Detecting Coarticulation in Sign Language using Conditional Random Fields

Ruiduo Yang and Sudeep Sarkar
Computer Science and Engineering Department
University of South Florida
4202 E. Fowler Ave. Tampa, FL 33620
{ryang,sarkar}@csee.usf.edu

Abstract

Coarticulation is one of the important factors that makes automatic sign language recognition a hard problem. Unlike in speech recognition, coarticulation effects in sign languages are over longer durations and simultaneously impact different aspects of the sign such as the hand shape, position, and movement. Due to this effect, the appearance of a sign, especially at the beginning and at the end, can be significantly different under different sentence contexts, which makes the recognition of signs in sentences hard. We advocate a two-step approach, where in the first step one segments the individual signs in a sentence and in the next step one recognizes the signs. In this work, we show how the first step, i.e. sign segmentation, can be performed effectively by using the conditional random fields (CRF) to directly detect the coarticulation points. The CRF approach does not make conditional independence assumptions about the observations and can be trained with fewer samples than Hidden Markov Models (HMMs). We validate our approach by demonstrating performance with American Sign Language (ASL) sentence level data and show that the CRF approach is 85% accurate in segmenting signs compared to 60% for the HMM approach at 0.1 false alarm rate.

1. Introduction

As recent reviews [4, 1] show, there is increased emphasis on recognition of continuous signs, i.e. recognition of signs embedded in sentences, rather than isolated sign language recognition. Most approaches use the Hidden Markov Model (HMM) at the sign or phoneme model. For example, Vogler *et al.* model American Sign Language (ASL) using context-dependent signs [9] and later using phonemes [8]. However, context dependent sign modeling requires more training data than isolated signs, usually orders of magnitude larger. Phoneme-based approach can be scalable because it works with common parts between signs, but the concept of phonemes is not yet linguistically well established in ASL. We advocate a two step approach for the recognition of continuous signs. In the first step, one segments the sentence into individual signs and then in the second step, one recognizes the signs, exploiting possibly grammar models. In this paper, we show we can effectively perform the first step, i.e. segmentation of signs in sentences.

Each frame in a sequence is represented using a motion snapshot based representation, capturing short term movement over few frames. To reduce the combinatorics of the segmentation process, we identify possible segmentation points to be those frames in the sentence that are most different from each other and different from sign frames; the identified frames are referred to as the key frames. We discuss how this is performed later in the paper. These key frames are then labeled using a Conditional random fields (CRF) model defined over signs and coarticulations.

Unlike a Hidden Markov Model (HMM) that is a generative model based on likelihoods and priors, the CRF is a discriminative model that directly computes the posterior label probabilities [3, 2]. HMM requires strict independence assumptions across multivariate features and conditional independence between observations, given the states, which is generally violated in sign languages, i.e. observations are not only dependent on the state but also on the past observations. The other disadvantage of using HMM is that the estimation of the observation parameters requires a large amount of training data.

Fig 1 depicts the essential differences between HMM and CRFs. Fig 1(a) shows the structure of HMM, where the directed links indicate the conditional likelihoods given the state and the state transition probabilities. The CRF model is shown in Fig 1(b); it is an undirected graph, allowing for arbitrary dependence among the nodes. Each given observation is associated with a state label. Two consecutive state labels as well as the state-observation pairs are jointly modeled. Our key frame based CRF approach works essentially similar to frame based CRF, except that the states are as-

signed to each key frame. Each key frame captures points in a sentence with distinctive short term motion.

CRF has been used successfully by [3] to label and segment text sequential data and used to segment images [2]. Recently Sminchisescu *et al.* [6] used CRF to recognize whole body human movement, not sign language. They reported CRF to outperform HMM, especially under a large context dependent situations. However, the movements considered by them are basically consecutive performances of single gestures with no "coarticulation" effects. Also unlike their approach, we do not use CRF for recognition, but rather for segmentation.

The rest of the paper is organized as follows. Section 2 describes the process to represent and identify key frames. The coarticulation segmentation algorithm is described in Section 3. Experimental results are shown in Section 4. We conclude with Section 5.

2. Representation and key frame extraction

We look at the problem of sign language recognition problem, without the use of any special gloves or markers – just plain one 2D color image sequence. Since motion is a primary manifestation of sign languages, for each instance of time, we first extract a representation that capture the spatio-temporal characteristics over a short duration temporal window. We term this representation as the motion snap-shot representation. Using this representation, we identify most "distinctive" frames in the sequences, i.e. frames that are most different from frames in signs in the database and other frames in the same sequence. These are the key frames, which are then labeled as either being a coarticulation key frame or not. The selection of a reduced set of key frames to label significantly reduces the combinatorics of the subsequent labeling process using CRFs.

2.1. Motion Snapshot Representation

Given the difficulty of reliably extracting and tracking hands in 2D images under varying lighting conditions, we opt for a representation that does not require the extraction or long term tracking of hands. One can observe that motion and relative spatial relationship between motion regions are two important manifestation of sign language as captured by 2D image sequences. We capture these two aspects by first computing short term motion tracks and then representing the spatial relationship between these tracks using a compact statistical representation called the Space of Probability Functions (SoPF).

To capture short term motion at each frame, we considered motion of salient corner points, as detected and tracked by the KLT (Kanade-Lucas-Tomasi) method [7]. However, instead of constructing trajectories over long sequence of frames, which results in lot of errors, we consider tracks only over three frames at a times, one before and one after the frame under consideration. We refer to these collection of short term tracks as motion snapshot. One such example is shown in Fig. 2. We then capture the spatial structure in this motion snapshot by considering the distribution of the horizontal and vertical distances between pairs of points on this trajectory; we compute the joint relational histogram of the displacement between all pairs of coordinates on the motion trajectory. We then represent these relational histograms, normalized to sum to one, as points in a space of probability functions (SoPF), like that used in [5]. The SoPF is constructed by performing principal component analysis of these relational histograms from a training set of images. The Euclidean distances between points in the SoPF is proportional to the Bhattacharya distance between the relational histograms. Thus for each frame in the sequence, we have a feature vector that are the coordinates of the motion snapshot representation in the SoPF space. A sign or sentence is then abstracted as a sequence of feature vectors $S = < \mathbf{s_1}, \cdots, \mathbf{s_n} >$.

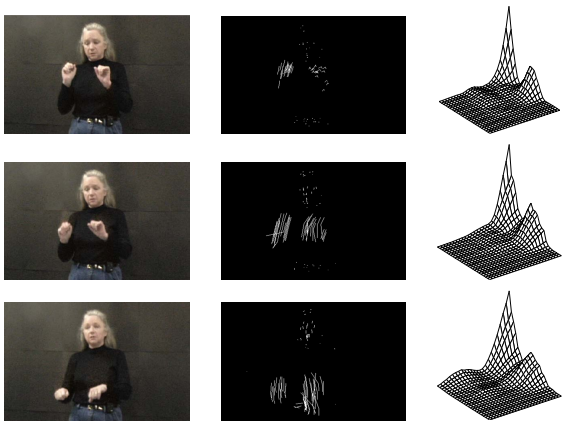
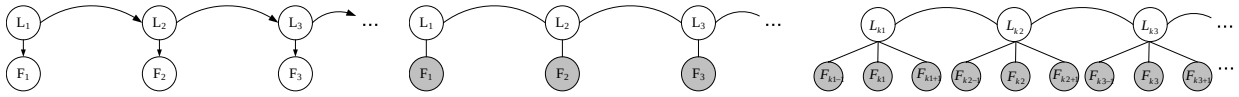


Figure 2. Left column: Three frames from the ASL sign "CAN". Middle column: Motion snapshot corresponding to each frame on the left. Right column: The corresponding relational histogram between points on the motion snapshot trajectory.

2.2. Identifying Key Frames

To segment a sentence into signs, we need not label all frames in the coarticulation between signs; labeling just one frame for each coarticulation would suffice to partition the sentence into individual signs. This helps us control the combinatorics of labeling step. So prior to labeling, we form these set of candidate frames by considering the most



(a) HMM (b) CRF (c) CRF with Key frame

Figure 1. (a) HMM defined with state and observation pairs using directed links. Multiple consecutive observations in any given sequence can be mapped onto the same state. (b) The CRF model uses pairwise probabilities over states and observations for each time instant. Each observation is associated with a state label. (c) Key frame CRF.

distinctive frames, which we terms as the key frames. The idea of key frames have been used in video database summarization [10], but not for sign language analysis. The specific definition of key frames is also different from the video summarization contexts in that we consider their distinctiveness with respect to a modelbase of signs. We define key frames to be the frames that are most different from each other in the same sequence *and* different from frames in a model database of signs of interest.

The model dataset of signs are continuous signs, segmented manually by excluding the coarticulation frames, from a training set of ASL sentences. Let us denote this training set of signs as $T = \{T^1, \cdots, T^k\}$ where T^k is the k-th model sign. Each sign is represented as sequence of motion snapshot distribution coordinates corresponding to its constituent frames, as discussed earlier, i.e. $T^i = < \mathbf{t}^{\mathbf{i}}_{\mathbf{1}}, \cdots, \mathbf{t}^{\mathbf{i}}_{\mathbf{n_i}} >$. We represent T also base on individual frames as $T = < \mathbf{t_1}, \cdots, \mathbf{t_N} >$, where N is the total number of frames in the sign modelbase. Let the given sentence that needs to be segmented be represented by $S = < \mathbf{s_1}, \cdots, \mathbf{s_n} >$ with n frames. We denote the choice of the keyframes by a 0-1 vector, $\mathbf{x}$, whose each component, $\mathbf{x}(i)$, correspond to the frames in S and 1 indicates that it is a key frame and 0 indicates otherwise. The distinctive nature of the key frames is defined based on two kinds of distances: the inter-key frame distances and the model set distances. We select the set of key frames K by maximizing the sum of the distances to other key frames and the average distances from the model set. In other words, the key frames are those frames that are most different from other key frame in the sentence and are also most different from frames in the sets of signs to be recognized and hence are most likely to be part coarticulations. Mathematically, this optimization criterion can be expressed as:

$$\Xi = \sum_{i \neq j} \mathbf{x}(i) d(\mathbf{s_i}, \mathbf{s_j}) \mathbf{x}(j) + \sum_i \mathbf{x}(i) \left(\frac{1}{N} \sum_{k=1}^{N} d(\mathbf{s_i}, \mathbf{t_k}) \right) \mathbf{x}(i) \tag{1}$$

where the function $d(\mathbf{z}, \mathbf{y})$ represents the Euclidean distance between the snapshot based feature coordinates, $\mathbf{z}$ and $\mathbf{y}$. The first term of Ξ above represents the distance between the selected key frames, represented by $\mathbf{x}$ and the second term represents the distance of the key frames to the sign model set. The final set of keyframes are given by $\mathbf{x}$ that maximizes Ξ, subject to the constraint that $\sum \mathbf{x}(i) = m$, where m is the number of key frames. This is an NP-hard problem, hence we consider its corresponding continuous valued version, where the components of $\mathbf{x}$ are relaxed to be any continuous value between 0 and 1. Each of component of this continuous valued vector, denoted here by $\mathbf{y}$, can be taken to represent as a soft confidence measure of the corresponding frame being a key frame. The associated constraint is modified to $\mathbf{y}^T \mathbf{y} = 1$. We then construct the approximate solution to the discrete version of the problem from the solution to the continuous version. The continuous version of the problem is given by

$$\arg\max_{\mathbf{y}} \mathbf{y}^T \mathbf{D} \mathbf{y} \text{ subject to } \mathbf{y}^T \mathbf{y} = 1 \tag{2}$$

where the $\mathbf{D}$ is a matrix diagonal entries are given by $\mathbf{D}(i,i) = \frac{1}{N} \sum_{k=1}^{N} d(\mathbf{s_i}, \mathbf{t_k})$ and the off diagonal entries are given by $\mathbf{D}(i,j) = d(\mathbf{s_i}, \mathbf{s_j})$. It is well known in linear algebra (Raleigh Ritz theorem) that the solution is given by the maximum eigenvector of the matrix $\mathbf{D}$. Given in optimal solution to the continuous problem, we construct an approximate solution to our discrete problem by selecting the frames corresponding to the local maxima of the components of the eigenvector whose indices actually correspond to time.

3. Conditional Random Fields (CRF)

We define a conditional random fields (CRF) [3] over the set of key frames, detected in a sentence, as a linear chain, where the observations are the key frame features $K = < \mathbf{k_1}, \cdots, \mathbf{k_m} >$ and the corresponding labels are $L = < L_1, \cdots, L_m >$, with $L_i \in \{COAR, SIGN\}$ – the possible labels: coarticulation or sign. $< L, K >$ is a conditional random fields when L, globally conditioned on K, obeys the Markov rule, as modeled by a linear graph.

$$P(L_i | K, L - \{L_i\}) = P(L_i | K, N(L_i)) \tag{3}$$

where $N(L_i)$ are the neighbors of L_i. Let us consider the linear chain graph G constructed by $< K,L >$, let $C(K,L)$ denote the set of cliques in G. By the fundamental theorem of random fields, the probability of a label sequence L given the observation sequence K can be represented as:

$$P(L|K) \propto exp^{\sum_{c \in C(K,L)} f_\theta F_\theta(c,K)} \quad (4)$$

where $\{F_\theta\}$ are the feature functions defined over all the cliques and $f = \{f_\theta\}$ are the parameters set, weighted the corresponding feature functions. In a linear chain graph, the cliques can be the pair of adjacent labels $< L_{t-1}, L_t >$ and the label-observation pair $< L_t, K_t >$. For adjacent labels $< L_{t-1}, L_t >$, we consider the feature function $\{F_\theta\}$ as a indicator function:

$$F_\theta(< L_{t-1}, L_t >, K) = \begin{cases} 0, if < L_{t-1}, L_t > \text{ does not correspond to this } \theta \\ 1, if < L_{t-1}, L_t > \text{ does correspond to this } \theta \end{cases} \quad (5)$$

For label-observation pair $< L_t, K_t >$, we consider the feature function $\{F_\theta\}$ as a linear function:

$$F_\theta(< L_t, K_t >, K) = K_t^i, \; K_t^i \text{ is the } i_{th} \text{ element of the feature vector } K_t, K_t^i \text{ corresponds to this } \theta \quad (6)$$

For example, consider the start of a sentence when the signer usually will lift the hands up. Let us denote the key frame of this action as K_0 and the corresponding label as *COAR*, then a penalty of assigning *COAR* to K_0 will be selected and then weighted by the corresponding f_θ. For a transition feature, similarly, suppose we have 2 adjacent key frames K_0 and K_1 that are labeled both as *COAR*, then a penalty of assigning $COAR - COAR$ to an edge will be selected and weighted by corresponding f_θ.

Note, unlike HMM where strict independence does not allow us to represent the relationship between the labels and observations across time, in CRF this can be represented over an arbitrary temporal window w as $< L_t, K_{t \pm w} >$, allowing us to easily embed context dependence and hence is flexible.

This CRF is trained using the model dataset sequences. We first detect the key frames for each of the model dataset sequences that are then manually labeled with COAR and SIGN labels. This forms the training set: $< L_d, K_d >, d \in 1,2,...,N_s$ where N_s is the size of the training database. The parameter set f is found by maximizing the log-likelihood:

$$\begin{aligned} L_f &= \sum_{d=1}^{N_s} log(P(L_d|K_d)) \\ &= \sum_{d=1}^{N_s} \sum_{c \in C(K,L)} (f_\theta F_\theta(c,K)) - log(Z_\theta(K)) \end{aligned} \quad (7)$$

where Z_θ is the normalization factor depend on the observation sequences. We use a gradient based approach with a random start point to seek the maximal point of 7. For inference with the trained model we use belief propagation (BP) over a chain structure [3].

4. Experimental Results

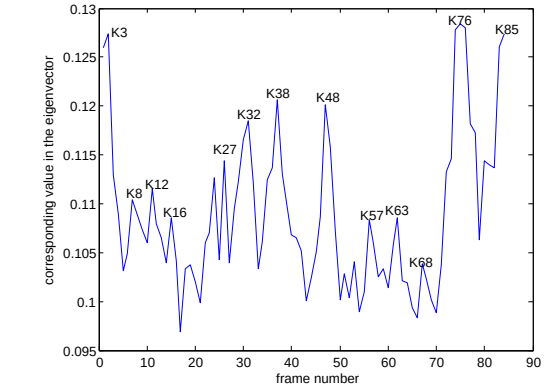
The dataset consists of 39 different signs, articulated in 25 different sentences and each of them is performed 5 times to capture some amount of variation across time. Some variation is also introduced by signing the same sentences differently, for example, the sentence 'if the plane is delayed, I'll be mad' in English will be signed as 'AIRPLANE POSTPONE AGAIN, MAD I' as well as 'AIRPLANE AGAIN POSTPONE, MAD I'. The frame rate is 30 fps and the resolution is 460 by 290.

These sentences are performed by the same signer, under a uniform background, the signer is wearing dark clothes. Although motion estimation could be hard under such a less textured scene, we are not trying to obtain the whole motion field, instead we only estimate the motion of corner points.

Four of the five instances of each sentence form the model data set of signs (training set) and the fifth sequence is the test sequence. One CRF is used to model all the sentences in the training set. For each frame, the KLT corner detection method usually generates 50 to 100 feature points. The relational histograms are constructed with 32×32 bins. To select key frames we detect the local maxima in the associated maximum eigenvector over sliding window of 7 frames. Fig 3 shows us the result of key frame detection. Note that we can have multiple key frames for each sign or coarticulation part. For example, in Fig 3, we have 2 key frames selected from the starting portion and end portion of the sign "GATE". Fig 4 shows us the plot of the detection rate versus the false alarm rate (ROC) for marked coarticulations. The total number of keyframes are around 500. We show the curves for the CRF with different spatial contexts, i.e. with window sizes 1, 3, and 5. As a baseline we compare these with a standard Hidden Markov Model (HMM) based formulation results. We note that the CRF based labeling significantly out performs the HMM based one.

5. Conclusions

We used conditional random fields (CRF) formulation along with the concept of key frames, capturing frames with the distinctive short term motion, to detect and label coarticulations points in a sign language sentence. The CRF has the advantage of directly modeling the posterior and can allow arbitrary dependence between the states and observations, which is desired for labeling a sequence with highly related context such as that exists in an ASL sentence. We experimental found that the CRF based approach significantly out performs an HMM based one. The detected coarticulations are then used to segment a continuous ASL sentence into isolated sign to ease the combinatorics of recognition of signs in continuous sentences.



(a)

(b)

Figure 3. The top figure shows the plot of the components of first eigenvector for a sentence, where key frame is detected by selecting the local maxima. Detected key frames are marked as either as sign or coarticulation

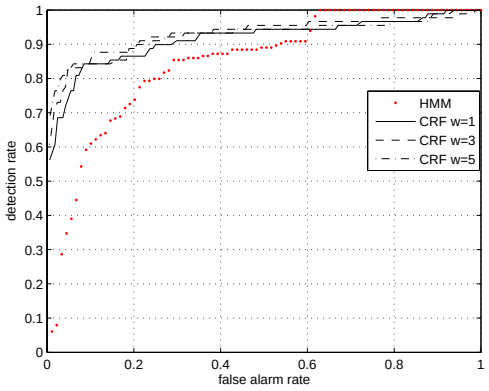


Figure 4. The ROC curve for detecting coarticulation using HMM and CRFs.

6 Acknowledgment

This work was supported in part by the National Science Foundation under grant IIS 0312993. Any Opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

References

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

[2] S. Kumar and M. Hebert. Man-made structure detection in natural images using a causal multiscale random field. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[3] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning.*, 2001.

[4] B. L. Loeding, S. Sarkar, A. Parashar, and A. I. Karshmer. Progress in automated computer recognition of sign language. In *ICCHP*, volume 3118, pages 1079–1087, 2004.

[5] I. Robledo and S. Sarkar. Representation of the evolution of feature relationship statistics: Human gait-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1323–1328, Oct 2003.

[6] C. Sminchisescu, A. Kanaujia, Z. Li, and D. N. Metaxas. Conditional random fields for contextual human motion recognition. In *10th IEEE International Conference on Computer Vision*, pages 1808–1815, 2005.

[7] C. Tomasi. and T. Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University Technical Report CMU-CS-91-132,, April 1991.

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

[9] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *International Conference on Computer Vision*, page 1998, 363–369.

[10] H. Zhong, M. Visontai, and J. Shi. Detecting unusual activity in video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*,, volume 2, pages 819–826, 2004.