Evaluation of Threshold Model HMMs and Conditional Random Fields for Recognition of Spatiotemporal Gestures in Sign Language

Daniel Kelly, John Mc Donald, Charles Markham
Computer Science Department
National University of Ireland, Maynooth
dankelly@cs.nuim.ie

Abstract

In this paper we evaluate the performance of Conditional Random Fields (CRF) and Hidden Markov Models when recognizing motion based gestures in sign language. We implement CRF, Hidden CRF and Latent-Dynamic CRF based systems and compare these to a HMM based system when recognizing motion gestures and identifying inter gesture transitions. We implement a extension to the standard HMM model to develop a threshold HMM framework which is specifically designed to identify inter gesture transitions. We evaluate the performance of this system, and the different CRF systems, when recognizing gestures and identifying inter gesture transitions.

1. Introduction

Recognizing gestures which appear in sign language is a challenging problem. Gestures lack a clear categorical structure and similar gestures can happen at various timescales. Another difficulty with recognizing gestures are inter gesture transitions which occur between valid gestures. For example, when performing hand gestures, the hands must move from the end point of the previous gesture to the start point of the next gesture. These inter gesture transition periods are called movement epenthesis [9] and are not part of either of the gesture. As such, an accurate recognition system must be able to distinguish between valid sign segments and movement epenthesis.

Hidden Markov Models (HMM) have been proposed as a solution to dealing with continuous gesture recognition without explicit segmentation. Starner et al. [14] and Bauer and Kraiss [2] model each word or subunit with a HMM and then train the HMMs with data collected from full sentences. A downside to this is that training on full sentence data may result in a loss in valid sign recognition accuracy due to the large variations in the appearance of all the possible movement epenthesis that could occur between two signs.

Wang et al. [18] also use HMMs to recognize continuous signs sequences with 92.8% accuracy, although signs were assumed to end when no hand motion occurred. Assan et al. [1] model the HMMs such that all transitions go through a single state, while Gao et al. [5] create separate HMMs that model the transitions between each unique pair of signs that occur in sequence. Vogler at al. [17] also use an explicit epenthesis modeling system where one HMM is trained for every two valid combinations of signs.

While these works have had promising results in gesture recognition and movement epenthesis detection, the training of such systems involves a large amount of extra data collection, model training and recognition computation due to the extra number of HMMs required to detect movement epenthesis.

More recently, there has been an increasing interest in using Conditional Random Fields (CRF), as an alternative to HMMs, for human gesture recognition. CRFs were first introduced by Lafferty et al [7] as a framework for building probabilistic models to segment and label sequence data. Sminchisescu et al. [13] use CRFs to classify 11 different human motion activities. As an extension to traditional CRFs, Hidden state conditional random field (HCRF) based gesture recognition systems have also been proposed. Wang et al. [19] use a HCRF framework to classify three different head gestures and six different arm gestures. Morency et al [11] expand on the work of Wang et al to develop a Latent-Dynamic Conditional Random Field which combines combine the strengths of CRFs and HCRFs by capturing both extrinsic dynamics and intrinsic sub-structure.

In this paper we discuss our threshold HMM framework which is specifically designed to identify movement epenthesis. We carry out performance evaluations to compare the threshold HMM system to CRF, HCRF, LD-CRF and standard HMM systems when recognizing motion based gestures and identifying movement epenthesis.

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

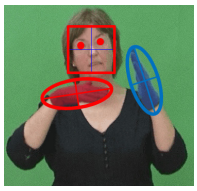


Figure 1. Extracted Features from Image

2. Feature Extraction

The gesture recognition evaluations we carry out in this work uses data extracted from video sequences of sign language sentences being performed by a fluent signer. For completeness, prior to discussing the evaluations we carry out on HMMs and CRFs, we briefly describe the feature tracking techniques implemented.

Tracking of the hands is performed by tracking colored gloves using the Mean Shift algorithm [4]. Face and eye positions are used as features for head movement recognition and also used as hand gesture cues. Face and eye detection is carried out using a cascade of boosted classifiers working with haar-like features proposed by Viola and Jones [15]. A set of public domain classifiers [10], for the face, left eye and right eye, are used in conjunction with the OpenCV implementation of the haar cascade object detection algorithm.

We define the raw features extracted from each image as follows; right hand position (RH_x, RH_y), left hand position (LH_x, LH_y), face position (FC_x, FC_y), face width (FW),left eye position (LE_x, LE_y) and right eye position (RE_x, RE_y).

To represent a gesture sequence such that it can be modeled by the HMM and CRF models, the gesture sequence must be defined as a set of observations. An observation O_t, is defined as an observation vector made at time t, where $O_t = \{o_1, o_2, ..., o_M\}$ and M is the dimension of the observation vector. A particular gesture sequence is then defined as $\Theta = \{O_1, O_2, ..., O_T\}$.

3. Hidden Markov Models and Hidden Conditional Random Fields

HMMs are generative models, assigning a joint probability to pairs of observations and labels. HMM parameters are typically trained to maximize the joint likelihood of training examples. To define a joint probability over observation and label sequences, a generative model needs to enumerate all possible observation sequences. HMMs typically require features appropriate for the particular recognition task and it is not practical to use feature vectors which are comprised of multiple interacting features. The main weakness of HMMs is the assumption of independence, which assumes that current observations are statistically independent of the previous observations. This is one of the main motivations for the use of CRFs. CRF use an exponential distribution to model the entire sequence given the observation sequence. This avoids the independence assumption between observations, and allows non-local dependencies between state and observations.

3.1. Hidden Markov Models

Hidden Markov Models (HMMs) are a type of statistical model and can model spatiotemporal information in a natural way. HMMs have efficient algorithms for learning and recognition, such as the Baum-Welch algorithm and Viterbi search algorithm [12].

A HMM is a collection of states connected by transitions. Each transition (or time step) has a pair of probabilities: a transition probability (the probability of taking a particular transition to a particular state) and an output probability (the probability of emitting a particular output symbol from a given state).

We use the compact notation $\lambda = \{A, B, \pi\}$ to indicate the complete parameter set of the model where A is a matrix storing transitions probabilities and a_{ij} denotes the probability of making a transition between states s_i and s_j. B is a matrix storing output probabilities for each state and π is a vector storing initial state probabilities.

HMMs can use either a set of discrete observation symbols or they can be extended for continuous observations signals. To calculate the probability of a specific observation O_t, we implement a probability density function of an M-dimensional multivariate gaussian (see Equation 1). Where μ is the mean vector and Σ is the covariance matrix.

$$\aleph(O_t|\mu, \Sigma) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} exp\big(-\tfrac{1}{2}(O_t-\mu)^T \Sigma^{-1}(O_t-\mu)\big) \quad (1)$$

3.1.1 HMM Threshold Model

Lee and Kim [8] proposed a single channel HMM threshold model using discrete observations to recognize a set of distinct gesture. We expand on the work of Lee and Kim to develop a HMM threshold model system which models continuous multidimensional sign language observations within a parallel HMM network to recognize two hand signs and identify movement epenthesis. A specific HMM, called a threshold model, is created to model movement epenthesis by calculating the likelihood threshold of an input gesture and provide a confirmation mechanism for provisionally matched gesture patterns. We denote each dedicated gesture HMM as λ_y. Each λ_y is used to calculate the likelihood that the input gesture is belonging to gesture class y. For a network of HMMs $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_Y\}$, a single threshold model $\overline{\lambda}$ is created. The threshold model $\overline{\lambda}$ is used to calculate the likelihood threshold for each of the dedicated gesture HMMs. It is not in the scope of this paper to describe the threshold model in detail and readers

should consult the works of Lee and Kim [8] and Kelly et al [6] for a more detailed discussion on the HMM threshold model technique.

3.2. Conditional Random Fields

CRFs are a framework based on conditional probability approaches for segmenting and labeling sequential data. The task is to learn a mapping of observations x to class labels $y \in Y$, where x is a m dimensional vector of local observations, $x = \{x_1, x_2, ..., x_m\}$, and each local observation x_j is represented by a feature vector $\phi(x_j) \in \Re^d$. A conditional model $p(y|x)$ is constructed from the paired observation and label sequences.

3.2.1 Hidden Conditional Random Fields

Wang et al [19] proposed a discriminative hidden-state approach for the recognition of human gestures. For any set of observations x they implement a set of hidden variables $s = \{s_1, s_2, ..., s_m\}$ which are not observed on training examples. Each s_j is a member of S where S is a finite set of possible parts in the model. Each s_j corresponds to a labeling of x_j with some member of S. A HCRF models the conditional probability of a class label given a set of observations by:

$$P(y|x,\theta) = \sum_s P(y,s|x,\theta) = \frac{\sum_s e^{\Psi(y,s,x;\theta)}}{\sum_{y' \in Y, s \in S^m} e^{\Psi(y',s,x;\theta)}} \tag{2}$$

The potential function $\Psi(y, s, x; \theta) \in \Re$, parameterized by θ, measures the compatibility between a label, a set of observations and a configuration of the hidden states.

3.2.2 Latent-Dynamic Conditional Random Fields

The CRF approach models the transitions between gestures, thus capturing extrinsic dynamics, but lacks the ability to represent internal sub-structure. Each Hidden-state Conditional Random Field models a single gesture label but cannot learn the dynamics between gesture labels. Morency et al. [11] propose a Latent-Dynamic Conditional Random Field (LDCRF) to combine the strengths of CRFs and HCRFs by capturing both extrinsic dynamics and intrinsic sub-structure. They define the latent conditional model as shown in Equation 6.

$$P(y|x,\theta) = \sum_s P(y|s,x,\theta) P(s|x,\theta) \tag{3}$$

$$P(s|x,\theta) = \frac{1}{Z(x,\theta)} \exp(\sum_k \theta_k \cdot F_k(s,x)) \tag{4}$$

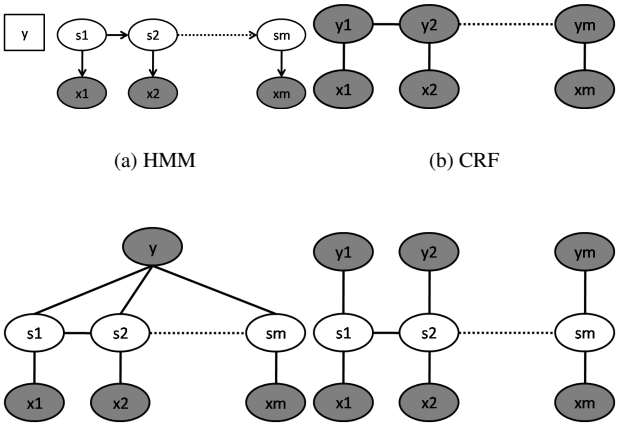
$$Z(x,\theta) = \sum_s \exp(\sum_k \theta_k \cdot F_k(s,x)) \tag{5}$$

Where F_k is defined as

$$F_k(s,x) = \sum_{j=1}^{m} f_k(s_{j-1}, s_j, x, j) \tag{6}$$

And each feature function f_k is either a state function $s_k(s_j, x, j)$ or a transition function $t_k(s_{j-1}, s_j, x, j)$.

Figure 2 illustrates the graphical models of HMM, CRF, HCRF and LDCRF.



(a) HMM (b) CRF

(c) HCRF (d) LDCRF

Figure 2. Comparison of HMM and different CRF models where grey circles denoted observed symbols.

4. HMM Threshold Model For Gesture Recognition

We develop a HMM threshold model system which models continuous multidimensional gesture observations within a HMM network to recognize motion based gestures and identify movement epenthesis. We now briefly describe this system.

4.1. HMM Training

We implement and train a dedicated HMM for each gesture to be recognized.

We denote each dedicated HMM as λ_y where $y \in Y$ and Y is the set gesture labels. Each HMM is trained using an automated HMM initialization and training technique, utilizing an iterative clustering, Baum Welch and Viterbi realignment process, proposed by Kelly et al [6].

A HMM threshold model, $\overline{\lambda}$ is then created using the network of trained HMMs λ_y (where $y \in Y$). The set of HMMs, to recognize the Y pre-trained gestures, is then denoted as $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_I, \overline{\lambda}\}$.

4.2. HMM Gesture Classification

Given an unknown sequence of gesture observations Θ, the goal is to accurately classify the gesture as a non-gesture or as one of the Y trained gestures. To classify the observations, the Viterbi algorithm is run on each model given the unknown observation sequences Θ, calculating the most likely state paths through each model y. The likelihoods of each state path, which we denote as $P(\Theta|\lambda_y)$, are also calculated. The sequence of observations can then be classified as y if the maximum likelihood $P_{ML}(\Theta|\lambda_y) \geq \Phi_y$, where the maximum likelihood is defined in Equation 7 and Φ_y is defined in Equation 11.

$$P_{ML}(\Theta|\lambda_y) = \max_y P(\Theta|\lambda_y) \tag{7}$$

$$\Phi_y = P(\Theta|\overline{\lambda})\Gamma_y \tag{8}$$

Where Γ_y is a constant scalar value used to tune the sensitivity of the system movement epenthesis gestures.

4.3. Parallel HMM Training

When recognizing two handed spatiotemporal gestures, a parallel HMM is required to model the left and right hands [16]. We implement a parallel HMM Threshold Model system which initializes and trains a dedicated parallel HMM denoted as $\lambda_y^{'} = \{\lambda_{Ly}, \lambda_{Ry}\}$ where λ_{Ly} and λ_{Ry} are HMMs which model the left and right hand gestures respectively.

The parallel HMMs are also trained using the same automated HMM initialization and training technique, utilizing an iterative clustering, Baum Welch and Viterbi realignment process, proposed by Kelly et al [6].

A weighting of ω_{Ly} and ω_{Ry} is applied to the left hand HMM and right hand HMM respectively, to account for variations in information held in each of the hands for a particular sign. The weighting applied in the system is based on a variance measure of the observation sequences. Using data from all observation sequences Θ_{Ly}^k and Θ_{Ry}^k, where $1 \leq k \leq K$, K is the total number of training examples and Θ_{Ly} and Θ_{Ry} are the left and right hand observations respectively. The variance of the left and right hand observations are calculated by calculating the variance of each observation dimension $\sigma_{Ly}^2[i]$ and $\sigma_{Ry}^2[i]$, where $0 \leq i \leq D$ and D is the dimension of the observation vectors. The left HMM weight, ω_{Ly}, and right HMM weight, ω_{Ry}, are then calculated as using Equation 9 where $\omega_{Ly} + \omega_{Ry} = 1$.

$$\omega_{Ly} = \frac{\sum_{i=0}^{D} \sigma_{Ly}^2[i]}{(\sigma_{Ly}^2[i] + \sigma_{Ry}^2[i]) \times D} \qquad \omega_{Ry} = \frac{\sum_{i=0}^{D} \sigma_{Ry}^2[i]}{(\sigma_{Ly}^2[i] + \sigma_{Ry}^2[i]) \times D} \tag{9}$$

A parallel HMM threshold model, $\overline{\lambda'} = \{\overline{\lambda_L}, \overline{\lambda_R}\}$ is then created using the network of trained parallel HMMs λ_y ($y \in Y$).

4.4. Parallel HMM Gesture Classification

To classify the parallel observations $\Theta^{'} = \{\Theta_L, \Theta_R\}$, the Viterbi algorithm is run on each model given the unknown observation sequences Θ_L and Θ_R, calculating the most likely state paths through each model y. The likelihoods of each state path, which we denote as $P(\Theta_L|\lambda_{Ly})$ and $P(\Theta_L|\lambda_{Ry})$, are also calculated. We calculate the overall likelihoods of a dedicated gesture and a movement epenthesis with the equations defined in Equations 10 and 11.

$$P(\Theta^{'}|\lambda_y^{'}) = P(\Theta_L|\lambda_{Ly})\omega_{Ly} + P(\Theta_R|\lambda_{Ry})\omega_{Ry} \tag{10}$$

$$\Phi_y^{'} = \frac{P(\Theta_L|\overline{\lambda_L})\Gamma_{Ly} + P(\Theta_R|\overline{\lambda_R})\Gamma_{Ry}}{2} \tag{11}$$

Where Γ_{Ly} and Γ_{Ry} are constant scalar values used to tune the sensitivity of the system to movement epenthesis. The sequence of observations can then be classified as y if $P_{ML}(\Theta^{'}|\lambda_y^{'}) \geq \Phi_y^{'}$, where $P_{ML}(\Theta^{'}|\lambda_y^{'})$ is the maximum likelihood defined as $\max_y P(\Theta^{'}|\lambda_y^{'})$.

5. CRFs For Gesture Recognition

Wang et al [19] and Morency et al [11] propose gesture recognition framework using HCRFs and LDCRFs respectively. We evaluate this same framework for the recognition of motion based gestures in sign language.

5.1. CRF Training

Similar to the works of Wang et al and Morency et al, we implement an objective function, shown in Equation 12, to train the parameters of each of the CRF models.

$$L(\theta) = \sum_{i=1}^{n} \log P(y_i|x_i, \theta) - \frac{1}{2\sigma^2}\|\theta\|^2 \tag{12}$$

Where n is the total number of training sequences. We implement a gradient ascent search to find the optimal parameter values, $\theta^* = \operatorname{argmax}_\theta L(\theta)$ using a Quasi-Newton optimization technique.

5.2. CRF Gesture Classification

Given an unknown sequence of gesture observations Θ, we calculate the conditional probability $P(y|\Theta, \theta)$ of each of the CRF, HCRF and LDCRF models for gesture labels $y \in Y$.

We classify a given observation sequence Θ as gesture class y if $P_{ML}(y|\Theta, \theta) > \Omega$, where Ω is a pre defined threshold value and $P_{ML}(y|\Theta, \theta)$ is the maximum likelihood defined as $\max_y P(y|\Theta, \theta)$. .

5.3. CRF Parallel Training

Similar to the parallel HMM system, we implement a parallel CRF model in order to recognize two handed spatiotemporal gestures. We apply to the same weighting technique, discussed in Section 4.3, to the parallel CRF models by calculating left CRF weights, ω_{Ly} and right CRF weights ω_{Ry}.

5.4. CRF Parallel Classification

Given a parallel observation sequence $\Theta^{'} = \{\Theta_L, \Theta_R\}$, we calculate the conditional probability $P(y|\Theta_L, \theta)$ and $P(y|\Theta_R, \theta)$ for each parallel CRF model. The parallel conditional probability is then defined in Equation 13.

$$P(y|\Theta^{'}, \theta) = P(y|\Theta_L, \theta)\omega_{Ly} + P(y|\Theta_R, \theta)\omega_{Ry} \quad (13)$$

We classify a given observation sequence $\Theta^{'}$ as gesture class y if $P(y|\Theta^{'}, \theta) > \Omega^{'}$, where $\Omega^{'}$ is a pre defined threshold value and $P_{ML}(y|\Theta^{'}, \theta)$ is the maximum likelihood defined as $\max_y P(y|\Theta^{'}, \theta)$..

6. Evaluation of Techniques

Wang et al [19] perform experiments to show that the HCRF model performs better at classifying head and arm gestures than CRFs and HMMs. In their experiments, the models were evaluated on their ability to classify a given segmented gesture sequence as one of a number of pre trained gestures but the models were not tested on non-gesture sequences. In order to evaluate and access the ability of a HCRF model to recognize gestures in sign language, the performance of the model must be evaluated when identifying non-gestures/epenthesis as well as being evaluated on the performance of classifying gestures.

Morency et al [11] perform experiments to evaluate the performance of the LDCRF model on three different data sets. The first data set was a head nod data set where the system was trained and tested on frames labeled as a head nod or labeled as not a head nod. The second data set, similar to the first data set, was trained and tested on positive and negative examples of heads nods. The final data set was an eye gaze data set, and the system was trained and tested on frames labeled as either an eye gaze-aversion gesture or a non gaze-aversion gesture. The LDCRF model was shown to out perform CRF, HDCRF and HMM based classifiers (as well as a support vector machine based classifier). From these experiments it is difficult to access whether or not the LDCRF model could be implemented to recognize a larger vocabulary of gestures or whether or not the LDCRF model could be used in a sign language based system. In the experiment Morency et al carry out, each of the gesture data set experiments were trained to recognize a single gesture with positive and negative examples of the gesture. In order to evaluate the LDCRF model for a sign language recognition system, the model should be tested on a larger vocabulary of gestures. In their experiments the gesture model was trained on positive and negative examples of the gesture. Training a model to recognize to recognize movement epenthesis in sign language is unfeasible due to the large number of possible epenthesis that can occur between signs.

The goal of this work is to evaluate the performance of the HMM threshold model and the different CRF models when recognizing motion based gestures and identifying epenthesis which occur in sign language. Since sign language communication is multimodal it involves not only hand gestures (i.e., manual signing) but also non-manual signals (NMS) conveyed through facial expressions, head movements, body postures and torso movements [3]. In order to evaluate the use of HMMs and CRFs in recognizing motion based gestures in sign language, we evaluate the models on two data sets; a manual signing data set (i.e. two handed motion based gestures) and a non-manual signal data set based on head motion gestures.

6.1. Manual Sign Experiments

The first data set we use to evaluate the models on is a set of two handed spatiotemporal hand gestures used in sign language. This data set consists of eight different manual signs extracted from videos of a fluent signer performing natural sign language sentences. Figure 3 illustrates an example of a signer performing each of the eight manual signs.

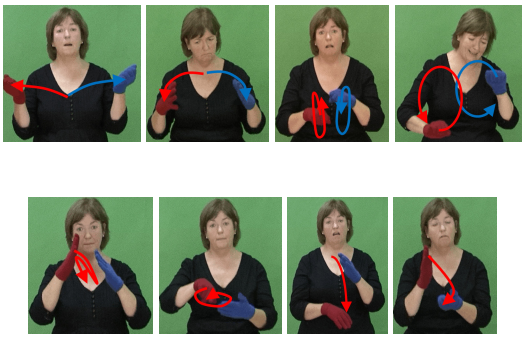


Figure 3. Example of the eight different signs the system was tested on (a) Newspaper, (b) A lot, (c) Bike, (d) Clean, (e) Paint, (f) Plate, (g) Lost, (h) Gone

In order recognize manual signs, we must extract two observation channels from the video streams. The two observation channels correspond to the left hand observations Θ_L and the right hand observations Θ_R. The observations Θ_L and Θ_R are combined into a parallel observation sequence $\Theta^{'}$ which is processed by the parallel models. We extract a set of observation sequences $\Delta_y^{'}$ from the video

sequences, where $y \in Y$, Y is the set of sign labels, $\Delta_y^{'} = \{\Theta_{1y}^{'}, ...\Theta_{Ty}^{'}\}$ and T is the number of sample observation sequences recorded for each gesture label y.

This set is then divided into a training set, $\Delta_y^{'\tau}$, and a test set, $\Delta_y^{'\zeta}$. A set of 10 training signs and a set of 10 test signs were recorded for each sign (A total of 160 gesture samples). The HMM, CRF, HCRF and LDCRF models were then trained on $\Delta_y^{'\tau}$.

An additional set of observations $\Delta_E^{'}$, which represents a collection of movement epenthesis, were also extracted from the video sequences to test the performance of the threshold model. For each valid sign, 10 movement epenthesis, that occurred before and after the valid sign in different sign language sentences, were recorded. An additional set of 20 random movement epenthesis were also recorded, resulting is a test set of 100 samples to evaluate the models on.

Before comparing the performance of the different models we first discuss the feature vectors used for the HMM and CRF models. Preliminary experiments show that the best performing feature vector for the HMM threshold model was the feature, $O = \{RP_x, RP_y, V_x, V_y, D_H\}$, which describes the position of the hands relative to the eyes, the direction of the movement of the hand and the distance between the two hands. The best performing feature vector for the three different CRF models was the feature vector $O = \{V_x, V_y\}$, which describes the direction of the movement of the hand. These are the feature vectors used for the evaluation of the HMM and CRF models.

To evaluate the performance of the models, we perform a ROC analysis on the different models and calculate the area under the curve (AUC) for each model. The classification of a gesture is based on a comparison of a model probability and a threshold value. In our ROC analysis of each model, we vary the threshold and create a confusion matrix for each of the thresholds. In the case of the HMM threshold model system, we vary the weighting of the threshold. When implementing the HCRF model and LDCRF model we vary the number of hidden states and also vary the window parameter ω. The window parameter defines the amount of past and future history to be used when predicting the state at time t such that long range dependencies can be incorporated. In our experiments we test each model on a two different groups of data. The first data group, which we denote as data set 1, is a set which includes all test sequences $\Delta_y^{'\zeta}$ and epenthesis sequences $\Delta_E^{'}$. The second data group, which we denote as data set 2, is a set which includes just the test sequences $\Delta_y^{'\zeta}$. Table 1 shows the AUC measurements of the traditional HMM model, the HMM threshold model and different variations of the CRF models.

The results show that the overall best performing model, with an AUC of 0.985, was the LDCRF model with 8 hidden states per label when tested on the data set 2. Since a sign language recognition system must be able to identify movement epenthesis as well as recognize gestures, the results of the tests performed on the data set 1 are more relevant to the evaluation of a sign language recognition model. The model which scores best when recognizing data set 1 is the HMM threshold model which has an AUC of 0.976. Although the HCRF and LDCRF perform better than the HMM threshold model when classifying gestures, the performance of both drop significantly when the epenthesis data is introduced. The performance of the HMM threshold model drops small amount compared to the relatively large drops of all the CRF models. This indicates that the HMM threshold model is more robust when classifying gestures and identifying epenthesis.

Table 1. Manual Signs: AUC Measurements for Different Models

Model	**Data Set** $1^{\dagger}$	**Data Set** $2^{\ddagger}$
HMM	0.902	0.943
Threshold HMM	0.976	0.977
CRF $\omega = 0$	0.833	0.876
CRF $\omega = 1$	0.794	0.828
HCRF $\omega = 0$, $S = 6$	0.909	0.944
HCRF $\omega = 1$, $S = 6$	0.957	0.983
HCRF $\omega = 2$, $S = 6$	0.944	0.971
HCRF $\omega = 0$, $S = 8$	0.947	0.965
HCRF $\omega = 1$, $S = 8$	0.934	0.968
LDCRF $\omega = 0$, $S^* = 1$	0.847	0.881
LDCRF $\omega = 0$, $S^* = 2$	0.806	0.842
LDCRF $\omega = 0$, $S^* = 3$	0.808	0.836
LDCRF $\omega = 0$, $S^* = 4$	0.863	0.901
LDCRF $\omega = 0$, $S^* = 8$	0.942	0.985
LDCRF $\omega = 1$, $S^* = 8$	0.899	0.928

$\dagger$ - Data Set which includes 100 epenthesis samples
$\ddagger$ - Data Set which does not include epenthesis samples
$*$ - S^* refers to number of hidden states per label for LDCRF

6.2. Head Gesture Experiments

The second data set we evaluate the HMM and CRF models on is a set of head movement gestures used to convey non manual information in sign language. The head gesture set consists of three different head movement gestures extracted from videos of a fluent signer performing natural sign language sentences.

A visual example of a signer performing each of the three different head movement gesture is in shown in Figure 6.2.

Similar to the manual sign experiments described in Section 6.1, observation sequences $\Delta_y = \{\Theta_{1y}, ...\Theta_{Ty}\}$ were extracted from the videos and divided into a training set, Δ_y^{τ}, and a test set, Δ_y^{ζ}. For the non-manual signal experiments, a set of 6 training signs and a set of 6 test signs were recorded for each sign (A total of 36 gesture samples). The

Figure 4. Example of the three different head movement gestures the system was tested on (a) Right Movement (b) Left Movement (c) Left Forward Movement

HMM models and all CRF models were then trained on Δ_y^τ.

An additional set of 25 other head gesture sequences Δ_E, outside of the training set, were also extracted from the video sequences to test the performance of the system when identifying movement epenthesis.

Preliminary experiments show that the best performing feature vector for the HMM models, when classifying head gestures, was a 2 dimensional vector $O = \{V_x, V_y\}$ describing the velocity of the head movement in the x and y directions.

To calculate the velocity vector of the head we use the mid point between the eyes and calculate the movement of the midpoint from frame to frame. As with the HMM models, the best performing feature vector for the CRF models was the 2 dimensional velocity vector $O = \{V_x, V_y\}$. These are the feature vectors used for the evaluation of the different models. Similar to the hand gesture experiments, we test the head gesture models on two data groups; data group 1 includes the gesture test sequences and the non gesture sequences, while data set 2 includes only the gesture test sequences.

We carry out a ROC analysis of the non-manual models using the same procedure described in Section 6.1. Table 2 shows the AUC measurements of models.

The results of this experiment repeat the same trend found in the results of the manual sign recognition experiment. The LDCRF model performs best when classifying gestures in data set 2. The recognition rate of the CRF models then decrease significantly when non-gestures are introduced. The best performing model for data set 1 is again the HMM threshold model with an AUC of 0.936. The difference between the data set 1 AUCs of the HMM threshold model and the 9 state LDCRF was 0.042. This result suggest that the HMM threshold model is a more robust model when recognizing the head movement gestures when epenthesis gestures are taken in to account.

Table 2. Non-Manual Signals: AUC Measurements for Different Models

Model	**Data Set** $1^\dagger$	**Data Set** $2^\ddagger$
HMM	0.873	0.901
Threshold HMM	0.936	0.947
CRF $\omega = 0$	0.736	0.768
CRF $\omega = 1$	0.527	0.545
HCRF $\omega = 0, S = 2$	0.698	0.801
HCRF $\omega = 1, S = 2$	0.786	0.911
HCRF $\omega = 2, S = 2$	0.702	0.816
HCRF $\omega = 0, S = 4$	0.784	0.927
HCRF $\omega = 1, S = 4$	0.719	0.811
HCRF $\omega = 0, S = 6$	0.743	0.850
HCRF $\omega = 1, S = 6$	0.736	0.893
HCRF $\omega = 0, S = 8$	0.715	0.838
HCRF $\omega = 1, S = 8$	0.708	0.788
LDCRF $\omega = 0, S^* = 3$	0.794	0.899
LDCRF $\omega = 1, S^* = 3$	0.763	0.880
LDCRF $\omega = 0, S^* = 6$	0.760	0.827
LDCRF $\omega = 1, S^* = 6$	0.717	0.791
LDCRF $\omega = 0, S^* = 9$	0.868	0.922
LDCRF $\omega = 1, S^* = 9$	0.837	0.901
LDCRF $\omega = 2, S^* = 9$	0.894	0.952
LDCRF $\omega = 3, S^* = 9$	0.795	0.861

† - Data Set which includes 25 non-gesture samples
‡ - Data Set which does not include non-gesture samples
$*$ - S^* refers to number of hidden states per label for LDCRF

The change in performance of the LDCRF from data set 2 to data set 1 was 0.058, while the change in performance of the HMM threshold model was only 0.011. This result suggests that the performance of the LDCRF would decrease more than that of the HMM threshold model when the number of epenthesis gestures introduced into the system increased.

7. Conclusion

In this paper we described our HMM threshold model system for identifying epenthesis and classifying motion based gestures in sign language. We evaluated the HMM threshold model and compared it to current models for recognizing human motion. HMMs, CRFs, HCRFs and LDCRFs have recently been implemented in current works for recognizing different human actions. We evaluate these techniques in the domain of sign language gesture recognition. In order to evaluate the performance of the models when recognizing sign language gestures, it was important to evaluate each model when identifying movement epenthesis as well as evaluating the performance of the

models when classifying gestures. We performed experiments on a data set of motion based manual signs and a data set of non-manual head motion gestures. In the hand gesture experiments and head gesture experiments, the best performing model was the LDCRF when tested on data set 2. The results of the experiments on data set 2 were consistent with previous experiments on HCRFs and LDCRFs which Wang et al [19] and Morency et al [11] who show that HCRFs and LDCRF perform better than the standard HMM model when classifying gestures. When data set 1 was introduced to the experiments, the performance of the standard HMM model, and all CRF models, dropped significantly in relation to the performance of HMM threshold model. The HMM threshold model performed best in both experiments, with movement epenthesis data, with an AUC of 0.976 and 0.936 for the hand gesture and head gesture evaluations respectively.

The contribution of this paper is that we have performed a full evaluation of the different CRF and HMM models when recognizing sign language gestures. We show that our HMM threshold model performs better than the HCRF and LDCRF models when identifying movement epenthesis and classifying gestures. The significance of this result is that, even though the assumption of independence is an inherent weakness in the HMM model, we have shown that a threshold HMM model, which is trained on appropriate features, can outperform the HCRF and LDCRF models when recognizing sign language gestures.

References

[1] M. Assan and K. Grobel. Video-based sign language recognition using hidden markov models. In *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 97–109, London, UK, 1998. Springer-Verlag.

[2] B. Bauer and K.-F. Kraiss. Towards an automatic sign language recognition system using subunits. In *GW '01: Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction*, pages 64–75, London, UK, 2002. Springer-Verlag.

[3] S. C., W. Ong, and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. PAMI*, 27(6):873–891, 2005.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2:142–149 vol.2, 2000.

[5] W. Gao, G. Fang, D. Zhao, and Y. Chen. Transition movement models for large vocabulary continuous sign language recognition. *IEEE FG 2004*, pages 553–558, May 2004.

[6] D. Kelly, J. McDonald, and C. Markham. Recognizing spatiotemporal gestures and movement epenthesis in sign language. In *IMVIP 2009*, 2009.

[7] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[8] H. K. Lee and J. H. Kim. An hmm-based threshold model approach for gesture recognition. *IEEE PAMI*, 21(10):961–973, 1999.

[9] J. R. Liddell, S.K. American sign language: The phonological base. *Sign Langauge Studies*, 64.

[10] L. A.-C. M. Castrillon-Santana, O. Deniz-Suarez and J. Lorenzo-Navarro. Performance evaluation of public domain haar detectors for face and facial feature detection. *VISAPP 2008*, 2008.

[11] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. pages 1–8, June 2007.

[12] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[13] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. volume 2, pages 1808–1815 Vol. 2, Oct. 2005.

[14] T. Starner, A. Pentland, and J. Weaver. Real-time american sign language recognition using desk and wearable computer based video. *IEEE PAMI*, 20(12):1371–1375, 1998.

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR, IEEE*, 1:511, 2001.

[16] C. Vogler and D. Metaxas. Parallel hidden markov models for american sign language recognition. In *In ICCV*, pages 116–122, 1999.

[17] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *Computer Vision and Image Understanding*, 81:358–384, 2001.

[18] C. Wang, S. Shan, and W. Gao. An approach based on phonemes to large vocabulary chinese sign language recognition. In *IEEE FG 2002*, page 411, Washington, DC, USA, 2002. IEEE Computer Society.

[19] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. volume 2, pages 1521–1527, 2006.