Hidden Conditional Random Fields for Gesture Recognition

Sy Bor Wang Ariadna Quattoni Louis-Philippe Morency David Demirdjian
Trevor Darrell
{sybor, ariadna, lmorency, demirdji, trevor}@csail.mit.edu

Computer Science and Artificial Intelligence Laboratory, MIT
32 Vassar Street, Cambridge, MA 02139, USA

Abstract

We introduce a discriminative hidden-state approach for the recognition of human gestures. Gesture sequences often have a complex underlying structure, and models that can incorporate hidden structures have proven to be advantageous for recognition tasks. Most existing approaches to gesture recognition with hidden states employ a Hidden Markov Model or suitable variant (e.g., a factored or coupled state model) to model gesture streams; a significant limitation of these models is the requirement of conditional independence of observations. In addition, hidden states in a generative model are selected to maximize the likelihood of generating all the examples of a given gesture class, which is not necessarily optimal for discriminating the gesture class against other gestures. Previous discriminative approaches to gesture sequence recognition have shown promising results, but have not incorporated hidden states nor addressed the problem of predicting the label of an entire sequence. In this paper, we derive a discriminative sequence model with a hidden state structure, and demonstrate its utility both in a detection and in a multi-way classification formulation. We evaluate our method on the task of recognizing human arm and head gestures, and compare the performance of our method to both generative hidden state and discriminative fully-observable models.

1. Introduction

With the potential for many interactive applications, automatic gesture recognition has been actively investigated in the computer vision and pattern recognition community. Head and arm gestures are often subtle, can happen at various timescales, and may exhibit long-range dependencies. All these issues make gesture recognition a challenging problem.

One of the most common approaches for gesture recognition is to use Hidden Markov Models (HMM) [19, 23], a powerful generative model that includes hidden state structure. More generally, factored or coupled state models have been developed, resulting in multi-stream dynamic Bayesian networks [20, 3]. However, these generative models assume that observations are conditionally independent. This restriction makes it difficult or impossible to accommodate long-range dependencies among observations or multiple overlapping features of the observations.

Conditional random fields (CRF) use an exponential distribution to model the entire sequence given the observation sequence [10, 9, 21]. This avoids the independence assumption between observations, and allows non-local dependencies between state and observations. A Markov assumption may still be enforced in the state sequence, allowing inference to be performed efficiently using dynamic programming. CRFs assign a label for each observation (e.g., each time point in a sequence), and they neither capture hidden states nor directly provide a way to estimate the conditional probability of a class label for an entire sequence.

We propose a model for gesture recognition which incorporates hidden state variables in a discriminative multi-class random field model, extending previous models for spatial CRFs into the temporal domain. By allowing a classification model with hidden states, no a-priori segmentation into substructures is needed, and labels at individual observations are optimally combined to form a class conditional estimate.

Our hidden state conditional random field (HCRF) model can be used either as a gesture class detector, where a single class is discriminatively trained against all other gestures, or as a multi-way gesture classifier, where discriminative models for multiple gestures are simultaneously trained. The latter approach has the potential to share useful hidden state structures across the different classification tasks, allowing higher recognition rates.

We have implemented HCRF-based methods for arm and head gesture recognition and compared their performance against both HMMs and fully observable CRF techniques.

In the remainder of this paper we review related work, describe our HCRF model, and then present a comparative evaluation of different models.

2. Related Work

There is extensive literature dedicated to gesture recognition. Here we review the methods most relevant to our work. For hand and arm gestures, a comprehensive survey was presented by Pavlovic *et al.* [16]. Generative models, like HMMs [19], and many extensions have been used successfully to recognize arm gestures [3] and a number of sign languages [2, 22]. Kapoor and Picard presented a HMM-based, real time head nod and head shake detector [8]. Fugie *et al.* also used HMMs to perform head nod recognition [6].

Apart from generative models, discriminative models have been used to solve sequence labeling problems. In the speech and natural language processing community, Maximum Entropy Markov models (MEMMs) [11] have been used for tasks such as word recognition, part-of-speech tagging, text segmentation and information extraction. The advantages of MEMMs are that they can model arbitrary features of observation sequences and can therefore accommodate overlapping features.

CRFs were first introduced by Lafferty *et al.* [10] and have been widely used since then in the natural language processing community for tasks such as noun coreference resolution [13], name entity recognition [12] and information extraction [4].

Recently, there has been increasing interest in using CRFs in the vision community. Sminchisescu *et al.* [21] applied CRFs to classify human motion activities (i.e. walking, jumping, etc); their model can also discriminate subtle motion styles like normal walk and wander walk. Kumar *et al.* [9] used a CRF model for the task of image region labeling. Torralba *et al.* [24] introduced Boosted Random Fields, a model that combines local and global image information for contextual object recognition.

Hidden-state conditional models have been applied successfully in both the vision and speech community. In the vision community, Quattoni [18] applied HCRFs to model spatial dependencies for object recognition in unsegmented cluttered images. In the speech community, it was applied to phone classification [7] and the equivalence of HMM models to a subset of CRF models was established. Here we extend and demonstrate HCRF's applicability to model temporal sequences for gesture recognition.

3. HCRFs: A Review

We will review HCRFs as described in [18]. We wish to learn a mapping of observations $\mathbf{x}$ to class labels $y \in \mathcal{Y}$, where $\mathbf{x}$ is a vector of m local observations, $\mathbf{x} = \{x_1, x_2, \ldots x_m\}$, and each local observation x_j is represented by a feature vector $\phi(x_j) \in \Re^d$.

An HCRF models the conditional probability of a class label given a set of observations by:

$$P(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{s}} P(y, \mathbf{s} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{s}} e^{\Psi(y, \mathbf{s}, \mathbf{x}; \theta)}}{\sum_{y' \in \mathcal{Y}, \mathbf{s} \in S^m} e^{\Psi(y', \mathbf{s}, \mathbf{x}; \theta)}} \quad (1)$$

where $\mathbf{s} = \{s_1, s_2, ..., s_m\}$, each $s_i \in S$ captures certain underlying structure of each class and S is the set of hidden states in the model. If we assume that $\mathbf{s}$ is observed and that there is a single class label y then the conditional probability of $\mathbf{s}$ given $\mathbf{x}$ becomes a regular CRF. The potential function $\Psi(y, \mathbf{s}, \mathbf{x}; \theta) \in \Re$, parameterized by θ, measures the compatibility between a label, a set of observations and a configuration of the hidden states.

Following previous work on CRFs [9, 10], we use the following objective function in training the parameters:

$$L(\theta) = \sum_{i=1}^{n} \log P(y_i \mid \mathbf{x}_i, \theta) - \frac{1}{2\sigma^2} ||\theta||^2 \quad (2)$$

where n is the total number of training sequences. The first term in Eq. 2 is the log-likelihood of the data; the second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp\left(\frac{1}{2\sigma^2}||\theta||^2\right)$. We use gradient ascent to search for the optimal parameter values, $\theta^* = \arg\max_\theta L(\theta)$. For our experiments we used a Quasi-Newton optimization technique [1].

4. HCRFs for Gesture Recognition

HCRFs—discriminative models that contain hidden states—are well-suited to the problem of gesture recognition. Quattoni [18] developed a discriminative hidden state approach where the underlying graphical model captured spatial dependencies between hidden object parts. In this work, we modify the original HCRF approach to model sequences where the underlying graphical model captures temporal dependencies across frames, and to incorporate long range dependencies.

Our goal is to distinguish between different gesture classes. To achieve this goal, we learn a state distribution among the different gesture classes in a discriminative manner. Generative models can require a considerable number of observations for certain gestures classes. In addition, generative models may not learn a shared common structure among gesture classes nor uncover the distinctive configuration that sets one gesture class uniquely against others. For example, the flip-back gesture used in the arm gesture experiments (see Figure 1) consists of four parts: 1) lifting one arm up, 2) lifting the other arm up, 3) crossing one arm over the other and 4) returning both arms to their starting position. We could use the fact that when we observe

the joints in a particular configuration (see FB illustration in Figure 1) we can predict with certainty the flip-back gesture. Therefore, we would expect that this gesture would be easier to learn with a discriminative model. We would also like a model that incorporates long range dependencies (i.e., that the state at time t can depend on observations that happened earlier or later in the sequence.) An HCRF can learn a discriminative state distribution and can be easily extended to incorporate long range dependencies.

To incorporate long range dependencies, we modify the potential function Ψ in Equation 1 to include a window parameter ω that defines the amount of past and future history to be used when predicting the state at time t. Here, $\Psi(y, \mathbf{s}, \mathbf{x}; \theta, \omega) \in \Re$ is defined as a potential function parameterized by θ and ω.

$$\begin{aligned}\Psi(y, \mathbf{s}, \mathbf{x}; \theta, \omega) &= \sum_{j=1}^{n} \varphi(\mathbf{x}, j, \omega) \cdot \theta_s[s_j] + \sum_{j=1}^{n} \theta_y[y, s_j] \\ &\quad + \sum_{(j,k)\in E} \theta_e[y, s_j, s_k] \qquad (3)\end{aligned}$$

The graph E is a chain where each node corresponds to a hidden state variable at time t; $\varphi(\mathbf{x}, j, \omega)$ is a vector that can include any feature of the observation sequence for a specific window size ω. (i.e. for window size ω, observations from $t - \omega$ to $t + \omega$ are used to compute the features.)

The parameter vector θ is made up of three components: $\theta = [\theta_e \ \theta_y \ \theta_\mathbf{s}]$. We use the notation $\theta_s[s_j]$ to refer to the parameters θ_s that correspond to state $s_j \in S$. Similarly, $\theta_y[y, s_j]$ stands for parameters that correspond to class y and state s_j and $\theta_e[y, s_j, s_k]$ refers to parameters that correspond to class y and the pair of states s_j and s_k.

The inner product $\varphi(\mathbf{x}, j, \omega) \cdot \theta_s[s_j]$ can be interpreted as a measure of the compatibility between the observation sequence and the state at time j at window size ω. Each parameter $\theta_y[y, s_j]$ can be interpreted as a measure of the compatibility between a hidden state k and a gesture y. Finally, each parameter $\theta_e[y, s_j, s_k]$ measures the compatibility between pairs of consecutive states j and k and the gesture y.

Given a new test sequence $\mathbf{x}$, and parameter values θ^* learned from training examples, we will take the label for the sequence to be:

$$\arg\max_{y\in\mathcal{Y}} P(y \mid \mathbf{x}, \omega, \theta^*). \qquad (4)$$

Since E is a chain, there are exact methods for inference and parameter estimation as both the objective function and its gradient can be written in terms of marginal distributions over the hidden state variables. These distributions can be computed using belief propagation [17].

5. Experiments

We conducted two sets of experiments comparing HMM, CRF, and HCRF models on head gesture and arm gesture datasets. The evaluation metric that we used for all the experiments was the percentage of sequences for which we predicted the correct gesture label.

5.1. Datasets

Head Gesture Dataset: To collect a head gesture dataset, pose tracking was performed using an adaptive view-based appearance model which captured the user-specific appearance under different poses [14]. We used the fast Fourier transform of the 3D angular velocities as features for gesture recognition.

The head gesture dataset consisted of interactions between human participants and an embodied agent [15]. A total of 16 participants interacted with a robot, with each interaction lasting between 2 to 5 minutes. Human participants were video recorded while interacting with the robot to obtain ground truth. A total of 152 head nods, 11 head shakes and 159 junk sequences were extracted based on ground truth labels. The junk class had sequences that did not contain any head nods or head shakes during the interactions with the robot. Half of the sequences were used for training and the rest were used for testing. For the experiments, we separated the data such that the testing dataset had no participants from the training set.

Arm Gesture Dataset: We defined six arm gestures for the experiments (see Figure 1). In the Expand Horizontally (EH) arm gesture, the user starts with both arms close to the hips, moves both arms laterally apart and retracts back to the resting position. In the Expand Vertically (EV) arm gesture, the arms move vertically apart and return to the resting position. In the Shrink Vertically (SV) gesture, both arms begin from the hips, move vertically together and back to the hips. In the Point and Back (PB) gesture, the user points with one hand and beckons with the other. In the Double Back (DB) gesture, both arms beckon towards the user. Lastly, in the Flip Back (FB) gesture, the user simulates holding a book with one hand while the other hand makes a flipping motion, to mimic flipping the pages of the book.

Users were asked to perform these gestures in front of a stereo camera. From each image frame, a 3D cylindrical body model, consisting of a head, torso, arms and forearms was estimated using a stereo-tracking algorithm [5]. Figure 5 shows a gesture sequence with the estimated body model superimposed on the user. From these body models, both the joint angles and the relative co-ordinates of the joints of the arms are used as observations for our experiments and were manually segmented into six arm gesture classes. Thirteen users were asked to perform these six gestures; an average of 90 gestures per class were collected.

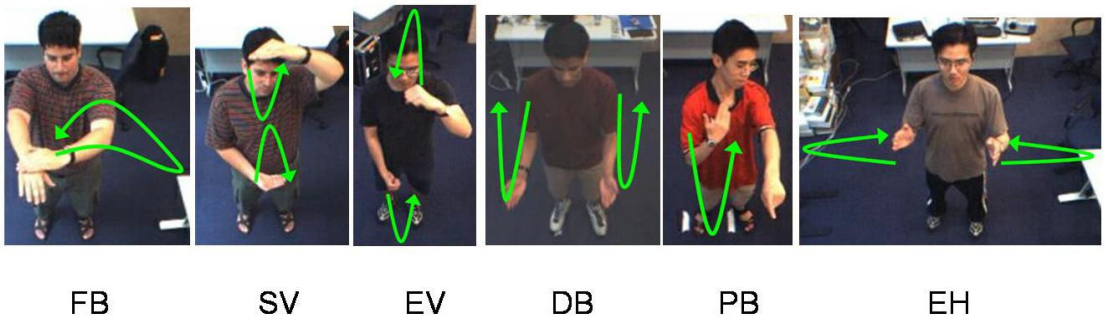

Figure 1. Illustrations of the six gesture classes for the experiments. Below each image is the abbreviation for the gesture class. These gesture classes are: FB - Flip Back, SV - Shrink Vertically, EV - Expand Vertically, DB - Double Back, PB - Point and Back, EH - Expand Horizontally. The green arrows are the motion trajectory of the fingertip and the numbers next to the arrows symbolize the order of these arrows.

5.2. Models

Figures 2, 3 and 4 show graphical representations of the HMM model, the CRF model, and the HCRF (multi-class) model used in our experiments.

HMM Model - As a first baseline, we trained a HMM model per class. Each model had four states and used a single Gaussian observation model. During evaluation, test sequences were passed through each of these models, and the model with the highest likelihood was selected as the recognized gesture.

CRF Model - As a second baseline, we trained a single CRF chain model where every gesture class had a corresponding state. In this case, the CRF predicts labels for each frame in a sequence, not the entire sequence. During evaluation, we found the Viterbi path under the CRF model, and assigned the sequence label based on the most frequently occurring gesture label per frame. We ran additional experiments that incorporated different long range dependencies (i.e. using different window sizes ω, as described in Section 4).

HCRF (one-vs-all) Model - For each gesture class, we trained a separate HCRF model to discriminate the gesture class from other classes. Each HCRF was trained using six hidden states. For a given test sequence, we compared the probabilities for each single HCRF, and the highest scoring HCRF model is selected as the recognized gesture.

HCRF (multi-class) Model - We trained a single HCRF using twelve hidden states. Test sequences were run with this model and the gesture class with the highest probability was selected as the recognized gesture. We also conducted experiments that incorporated different long range dependencies in the same way as described in the CRF experiments.

For the HMM model, the number of Gaussian mixtures and states were set by minimizing the error on training data, and for hidden state models the number of hidden states was set in a similar fashion.

Models	Accuracy (%)
HMM $\omega = 0$	65.33
CRF $\omega = 0$	66.53
CRF $\omega = 1$	68.24
HCRF (multi-class) $\omega = 0$	71.88
HCRF (multi-class) $\omega = 1$	85.25

Table 1. Comparisons of recognition performance (percentage accuracy) for head gestures.

6. Results and Discussion

For the training process, the CRF models for the arm and head gesture dataset took about 200 iterations to train. The HCRF models for the arm and head gesture dataset required 300 and 400 iterations for training respectively.

Table 1 summarizes the results for the head gesture experiments. The multi-class HCRF model performs better than the HMM and CRF models at a window size of zero. The CRF has slightly better performance than the HMMs for the head gesture task, and this performance improved with increased window sizes. The HCRF multi-class model made a significant improvement when the window size was increased, which indicates that incorporating long range dependencies was useful.

Table 2 summarizes results for the arm gesture recognition experiments. In these experiments the CRF performed better than HMMs at window size zero. At window size one, however, the CRF performance was poorer; this may be due to overfitting when training the CRF model parameters. Both multi-class and one-vs-all HCRFs perform better than HMMs and CRFs. The most significant improvement in performance was obtained when we used a multi-class HCRF, suggesting that it is important to jointly learn the best discriminative structure.

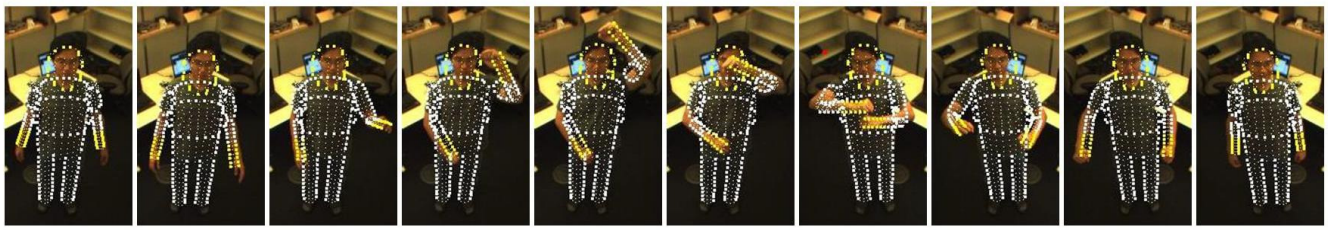

Figure 5. Sample image sequence with the estimated body pose superimposed on the user in each frame.

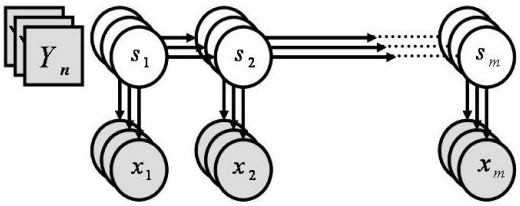

Figure 2. HMM model

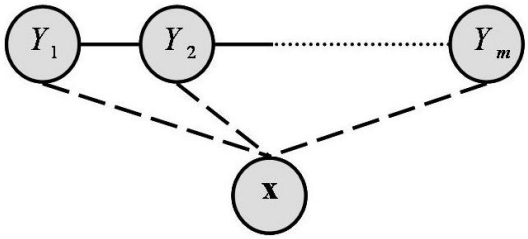

Figure 3. CRF Model

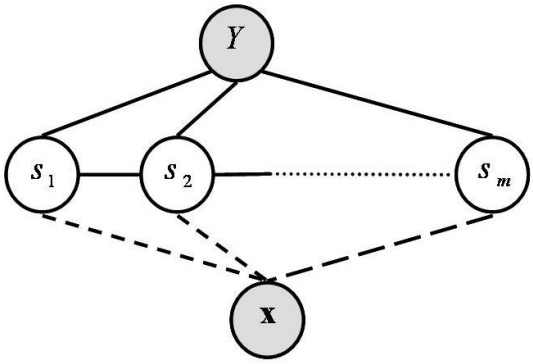

Figure 4. HCRF Model

Figure 6 shows the distribution of states for different gesture classes learned by the best performing model (multi-class HCRF). This graph was obtained by computing the Viterbi path for each sequence (i.e. the most likely assignment for the hidden state variables) and counting the number of times that a given state occurred among those sequences. As we can see, the model has found a unique distribution of hidden states for each gesture, and there is a significant amount of state sharing among different gesture classes. The state assignment for each image frame of various gesture classes is illustrated in Figure 7. Here, we see that body poses that are visually more unique for a gesture class are assigned very distinct hidden states, while body poses common between different gesture classes are assigned the same states. For example, frames of the FB

Models	Accuracy (%)
HMM $\omega = 0$	84.22
CRF $\omega = 0$	86.03
CRF $\omega = 1$	81.75
HCRF (one-vs-all) $\omega = 0$	87.49
HCRF (multi-class) $\omega = 0$	91.64
HCRF (multi-class) $\omega = 1$	93.81

Table 2. Comparisons of recognition performance (percentage accuracy) for body poses estimated from image sequences.

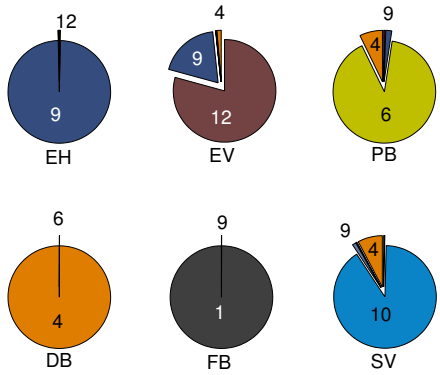

Figure 6. Graph showing the distribution of the hidden states for each gesture class. The numbers in each pie represent the hidden state label, and the area enclosed by the number represents the proportion.

Models	Accuracy (%)
HCRF $\omega = 0$	86.44
HCRF $\omega = 1$	96.81
HCRF $\omega = 2$	97.75

Table 3. Experiment on 3 arm gesture classes using the multi-class HCRF with different window sizes. The 3 different gesture classes are: EV-Expand Vertically, SV Shrink Vertically and FB - Flip Back. The gesture recognition accuracy increases as more long range dependencies are incorporated.

gesture are uniquely assigned a state of one while the SV and DB gesture class have visibly similar frames that share the hidden state four.

The arm gesture results with varying window sizes are shown in Table 3. From these results, it is clear that incorporating some amount of contextual dependency is important, since the HCRF performance improved with increasing window size.

7. Conclusion

In this work we presented a discriminative hidden-state approach for gesture recognition. Our proposed model combines the two main advantages of current approaches to gesture recognition: the ability of CRFs to use long range dependencies, and the ability of HMMs to model latent structure. By regarding the sequence label as a random variable we can train a single joint model for all the gestures and share hidden states between them. Our results have shown that HCRFs outperform both CRFs and HMMs for certain gesture recognition tasks. For arm gestures, the multi-class HCRF model outperforms HMMs and CRFs even when long range dependencies are not used, demonstrating the advantages of joint discriminative learning.

References

[1] Quasi-newton optimization toolbox in matlab.

[2] M. Assan and K. Groebel. Video-based sign language recognition using hidden markov models. In *Int'l Gest Wksp: Gest. and Sign Lang.*, 1997.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR*, 1996.

[4] A. Culotta and P. V. amd A. Callum. Interactive information extraction with constrained conditional random fields. In *AAAI*, 2004.

[5] D. Demirdjian and T. Darrell. 3-d articulated pose tracking for untethered deictic reference. In *Int'l Conf. on Multimodal Interfaces*, 2002.

[6] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of 13th IEEE International Workshop on Robot and Human*

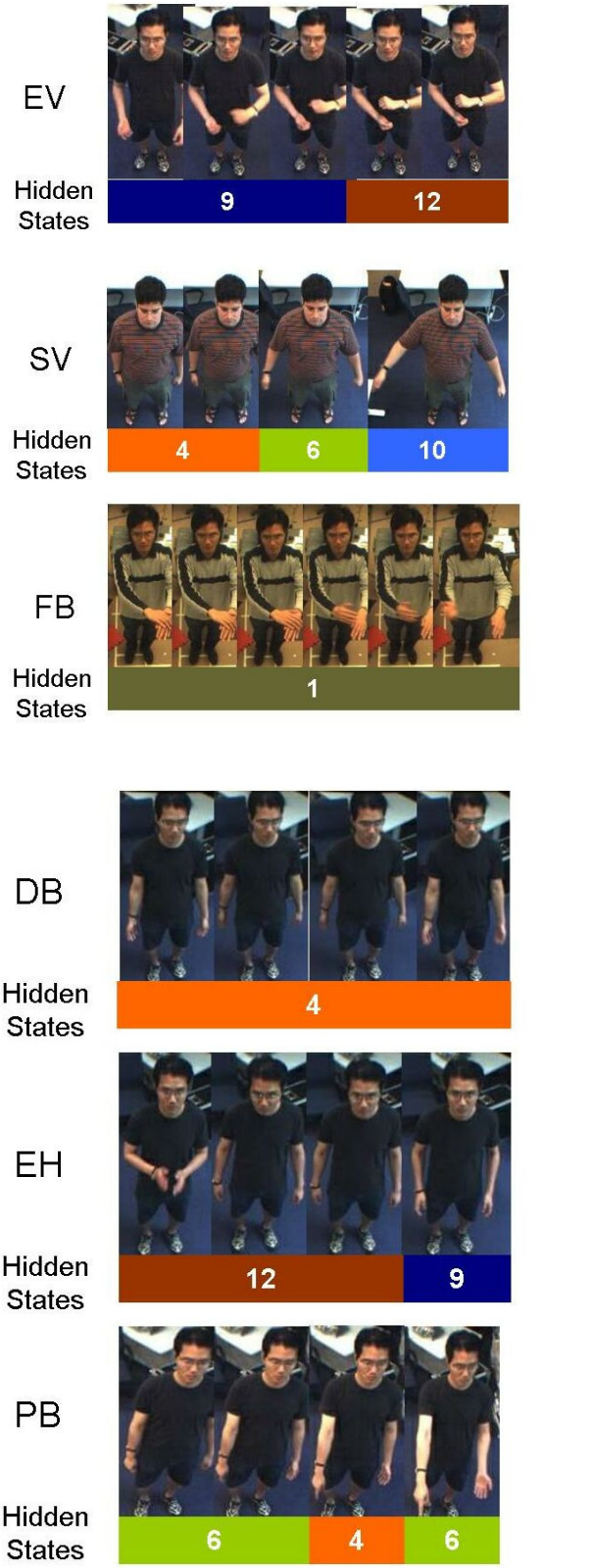

Figure 7. Articulation of the six gesture classes. The first few consecutive frames of each gesture class are displayed. Below each frame is the corresponding hidden state assigned by the multi-class HCRF model.

Communication, RO-MAN 2004, pages 159–164, September 2004.

[7] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. Hidden conditional random fields for phone classification. In *INTERSPEECH*, 2005.

[8] A. Kapoor and R. Picard. A real-time head nod and shake detector. In *Proceedings from the Workshop on Perspective User Interfaces*, November 2001.

[9] S. Kumar and M. Herbert. Discriminative random fields: A framework for contextual interaction in classification. In *ICCV*, 2003.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.

[11] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, 2000.

[12] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *CoNLL*, 2003.

[13] A. McCallum and B. Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. In *IJCAI Workshop on Information Integration on the Web*, 2003.

[14] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *CVPR*, 2003.

[15] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *ICMI*, 2005.

[16] V. I. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction. In *PAMI*, volume 19, pages 677–695, 1997.

[17] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[18] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *NIPS*, 2004.

[19] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 77, pages 257–286, 2002.

[20] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell. Visual speech recognition with loosely synchronized feature streams. In *ICCV*, 2005.

[21] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *Int'l Conf. on Computer Vision*, 2005.

[22] T. Starner and A. Pentland. Real-time asl recognition from video using hidden markov models. In *ISCV*, 1995.

[23] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *Int'l Wkshp on Automatic Face and Gesture Recognition*, 1995.

[24] A. Torralba, K. Murphy, and W. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2004.