Conditional Models for Contextual Human Motion Recognition

Cristian Sminchisescu[1,2,3] **Atul Kanaujia**[3] **Zhiguo Li**[3] **Dimitris Metaxas**[3]
[1]TTI-C, 1497 East 50th Street, Chicago, USA
[2]Department of Computer Science, University of Toronto, Canada, *crismin@cs.toronto.edu*
[3]Department of Computer Science, Rutgers University, USA, {*kanaujia,zhli,dnm*}*@cs.rutgers.edu*

Abstract

We present algorithms for recognizing human motion in monocular video sequences, based on discriminative Conditional Random Field (CRF) and Maximum Entropy Markov Models (MEMM). Existing approaches to this problem typically use generative (joint) structures like the Hidden Markov Model (HMM). Therefore they have to make simplifying, often unrealistic assumptions on the conditional independence of observations given the motion class labels and cannot accommodate overlapping features or long term contextual dependencies in the observation sequence. In contrast, conditional models like the CRFs seamlessly represent contextual dependencies, support efficient, exact inference using dynamic programming, and their parameters can be trained using convex optimization. We introduce conditional graphical models as complementary *tools for human motion recognition and present an extensive set of experiments that show how these typically outperform HMMs in classifying not only diverse human activities like walking, jumping, running, picking or dancing, but also for discriminating among subtle motion styles like normal walk and wander walk.*

Keywords: *Markov random fields, discriminative models, Hidden Markov Models, human motion recognition, multiclass logistic regression, feature selection, conditional models, optimization.*

1 Introduction

Tracking and recognizing human motion in natural environments provides a basic infrastructure for the advancement of several technologies that enable adaptive visual assistants for intelligent human-computer interfaces or systems for entertainment, surveillance and security. Tracking is complex due to the large variability in the shape and articulation of the human body, the presence of clothing or fast motions. Highly variable lighting conditions or occlusion from other people or objects further complicate the problem.

Motion class recognition on the other hand is challenging because human motion lacks a clear categorical structure: the motion can be often classified into several categories simultaneously, because some activities have a natural compositional (or concurrent) structure in terms of basic action units (run and hand-wave, walk and shake hands while involved in a conversation with somebody known) and because even the transition between simple activities naturally has temporal segments of ambiguity and overlap.

Perhaps most importantly, similar motions can happen at various timescales and because they often exhibit long-term dependencies, *context needs to be considered for correct classification*. For instance, the motion class may be hard to predict at a particular point in time using only the previous state and the current image observation alone, but may be less ambiguous if several neighboring states or observations possibly both backward and forward in time are considered. However, this behavior would be hard to model using a Hidden Markov Model (HMM) [20], where stringent independence assumptions among observations are required in order to ensure computational tractability (notably, conditional independence of observations given the class labels).

HMMs and more generally the class of stochastic grammars, are generative models that define a joint probability distribution $p(\mathbf{X}, \mathbf{R})$ over observations $\mathbf{R}$ and motion label sequences $\mathbf{X}$, and use Bayes rule to compute $p(\mathbf{X}|\mathbf{R})$. In order to *model the observation* process and enumerate all possible sequences of observations, generative models need to assume them as being atomic and independent. Therefore they can't accommodate multiple overlapping features of the observation or long-range dependencies among observations at multiple time steps, because the inference problem for such models becomes intractable. Arguably, another inconvenient of using generative models like HMMs stems from their indirection: they use a joint model to solve a conditional problem thus focusing on modeling the observations that at runtime are fixed anyway. Even if the generative model were accurate, this approach could be non-economical in cases where the underlying generative model may be quite complex, but the motion class *conditioned on the observation* (or the boundary between classes) is nevertheless simple.

In this paper we advocate a *complementary* discriminative approach to human motion recognition based on extensions to Conditional Random Fields (CRF) [14] and Maximum Entropy Markov Models (MEMM) [16]. A CRF conditions on the observation without modeling it, therefore it avoids independence assumptions and can accommodate long range interactions among observations at different timesteps. Our approach is based on non-locally defined, multiple features of the observation, represented as log-linear models, that can be seen as a generalization of logistic regression to account for

correlations among successive class labels. Inference can be performed efficiently using dynamic programming and training the parameters is based on a convex problem, with guaranteed global optimality. We demonstrate the algorithms on the task of both recognizing broader classes of human motions like walking, running, jumping, conversation or dancing, but also for finely discriminating among motion styles like slow walk or wander walk. We compare against HMMs and demonstrate that the conditional models can significantly improve recognition performance in tests that use both features extracted from 3d reconstructed joint angles, but also in recognition experiments that use feature descriptors extracted directly from image silhouettes.

1.1 Related Work

The research devoted to human motion recognition is extensive, accounting for its clear social and technological importance. We refer to [2, 9, 5] for comprehensive surveys and here aim only at a brief literature overview. HMMs [20] and their various extensions have been successfully used for recognizing human motion based on both 2d observations [26, 6, 7, 10] and 3d observations [29, 21]. Generative approaches to tracking and motion classification have also been proposed by [4, 17], where the variation within each motion class is represented as an auto-regressive process, or a linear dynamical system, whereas learning and inference are based on Condensation and variational techniques, respectively. Black & Jepson [3] model motion as a trajectory through a state space spanned by optical flow fields and infer the activity class based on propagating multiple hypotheses using a Condensation filter. Fablet & Bouthemy [8] present a powerful approach to recognition using multiscale Gibbs models. Shi *et al* [22] employ a P-net and a discrete Condensation algorithm in order to better take into account sequential activities that include parallel streams of action.

We are not aware of conditional approaches previously applied to human motion recognition (temporal chains), but discriminative models have been successfully demonstrated in spatial inference, for the propose of detecting man-made or natural structures in images [13, 11, 28, 18]. There is a relation between CRFs/MEMMs and the so-called 'sliding window' methods [19] that essentially predict the current state label independently by considering a window centered at the current observation. However such methods do not account for correlations between neighboring, temporal state labels, as common in motion recognition problems.

2 Conditional Random Fields

We work with graphical models with a linear chain structure, as shown in fig. 1. These have discrete temporal states x_t, here discrete motion class labels $x \in \mathcal{X} = \{1, 2, \ldots, c\}$, $t = 1 \ldots T$, prior $p(x_1)$, observations $\mathbf{r}_t$, with $\dim(\mathbf{r}) = r$. For notational compactness, we also consider joint states $\mathbf{X}_t = (x_1, x_2, \ldots, x_t)$ or joint observations $\mathbf{R}_t = (\mathbf{r}_1, \ldots, \mathbf{r}_t)$. Sometimes we drop the subscript, *i.e.* $\mathbf{X}_T = \mathbf{X}$ and $\mathbf{R}_T = \mathbf{R}$, for brevity.

Let $G = (V, E)$ be a graph and $\mathbf{X}$ being indexed by the vertices of G, say x_i. A pair $(\mathbf{X}, \mathbf{R})$ is called a Conditional Random Field (CRF) [14], if when conditioning on $\mathbf{R}$, the variables x_i obey the Markov property w.r.t. the graph: $p(x_i|\mathbf{R}, \mathbf{X}_{V-\{i\}}) = p(x_i|\mathbf{R}, \mathbf{X}_{\mathcal{N}_i})$, where $\mathcal{N}_i$ is the set of neighbors of node i and $\mathbf{X}_{\mathcal{N}_i}$ is the joint vector of variables in the subscript set. Let $C(\mathbf{X}, \mathbf{R})$ be the set of maximal cliques of G. Using the Hammersley Clifford theorem [12], the distribution over joint labels $\mathbf{X}$ given observations $\mathbf{R}$ and parameters $\boldsymbol{\theta}$, can be written as an expansion:

$$p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{R}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{R})} \prod_{c \in C(\mathbf{X},\mathbf{R})} \phi^c_{\boldsymbol{\theta}}(\mathbf{X}_c, \mathbf{R}_c) \qquad (1)$$

where $\phi^c_{\boldsymbol{\theta}}$ is the positive-valued potential function of clique c, and $Z_{\boldsymbol{\theta}}(\mathbf{R})$ is the *observation dependent* normalization:

$$Z_{\boldsymbol{\theta}}(\mathbf{R}) = \sum_{\mathbf{X}} \prod_{c \in C(\mathbf{X},\mathbf{R})} \phi^c_{\boldsymbol{\theta}}(\mathbf{X}_c, \mathbf{R}_c) \qquad (2)$$

For a linear chain (first-order state dependency), the cliques include pairs of neighboring states (x_{t-1}, x_t), whereas the connectivity among observations is unrestricted, as these are known and fixed (see fig. 1b). Therefore, arbitrary clique structures that include complex observation dependencies do not complicate inference. For a model with T timesteps, the CRF in (1) can be rewritten in terms of exponentiated feature functions $F_{\boldsymbol{\theta}}$, computed in terms of weighted sums over the features of the cliques, *c.f.* (3) and (12):[1]

$$p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{R}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{R})} \exp\Big(\sum_{t=1}^{T} F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{R})\Big) \qquad (3)$$

$$Z_{\boldsymbol{\theta}}(\mathbf{R}) = \sum_{\mathbf{X}} \exp\Big(\sum_{t=1}^{T} F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{R})\Big) \qquad (4)$$

Assuming a fully labeled training set $\{\mathbf{X}^d, \mathbf{R}^d\}_{d=1\ldots D}$, the CRF parameters can be obtained by optimizing the conditional log-likelihood:

$$\mathcal{L}_{\boldsymbol{\theta}} = \sum_{d=1}^{D} \log p_{\boldsymbol{\theta}}(\mathbf{X}^d|\mathbf{R}^d) = \qquad (5)$$

$$= \sum_{d=1}^{D} \Big(\sum_{i=t}^{T} F_{\boldsymbol{\theta}}(x^d_t, x^d_{t-1}, \mathbf{R}^d) - \log Z_{\boldsymbol{\theta}}(\mathbf{R}^d)\Big) \qquad (6)$$

[1] We use a model with tied parameters $\boldsymbol{\theta}$ across all cliques, in order to seamlessly handle models of arbitrary size, *i.e.*, sequences of arbitrary length.

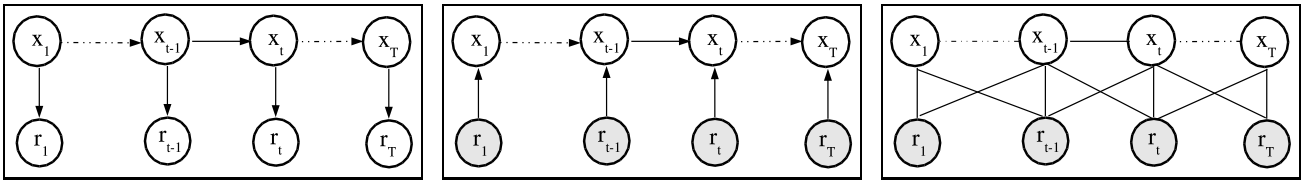

Figure 1: *(a, Left)* A generative Hidden Markov Model represents $p(\mathbf{r}_t|x_t)$ and $p(x_t|x_{t-1})$ and requires a probabilistic inversion to compute $p(\mathbf{X}_T|\mathbf{R}_T)$ using Bayes rule. Modeling long range dependencies among temporal observations is not tractable. *(b, Middle)* A directed conditional model (*e.g.* a Maximum Entropy Markov Model) represents $p(x_t|x_{t-1}, r_t)$ or, more generally, a locally normalized conditional distribution based on the previous state, and a past observation window of arbitrary size. Shadowed nodes indicate that the model conditions on the observation without modeling it. But the local normalization may face label-bias problems (see text). *(c, Right)* A Conditional Random Field accommodates arbitrary overlapping features of the observation. Here we show a model based on a context of 3 observation timesteps, but the dependencies can be arbitrarily long-range. Generally, the architecture does not rule out an on-line system, where long-range dependencies from the current state can be restricted only towards past observations.

In practice, we often regularize the problem by optimizing a penalized likelihood: $\mathcal{L}_{\boldsymbol{\theta}} + \mathcal{R}_{\boldsymbol{\theta}}$, either using soft (ridge) feature selection: $\mathcal{R}_{\boldsymbol{\theta}} = -||\boldsymbol{\theta}||^2$, or a more aggressive Jeffrey prior: $\mathcal{R}_{\boldsymbol{\theta}} = -\log||\boldsymbol{\theta}||$.

Likelihood maximization can be performed using a gradient ascent (*e.g.* BFGS [15]) method:

$$\frac{d\mathcal{L}_{\boldsymbol{\theta}}}{\mathbf{d}\boldsymbol{\theta}} = \sum_{d=1}^{D}\Big(\sum_{t=1}^{T}\frac{dF_{\boldsymbol{\theta}}(x_t^d, x_{t-1}^d, \mathbf{R}^d)}{\mathbf{d}\boldsymbol{\theta}} - \tag{7}$$

$$- \sum_{\mathbf{X}} p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{R}^d)\sum_{t=1}^{T}\frac{dF_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{R}^d)}{\mathbf{d}\boldsymbol{\theta}}\Big) \tag{8}$$

For discrete-valued chain models with state dependencies acting over a short range, the observation dependent normalization can be computed efficiently by matrix / tensor multiplication. For a bigram model, we work with the matrix of size $c \times c$, containing all possible assignments of pairs of neighboring states to class labels:[2]

$$M_t(\mathbf{R}) = [\exp(F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{R}))], x_t, x_{t-1} \in \mathcal{X} \tag{9}$$

Then the observation dependent normalization factor can be computed as:

$$Z_{\boldsymbol{\theta}}(\mathbf{R}) = \Big(\prod_{t=1}^{T+1} M_t(\mathbf{R})\Big)_{start,stop} \tag{10}$$

where we have added two dummy start and stop states $x_0 = start$ and $x_{T+1} = stop$ and the subscript indicates the particular entry of the matrix product [14].

The conditional probability of a class label sequence is:

$$p_{\boldsymbol{\theta}}(\mathbf{X}|\mathbf{R}) = \frac{\prod_{t=1}^{T+1}\exp(F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{R}))}{Z_{\boldsymbol{\theta}}(\mathbf{R})} \tag{11}$$

[2]Longer range state interactions be accommodated, *e.g.*, a trigram model by working with a tensor of size c^3.

The potential functions at pairs of neighboring sites can be chosen as:

$$F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{R}) = \psi_{\boldsymbol{\theta}}(x_t, \mathbf{R}) + \psi_{\boldsymbol{\theta}}(x_t, x_{t-1}) \tag{12}$$

where $\psi_{\boldsymbol{\theta}}$ are linear models:

$$\psi_{\boldsymbol{\theta}}(x_t, \mathbf{R}) = \sum_{a=1}^{A}\lambda_a f_a(x_t, \mathbf{R}) \tag{13}$$

$$\psi_{\boldsymbol{\theta}}(x_t, x_{t-1}) = \sum_{b=1}^{B}\beta_b g_b(x_t, x_{t-1}) \tag{14}$$

with parameters $\boldsymbol{\theta} = \{(\lambda_a, \beta_b), a = 1 \ldots A, b = 1 \ldots B\}$, to be estimated, and preset feature functions f_a, g_b based on conjunctions of simple rules. For instance, given a temporal context window of size $2W + 1$ (observations) around the current observation, the combined observation-label feature function is: $f_a(x_t, \mathbf{R}) = \mathbb{I}[x_t = m]\mathbf{r}_{t-j}[i], m \in \mathcal{X}, i \in \{1 \ldots r\}, j \in [-W, W]$, for a total of $A = c \times (2W + 1) \times r$ feature functions ($\mathbb{I}$ is the indicator function). Intuitively, the features encode correlations among motion classes and components of the observation vector forward or backward in time. The features that model inter-label dependencies are: $g_b(\mathbf{x}_t, \mathbf{x}_{t-1}) = \mathbb{I}[x_t = m_1 \wedge x_{t-1} = m_2], m_1, m_2 \in \mathcal{X}$, for a total of $B = c^2$ functions.

CRFs are convenient because, as for HMMs, inference can be performed efficiently using dynamic programming. Learning the model parameters leads to a convex problem with guaranteed global optimality [14]. We solve this optimization using a limited-memory variable-metric gradient ascent (BFGS) method [15] that converges in a couple of hundred iterations in most of our experiments (see fig. 4).

Directed Conditional Models. Maximum Entropy Markov Models (MEMM): An alternative approach to conditional modeling is to use a directed model [16] as shown in

fig. 1b. This requires a locally normalized representation for $p(x_t|x_{t-1}, \mathbf{r}_t)$. Inference can be performed efficiently using a dynamic programming procedure based on recursive Viterbi steps: $\alpha_t(x) = \sum_{x' \in \mathcal{X}} \alpha_{t-1}(x) \cdot p(x|x', \mathbf{r}_t)$ where $\alpha_t(x)$ computes the probability of being in state x at time t, given the observation sequence up to time t. Similarly, the backward procedure computes β_t as the probability of starting from state x at time t, given the observation sequence after time t as: $\beta_t(x') = \sum_{x \in \mathcal{X}} p(x|x', \mathbf{r}_t) \cdot \beta_{t+1}(x)$. The conditional distribution $p(x_t|x_{t-1}, \mathbf{r}_t)$ can be modeled as a log-linear model expressed in terms of feature functions $F_{\boldsymbol{\theta}}$ as in (12), (13) and (14):

$$p(x_t|x_{t-1}, \mathbf{r}_t) = \frac{1}{Z(x_{t-1}, \mathbf{r}_t)} \exp(F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{r}_t)) \quad (15)$$

where $Z(x_{t-1}, \mathbf{r}_t) = \sum_{x_t} F_{\boldsymbol{\theta}}(x_t, x_{t-1}, \mathbf{r}_t)$.

It is worth noticing that CRFs solve a problem that exists in MEMMs [16, 14], called the label-bias problem. This problem arises because such models are locally normalized. (MEMMs still have a non-linear decision surface because the local normalization depends on the state.) The per-state normalization requirement implies that the current observation is only able to select what successor state is selected, but not the probability mass transfered to that state, causing biases towards states with low-entropy transitions. In the limit, the current observation is effectively ignored for states with single outgoing transitions. In order to avoid this effect, a CRF employs an undirected graphical model that defines a single log-linear distribution over the joint vector of an entire class label sequence given a particular observation sequence (thus the model has a linear decision surface). By virtue of the global normalization, entire state sequences are accounted for at once, and this allows individual states to boost or damp the probability mass transfered to their successive states.

3 Experiments

We run a variety of recognition experiments based on both 2d features derived from image silhouettes and based on reconstructed 3d human joint angles.

Training Set: To gather image training data, we use Maya (Alias Wavefront), with realistically rendered computer graphics human surface models that we animate using human motion capture [1]. This database is annotated by activity class (with each individually sequence supplementary sub-segmented by activity type) and this information can be used to generate a labeled training set on which we perform segmentation and classification. Our 3d human state representation is based on an articulated skeleton with spherical joints, and has 56 d.o.f. including global translation. Our database consists of 8000 samples that involve various human activities including walking, running, turns, jumps, gestures in conversations and dancing. Some insight into the structure of the database is given in fig. 2, whereas image samples from our motion test sequences are shown in fig. 3.

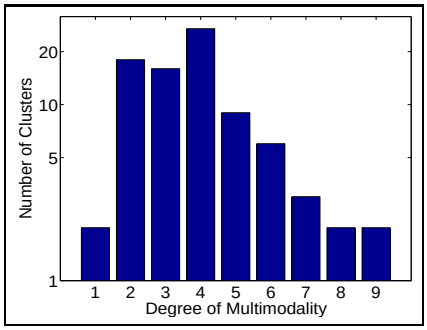

Figure 2: Analysis of the degree of ambiguity in the motion class labeling for our database, under moderate input (silhouette) perturbations. We cluster the input silhouette feature vectors into 80 clusters, count the number of different motion labels that fall within each, and histogram those.

Image Features: We work with silhouettes that we obtain using a combination of statistical background subtraction and motion segmentation [25]. As image descriptors, we use 50-dimensional histograms of combined shape context and pairwise edge features extracted at a variety of scales on the silhouette [25]. This representation is semi-local, rich and has been effectively demonstrated in many applications, including texture recognition or pose prediction. The representation is based on *overlapping features of the observation*. Therefore the elements of the silhouette feature vector are generally not independent. However, due to its conditional structure, a CRF flexibly accommodates this representation without modeling assumption violations.

We run several tests in order to compare the CRF model described in §2 with a HMM and a MEMM (see [24] for additional experiments). The HMM we use is a fully ergodic model based on Gaussian emission probabilities having full covariance matrix for each state. The parameters of the model (the emission probability density, the state transition matrix) are learned from training data [20] using Maximum Likelihood. We also learn a variety of CRFs that model long-range dependencies between observations to various degrees, *i.e.* windows $W = \{0, 1, 3\}$, meaning that we considered contexts of observations of size 0,3 and 7 centered at the current observation.[3] Fig. 4 gives some insight into the learning procedure for CRFs and the distribution of estimated coefficients for our feature functions. Training is more expensive for CRFs, ranging from 30 minutes to several hours for

[3]For the experiments, we only consider baseline models, arguably, more complex HMMs or CRFs can be used. Nevertheless, most of the technology previously used to construct sophisticated HMMs including layering or left-right models can be directly applied to build CRF counterparts (*e.g.* left-right implementations can be obtained by setting some of the β parameters in (14) to zero; one can build a separate left-right model for each motion class, *etc.*). None of the models is thus disadvantaged by *not* using such features.

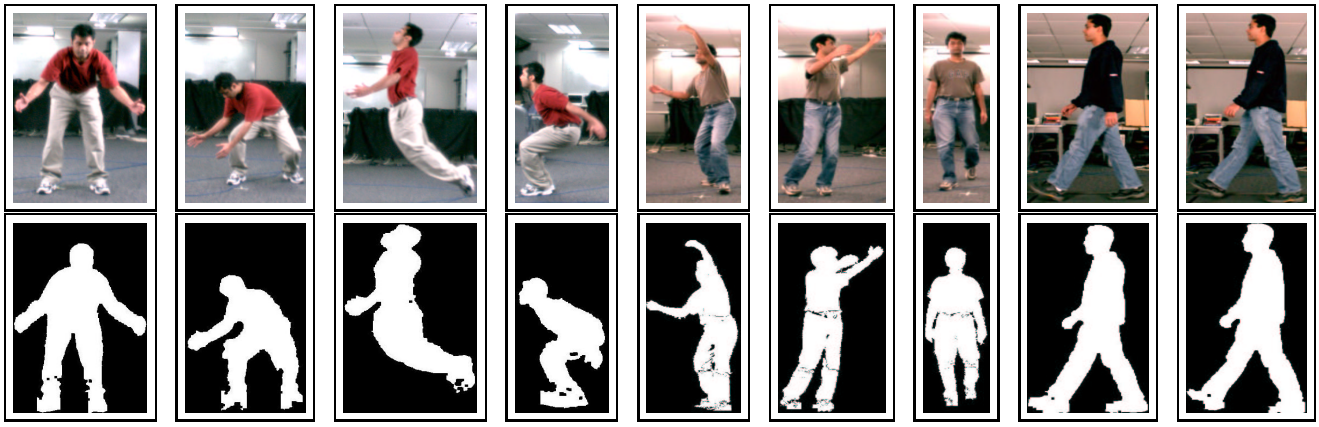

Figure 3: Sample images (*top row*) and silhouettes (*bottom row*) of typical motions that are recognized: walking, running, bending and picking, dancing, *etc*. Notice that the silhouettes are of average quality and potentially quite different from the ones we trained on and which were artificially generated. Nevertheless, we observe that the recognition accuracy is usually robust to these factors.

models having longer windows of observations (as opposed to seconds for HMMs, or minutes for MEMMs), on a standard desktop PC. Inference is about as fast for all models in the order of seconds for sequences of several hundred frames.

Recognition Experiments based on 2d features: We test our algorithms on both real data (table 2; we work with 7 motion labels, and we test on 1029 frames) and artificial data (table 1; we work with 11 motion labels and we test on 2536 frames, and use a CRF with $W = 1$) and evaluate their recognition performance not only w.r.t. broader classes of motion like running, walking or dancing, but also w.r.t. to finer styles like normal walk, wander walk or slow walk (tables 3,4,5; we work with 4 motion labels, we test on 700 frames and use a CRF with $W = 1$). It is noticeable that the CRF typically outperforms the MEMM and the HMM in most test cases.

In table 2 we show an extensive set of experiments for different motion labels and models. The CRFs learned using larger window contexts generally outperform the HMM, with the exception of the jump, which the CRF confuses with the motion of picking, or of grabbing something from the floor. CRFs also show significantly better and stabler recognition performance in the presence of larger variability w.r.t. the training set (*e.g*. the test set denoted HWSW has input silhouettes that are significantly different from the ones on the training set). It is also important to notice how increasing the context of the current observation improves recognition and changes the inferred distribution of class labels. In fig. 5 we show how a larger observation context can improve recognition results by as much as 70%.

In tables 3, 4 and 5, we analyze the recognition performance w.r.t. viewpoint and finer motion differences. For the experiments shown in table 3, we have selected a viewpoint that is somewhat uninformative with respect to the motion. As a consequence, the recognition rates are not high, often the normal walk and the wander walk are confused.

	NW	WW	SW	R
CRF *W=0*	38.9	65	86.5	100
CRF *W=3*	100	45	100	100
MEMM	16.31	64.5	50.5	75
HMM	0	76.5	44.3	100

Table 3: Recognition accuracy for a 45^o viewpoint. NW / WW / SW = Normal / Wander / Slow Walk; R = Run.

In table 4, the recognition is generally improved (the side viewpoint appears quite informative in disambiguating running from anything else), but the MEMM and the HMM have difficulty in accounting for long-range observation dependencies that appear useful in discriminating different styles of walking.

	NW	WW	SW	R
CRF *W=0*	79.62	100	51	100
CRF *W=3*	100	100	100	100
MEMM	59.25	96.57	53	100
HMM	80	100	33	100

Table 4: Recognition accuracy for a side viewpoint. NW / WW / SW = Normal / Wander / Slow Walk; R = Run.

In table 5, we show recognition results for motions seen from a challenging frontal viewpoint. The wander walk tends to be the easiest to discriminate, presumably because it produces informative sideways variations in the frontally projected silhouette. CRF's contextual power helps improving performance, which nevertheless remains low, as it often confuses the normal and slow walks.

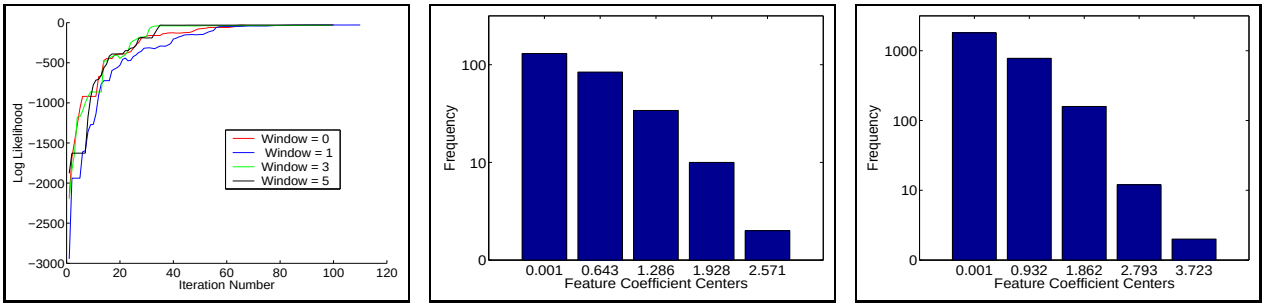

Figure 4: *(Left)* plots the data conditional log-likelihood as per eq. (5), versus the iteration number for various observation windows ($W = 0, \ldots, 5$). Notice that all models converge equally fast, in about 100 iterations. *(Middle)* and *(right)* plots show histograms of parameters $\boldsymbol{\theta}$ corresponding to $W = 0$ and $W = 5$. Many parameters are small because we use a ridge penalized likelihood. Notice an increase in the range of parameters for models that use a larger context.

	C	FR	FWT	JLT	PD	RLT	SR	SW	SWF	WF	WS
CRF	72.8	100	100	100	100	100	100	100	100	100	100
MEMM	100	40	100	5.2	100	100	90.5	98.14	100	91.4	100
HMM	1.4	100	2.5	1.7	87.41	93.75	100	100	100	100	100

Table 1: Comparisons of recognition performance (percentage accuracy) for synthetically generated silhouette input features. C = Conversation, FR = Run seen Frontally, FWT = Walk and Turn seen Frontally, JLT = Jogging and Left Turn, PD = Professional Dance, RLT = Run and Turn Left,SR = Run seen from a Side, SW = Walk seen from a Side, SWF = Slow Walk seen Frontally, SWS = Slow Walk seen from a Side, WF = Wander walk seen Frontally, WS = Wander walk seen from a Side.

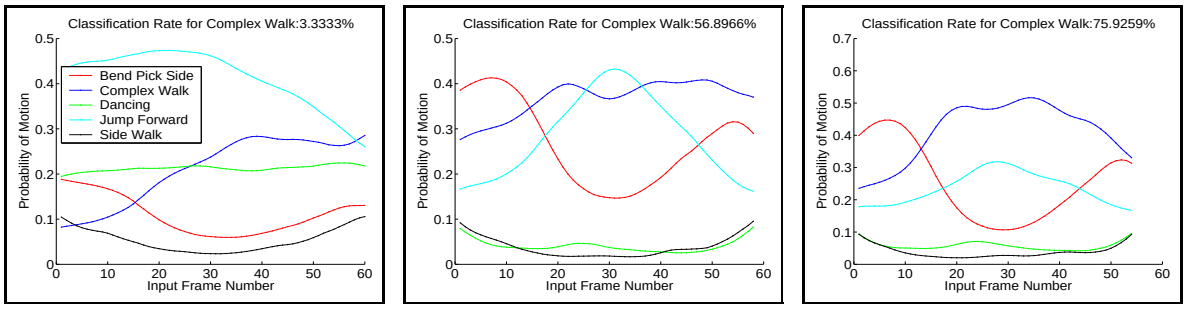

Figure 5: (Best viewed in color) The use of context in CRFs significantly improves the recognition performance (2d testing on feature vectors based on silhouettes extracted from real images). *Left* plots shows the distribution over motion class labels when we use only the current observation (*i.e.* no context $W = 0$) whereas the *middle* and *right* plots use contexts of size $W = 1$ and $W = 3$ respectively (3 and and 7 observation timesteps centered at the current one). A HMM tested on the same sequence entirely mis-classifies the complex walk (motion towards the camera, turning and walking back – with low accuracy of about 1.5%), which is close to the performance of a CRF with no context (*left* plot).

Recognition based on reconstructed 3d joint angle features: In table 6 we give motion recognition results based on reconstructed 3d joint angle features [1], as opposed to directly based on image silhouette features (we use various motions for a total of 1200 frames for testing). We directly use the human motion capture output as opposed to the 3d reconstruction results from an algorithm like [25], because often multiple 3d trajectories are plausible give an image sequence [23]. Therefore probabilistically correct recognition in this context would be more complex, as a recognizer may have to consider different 3d input hypotheses and not just one. The CRFs based on larger contexts have generally better performance than the HMM (see also fig. 6 and fig. 7), except for conversations which are sometimes confused with dancing (see fig. 6). This is not entirely surprising given that both of these activities involve similar, ample arm movements. The occasional drop in the performance of CRFs could be caused by insufficient training data. MEMMs can outperform CRFs in problems where their non-linear decision boundary is more adequate than the linear CRF one.

	CW	D1	D2	BPS	LVSW	HVSW	JF
CRF $W = 0$	100	37	100	100	100	100	16
CRF $W = 1$	100	42	96	100	100	100	27
CRF $W = 3$	100	56.44	90.8	100	100	100	28
HMM	100	39	90	76	98.02	17	58

Table 2: Comparisons of recognition performance (percentage accuracy) for silhouettes extracted in real image sequences. CW = Complex Walk of a person coming towards the camera, turning and walking back, D1 = classical Dancing, D2 = modern Dancing, BPS = Bending and Picking seen from a Side, LVSW = Walking seen from a Side, silhouettes having Lower Variability w.r.t. the training set, HVSW = Walking seen from a Side, silhouettes having significantly Higher Variability w.r.t. the training set, JF = Jump Forward. The CRF with longer range dependencies generally does better, but seems to confuse the jump with the pick-up. These motions indeed have similar parts, especially given that translation information is not used in the silhouette representation (but an object centered coordinate system for features). Notice that CRF does significantly better in the presence of larger variability w.r.t. the training set (*e.g.* HVSW), which has been also noticed in [14].

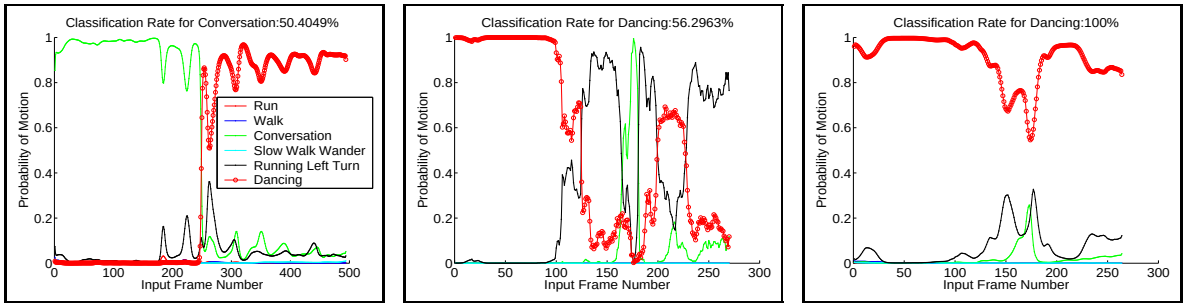

Figure 6: (Best viewed in color) The distribution over class labels for recognition experiments based on 3d joint angle observations. *Left* plot shows motion class distributions for a conversation test set. Even a CRF that uses context partly confuses conversation and dancing, presumably because both classes involve ample arm movements that are similar. *Middle* shows recognition results for a dancing test sequence, based on a CRF with no context ($W = 0$). *Right* shows how a CRF with context $W = 3$ improves the recognition performance for dancing by 43% w.r.t. the CRF with no context ($W = 0$, *middle*).

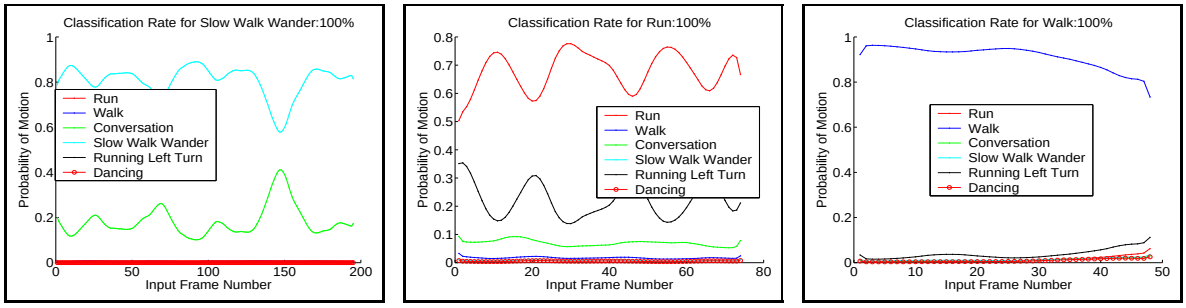

Figure 7: (Best viewed in color) *Left* plot shows increased recognition performance for a slow walk wander motion, here a CRF with $W = 3$ improved recognition accuracy by 100% w.r.t. to a a CRF with no context $W = 0$ (see table 6). *Middle* and *right* plots show good CRF recognition accuracy for running and walking.

4 Conclusions

We have presented a framework for human motion recognition, that unlike existing generative approaches based on HMM, is discriminative and based on Conditional Random Fields and Maximum Entropy Markov Models. These complement the popular HMMs and can be used in tandem with them in recognition systems. By virtue of their conditional structure, the models can accommodate arbitrary overlapping features of the observation as well as long-term contextual dependencies among observations at different timesteps. This wouldn't be possible in a HMM where strict independence assumption among observations are required in order to ensure tractability. Similarly to HMMs, inference in the conditional

	NW	WW	SW	R
CRF *W=0*	30.5	100	100	22
CRF *W=3*	36.1	100	96	21.5
MEMM	34	91.5	96	16.25
HMM	14.51	80.60	81	0

Table 5: Recognition accuracy for a frontal viewpoint. NW / WW / SW= Normal / Wander / Slow Walk; R = Run.

	R	W	SWW	RTL	C	D
CRF $W = 0$	100	100	0	100	60	56.29
CRF $W = 3$	100	100	100	100	50.40	100
MEMM	100	100	19.9	100	79.8	100
HMM	100	68.5	0	100	82.5	89

Table 6: Recognition accuracy based on 3d joint angle features. R = Running, W = Walking, SWW = Slow Walk Wandering, RTL = Run and Turn Left, C = Conversation, D = Dancing. The accuracy of CRF with long-range dependencies is generally better, however it seems to confuse conversation and dancing, as can be seen in fig. 6. This is not surprising given that both activities involve sometimes similar arm movements. Notice also how the context helped boosting the recognition performance for SWW in fig. 7.

models can be performed efficiently using dynamic programming, whereas the training procedure for the parameters is based on convex optimization. We have demonstrated the algorithms for the recognition of a variety of human motions including walking, running, bending or dancing where we observed that CRFs significantly improved recognition performance over MEMMs, that in turn, typically outperformed competing HMMs.

Future Work: Inference and learning with CRFs provides an avenue for many associated research problems. It would be interesting to systematically investigate how long-range should the observation dependency be for optimal recognition performance, as well as recognition based on different selections of features. The number of possible feature combinations can be large, so efficient methods for feature selection or feature induction are necessary. In this work we use a model with first order state dependency (a bigram) but it would be interesting to study longer range state dependencies, *e.g*. trigrams. All these extensions are straightforward to include in a CRF.

References

[1] CMU Human Motion Capture DataBase. Available online at http://mocap.cs.cmu.edu/search.html, 2003.

[2] J. Aggarwal and Q. Cai. Human Motion Analysis: A Review. *CVIU*, 73(3):428–440, 1999.

[3] M. Black and A. Jepson. A probabilistic framework for matching temporal trajectories:Condensation-based recognition of gestures and expressions. In *ECCV*, 1998.

[4] A. Blake, B. North, and M. Isard. Learning Multi-Class Dynamics. *NIPS*, 11:389–395, 1999.

[5] A. Bobick and J. Davis. The recognition of human movement using temporal templates. In *PAMI*, 2001.

[6] M. Brand, N. Oliver, and A. Pentland. Coupled Hidded Markov models for complex action recognition. In *CVPR*, 1996.

[7] C. Bregler. Learning and recognizing human dynamics in video sequences. In *CVPR*, 1997.

[8] R. Fablet and P. Bouthemy. Non-parametric motion recognition using temporal multiscale Gibbs models. In *CVPR*, 2001.

[9] D. Gavrila. The Visual Analysis of Human Movement: A Survey. *CVIU*, 73(1):82–98, 1999.

[10] S. Gong and T. Xing. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, 2003.

[11] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[12] M. Jordan. *Learning in graphical models*. MIT Press, 2001.

[13] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction and classification. In *ICCV*, 2003.

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

[15] A. McCallum. Efficiently inducing features of conditional random fields. In *UAI*, 2003.

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

[17] V. Pavlovic and J. Rehg. Imact of dynamic model learning on the classification of human motion. In *CVPR*, 2000.

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

[19] N. Quian and T. Sejnowsky. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Bio.*, 1988.

[20] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[21] D. Ramanan and D. Forsyth. Automatic annotation of everyday movements. In *NIPS*, 2003.

[22] Y. Shi, Y. Huang, D. Minnen, A. Bobick, and I. Essa. Propagation networks for recognition of partially ordered sequential action. In *CVPR*, 2004.

[23] C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems. In *CVPR*, volume 2, pages 608–615, Washington D.C., 2004.

[24] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for human motion recognition. Technical Report CSRG-517, University of Toronto, March 2005.

[25] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, 2005.

[26] T. Starner and A. Pentland. Real-time ASL recognition from video using Hidden Markov Models. In *ISCV*, 1995.

[27] C. Stauffer and E. Grimson. Learning patterns of activity using real-time tracking. In *PAMI*, 2000.

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

[29] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of ASL. In *CVIU*, 2001.