Modeling Signs Using Functional Data Analysis

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

Kuntal Sengupta
AuthenTec Inc
709 S. Harbor City Blvd
Melbourne FL 32901, USA
kuntal.sengupta@authentec.com

Abstract

We present a functional data analysis (FDA) based method to statistically model continuous signs of the American Sign Language (ASL) for use in the recognition of signs in continuous sentences. We build models in the Space of Probability Functions (SoPF) that captures the evolution of the relationships among the low-level features (e.g. edge pixels) in each frame. The distribution (histogram) of the horizontal and vertical displacements between all pairs of edge pixels in an image frame forms the relational distributions. We represent these sequence of relational distributions, corresponding to the sequence of image frames in a sign, as a sequence of points in a multi-dimensional space, capturing the salient variations in these relational distributions over time; we call this space the SoPF. Each sign model consists of a mean sign function and covariance functions, capturing the variability of each sign in the training set. We use functional data analysis to arrive at this model. Recognition and sign localization is performed by correlating this statistical model with any given sentence. We also present a method to infer and learn sign models, in an unsupervised manner, from sentence samples containing the sign; there is no need for manual intervention.

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.

Previous works in sign language have been mostly in the recognition of static gestures, e.g. [2, 21, 13] and isolated signs, e.g. [19]. Yeasin and Chaudhuri [20] had worked on dynamic hand gestures. Bobick and Wilson [1] had proposed a state-based approach to model gestures. Starner and Pentland [11] were the first to seriously consider *continuous* sign recognition. Using Hidden Markov Model (HMM) based representations, they achieved near perfect recognition with sentences of fixed structure, i.e. containing personal pronoun, verb, noun, adjective, personal pronoun in that order. Vogler and Metaxas [15, 16, 17] 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 wide use of HMM is also seen in other sign language recognizers.

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 [15, 16, 17] have used 3D magnetic tracking system; Starner and Pentland [11] have used colored gloves while Ma *et.al.* [5, 18] have used Cybergloves. In this paper, we restrict ourselves to plain color images, without the use of any augmenting wearable devices.

There are two kinds of information that can be used for recognition, viz. manual and non-manual. The manual information relates to the hand motion or shape, while the non-manual information relates to the facial expressions, head movement, or torso movement. Here we use the manual information from hand motion. The hand motion is first modeled using relational distributions, which are efficiently represented as points in the Space of Probability functions (SoPF). The points are then transformed into smooth curves that are registered and trained to form a unique model for a sign using Functional Data Analysis.

2. Data Set

A vital component in ASL recognition research is the data set used in the study. The largest corpus used in ASL recognition contains a vocabulary of around 50 signs, embedded in approximately 500 sentences [15, 16, 17]. Only recently has there been a concerted effort in systematically constructing a common ASL corpus for public dissemination. At Boston University, Neidle *et al.* [6] have created such a dataset using SignStream, which is a system for linguistic annotation, storage, and retrieval of ASL and other forms of gestural communication. This dataset also had no wearable aids, but the video was sampled too coarsely. On an average there were only 5.8 frames per sign. So, we had to do our own data collection.

Setting the realistic long term goal of automated ASL recognition, but in a constrained domain, we selected the sentences that would be used while communicating with deaf people at airports. Data was collected and ground truthed by an ASL interpreter. A color video camera was used. The background was kept plain. The dataset has 39 distinct signs forming 25 sentences. There are 10 to 12 instances of each of the sentences. The details of this data is available in [7].

3. Relational Distributions and Space of Probability Functions

In most of the previous works in *continuous* ASL, detection and tracking of hand have been simplified using colored gloves [12] or magnetic markers [15]. Even other sign language recognizers have used colored gloves or data gloves. Only recently there has been effort to extract information and to track directly from color images, without the use of special devices [19], but it has only been used for *isolated* sign recognition. As we shall see, our representation does not require tracking of hands. We would like these representations to be somewhat robust to low-level errors. We use the Canny edge pixels of each video frame as the low-level primitives.

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. Such representations have been successfully used for modeling periodic motion in the context of identification of a person from gait [14] and non-periodic motion in the context of sign recognition [7]. Here, we use it to build statistical models for non-periodic motion in ASL signs. Primitive level statistical distributions, such as orientation histograms, have been used for gesture recognition [3]. However, the only uses of relational histograms that we are aware of are by Huet and Hancock [4], who used it to model line distributions in the context of image database indexing. The novelty of relational distributions lies in that it offers a strategy for incorporating dynamic aspects.

We refer the reader to [14] for the details of the representation. Here we just sketch the essentials. 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*.

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 Probability Functions (SoPF).

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 'CAN'. 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 Probability Functions (SoPF).

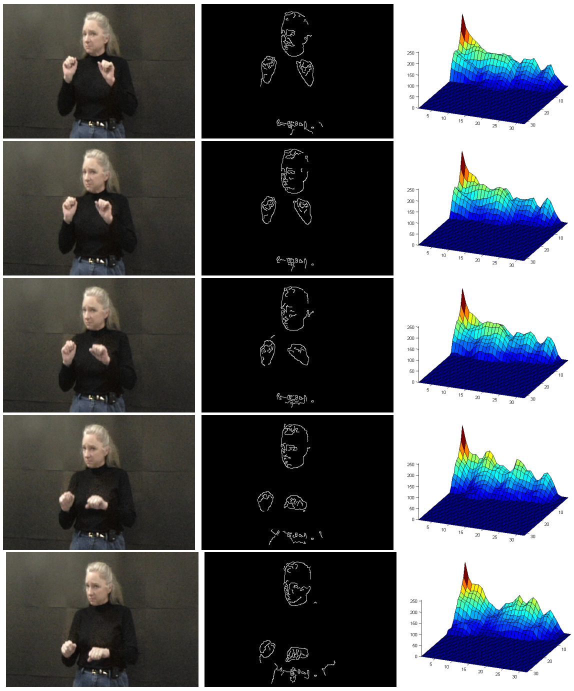


Figure 1. Variations in relational distributions with motion. The left column shows the image frames in the sign 'CAN'. The middle column shows the edge pixels, and the right column shows the relational distributions

We use the square root function so that we arrive at a space where the distances are not arbitrary ones but 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 [14]. Given a set of relational distributions, $\{P(\mathbf{a_k}, t_i) \mid i = 1, ..., T\}$, the SoPF 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. Here, most of the variation is captured by the eigen vectors associated with the top 20 (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 Probability Functions.

4. Supervised Learning using Functional Data Analysis

In the first learning scenario, we use sign samples that are manually segmented from sentences. Each sign sample consists of a sequence of SoPF coordinates. Each coordinate sequence can be looked upon as samples of a smooth curve, or function, in the SoPF space. We arrive at the underlying smooth functional representation for each sign sample using B-spline interpolation [9, 10]. This converts each training sequence into functional data, which are then smoothed and registered to arrive at a single statistical functional model [9, 10]. The specific steps involved are as follows:

1. Each training sequence of SoPF coordinates are time-normalized by linearly interpolated resampling mapped to a fixed time period, which is chosen to be the mean length of all the sequences. For further manipulation, the normalized data is again resampled at a 20 times finer resolution than the original data.

2. All the time-normalized, discretely sampled, sequences are then together turned into a *functional data object*, which represents the underlying sequences of continuous functions in terms of basis functions (B-splines, in our experiments) and the coefficients required to reconstruct the observed data. The functional data of the i^{th} sequence at time t is represented by

$$x_i(t) = \sum_{k=1}^{N} \alpha_{ik}\phi_k(t) \tag{2}$$

where $\phi_1, \phi_2, \phi_3, ..., \phi_N$ are the N basis functions. The coefficients, α_{ik}, determining the above expansion are obtained by minimizing the sum of squares of the difference of the discrete data, d_{ij}, where j=1, 2,..., n represent the n sampling(observation) points, to the corresponding values of x_i, i.e.

$$SSE(d_i, \alpha) = \sum_{j=1}^{n}[d_{ij} - \sum_{k=1}^{N} \alpha_{ik}\phi_k(t_j)]^2 \tag{3}$$

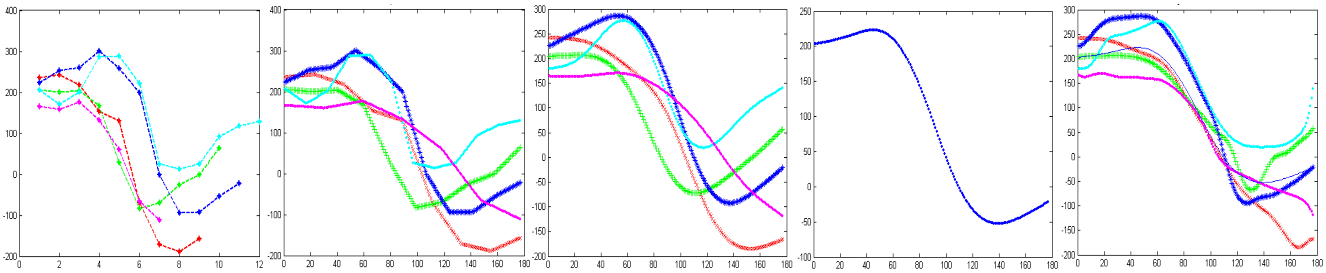
is minimized for the i^{th} sequence of the data. The number of basis function in the B-Spline representation, N, can be determined by $N = N_R + N_D + 4$, where the N_R represents the required resolution, i.e the minimum number of features or events needed to be present in the observation. N_D is the highest order of derivative that needs to be retained in the observation. In our experiments, we have used cubic B-Splines and considered N_R to be 10 and N_D to be 6.

3. The functional data, x_i, represented above is further smoothed by minimizing the following penalty criterion:

$$PSSE = \int [x_i(t) - z_i(t)]^2 dt + \lambda PR(z_i) \tag{4}$$

where z_i is the smoothed form of the data and the last term on the right side of the equation is for *penalizing the roughness* of z_i. $PR(z_i)$ can be defined as the integral of square of the second derivative of z_i, i.e.,

$$PR(z_i) = \int [z_i''(t)]^2 dt \tag{5}$$



(a) (b) (c) (d) (e)

Figure 2. Supervised learning of word models. (a) shows the plots of just the first dimension of SoPF representation w.r.t time, of five instances of the sign 'CAN'. (b) shows the interpolated data. (c) shows the smoothed data, (d) shows the mean of the smoothed data. And (e) shows the registered curves.

The amount of smoothing can be controlled by varying the value of the smoothing parameter λ.

4. The mean, $\mu(t)$, of the smoothed sequences is then computed.

5. Each of the smoothed curves is registered to the mean curve, $\mu(t)$, by estimating a warping function, $h_i(t)$, for each of them so that the registered curves, $r_i(t) = z_i[h_i(t)]$, minimize a global criterion:

$$REGSSE = \sum_{i=1}^{M} \int_T [r_i(t) - \mu(t)]^2 dt \qquad (6)$$

where M is the number of curves and T is interval over which the curves are registered. The process of registration is then done iteratively till a convergence criterion is reached. We use the convergence criterion of 0.01 and an iteration limit of 5 iterations.

6. The covariance is computed at each of the points in the time axis. Mean and covariance functions together form a model of each sign. Both the mean and covariance are computed in the same way as for any other statistical observations, from all the replications of observation at each time instant in the functional data object.

For more detailed discussion on the above processes, we refer the reader to [9, 10]. The code at [8] was used for our experiments.

Figure 2 illustrates the above modeling process using just the first dimension in the SoPF representation of each sign. We conduct the actual analysis in a 20 dimensional SoPF space, however, the figure is sufficient to illustrate how the traces are simultaneously registered and mean representation is extracted. In addition to the mean, we also store the covariances among the 20 dimensions at each time instance, i.e. we also store a multidimensional *covariance function*.

5. Unsupervised Learning of Sign Models

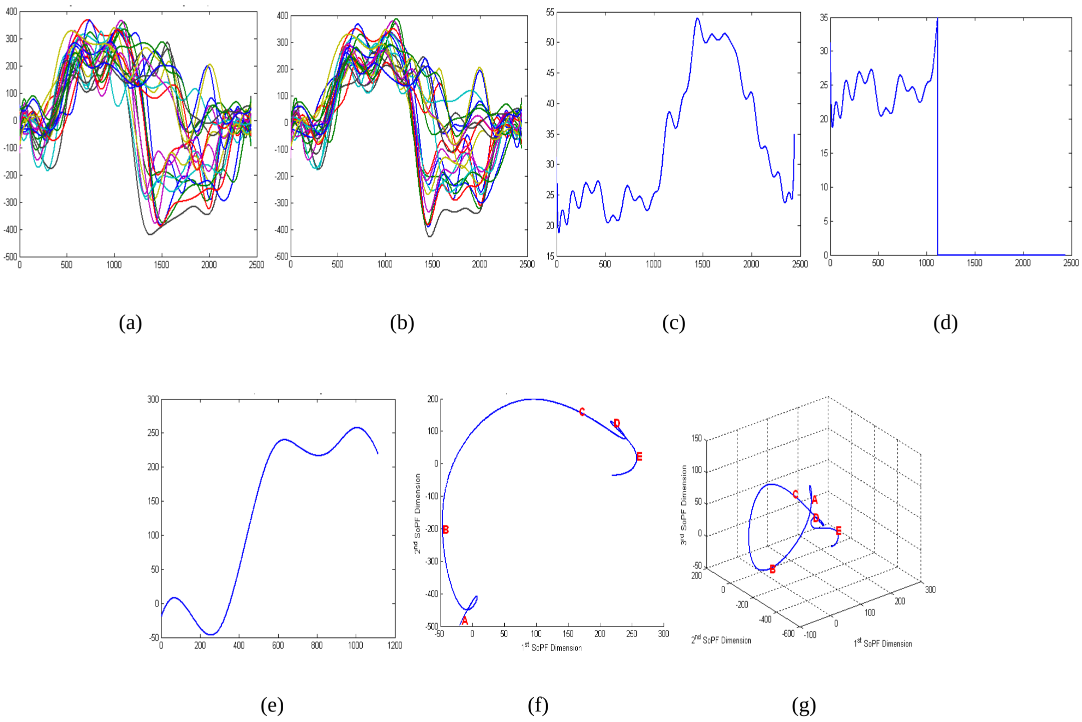
Is it possible to learn a sign model without supervision or requiring manual segmentation of words in the training dataset? In this section, we outline an approach, again based on functional data analysis, for this task. The data consists of many ASL sentences, each consisting of different signs of similar temporal duration, but with the constraint that all contain one common sign. We can automatically generate the model for this common sign. For this, instead of conducting functional data analysis at the word level, we consider the *whole* sentence. The outline of the steps are as follows:

1. We build the mean and covariance function representations from the functional data object for the set of training sentences, using the steps outlined in Section 4. Of course, the registration will be of poor quality since, the sentences contain different signs. However, the registration should be good over the part of sentences containing the common word.

2. The trace of the covariance matrix for each time instant forms a measure of the variability of the registration among the sentences.

3. The portion of the mean and covariance functions over which the variability is low form the model of the common sign. We can additionally use prior knowledge, if available, about the possible location of the common word to prune out residual ambiguities.

The process is illustrated in Figure 3, again using just one of the 20 dimensions of the SoPF representation. The common sign in the 23 sentences training set corresponds to the sign 'IDPAPERS'. We have 12 instances of the sentence 'IDPAPERS_WHERE' and 11 instances of the sentence 'IDPAPERS_TABLE'. We see the variance in the registered curves is low towards the first half of the sentence, when the common sign occurs. The variance is also low towards the end



(a) (b) (c) (d)

(e) (f) (g)

Figure 3. Unsupervised learning of sign models. (a) shows the plot of the first dimension of the SoPF representation for sentences with one common word 'IDPAPERS', smoothed with smoothing parameter(λ)= 0.1. (b) shows the registered curves for the same set of sentences. (c) shows the variation of the standard deviation of the registered sentences. (d) shows the relevant time period indicating the common sign. (e) shows first SoPF dimension of the mean representation of the model formed for the sign 'IDPAPERS'. (f) and (g) respectively show the plots of the first vs. second SoPF dimension, and the first vs. second vs. third SoPF dimension of the mean of the learnt model for the same sign.

because of end-of-sentence coarticulation, i.e. all the sentences end in a common stance. This is easy to filter out based on prior knowledge of the common stance, or by simply ignoring the last few frames.

6. Recognition

The models created above for each of the signs are used for recognizing the signs in *continuous* ASL sentences. At present, we use a simple correlation based recognition process. Any given test sequence is turned into a functional data object, in much the same way as in the model formation process. The relational distribution of each frame of the test sequence is represented as a point in the 20 dimensional SoPF. The test sentence traces a curve in the SoPF space. That curve is interpolated in the same way as in the case of the training data, and then converted to functional data, using the same B-spline basis functions. Then the test data is smoothed to remove irrelevant features. The smoothing parameter is kept same as in the training set, i.e. 0.1.

Now each sign model is matched to the test sentence by correlation. The distance is calculated by summing up the distance of each point of the mean curve of the sign from the test sentence curve, and then normalizing the sum by the sign's length. Note that one of the property of the SoPF is that Euclidean distances in this space correspond to Bhattacharya distance between the corresponding relational distributions [14]. The sign is said to be located at the point of minimum correlational distance. The value of the minimum correlation is a measure of distance of the sign model to the sentence.

7. Actual Recognition Experiments

The data set used for the experiments consists of 16 signs forming 10 sentences, with two to three signs per sentence. The average length of the test sentences was 90 frames. First, we present supervised learning (Section 4) results. Each learnt mean functional data model is correlated with the functional data object constructed from the test sentence. The model signs are sorted based on the minimum correlational distance; the sign with smaller minimum correlational distance is more likely to be present in the sentence. To compute recognition rates, we consider if the correct sign occurs within the top n matches in a n-sign sentence. In this way, the recognition rate found was to be 57%. If we consider the top $n+1$ matches, the recognition rate increases to 69%. We note that these rates are for a very simple recognition strategy; we expect the rates to be higher for better recognition strategies. As the focus of this paper is in modeling, we did not yet explore more complicated strategies.

The correlation based recognition is good at localizing signs in a sentence. For most of the signs in the sentences, the location is found near to the actual position. Signs were located with about 92% accuracy. We define the error rate as the difference of the actual starting frame number of the sign to computed starting frame number, normalized by the total number of frames in the sentence.

For unsupervised modeling, we considered four signs, viz. 'WHERE', 'SUITCASE', 'FINISH' and 'IDPAPERS'. The built models were tested on sentences not used in training. Four out of six possible occurances of the above words in the test data were located with 82% or higher localization accuracy.

8. Conclusions and Future Work

This papers presents a functional approach for supervised and unsupervised modeling of the signs of American Sign Language as smooth curves with variance at each point in the curve, in a multidimensional space. The approach uses plain video data that does not use any wearable aids like data gloves, magnetic trackers etc., as its input. Instead it relies on inter-feature relational distribution in any image frame. We are presently working on automating the thresholds used in the above process of self-learning of signs, and using sentences with common signs at different locations. The use of dynamic time warping while matching the sign and the use of covariance while finding the distance from the sentence, can significantly improve the recognition rate. Also the above approach has to be tried on a dataset having more number of signs and sentences.

Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant No. IIS 0312993.

References

[1] A. Bobick and A. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1325 – 1337, December 1997.

[2] Y. Cui and J. Weng. Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, 78(2):157 – 176, May 2000.

[3] W. Freeman and M. Roth. Orientation histograms for hand and gesture recognition. In *International Workshop on Face and Gesture Recognition*, pages 296 – 301. 1995.

[4] A. Huet and E. Hancock. Line pattern retrieval using relational histograms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(13):1363 – 1370, 1999.

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

[6] C. Neidle, S. Sclaroff, and V. Athisos. A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311 – 320, November 2001.

[7] A. S. Parashar. Representation and interpretation of manual and non-manual information for automated American Sign Language recognition, Master's thesis, Department of Computer Science Engineering, University of South Florida, 2003.

[8] J. Ramsay. Matlab,R and S-PLUS Functions for Functional Data Analysis. ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/ .

[9] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, 1997.

[10] J. Ramsay and B. Silverman. *Applied Functional Data Analysis*. Springer, 2002.

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

[12] T. Starner and A. Pentland. Visual recognition of American Sign Language using hidden Markov models, Master's thesis, MIT, Media Lab., 1995.

[13] J. Triesch and C. von der Malsburg. Robust classification of hand postures against complex backgrounds. In *International Conference on Automatic Face and Gesture Recognition*, pages 170 – 175. 1996.

[14] I. R. Vega. *Motion model based on statistics of feature relations: Human Identification From Gait*. PhD thesis, Department of Computer Science Engineering, University of South Florida, 2002.

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

[16] C. Vogler and D. Metaxas. Parallel hidden Markov models for American Sign Language recognition. In *International Conference on Computer Vision*, pages 116 – 122. 1999.

[17] 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.

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

[19] 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*, 24:168 – 185, August 2002.

[20] M. Yeasin and S. Chaudhuri. Visual understanding of dynamic hand gestures. *Pattern Recognition*, 33(11), 2000.

[21] M. Zhao and F. K. H. Quek. RIEVL: recursive induction learning in hand gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1174 – 1185, 1998.