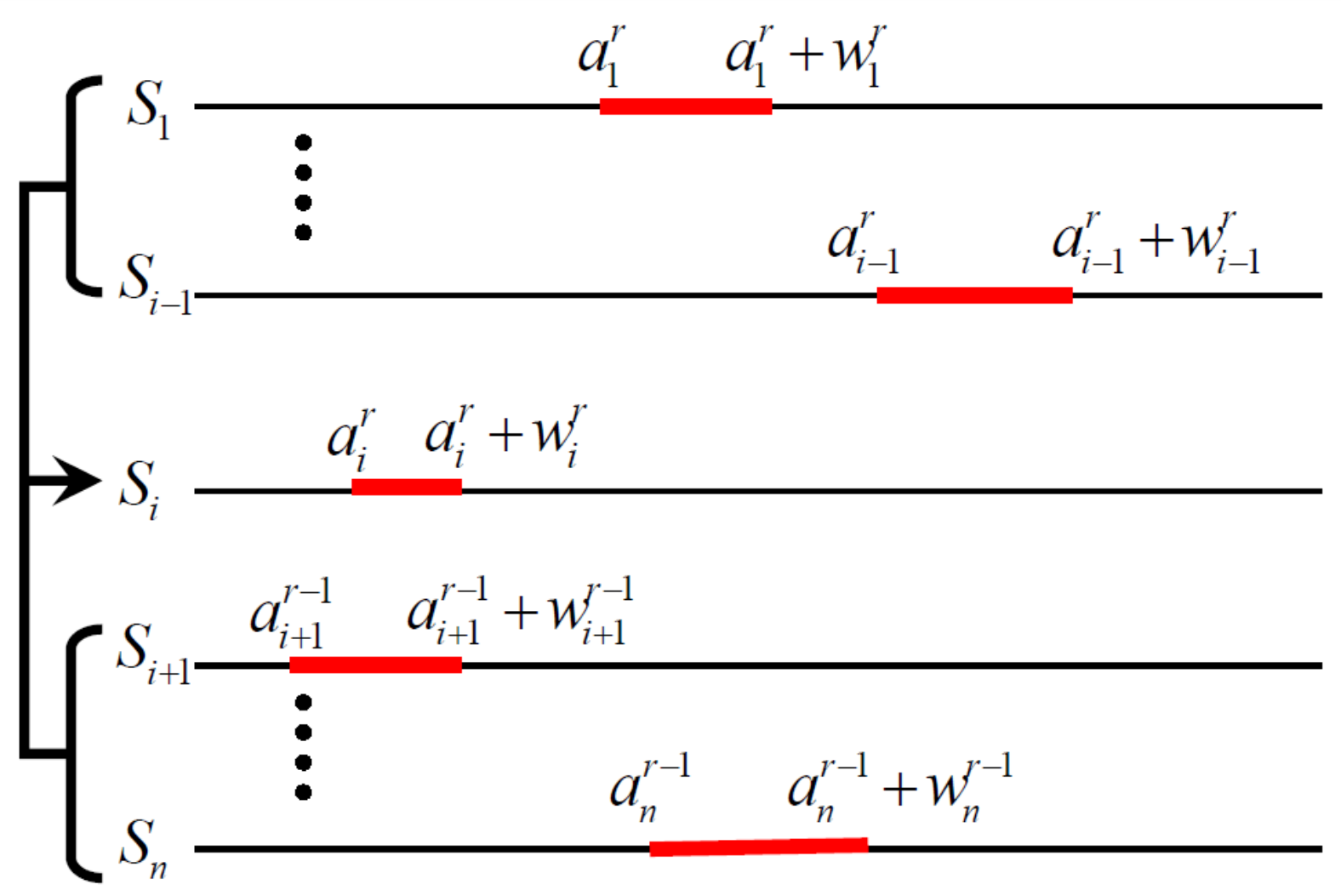
$$g(\theta) = \exp\left(-\beta \sum_{i=1}^{n} \sum_{j=1}^{n} d(\vec{s}_{a_i}^{\,w_i}, \vec{s}_{a_j}^{\,w_j})\right)$$

→ Warp distance between substrings

Conditional probability over the parameters

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

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



Problem Solution: Iterated Conditional Modes (ICM)

```
comment: Chooses $(a_1, w_1, \cdots, a_n, w_n)$ 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$(\{a_1, w_1, \cdots, a_n, w_n\}) == 0$
```

Some Extracted Signemes

(a) Buy

(b) Cant

(c) Move

(d) Passport

(e) Security

(f) Ticket

(g) Table

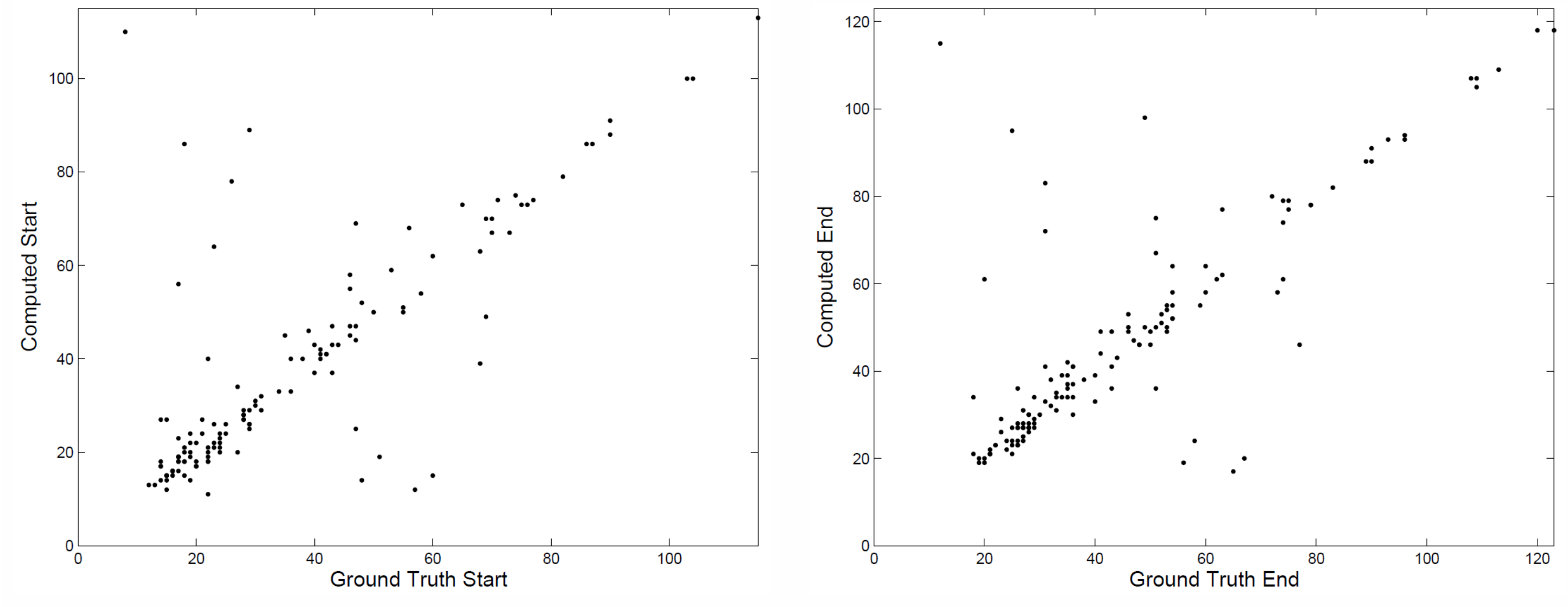
(h) Future

(i) Time

(j) Depart

http://marathon.csee.usf.edu/ASL/SignemeExtraction.html

Computed vs. Ground Truth Start and End Points



(a) Video Start Point Estimation

(b) Video End Point Estimation

Conclusions

- Extracted sign segments (signemes) match ground truth.
- No need for sign glosses.
- Can be used for automated generation of training data
- Demonstrated it on audio data too.

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