Hidden Conditional Random Fields

Ariadna Quattoni, Sybor Wang,
Louis-Philippe Morency, Michael Collins,
and Trevor Darrell

Abstract—We present a discriminative latent variable model for classification problems in structured domains where inputs can be represented by a graph of local observations. A hidden-state Conditional Random Field framework learns a set of latent variables conditioned on local features. Observations need not be independent and may overlap in space and time.

Index Terms—Object recognition, model, supervised learning, classification.

1 Introduction

It is well-known that models which include latent or hidden-state structure may be more expressive than fully observable models, and can often find relevant substructure in a given domain. Hidden Markov Models (HMMs) and Dynamic Bayesian Networks use hidden state to model observations and have a clear generative probabilistic formulation.

A limitation of generative models is that observations are assumed to be independent given the values of the latent variables. Accurately specifying such a generative model may be challenging, particularly in cases where we wish to incorporate long range dependencies in the model and allow hidden variables to depend on several local features. These observations led to the introduction of discriminative models for sequence labeling, including MEMM's [14], [19] and Conditional Random Fields (CRFs) [12]. CRFs are powerful discriminative models, which can incorporate essentially arbitrary feature-vector representations of the observed data points, and have been widely used in the natural language processing community.

CRFs are limited in that they cannot capture intermediate structures using hidden-state variables. In this paper, we propose a new model for classification based on CRFs augmented with latent state, which we call Hidden-state Conditional Random Fields (HCRFs). HCRFs use intermediate hidden variables to model the latent structure of the input domain; they define a joint distribution over the class label and hidden state labels conditioned on the observations, with dependencies between the hidden variables expressed by an undirected graph. The result is a model where inference and parameter estimation can be carried out using standard graphical model algorithms. In this paper, we first describe the HCRF model, then describe experiments that demonstrate the ability of HCRFs to outperform generative hidden-state and discriminative fully-observable models on object and gesture recognition tasks.

• *The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge MA 02139-4309.*
E-mail: {ariadna, mcollins, trevor}@csail.mit.edu, {sybor, lmorency}@mit.edu.

Manuscript received 9 May 2006; revised 30 Oct. 2006; accepted 23 Jan. 2007; published online 6 Mar. 2007.
Recommended for acceptance by A. Rangarajan.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0361-0506.
Digital Object Identifier no. 10.1109/TPAMI.2007.1124.

2 Related Work

A complete review of related work is beyond the scope of this paper; here, we discuss previous related work on object and gesture recognition using generative and discriminative learning approaches. There is an extensive literature dedicated to gesture recognition. Generative models have been used successfully to recognize arm gestures [2] and a number of sign languages [1], [21]. 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 [5]. For a comprehensive survey of hand and arm gesture recognition see Pavlovic et al. [17].

In computer vision, CRFs have been applied to the task of detecting man-made structures in natural images and have been shown to outperform Markov Random Fields (MRFs) [11]. Sminchisescu [20] applied CRFs to classify human motion activity and demonstrated their model was more accurate than MEMMs. Torralba et al. [22] introduced Boosted Random Fields, a model that combines local and global image information for object recognition.

Our latent discriminative approach for object recognition is related to the work of Kumar and Herbert [11], [10], who train a discriminative model using fully-labeled data where each image region is assigned a part label from a discrete set of object parts. A CRF is trained and detection and segmentation are performed by finding the most likely labeling of the image under the learned model. The main difference between our approach and Kumar's is that we do not assume that the part assignment variables are fully observed, instead regarding them as latent variables. Incorporating hidden variables allows use of training data not explicitly labeled with part (hidden-state) structure.

Another related model is presented in [24], which builds a discriminative classifier based on a part-based feature representation. The main difference between their approach and ours is that we do not perform a preselection of discriminative parts. In parallel to our work on object recognition [18], [6] developed a hidden-state CRF model for phone recognition. A similiar model for natural language parsing is shown in [9].

Our models are related to hidden Markov random fields (HMRFs) [7]. Both HMRFs and HCRFs employ a layer of latent variables with an undirected graph specifying dependencies between those variables. However, there is the important difference that HMRFs model a joint distribution over latent variables and observations, whereas HCRFs are a discriminative model.

3 Hidden Conditional Random Fields

We assume a task where we wish to predict a label y from an input $\mathbf{x}$. Each y is a member of a set $\mathcal{Y}$ of possible labels and each vector $\mathbf{x}$ is a vector of local observations $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$.[1]

Each local observation x_j is represented by a feature vector $\phi(x_j) \in \Re^d$, where d is the dimensionality of the representation. Our training set consists of labeled examples $(\mathbf{x}_i, y_i)$ for $i = 1 \ldots n$, where each $y_i \in \mathcal{Y}$, and each $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,m}\}$. For example, in gesture recognition, $x_{i,j}$ might correspond to the jth frame of the ith video sequence and in the object recognition case it might correspond to the jth local feature of the ith image.

1. The number of local observations m can vary across examples; for convenience of notation, we omit dependence on the example index and simply refer to the number of observations as m in each case. In reality, m will vary across examples but this leads to minor changes to the model.

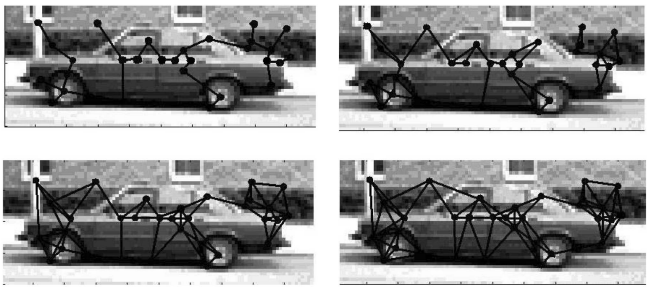


Fig. 1. Images show minimum spanning tree, 1-lattice (top), 2-lattice and 3-lattice (bottom) over detected features. Each circle corresponds to a local feature x_i and an edge between two circles i, j signifies a dependency between the corresponding hidden variables h_i and h_j in the model.

For any example $\mathbf{x}$, we also assume a vector of latent variables $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$, which are not observed on training examples, and where each h_j is a member of a finite set $\mathcal{H}$ of possible hidden labels in the model. Intuitively, each h_j corresponds to a labeling of x_j with some member of $\mathcal{H}$, which may correspond to "part" or "subgesture" structure in an observation. Given these definitions of labels y, observations $\mathbf{x}$, and latent variables $\mathbf{h}$, we define a conditional probabilistic model

$$P(y, \mathbf{h} \mid \mathbf{x}, \theta) = \frac{e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y', \mathbf{h}} e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}}, \tag{1}$$

where θ are the parameters of the model, and $\Psi(y, \mathbf{h}, \mathbf{x}; \theta) \in \Re$ is a potential function parameterized by θ. The model gives the following form for $P(y \mid \mathbf{x}, \theta)$:

$$P(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} P(y, \mathbf{h} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)}}{\sum_{y', \mathbf{h}} e^{\Psi(y', \mathbf{h}, \mathbf{x}; \theta)}}. \tag{2}$$

Given a new test example $\mathbf{x}$ and parameter values θ^* induced from a training set, we will take the label for the example to be $\arg\max_{y \in \mathcal{Y}} P(y \mid \mathbf{x}, \theta^*)$. Following previous work on CRFs [12], [11], we use the following objective function to estimate the parameters:

$$L(\theta) = \sum_i \log P(y_i \mid \mathbf{x}_i, \theta) - \frac{1}{2\sigma^2} \|\theta\|^2. \tag{3}$$

The first term in (3) is the log-likelihood of the data. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp(-\frac{1}{2\sigma^2} \|\theta\|^2)$. We use gradient ascent to search for the optimal parameter values, $\theta^* = \arg\max_\theta L(\theta)$, under this criterion. Note that, in general, $L(\theta)$ will be nonconvex, having multiple local minima, so the optimization method will in practice reach a local optimum of this function.

We encode structural constraints with an undirected graph structure, where the hidden variables $\{h_1, \ldots, h_m\}$ correspond to vertices in the graph. The set of graph edges $(j, k) \in E$ correspond

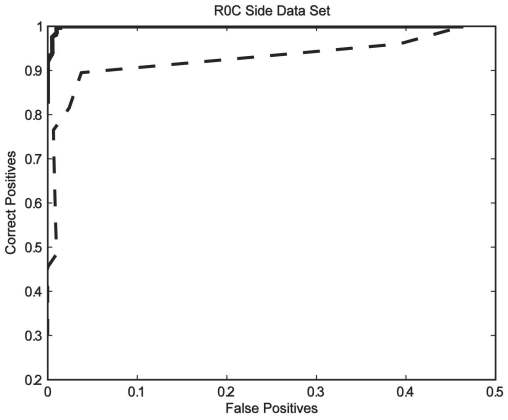


Fig. 3. ROC curves of car-side data set for models with different amounts of connectivity. The dotted line corresponds to a model with no connectivity. The solid lines correspond to models with minimum spanning tree, 2-lattice, and 3-lattice connectivity (these three models have near identical curves, so the solid lines are effectively superimposed).

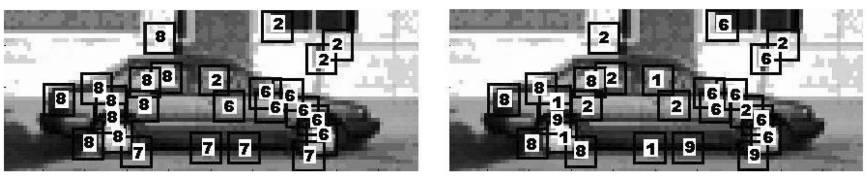
to links between variables h_j and h_k. The graph E can be defined arbitrarily; intuitively, it should capture any domain specific knowledge that we have about the structure of $\mathbf{h}$. In our object recognition task, it is a local mesh that encodes spatial consistency between local appearance features, while in our gesture recognition task it is a chain that captures temporal dynamics.

We define Ψ to take the following form:

$$\begin{aligned} \Psi(y, \mathbf{h}, \mathbf{x}; \theta) = &\sum_{j=1}^{m} \sum_{l \in L_1} f_{1,l}(j, y, h_j, \mathbf{x}) \theta_{1,l} \\ &+ \sum_{(j,k) \in E} \sum_{l \in L_2} f_{2,l}(j, k, y, h_j, h_k, \mathbf{x}) \theta_{2,l}, \end{aligned} \tag{4}$$

where L_1 is the set of node features, L_2 the set of edge features, $f_{1,l}$, $f_{2,l}$ are functions defining the features in the model, and $\theta_{1,l}$, $\theta_{2,l}$ are the components of θ, corresponding to node and edge parameters. The f_1 features depend on single hidden variable values in the model; the f_2 features can depend on pairs of values. Note that Ψ is linear in the parameters θ and the model in (1) is a log-linear model. Moreover, the features respect the structure of the graph, in that no feature depends on more than two hidden variables h_j, h_k, and if a feature does depend on variables h_j and h_k there must be an edge (j, k) in the graph E.

Assuming that the edges in E form a tree, and that Ψ takes the form in (4), then exact methods exist for inference and parameter estimation in the model. This follows because belief propagation can be used to calculate the following quantities in $O(|E||\mathcal{Y}||\mathcal{H}|^2)$ time:



(a) (b)

Fig. 2. Viterbi assignments of hidden states to local image patches for (a) minimum spanning tree and (b) unconnected model.

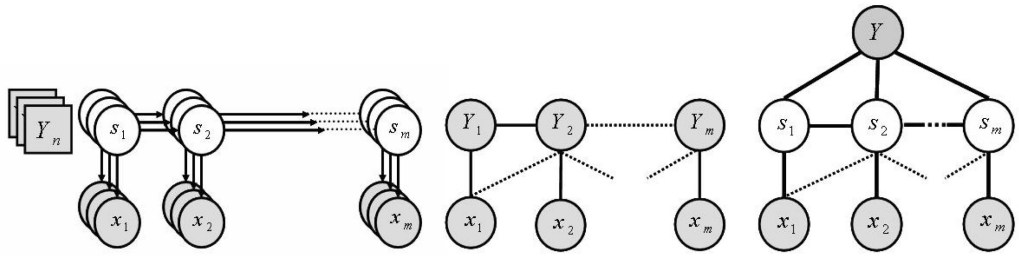


Fig. 4. Models used for comparative experiments on the gesture recognition task, Y is the gesture label and S the hidden state labels. (a) This figure shows a "stack of HMMs" model where a separate HMM is trained for each gesture class, (b) shows a CRF model, and (c) the proposed HCRF model.

$$\forall y \in \mathcal{Y}, \quad Z(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{h}} \exp\{\Psi(y, \mathbf{h}, \mathbf{x}; \theta)\}$$
$$\forall y \in \mathcal{Y}, j \in 1 \ldots m, a \in \mathcal{H},$$
$$P(h_j = a \mid y, \mathbf{x}, \theta) = \sum_{\mathbf{h}: h_j = a} P(\mathbf{h} \mid y, \mathbf{x}, \theta)$$
$$\forall y \in \mathcal{Y}, (j,k) \in E, a, b \in \mathcal{H},$$
$$P(h_j = a, h_k = b \mid y, \mathbf{x}, \theta) = \sum_{\mathbf{h}: h_j = a, h_k = b} P(\mathbf{h} \mid y, \mathbf{x}, \theta).$$

The first term $Z(y \mid \mathbf{x}, \theta)$ is a partition function defined by a summation over the $\mathbf{h}$ variables. Terms of this form can be used to calculate $P(y \mid \mathbf{x}, \theta) = Z(y \mid \mathbf{x}, \theta) / \sum_{y'} Z(y' \mid \mathbf{x}, \theta)$. Hence, inference —calculation of $\arg\max P(y \mid \mathbf{x}, \theta)$—can be performed efficiently in the model. The second and third terms are marginal distributions over individual variables h_j or pairs of variables h_j, h_k corresponding to edges in the graph. The gradient of $L(\theta)$ can be defined in terms of these marginals and can therefore be calculated efficiently. If E contains cycles then approximate methods, such as loopy belief-propagation, may be necessary for inference and parameter estimation.

We estimate parameters $\theta^* = \arg\max L(\theta)$ from a training set using a quasi-Newton gradient ascent method. We now describe how the gradient of $L(\theta)$ can be computed. The likelihood term due to the ith training example is:

$$L_i(\theta) = \log P(y_i \mid \mathbf{x}_i, \theta) = \log\left(\frac{\sum_{\mathbf{h}} e^{\Psi(y_i, \mathbf{h}, \mathbf{x}_i; \theta)}}{\sum_{y', \mathbf{h}} e^{\Psi(y', \mathbf{h}, \mathbf{x}_i; \theta)}}\right). \quad (5)$$

We first consider derivatives with respect to the parameters $\theta_{1,l}$ corresponding to features $f_{1,l}(j, y, h_j, \mathbf{x})$ that depend on single hidden variables. Taking derivatives gives

$$\begin{aligned}
\frac{\partial L_i(\theta)}{\partial \theta_{1,l}} &= \sum_{\mathbf{h}} P(\mathbf{h} \mid y_i, \mathbf{x}_i, \theta) \frac{\partial \Psi(y_i, \mathbf{h}, \mathbf{x}_i; \theta)}{\partial \theta_{1,l}} \\
&\quad - \sum_{y', \mathbf{h}} P(y', \mathbf{h} \mid \mathbf{x}_i, \theta) \frac{\partial \Psi(y', \mathbf{h}, \mathbf{x}_i; \theta)}{\partial \theta_{1,l}} \\
&= \sum_{\mathbf{h}} P(\mathbf{h} \mid y_i, \mathbf{x}_i, \theta) \sum_{j=1}^{m} f_{1,l}(j, y_i, h_j, \mathbf{x}_i) \\
&\quad - \sum_{y', \mathbf{h}} P(y', \mathbf{h} \mid \mathbf{x}_i, \theta) \sum_{j=1}^{m} f_{1,l}(j, y', h_j, \mathbf{x}_i) \\
&= \sum_{j,a} P(h_j = a \mid y_i, \mathbf{x}_i, \theta) f_{1,l}(j, y_i, a, \mathbf{x}_i) \\
&\quad - \sum_{y', j, a} P(h_j = a, y' \mid \mathbf{x}_i, \theta) f_{1,l}(j, y', a, \mathbf{x}_i).
\end{aligned}$$

It follows that $\frac{\partial L_i(\theta)}{\partial \theta_{1,l}}$ can be expressed in terms of components $P(h_j = a \mid \mathbf{x}_i, \theta)$ and $P(y \mid \mathbf{x}_i, \theta)$, which can be calculated using belief propagation, provided that the graph E forms a tree structure.[2]

A similar calculation gives

$$\begin{aligned}
\frac{\partial L_i(\theta)}{\partial \theta_{2,l}} &= \sum_{(j,k) \in E, a, b} P(h_j = a, h_k = b \mid y_i, \mathbf{x}_i, \theta) f_{2,l}(j, k, y_i, a, b, \mathbf{x}_i) \\
&\quad - \sum_{y', (j,k) \in E, a, b} P(h_j = a, h_k = b, y' \mid \mathbf{x}_i, \theta) f_{2,l}(j, k, y', a, b, \mathbf{x}_i),
\end{aligned}$$

hence $\partial L_i(\theta) / \partial \theta_{2,l}$ can also be expressed in terms of expressions that can be calculated using belief propagation.

4 Experiments

We explored the performance of our HCRF model on both object and gesture recognition tasks. In the object recognition experiments, we measured the effect of different degrees of connectivity in the mesh of local observations. In the gesture recognition experiments, we measured the effect of adding longer range dependencies in the model.

In our experiments, we use a restricted form of Ψ, where observations interact only with the hidden states,

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j \phi(x_j) \cdot \theta(h_j) + \sum_j \theta(y, h_j) + \sum_{(j,k) \in E} \theta(y, h_j, h_k), \quad (6)$$

where $\theta(h_j) \in \Re^d$ for $h_j \in \mathcal{H}$ is a parameter vector corresponding to the jth latent variable. The inner-product $\phi(x_j) \cdot \theta(h_j)$ can be interpreted as a measure of the compatibility between observation x_j and hidden-state h_j, the parameter $\theta(y, h_j) \in \Re$ for $h_j \in \mathcal{H}$, $y \in \mathcal{Y}$ can be interpreted as a measure of the compatibility between latent variable h_j and category label y, and each parameter $\theta(y, h_i, h_j) \in \Re$ for $y \in \mathcal{Y}$, and h_i, $h_j \in \mathcal{H}$ measures the compatibility between an edge with labels h_i and h_j and the label y.

4.1 Experiments on Object Recognition

In the object recognition domain, patches $x_{i,j}$ in each image are obtained using the SIFT detector [13], each patch $x_{i,j}$ is then represented by a feature vector $\phi(x_{i,j})$ that incorporates a combination of SIFT descriptor and relative location and scale features. We used a set $\mathcal{H}$ of hidden variables of size 10 for all experiments in this section. We assume that parts conditioned on proximate observations are likely to be dependent, as expressed in the neighborhood

2. Note that the terms $P(h_j = a, y' \mid \mathbf{x}_i, \theta)$ can be expressed as a product of terms $P(h_j = a \mid y', \mathbf{x}_i, \theta) \times P(y' \mid \mathbf{x}_i, \theta)$; these latter two terms can be calculated using belief propogation.

graph structure. We normally define proximity in terms of distance on the image plane but more generally it could include other attributes.

The graph E encodes the amount of connectivity between the hidden variables h_j. Intuitively, E determines the ability of our model to capture conditional dependencies between part assignments. Such dependencies can be encoded using n-neighbor lattices over local observations. Increasing connectivity leads, however, to an increase in the computational complexity of performing inference in such models. If E contains no edges the potential function for our model reduces to

$$\Psi(y, \mathbf{h}, \mathbf{x}; \theta) = \sum_j \phi(x_j) \cdot \theta(h_j) + \sum_j \theta(y, h_j). \quad (7)$$

This graph may be too poor to capture important dependencies between part assignments. Another option for defining E is to use a minimum spanning tree (MST) where the weights on the edges used to derive the MST are the distances between the corresponding image patches. The advantage of using such a graph is that because E contains no cycles, and Ψ takes the form in (4), we can perform exact inference on E as described above.

More generally, we can define E to be an n-lattice over the local observations. We build an n-neighbor lattice by linking every node to its n closest nodes, (i.e., the nodes that correspond to the n closest local observations). When E contains cycles computing exact inference becomes intractable so we need to resort to approximate methods; we use loopy belief-propagation.

We evaluated the effect of different neighborhood structures on recognition performance in a simple object category recognition task. We report results for the UIUC car-side data set. Given a neighborhood structure for our model, we trained a binary classifier to distinguish between a category and a background set formed from the remaining UIUC images.

For the first experiment, we defined E to be an unconnected graph (i.e., a graph with no edges), for the second a minimum spanning tree, for the third a 2-lattice, and for the fourth a 3-lattice, as shown in Fig. 1. For the first and second experiments, gradient ascent was initialized randomly, while for the third and fourth experiments, we used the minimum spanning tree solution as initial parameters. Fig. 3 shows ROC curves and associated equal error rates for the four variants of the model. From this figure, we observe a significant improvement in performance when the model incorporates some degree of dependency between the latent variables. Fig. 2 shows the most likely assignment of parts to features for the minimum spanning tree model and the unconnected model for an example in which the former gives a correct classification but the latter fails to do so. Both models appear to rely fairly strongly on the location features of each patch, as opposed to appearance features. However, the model with the MST structure shows a smoother assignment of the hidden-variable values: Nearby nodes in the graph tend to have the same value.

TABLE 1
Comparison of Recognition Performance (Percentage Accuracy) for Body Poses Estimated from Image Sequences on a 6-Way Classification Task

Arm Gesture	Avg. Accuracy(%)
HMM $\omega = 0$	84.83
CRF $\omega = 0$	86.03
CRF $\omega = 1$	81.75
HCRF (one-vs-all) $\omega = 0$	87.49
HCRF (multiclass) $\omega = 0$	91.64
HCRF (multiclass) $\omega = 1$	93.81
HCRF (multiclass) $\omega = 2$	93.07
HCRF (multiclass) $\omega = 3$	92.50

For this type of task, the minimum spanning tree model shows equivalent recognition performance to the models that use more densely connected graphs. Thus, it is clear that the minimum spanning tree can encode sufficient dependency constraints for certain categories. In [18], we conducted experiments comparing our model to a standard generative latent variable model [4] and found the average equal error rate of our model over all classes to be 96 percent and the one of the generative approach to be 92 percent when evaluated on the Caltech-4 data set.

4.2 Experiments on Gesture Recognition

We also explored our HCRF model on body and head gesture recognition, using motion features as the input representation. The task was to map a chain of observed motion features to a label denoting one of six possible gestures underlying the sequence. We evaluated HCRFs with varying levels of long range dependencies, and compared performance to baseline CRF and HMM models. Fig. 4 shows graphical representations of the HCRF, HMM, and CRF models used in our experiments.

In a first set of HCRF experiments, we trained HCRF models in a "one-versus-all" approach. For each gesture class, we first trained a separate HCRF model to discriminate the gesture class from other classes. For a given test sequence, we compared the probabilities given by each of the two-class HCRFs, and the highest scoring model was selected as the recognized gesture. In a second set of HCRF experiments, we trained a single joint multiclass HCRF to recognize all classes. Test sequences were run with this model and the gesture class with the highest probability was selected as the recognized gesture.

In the CRF experiments, each training or test sequence $\{x_1, x_2, x_3 \ldots x_m\}$ is associated with a sequence of labels $\{y_1, y_2, y_3, \ldots y_m\}$. In training data, the label sequences were taken to be the target label y for the gesture example repeated

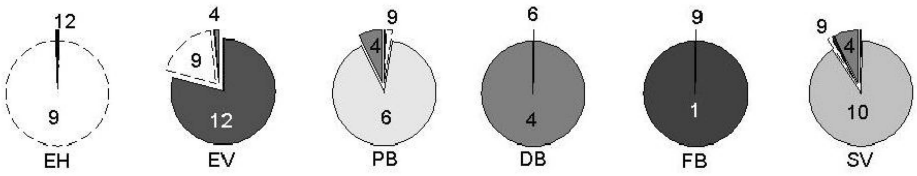


Fig. 5. Graph showing the distribution of the hidden states for each gesture class. The numbers in each pie represent the hidden state label, and the area enclosed by the number represents the proportion. EH, EV, ... SV are labels for the six different gesture types, see [23] for details.

TABLE 2
Comparison of Recognition Performance for Head Gestures

Models	Accuracy (%)
HMM $\omega = 0$	67.08
CRF $\omega = 0$	66.53
CRF $\omega = 1$	68.24
HCRF (multi-class) $\omega = 0$	71.88
HCRF (multi-class) $\omega = 1$	85.25

m times. For test examples, the most likely sequence of labels was decoded; the final label assigned to the test example was taken to be the label which appeared most frequently in the decoded sequence.

In both the HCRF and CRF models, we conducted experiments that incorporated different long range dependencies. To incorporate long range dependencies in the CRF and HCRF models, we add a window parameter ω that defines the amount of past and future history to be used when predicting the state at time t ($\omega = 0$ indicates only the current observation is used).

The HMMs were trained using maximum-likelihood estimation. We ran experiments with six hidden states for the one-versus-all HCRFs and 12 for the multiclass HCRFS; which states were shared among all the classes. For the HMM model, we used four hidden states for each class: These states were not shared among the different classes. The choice of four states was found to optimize performance on the test data (we tested values of 2, 4, 6, 8, 10, and 12 hidden states) and, the HMM results are therefore an upper bound on possible performance. We similarly optimized the number of Gaussian mixture components with respect to test data performance.

We ran experiments in two domains: arm and head gestures. In the arm gesture domain, we used a data set of gestures defined for a virtual manipulation task as described in [23].

From each observation of a user interacting with the system, a 3D cylindrical body model, consisting of a head, torso, arms, and forearms was estimated using a stereo-tracking algorithm [3]. From these body models, both the joint angles and the relative coordinates of the joints of the arms were used as observations for our experiments. Thirteen users were asked to perform these six gestures, an average of 90 gestures per class were collected.

Table 1 summarizes results for the arm gesture recognition experiments. In these experiments, the CRF performed better than HMMs at window size 0. At window size 1, however, the CRF performance was poorer. Both multiclass and one-versus-all HCRFs perform better than HMMs and CRFs. The most significant improvement in performance was obtained when we used a multiclass HCRF, suggesting that it is important to jointly learn the best discriminative structure.

It is surprising that increasing the window size from 0 to 1 degrades CRF performance since one would not expect that adding contextual features could harm the predictive power of the model. This performance drop may be caused by overfitting since adding contextual features increases the number of parameters of the model.

From the results in Table 1, we can see that incorporating some degree of long range dependencies is important since the HCRF performance improved when the window size was increased from 0 to 1. However, we also see that further increasing the window size did not improve performance.

Fig. 5 shows the distribution of states for different gesture classes learned by the best performing model (multiclass HCRF). 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.

We also conducted experiments with a head gesture data set obtained using the pose tracking system of [15]. A fast Fourier transform of the 3D angular velocities of users' head motion was used as input features. The data consisted of interactions between human participants and a robotic character [16]. A total of 16 participants interacted with a robot, with each interaction lasting between 2 to 5 minutes.

Table 2 summarizes the results for the head gesture experiments. The multiclass HCRF model performs better than the HMM and CRF models at a window size of 0. The HMM and CRF models have similar performance for the head gesture task. The HCRF multiclass model made a significant improvement when the window size was increased, which indicates that incorporating long range dependencies was useful.

5 Summary and Conclusions

We have developed a discriminative hidden-state model and demonstrated its utility on visual recognition tasks. Our model combines the ability of CRFs to use complex features of the input and the ability of HMMs to learn latent structure.

Our results have shown that HCRFs outperform both CRFs and HMMs for certain gesture recognition tasks. For arm gestures, the multiclass HCRF model outperforms HMMs and CRFs even when long range dependencies are not used, demonstrating the advantages of joint discriminative learning. For the object recognition data set, our results have shown that incorporating dependencies between latent variables is important and that the minimum spanning tree formulation can be a good approximation to more highly connected models.

References

[1] M. Assan and K. Groebel, "Video-Based Sign Language Recognition Using Hidden Markov Models," *Proc. Int'l Conf. Gesture Workshop: Gesture and Sign Language*, 1997.

[2] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996.

[3] D. Demirdjian and T. Darrell, "3-D Articulated Pose Tracking for Untethered Deictic Reference" *Proc. Int'l Conf. Multimodal Interfaces*, 2002.

[4] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003.

[5] S. Fujie, Y. Ejiri, K. Nakajima, Y. Matsusaka, and T. Kobayashi, "A Conversation Robot Using Head Gesture Recognition as Para-Linguistic Information," *Proc. 13th IEEE Int'l Workshop Robot and Human Comm.*, pp. 159-164, Sept. 2004.

[6] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden Conditional Random Fields for Phone Classification," *Proc. INTERSPEECH*, 2005.

[7] A.K.H. Kuensch and S. Geman, "Hidden Markov Random Fields," *Annals of Appied Probability*, vol. 5, 2005.

[8] A. Kapoor and R. Picard, "A Real-Time Head Nod and Shake Detector," *Proc. Workshop Perspective User Interfaces*, Nov. 2001.

[9] T. Koo and M. Collins, "Hidden-Variable Models for Discriminative Reranking," *Proc. IEEE Conf. Empirical Methods for Natural Language Processing*, 2005.

[10] S. Kumar and M. Hebert, "Multiclass Discriminative Fields for Parts-Based Object Detection," *Proc. Snowbird Learning Workshop*, 2004.

[11] S. Kumar and M. Herbert, "Discriminative Random Fields: A Framework for Contextual Interaction in Classification," *Proc. IEEE Int'l Conf. Computer Vision*, 2003.

[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data," *Proc. IEEE Int'l Conf. Machine Learning*, 2001.

[13] D. Lowe, "Object Recognition from Local Scale-Invariant Features," *Proc. IEEE Conf. Int'l Conf. Computer Vision*, 1999.

[14] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," *Proc. IEEE Conf. Empirical Methods for Natural Language Processing*, 2000.

[15] L.-P. Morency, A. Rahimi, and T. Darrell, "Adaptive View-Based Appearance Model," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.
[16] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Contextual Recognition of Head Gestures," *Proc. Int'l Conf. Multimodeal Interfaces,* 2005.
[17] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, pp. 677-695, 1997.
[18] A. Quattoni, M. Collins, and T. Darrell, "Conditional Random Fields for Object Recognition," *Proc. IEEE Conf. Neural Information Processing Systems,* 2004.
[19] A. Ratnaparkhi, "A Maximum Entropy Part-of-Speech Tagger," *Proc. IEEE Conf. Empirical Methods for Natural Language Processing,* 1996.
[20] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional Models for Contextual Human Motion Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.
[21] T. Starner and A. Pentland, "Real-Time ASL Recognition from Video Using Hidden Markov Models," *Proc. IEEE Symp. Computer Vision,* 1995.
[22] A. Torralba, K. Murphy, and W. Freeman, "Contextual Models for Object Detection Using Boosted Random Fields," *Proc. IEEE Conf. Neural Information Processing Systems,* 2004.
[23] S. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden Conditional Random Fields for Gesture Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.
[24] M. Yang, D. Roth, and N. Ahuja, "Learning to Recognize 3D Objects with Snow," *Proc. IEEE European Conf. Computer Vision,* 2000.