# Learning Discriminative Sequence Models from Partially Labelled Data for Activity Recognition

Tran The Truyen[1], Hung H. Bui[2,⋆], Dinh Q. Phung[1], and Svetha Venkatesh[1]

[1] Department of Computing, Curtin University of Technology
GPO Box U1987 Perth, Western Australia 6845, Australia
thetruyen.tran@postgrad.curtin.edu.au,
{d.phung,s.venkatesh}@curtin.edu.au
[2] Artificial Intelligence Center, SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025-3493, USA
bui@ai.sri.com

**Abstract.** Recognising daily activity patterns of people from low-level sensory data is an important problem. Traditional approaches typically rely on generative models such as the hidden Markov models and training on fully labelled data. While activity data can be readily acquired from pervasive sensors, e.g. in smart environments, providing manual labels to support fully supervised learning is often expensive. In this paper, we propose a new approach based on partially-supervised training of discriminative sequence models such as the conditional random field (CRF) and the maximum entropy Markov model (MEMM). We show that the approach can reduce labelling effort, and at the same time, provides us with the flexibility and accuracy of the discriminative framework. Our experimental results in the video surveillance domain illustrate that these models can perform better than their generative counterpart (i.e. the partially hidden Markov model), even when a substantial amount of labels are unavailable.

**Keywords:** activity recognition, discriminative models, partially labelled data, indoor video surveillance, conditional random fields, maximum entropy Markov models.

## 1 Introduction

An important task in human activity recognition from low-level noisy sensory data is segmenting the data streams and labeling them with meaningful sub-activities. The labels can then be used to facilitate data indexing and organisation, to recognise higher levels of semantics, and to provide useful context for intelligent assistive agents. To handle the uncertainty inherent in the data, current approaches to activity recognition typically employ probabilistic models

⋆ Hung Bui is supported by the Defense Advanced Research Projects Agency (DARPA), through the Department of Interior, NBC, Acquisition Services Division, under Contract No. NBCHD030010.

T.-B. Ho and Z.-H. Zhou (Eds.): PRICAI 2008, LNAI 5351, pp. 903–912, 2008.

such as the hidden Markov models (HMMs) and variants [1,2,7]. These models are essentially generative, i.e. they model the relation between the activity sequence $\mathbf{x}$ and the observable data stream $\mathbf{o}$ via the joint distribution $P(\mathbf{x}, \mathbf{o})$. However, it is often difficult to capture complex dependencies in the observation sequence $\mathbf{o}$, as typically, simplifying assumptions need to be made so that the conditional distribution $P(\mathbf{o}|\mathbf{x})$ is tractable. This limits the choice of features that one can use to encode multiple data streams. In addition, as we are often interested in finding the most probable activity sequence $\mathbf{x}^* = \arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{o})$, it is more natural to model $P(\mathbf{x}|\mathbf{o})$ directly.

Thus the discriminative model $P(\mathbf{x}|\mathbf{o})$ is more suitable to specify how an activity $\mathbf{x}$ would evolve *given* that we already observe a sequence of observations $\mathbf{o}$. With appropriate use of contextual information, the discriminative models can represent arbitrary, dynamic long-range interdependencies which are highly desirable for segmentation tasks.

Moreover, whilst capturing unlabeled sensor data for training is cheap, obtaining labels in a fully supervised setting often requires expert knowledge and is time consuming. In many cases we are certain about some particular labels, for example, in surveillance data, when a person enters a room or steps on a pressure mat. Other labels (e.g. other activities that occur inside the room) are left unknown. Therefore, it is more desirable to employ the partially-supervised approach in that some labels are missing in the training data. Specifically, we consider two recent discriminative models, namely, the undirected Conditional Random Fields (CRFs) [3], (Figure 1(b)) and the directed Maximum Entropy Markov Models (MEMMs) [5] (Figure 1(a)). As the original models require full labels, we provide a treatment of incomplete data for the CRFs and the MEMMs. The treatment mainly contributes to the main novelty of this paper despite the fact that there have been recent attempts to apply discriminative models for activity recognition, [4,9,10,6,11]. Note that the work in [6] also investigates hidden variables in modelling activity, this is for discovering latent aspects rather than for reducing labelling effort.

We provide experimental results in the video surveillance domain where we compare the performance of the proposed models and the equivalent partially hidden Markov models (PHMMs) [8] (Figure 1(c)) in learning and segmenting human indoor movement patterns. Out of three data sets studied, a common behaviour is that the HMM is outperformed by the discriminative counterparts even when a large portion of labels are missing. Providing contextual features for the models increases the performance significantly.

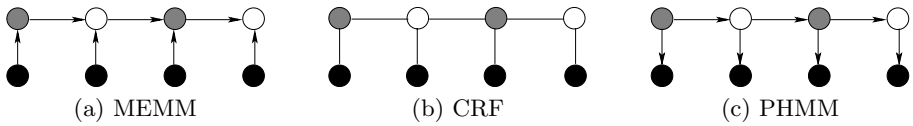


**Fig. 1.** (a,b): The partially labelled discriminative models, and (c): partially hidden Markov models. Filled circles and bars are data observations, empty circles are hidden labels, shaded circles are the visible labels.

The remainder of the paper is organised as follows. Section 2 provides background on CRFs and MEMMs. Section 3 describes learning discriminative models under missing labels. The paper then describes implementation and experiments and presents results in Section 4. The final section summarises major findings and further work.

## 2 Background

This section briefly reviews the MEMMs and the CRFs for sequence modelling. Given a data sequence $\mathbf{o}$ of length $T$, the MEMMs define the conditional distribution of the activity labelling $\mathbf{x}$ as follows

$$P(\mathbf{x}|\mathbf{o}) = P(\mathbf{x}_1|\mathbf{o}) \prod_{t=2}^{T} P(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{o}), \text{ where,} \tag{1}$$

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{o}) = \frac{1}{Z(\mathbf{o}, \mathbf{x}_{t-1})} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{o}, \mathbf{x}_{t-1}, \mathbf{x}_t)) \ , \tag{2}$$

where $Z(\mathbf{o}, \mathbf{x}_{t-1}) = \sum_{\mathbf{x}_t} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{o}, \mathbf{x}_{t-1}, \mathbf{x}_t))$. The functions $\mathbf{f}(\mathbf{o}, \mathbf{x}_{t-1}, \mathbf{x}_t)$ are the features that capture the statistics of the observational data and the activities and their transition at time $t$. The parameters $\mathbf{w}$ are the weights associated with the features and are estimated through training.

Thus, a MEMM is a directed Markov chain conditioned on the observational data $\mathbf{o}$. In supervised training, all activity labels $\{\mathbf{x}_t\}_{t=1}^{T}$ are given, so only local classifiers $P(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{o})$ are learnt. During inference time, however, since no history labels are given for those local classifiers, Viterbi decoding must be used for simultaneous labelling. Since learning is conditioned on the previous labels, if the previous labels allows only limited transition to the current labels, a phenomenon known as *label-bias* will occur.

The CRFs, on the other hand, do not suffer from this drawback as they model the activity sequence entirely

$$P(\mathbf{x}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})} \prod_{t=2}^{T} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{o}, \mathbf{x}_{t-1}, \mathbf{x}_t)) \ , \tag{3}$$

where $Z(\mathbf{o}) = \sum_{\mathbf{x}} \prod_{t=2}^{T} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{o}, \mathbf{x}_{t-1}, \mathbf{x}_t))$. Since the computation of $Z(\mathbf{o})$ has the standard sum-product form, we can use dynamic programming at the cost of $\mathcal{O}(T)$ time. Thus, a CRF is a undirected Markov chain conditioned on the observational data $\mathbf{o}$.

Fully supervised learning in the CRFs and MEMMs typically maximises the conditional log-likelihood[1] $\mathcal{L}(\mathbf{w}) = \log P(\mathbf{x}|\mathbf{o})$.

[1] For multiple iid data instances, we should write $\mathcal{L}(\mathbf{w}) = \sum_x \tilde{P}(\mathbf{o}) \log P(\mathbf{x}|\mathbf{o})$ where $\tilde{P}(\mathbf{o})$ is the empirical distribution of training data, but we drop this notation for clarity.

# 3 Learning Discriminative Models from Partially Labelled Data

In our partially labelled discriminative models, the label sequence $\mathbf{x}$ consists of a visible component $\mathbf{v}$ (e.g. labels that are provided manually, or are acquired automatically by reliable sensors) and a hidden part $\mathbf{h}$ (labels that are left unspecified or those we are unsure), that is $\mathbf{x} = (\mathbf{v}, \mathbf{h})$. The joint distribution of all visible variables $\mathbf{v}$ is therefore given as

$$P(\mathbf{v}|\mathbf{o}) = \sum_{\mathbf{h}} P(\mathbf{v}, \mathbf{h}|\mathbf{o}) = \sum_{\mathbf{h}} P(\mathbf{x}|\mathbf{o}) \ . \tag{4}$$

To learn the model parameters that are best explained by the data, we maximise the penalised log-likelihood

$$\Lambda(\mathbf{w}) = \mathcal{L}(\mathbf{w}) - \frac{1}{2\sigma^2}\|\mathbf{w}\|^2 \ ,$$

where $\mathcal{L}(\mathbf{w}) = \log P(\mathbf{v}|\mathbf{o})$. The regularisation term is needed to avoid over-fitting when only limited data is available for training. For simplicity, the parameter $\sigma$ is shared among all dimensions and is selected experimentally.

As with incomplete data, an alternative to maximise the log-likelihood is using the EM algorithm whose Expectation (E-step) at step $j$ is to calculate the quantity

$$Q(\mathbf{w}^j, \mathbf{w}) = \sum_{\mathbf{h}} P(\mathbf{h}|\mathbf{v}, \mathbf{o}; \mathbf{w}^j) \log P(\mathbf{h}, \mathbf{v}|\mathbf{o}) \ , \tag{5}$$

and the Maximisation (M-step) maximises the concave lower bound of the log-likelihood $Q(\mathbf{w}^j, \mathbf{w}) - \frac{1}{2\sigma^2}||\mathbf{w}||^2$ with respect to $\mathbf{w}$. Unlike Bayesian networks, the log-linear models do not yield closed form solutions in the the M-step. However, as the function $Q(\mathbf{w}^j, \mathbf{w})$ is concave, it is still advantageous to optimise with efficient Newton-like algorithms.

## 3.1 Learning MEMMs

Directed models like the MEMMs are important in activity modeling because they naturally encode the state transitions given the observations. As we are free to encode arbitrary information exacted from the whole sequence $\mathbf{o}$ to the local distribution, we use a sliding window $\Omega_t$ of size $s$ centred at the current time $t$ to capture the local context of the observation. The joint incomplete distribution is therefore

$$P(\mathbf{v}|\mathbf{o}) = \sum_{\mathbf{h}} P(\mathbf{x}_1|\Omega_1) \prod_{t=2}^{T} P(\mathbf{x}_t|\Omega_t, \mathbf{x}_{t-1}) \ . \tag{6}$$

Since this is a standard sum-product problem, dynamic programming can be used to solve in $\mathcal{O}(T)$ time.

In learning of MEMMs using EM, the E-step is to calculate

$$Q(\mathbf{w}^j, \mathbf{w}) = \sum_t \sum_{\mathbf{h}_{t-1}} P(\mathbf{h}_{t-1}|\mathbf{v}, \Omega_t^j) \sum_{\mathbf{h}_t} P(\mathbf{h}_t|\mathbf{h}_{t-1}, \Omega_t^j) \log P(\mathbf{h}_t|\mathbf{h}_{t-1}, \Omega_t) \ . \quad (7)$$

and the M-step is to solve the zeroing gradient equation

$$\nabla Q(\mathbf{w}^j, \mathbf{w}) = \sum_t \sum_{\mathbf{h}_{t-1}} P(\mathbf{h}_{t-1}|\mathbf{v}, \Omega_t^j) \left\{ \sum_{\mathbf{h}_t} P(\mathbf{h}_t|\mathbf{h}_{t-1}, \Omega_t^j) \mathbf{f}(\mathbf{h}_{t-1}, \mathbf{h}_t, \Omega_t) - \right.$$
$$\left. \sum_{\mathbf{x}_t} P(\mathbf{x}_t|\mathbf{h}_{t-1}, \Omega_t) \mathbf{f}(\mathbf{h}_{t-1}, \mathbf{x}_t, \Omega_t) \right\} \ .$$

Computation of the EM reduces to that of marginals and state transition probabilities, which can be carried out efficiently in the Markov chain framework using dynamic programming.

### 3.2 Learning CRFs

From Eq. 3, we have

$$P(\mathbf{v}|\mathbf{o}) = \frac{1}{Z(\mathbf{o})} \sum_{\mathbf{h}} \exp(\mathbf{w}^\top \mathbf{f}(\mathbf{o}, \mathbf{x}_{t-1}, \mathbf{x}_t)) \ . \quad (8)$$

In this case, the complexity of computing $P(\mathbf{v}|\mathbf{o})$ is the same as that of computing the partition function $Z(\mathbf{o})$ up to a constant.

For the partially labelled CRFs, the gradient of incomplete likelihood reads

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial \mathbf{w}_k} = \sum_t \left( \sum_{\mathbf{h}_{t-1}, \mathbf{h}_t} P(\mathbf{h}_{t-1}, \mathbf{h}_t|\mathbf{v}, \mathbf{o}) f_k(\mathbf{h}_{t-1}, \mathbf{h}_t, \mathbf{v}, \mathbf{o}) - \right.$$
$$\left. \sum_{\mathbf{x}_{t-1}, \mathbf{x}_t} P(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{o}) f_k(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{o}) \right) \ . \quad (9)$$

Zeroing the gradient does not yield an analytical solution, so typically iterative numerical methods such as conjugate gradient and Newton methods are needed. The gradient of the lower bound in the EM framework of (5) is similar to (9), except that the pairwise marginals $P(\mathbf{h}_{t-1}, \mathbf{h}_t|\mathbf{v}, \mathbf{o})$ are now replaced by the marginals of the previous EM iteration $P(\mathbf{h}_{t-1}, \mathbf{h}_t|\mathbf{v}, \mathbf{o}^j)$. The pairwise marginals $P(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{o})$ can be computed easily using a forward pass and a backward pass in the standard message passing scheme on the chain.

### 3.3 Comparison with the PHMMs

The main difference between the models described in this section (Figure 1(a,b)) and the PHMMs [8] (Figure 1(c)) is the conditional distribution $P(\mathbf{x}|\mathbf{o})$ in discriminative models compared to the joint distribution $P(\mathbf{x}, \mathbf{o})$ in the PHMMs.

The data distribution of $P(\mathbf{o})$ and how $\mathbf{o}$ is generated are not of concern in the discriminative models. In the PHMMs, on the contrary, the observation point $\mathbf{o}_t$ is presumably generated by the parent label node $\mathbf{x}_t$, so care must be taken to ensure proper conditional independence among $\{\mathbf{o}_t\}_{t=1}^{T}$. This difference has an implication that, while the discriminative models may be good to encode the output labels directly with arbitrary information extracted from the whole observation sequence $\mathbf{o}$, the PHMMs better represent $\mathbf{o}$ when little information is associated with $\mathbf{x}$. For example, when $\mathbf{x}$ is totally missing, $P(\mathbf{o}) = \sum_{\mathbf{x}} P(\mathbf{x}, \mathbf{o})$ is still modeled in the PHMMs and provides useful information.

## 4 Experiments and Results

Our task is to infer the activity patterns of a person (the actor) in a video surveillance scene. The observation data is provided by static cameras while the labels, which are activities such as *'go-from-A-to-B'* during the time interval $[t_a, t_b]$ (see Table 1), are recognised by the trained models.

### 4.1 Setup and Data

The surveillance environment is a $4 \times 6m^2$ dining room and kitchen (Figure 2). Two static cameras are installed to capture the video of the actor making some meals. There are six landmarks which the person can visit during the meals: door, TV chair, fridge, stove, cupboard, and dining chair.

We study three scenarios corresponding to the person making a short meal (denoted by SHORT_MEAL), having a snack (HAVE_SNACK), and making a normal meal (NORMAL_MEAL). Each scenario comprises of a number of primitive activities as listed in Table 1. The association between scenarios and their primitive activities are: SHORT_MEAL = {1,2,3,4,11}, HAVE_SNACK = {2,5,6,7,8}, and NORMAL_MEAL = {1,2,4,9,10,11,12}. The SHORT_MEAL data set has 12 training and 22 testing video sequences; and each of the HAVE_SNACK

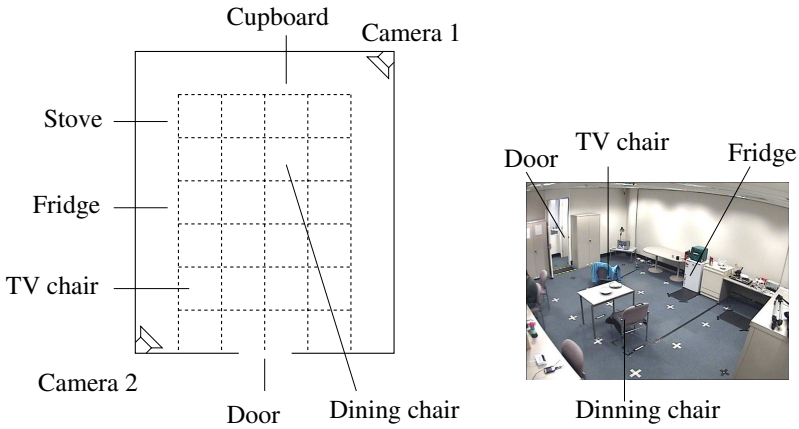


**Fig. 2.** The environment and scene viewed from one of the two cameras

**Table 1.** The primitive activities (the labels)

| Activity | Landmarks | Activity | Landmarks |
|---|---|---|---|
| 1 | Door→Cupboard | 7 | Fridge→TV chair |
| 2 | Cupboard→Fridge | 8 | TV chair→Door |
| 3 | Fridge→Dining chair | 9 | Fridge→Stove |
| 4 | Dining chair→Door | 10 | Stove→Dining chair |
| 5 | Door→TV chair | 11 | Fridge→Door |
| 6 | TV chair→Cupboard | 12 | Dining chair→Fridge |

and NORMAL_MEAL data sets consists of 15 training and 11 testing video sequences. For each raw video sequence captured, we use a background subtraction algorithm to extract a corresponding discrete sequence of coordinates of the person based on the person's bounding box. The training sequences are partially labeled, indicated by the portion of missing labels $\rho$. The testing sequences provide the ground-truth for the algorithms. The sequence length ranges from $T = 20-60$ and the number of labels per sequence is allowed to vary as $T*(1-\rho)$ where $\rho \in [0, 100\%]$.

We apply standard evaluation metrics such as precision $P$, recall $R$, and the $F1$ score given as $F1 = 2 * P * R/(P + R)$ on a per-token basis.

### 4.2 Feature Design and Contextual Extraction

Features are crucial components of the model as they tie raw observation data with semantic outputs (i.e. the labels). The features need to be discriminative enough to be useful, and at the same time, they should be as simple and intuitive as possible to reduce manual labour. The current raw data extracted from the video contains only $(X, Y)$ coordinates. From each coordinate sequences, at each time slice $t$, we extract a vector of five elements from the observation sequence $g(\mathbf{o}, t) = (X_t, Y_t, u_{X_t}, u_{Y_t}, s_t = \sqrt{u_{X_t}^2 + u_{Y_t}^2})$, which correspond to the $(X, Y)$ coordinates, the $X$ & $Y$ velocities, and the speed, respectively. Since the extracted coordinates are fairly noisy, we use the average velocity measurement within a time interval of small width $w$, i.e. $u_{X_t} = (X_{t+w/2} - X_{t-w/2})/w$. Typically, these observation-based features are real numbers and are normalised so that they have a similar scale.

We decompose the feature set $\{f_k(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{o})\}$ into two subsets: the *state-observation* features $f_{l,m,\epsilon}(\mathbf{o}, \mathbf{x}_t) := \mathbb{I}[\mathbf{x}_t = l]\mathbf{h}_m(\mathbf{o}, t, \epsilon)$ and the *state-transition* features $f_{l_1,l_2}(\mathbf{x}_{t-1}, \mathbf{x}_t) := \mathbb{I}[\mathbf{x}_{t-1} = l_1]\mathbb{I}[\mathbf{x}_t = l_2]$, where $m = 1..5$ and $\mathbf{h}_m(\mathbf{o}, t, \epsilon) = g_m(\mathbf{o}, t + \epsilon)$ with $\epsilon = -s_1, ..0, ..s_2$ for some positive integers $s_1$, $s_2$. The state-observation features in thus incorporate neighbouring observation points within a sliding window of width $s = s_1 + s_2 + 1$.

To have a rough idea of how the observation context influences the performance of the models, we try different window sizes $s$ (see Equation (1)). The experiments show that incorporating the context of observation sequences does help to improve the performance significantly (see Figure 3). We did not try exhaustive searches

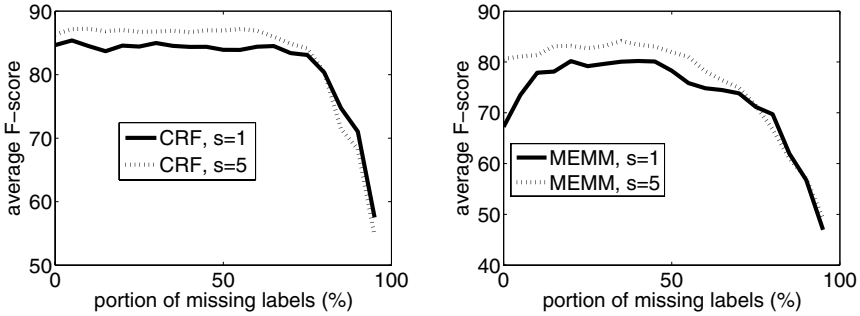


**Fig. 3.** The role of context (SHORT_MEAL), $s$: the window size to extract observation data. (a) CRFs, (b) MEMMs. In all figures, the x-axis: the portion of missing labels (%) and the y-axis: the averaged F-score (%) over all states and over 10 repetitions.

for the best context size, nor did we implement any feature selection mechanisms. As the number of features scales linearly with the context size as $K = 5s|Y| + |Y|^2$, where $s$ can be any integer between 1 and $T$, where $T$ is the sequence length, clearly a feature selection algorithm is needed when we want to capture long range correlation. For the practical purposes of this paper, we choose $s = 5$ for both CRFs and MEMMs. Thus in our experiments, CRFs and MEMMs share the same feature set, making the comparison between the two models consistent.

### 4.3 Performance of Models

To evaluate the performance of discriminative models against the equivalent generative counterparts, we implement the PHMMs (Figure 1(c)). The features extracted from the sensor data for the PHMMs include the discretised position and velocity. These features are different from those used in discriminative models in that discriminative features can be continuous. Thus the feature set used by PHMMs is different from those shared by CRFs and MEMMs. Although the difference may raise the concern about the compatibility of these models, it is precisely where discriminative models are more flexible as they have no difficulty selecting features.

To train discriminative models, we implement the non-linear conjugate gradient (CG) of Polak-Ribière and the limited memory quasi-Newton L-BFGS. After several pilot runs, we select the L-BFGS to optimise the objective function in (5) directly. In the case of MEMMs, the regularised EM algorithm is chosen together with the CG. The algorithms stop when the rate of convergence is less than $10^{-5}$. The regularisation constants are empirically selected as $\sigma = 5$ in the case of CRFs, and $\sigma = 20$ in the case of MEMMs.

For the PHMMs, it is observed that the initial parameter initialisation is critical to learn the correct model. Random initialisations often result in very poor performance. This is unlike the discriminative counterparts in which all initial parameters can be trivially set to zeros (equally important).

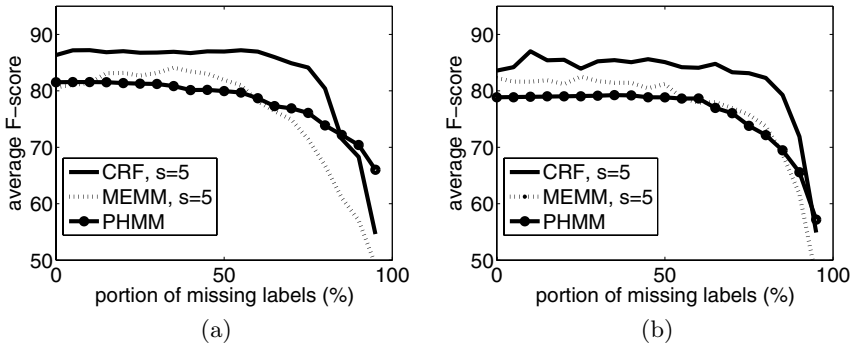


**Fig. 4.** Average performance of models (a: SHORT_MEAL, b: NORMAL_MEAL). x-axis: portion of missing labels (%) and y-axis: the averaged F-score (%) over all states and 10 repetitions.

Overall in our experiments (Figure 4) the generative PHMMs are outperformed by the discriminative counterparts in all cases given sufficient labels. This clearly matches the theoretical differences between these models in that when there are enough labels, richer information can be extracted in the discriminative framework, i.e. modeling $P(\mathbf{x}|\mathbf{o})$ is more suitable. On the other hand, when only a few labels are available, the unlabeled data is important so it makes sense to model and optimise $P(\mathbf{o}, \mathbf{x})$ as in the generative framework. On all data sets, the CRFs outperform the other models. These behaviours are consistent with the results reported in [3] in the fully observed setting. MEMMs are known to suffer from the label-bias problem [3], thus their performance does not match that of CRFs, although MEMMs are better than HMMs given enough training labels. In the HAVE_SNACK data set, the performance of MEMMs is surprisingly good.

A striking fact about the globally normalised CRFs is that the performance persists until most labels are missing. This is clearly a big time and effort saving for the labeling task.

## 5 Conclusions and Further Work

In this work, we have presented a partially-supervised framework for activity recognition on low-level noisy data from sensors using discriminative models. We illustrated the appropriateness of the discriminative models for segmentation of surveillance video into sub-activities. As more flexible information can be encoded using feature functions, the discriminative models can perform significantly better than the equivalent generative HMMs even when a large portion of the labels are missing. CRFs appear to be a promising model as the experiments show that they consistently outperform other models in all three data sets. Although less expressive than CRFs, MEMMs are still an important class of models as they enjoy the flexibility of the discriminative framework and enable online recognition as in directed graphical models.

Our study shows that primitive and intuitive contextual features work well in the area of video surveillance. However, to obtain the optimal context and to make use of the all information embedded in the whole observation sequence, a feature selection mechanism remains to be designed in conjunction with the models and training algorithms presented in this paper.

## References

1. Brand, M., Oliver, N., Pentland, A.: Coupled hidden Markov models for complex action recognition. In: Proc. CVPR, pp. 994–999 (1997)
2. Bui, H.H., Venkatesh, S., West, G.: Policy recognition in the abstract hidden Markov model. Journal of Articial Intelligence Research 17, 451–499 (2002)
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. ICML, pp. 282–289 (2001)
4. Liao, L., Fox, D., Kautz, H.: Location-Based Activity Recognition using Relational Markov Networks. In: Proc. IJCAI, pp. 773–778 (2005)
5. McCallum, A., Freitag, D., Pereira, F.: Maximum Entropy Markov models for information extraction and segmentation. In: Proc. ICML, pp. 591–598 (2000)
6. Morency, L.P., Quattoni, A., Darrell, T.: Latent-Dynamic Discriminative Models for Continuous Gesture Recognition. In: Proc. CVPR, pp. 1–8 (2007)
7. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. CVIU 96, 163–180 (2004)
8. Scheffer, T., Wrobel, S.: Active learning of partially labelled Markov models. In: Active Learning, Database Sampling, Experimental Design: Views on Instance Selection, Workshop at ECML 2001/PKDD 2001 (2001)
9. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. CVIU 104, 210–220 (2006)
10. Vail, D.L., Veloso, M.M., Lafferty, J.D.: Conditional random fields for activity recogntion. In: Proc. AAMAS (2007)
11. Wu, T., Lian, C., Hsu, J.Y.: Joint Recognition of Multiple Concurrent Activities using Factorial Conditional Random Fields. In: AAAI Workshop on Plan, Activity, and Intent Recognition (2007)
12. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden Markov models. In: Proc. CVPR (1992), pp. 379–385 (1992)