2017 IEEE International Conference on Robotics and Automation (ICRA)
Singapore, May 29 - June 3, 2017

Minimum Uncertainty Latent Variable Models for Robot Recognition of Sequential Human Activities

Fei Han[1], Christopher Reardon[2], Lynne E. Parker[3], and Hao Zhang[1]

***Abstract*— Recognition of sequential human activities, such as "sitting down" and "standing up", is a common but challenging problem in human-robot interaction, which requires modeling their underlying temporal patterns. Although previous sequence modeling methods, such as Hidden Conditional Random Fields (HCRFs), demonstrated satisfactory recognition accuracy, they do not explicitly model the uncertainty in underlying temporal patterns, which can provide valuable information to characterize sequential activities. To address this problem, we introduce a novel *Minimum Uncertainty* HCRF (MU, or μHCRF). Different from traditional HCRF-based techniques that only utilize the negative log-likelihood of the categories' conditional probability as the loss function, the proposed μ-HCRF also introduces a regularization term to model the underlying temporal pattern of the latent variables. As another theoretical contribution, we provide a derivation to show that the formulated problem has a closed-form solution, and prove that inference of the proposed μHCRF is tractable. Extensive empirical study is performed to evaluate our approach, using four public benchmark datasets. Experimental results have shown that our μHCRFs outperform previous techniques and achieve state-of-the-art performance on human activity recognition, especially on sequential activities.**

I. Introduction

Human activity recognition is an important research topic that has a wide variety of real-world applications in service robotics, human-robot interaction, and human-robot teaming. However, human activity recognition from visual perception is a challenging problem due to illumination changes, camera motion, background clutter, and diversity of human motion and appearance. In particular, robot recognition of *sequential* activities introduces additional challenges, including the requirement of modeling their underlying temporal patterns. For example, it is generally impossible to separate "standing up" and "sitting down" from a single image, because humans may exhibit the same pose in the images for both activities; therefore, modeling temporal patterns is necessary.

A popular algorithm to encode human activity's temporal patterns is the Conditional Random Field (CRF) [1] method, which is a discriminative graphical model that avoids encoding the distribution of the input. However, CRFs are limited in that they lack the capability to combine latent variables that can capture underlying patterns within the observation [2]. For example, a robot coach may need to model a complex activity "tennis serve," where atomic temporal motion patterns, such as "ball tossing" and "racquet swinging," are unknown, and thus must be modeled using latent variables. To address this problem, Hidden Conditional Random Fields (HCRFs) [2] were used to combine CRF model's strengths with latent variables. Due to the latent variable's capability to model temporal patterns of a sequence, HCRF methods are widely applied in sequence labeling, such as human activity recognition.

These previous HCRFs did not well model the uncertainty in the latent temporal pattern; therefore, the latent variables that encode temporal patterns are eliminated either by summation in HCRFs based on maximum likelihood estimation (MLE) [2] or through maximization in max-margin (MM) HCRFs [3]. The latent temporal pattern often provides useful information for improving prediction accuracy [4], [5]. Also, there are many scenarios in which a robot must confidently understand the latent temporal pattern itself. For example, a robot coach teaching "tennis serve" must maximize its confidence on the chronological order of "ball tossing" and "racquet swinging" in the temporal-motion pattern.

To address this critical problem, we propose a new HCRF-based method to identify sequential human activities through introducing a novel regularization term to the traditional loss function, which models their latent temporal structure. Since our approach is capable of modeling the uncertainty of latent variables, we name this method *Minimum-Uncertainty* HCRF (MU, or μHCRF). The contributions of this paper include:

- We introduce a novel regularization to the conventional negative log-likelihood loss, which captures the uncertainty in latent underlying temporal patterns and greatly improves recognition accuracy of sequential activities.
- We derive a theoretical proof that the formulated objective function has a closed form and the inference process of μHCRFs is trackable.

The rest of the paper is structured as follows. Section II reviews the related work. In Section III, we present the formulation of HCRFs. Then, our new μHCRF is introduced in Section IV. After presenting experimental results in Section V, we conclude the paper in Section VI.

II. Related Work

General reviews of activity recognition are conducted in [6], [7]. In the following, we discuss previous techniques that focus on modeling temporal patterns for sequential activity recognition, which can be grouped into two categories.

The first category uses space-time features to model temporal patterns of sequential activities. Laptev [8] applied histogram of spatial gradient (HOG) and optic flow (HOF) to

[1]Fei Han and Hao Zhang are with the Human-Centered Robotics Lab in the Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA. {fhan,hzhang}@mines.edu.

[2]Christopher Reardon is with the US Army Research Laboratory, Adelphi, MD 20783, USA. creardon@army.mil.

[3]Lynne E. Parker is with the Distributed Intelligence Laboratory in the Department of Electrical Engineering and Computer Science, University of Tennessee, TN37996, USA. leparker@utk.edu.

describe local motion and appearance patterns in space-time neighborhoods of the detected interest points. Dollar et al. [9] described spatio-temporal features by concatenating the gradients around the interest point's space-time neighborhoods. Recently, Wang et al. [10] introduced the motion boundary histograms (MBH) to describe temporal motion variations. Some other spatio-temporal features were also introduced in [11], [12], [13], [14]. The features are often used to formulate a bag-of-words (BoW) model to represent human activities. Although satisfactory activity recognition performance has been reported, space-time features encode temporal patterns only within a short period of time, in general.

The second category to recognize sequential activities is based on dynamic graphical models, which are able to represent temporal structures that extend over long periods of time. In a generative setting, Dynamic Bayesian Networks (DBNs) [15] are a popular method for sequence modeling, because they exploit structures in the problem to compactly represent distributions over multiple state variables. Hidden Markov Models (HMMs) [16] and their extensions [17], a special case of DBNs, are a classical method for sequential activity recognition. In a discriminative setting, CRFs [1], HCRFs [2] and their extensions [3], [5] are the most widely used approaches for modeling activity temporal structures. Although previous dynamic models use hidden variables to model the latent temporal pattern, the certainty in this pattern is not well studied. We address this issue for HCRFs.

Our work falls into the second category focusing on developing learning models to address the problem of sequential activity recognition. Different from previous HCRF methods, we introduce a novel concept of modeling the uncertainty of the underlying temporal pattern to improve the sequence labeling performance. This is achieved through introducing a new term to regularize the traditional negative log-likelihood loss, which results in a close-form objective function with a trackable inference process.

III. HCRFs for Activity Recognition

In this section, we introduce the background of traditional HCRFs, and discuss how sequential activity recognition can be modeled using HCRFs.

A. Hidden Conditional Random Fields

When using HCRFs for supervised multi-class classification, the goal is to learn a mapping $f : \mathcal{X} \mapsto \mathcal{Y}$ from a set of i.i.d. training data $\mathcal{D} = \{(\boldsymbol{x}^i, y^i), i = 1, \ldots, N\}$, to predict a class label $y^i \in \mathcal{Y}$ for an observation $\boldsymbol{x}^i \in \mathcal{X}$. Each observation is a vector of M attributes $\boldsymbol{x}^i = \{x_1, \ldots, x_M\}$, where $x_j \in \mathbb{R}$, $j = 1 \ldots M$, is an attribute extracted from visual data. The HCRF model also defines a vector of latent variables $\boldsymbol{h} = \{h^1, \ldots, h^N\}$, where $h^i \in \mathcal{H}$, corresponds to a hidden label that is associated with the observation $\boldsymbol{x}^i$.

The HCRF is defined on an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, whose nodes satisfy $\mathcal{V} = \{\boldsymbol{x} \cup \boldsymbol{h} \cup y\}$. The HCRF graph is annotated with a set of real-valued potentials $\boldsymbol{\psi}(\boldsymbol{D}; \boldsymbol{\theta}) = \{\psi_1(\boldsymbol{D}_1; \theta_1), \ldots, \psi_P(\boldsymbol{D}_P; \theta_P)\}$, where $\boldsymbol{D}$ is the scope of the potential ψ that satisfies $\boldsymbol{D} \subseteq \mathcal{V}$ and $\boldsymbol{D} \not\subseteq \boldsymbol{x}$, θ is the parameter, and P is the number of potentials. The HCRF network is connected with undirected edges $\mathcal{E} = \{v_i - v_j : \{v_i, v_j\} \subseteq \boldsymbol{D}_k; \forall i \neq j, k = 1, \ldots, P\}$. HCRFs encode the following conditional distribution:

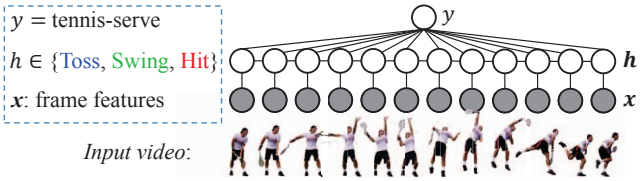


Fig. 1. An illustrative example of utilizing HCRFs to model a sequential "tennis-serve" activity. Each input $\boldsymbol{x}$ is a vector features extracted from a frame; each frame is associated with a latent variable h, which represents a primitive motion (e.g., "Toss", "Swing" and "Hit"); and the entire sequence has a single output y, encoding the sequential activity's label. The latent variables form a chain to model the temporal pattern of primitive motions.

$$P(y|\boldsymbol{x};\theta) = \frac{1}{Z(\boldsymbol{x};\theta)}\tilde{P}(y|\boldsymbol{x};\theta) = \frac{1}{Z(\boldsymbol{x};\theta)}\sum_{\boldsymbol{h}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) \quad (1)$$

where $\tilde{P}(y|\boldsymbol{x};\boldsymbol{\theta})$ is called the *unnormalized measure* that is represented by a product of potentials, i.e., $\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\theta) = \prod_i \psi_i(\boldsymbol{D}_i;\theta_i)$, and each potential $\psi_i(\boldsymbol{D}_i;\theta_i)$ must capture some domain knowledge about the structure of the latent variables; $Z(\boldsymbol{x};\boldsymbol{\theta})$ is the *partition function* that is computed by $Z(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{y\in\mathcal{Y},\boldsymbol{h}\in\mathcal{H}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})$.

To learn the model parameters $\boldsymbol{\theta}$, HCRFs minimize a loss function $\mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta})$ defined as the negative log-likelihood of the conditional probability presented in Eq. (1):

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} -\log P(y|\boldsymbol{x};\boldsymbol{\theta}) \quad (2)$$

The formulated optimization problem can be solved using off-the-shelf solvers, including Bundle Cutting Plane (BCP) [18] and Non-convex Regularized Bundle Method (NRBM) [19]. After learning the optimal parameters $\boldsymbol{\theta}^*$ during training, given a new observation $\boldsymbol{x}$ during testing, inference is performed by picking the category label that minimizes the loss function, i.e., $y^* = \operatorname{argmin}_{y\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x}, y; \boldsymbol{\theta}^*)$.

B. Sequential Activity Modeling Using HCRFs

We formulate human activity recognition as a sequence classification problem, which requires to model time information to distinguish challenging sequential human activities (e.g., "standing up" and "sitting down"). An illustrative example is provided in Figure 1 to model and recognize the "tennis-serve" activity from a video clip that contains a sequence of frames. Each input $\boldsymbol{x} \in \mathcal{X}$ is a feature vector from a frame; each frame is associated with a latent variable $h \in \mathcal{H}$, which represents a primitive motion (e.g., "Toss"); and the entire video has a single output y, representing the sequential activity's label. The latent variables form a linear chain to model time dependency of the frames. Let $\mathbb{1}(\cdot)$ be an indicator function, and y' and (h', h'') be the assignments to the label and hidden variables, respectively. Then, following

[2], [20], [3], our singleton observation potential $\psi_O(\cdot)$ at time (frame) t is defined as:

$$\psi_O(\boldsymbol{D};\theta) = \exp(\theta_{h_t,x}\phi(h,\boldsymbol{x})) = \exp(\theta_{h,x}\boldsymbol{x}\mathbb{1}(h_t = h'))$$

which models the compatibility between the input $\boldsymbol{x}$ and the latent motion variable h. The singleton label potential $\psi_L(\cdot)$ at time t is defined as:

$$\psi_L(\boldsymbol{D};\theta) = \exp(\theta_{h,y}\phi(h_t,y)) = \exp(\theta_{h,y}\mathbb{1}(h_t{=}h')\mathbb{1}(y{=}y'))$$

which models the compatibility between the activity label y and the latent motion h. The pairwise transition potential $\psi_T(\cdot)$ from time t to $t{+}1$ is defined as:

$$\begin{aligned}\psi_T(\boldsymbol{D};\theta) &= \exp(\theta_{h,h,y}\phi(h_t,h_{t+1},y)) \\ &= \exp(\theta_{h,h,y}\mathbb{1}(h_t{=}h')\mathbb{1}(h_{t+1}{=}h'')\mathbb{1}(y{=}y'))\end{aligned}$$

which models the compatibility between activity label y and a pair of latent motion variables (h_t, h_{t+1}), e.g., how likely a video with activity label y' contains a consecutive pair of motions h' and h'' in this case.

In real-world scenarios, human motions are continuous, and the transition from one pose to another is gradual, and the point at which one pose ends and another begins is not always distinct, as shown in Figure 1. Thus, values of hidden variables are typically uncertain and noisy. To this end, an approach that is able to minimize such uncertainty in latent variables is important to improve the accuracy of sequential activity recognition.

IV. Our μHCRF Approach

In this section, we discuss the new regularizer introduced in our μHCRF approach. Then, we prove that the formulated objective function can have a closed form, and that inference of μHCRFs is tractable.

A. Model Formulation

As presented in Eq. (2), the loss function $\mathcal{L}$ of traditional HCRF models ignores the uncertainty in the latent variables, which however is important to recognize sequential activities. To model this temporal uncertainty, our μHCRF approach introduces a novel regularizer $\mathcal{R}$ based on the entropy of the latent variables, resulting in a new regularized optimization problem as follows:

$$\operatorname*{argmin}_{\boldsymbol{\theta}} -\log P(y|\boldsymbol{x};\boldsymbol{\theta}) + \gamma H(P(\boldsymbol{h}|y,\boldsymbol{x};\boldsymbol{\theta}) \tag{3}$$

where the regularization term $\mathcal{R} = H(P(\boldsymbol{h}|y,\boldsymbol{x};\boldsymbol{\theta})$ is the entropy of the latent variables, which models their uncertainty in these variables. γ is a trade-off hyperparameter balancing the effect between the loss function and regularization term. In our implementation, we use a general entropy named the Kapur entropy [21] as our regularization term. Given discrete random variables $\boldsymbol{z}$, the Kapur entropy of order α and type β is defined as:

$$H_{\alpha,\beta}(P(\boldsymbol{z})) = \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{z}} P(\boldsymbol{z})^{\alpha+\beta-1}}{\sum_{\boldsymbol{z}} P(\boldsymbol{z})^{\beta}} \tag{4}$$

where $\alpha{\neq}1$, $\alpha{>}0$, $\beta{>}0$, and $\alpha{+}\beta{-}1{>}0$. If $\alpha{\to}0$, $\beta{=}1$, the Kapur entropy becomes the Hartley function [22], i.e., $H_{0,1}(P(\boldsymbol{z})) = \log K$, where K is the number of variables in $\boldsymbol{z}$ with a positive probability. In the limit $\alpha{\to}1$ and $\beta{=}1$, the Kapur entropy converges to the Shannon entropy. When $\alpha\to\infty$ and $\beta = 1$, we can obtain a quantity analogous to the Chebyshev norm, i.e., $H_{\infty,1}(P(\boldsymbol{z})) = -\log\max_{\boldsymbol{z}} P(\boldsymbol{z})$. The Kapur entropy is not convex, in general.

In the following, we prove that the regularized optimization problem in Eq. (3) can have a closed form as a theorem. First, we present a lemma:

Lemma 1: For finite discrete random variable $\boldsymbol{z}$, the Kapur entropy satisfies: $H_{\alpha,\beta}(\tilde{P}(\boldsymbol{z})) = H_{\alpha,\beta}(P(\boldsymbol{z})) - \log Z$, where $\alpha \neq 1$, $\alpha > 0$, $\beta > 0$, $\alpha+\beta-1 > 0$, $\tilde{P}(\boldsymbol{z})$ is the unnormalized measure of $P(\boldsymbol{z})$, and $Z = \sum_{\boldsymbol{z}} \tilde{P}(\boldsymbol{z})$ is the partition function.

Proof: Under the Kapur entropy constraints $\alpha \neq 1$, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta - 1 > 0$, we obtain the following:

$$\begin{aligned}H_{\alpha,\beta}(\tilde{P}(\boldsymbol{z})) &= \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{z}}\tilde{P}(\boldsymbol{z})^{\alpha+\beta-1}}{\sum_{\boldsymbol{z}}\tilde{P}(\boldsymbol{z})^{\beta}} \\ &= \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{z}}(P(\boldsymbol{z})\cdot Z)^{\alpha+\beta-1}}{\sum_{\boldsymbol{z}}(P(\boldsymbol{z})\cdot Z)^{\beta}} \\ &= \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{z}}P(\boldsymbol{z})^{\alpha+\beta-1}}{\sum_{\boldsymbol{z}}P(\boldsymbol{z})^{\beta}} + \frac{1}{1-\alpha}\log Z^{\alpha-1} \\ &= H_{\alpha,\beta}(P(\boldsymbol{z})) - \log Z\end{aligned}$$

■

Theorem 1: The optimization problem formulated by the μHCRF method in Eq. (3) has a closed form solution when $\gamma = 1$.

Proof: Because the normalization factor $Z(\boldsymbol{x};\boldsymbol{\theta})$ in Eq. (1) is a constant, when $\gamma = 1$, the optimization problem in Eq. (3) is equivalent to:

$$\operatorname*{argmin}_{\boldsymbol{\theta}} -\log P(y|\boldsymbol{x};\boldsymbol{\theta}) + H(P(\boldsymbol{h}|y,\boldsymbol{x};\boldsymbol{\theta}) + Z(\boldsymbol{x};\boldsymbol{\theta}) \tag{5}$$

Under the Kapur entropy constraints $\alpha \neq 1$, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta - 1 > 0$, we obtain the following:

$$\begin{aligned}&H_{\alpha,\beta}(P(\boldsymbol{h}|y,\boldsymbol{x};\boldsymbol{\theta})) - \log P(y|\boldsymbol{x};\boldsymbol{\theta}) - \log Z(\boldsymbol{x};\boldsymbol{\theta}) \\ &= \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{h}}P(\boldsymbol{h}|y,\boldsymbol{x};\boldsymbol{\theta})^{\alpha+\beta-1}}{\sum_{\boldsymbol{h}}P(\boldsymbol{h}|y,\boldsymbol{x};\boldsymbol{\theta})^{\beta}} \\ &\quad - \log P(y|\boldsymbol{x};\boldsymbol{\theta}) - \log Z(\boldsymbol{x};\boldsymbol{\theta}) \quad \text{[via entropy definition]} \\ &= \frac{1}{1-\alpha}\log\left(\frac{\sum_{\boldsymbol{h}}\left(\frac{P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})}{P(y|\boldsymbol{x};\boldsymbol{\theta})}\right)^{\alpha+\beta-1}}{\sum_{\boldsymbol{h}}\left(\frac{P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})}{P(y|\boldsymbol{x};\boldsymbol{\theta})}\right)^{\beta}}\right) \\ &\quad - \log P(y|\boldsymbol{x};\boldsymbol{\theta}) - \log Z(\boldsymbol{x};\boldsymbol{\theta}) \quad \text{[via Bayes rule]} \\ &= \frac{1}{1-\alpha}\log\left(\frac{\sum_{\boldsymbol{h}}P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\alpha+\beta-1}}{\sum_{\boldsymbol{h}}P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\beta}}\cdot\frac{P(y|\boldsymbol{x};\boldsymbol{\theta})^{1-\alpha-\beta}}{P(y|\boldsymbol{x};\boldsymbol{\theta})^{-\beta}}\right) \\ &\quad - \log P(y|\boldsymbol{x};\boldsymbol{\theta}) - \log Z(\boldsymbol{x};\boldsymbol{\theta}) \\ &= \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{h}}P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\alpha+\beta-1}}{\sum_{\boldsymbol{h}}P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\beta}} \\ &\quad + \left(\frac{1}{1-\alpha}\log P(y|\boldsymbol{x};\boldsymbol{\theta})^{1-\alpha} - \log P(y|\boldsymbol{x};\boldsymbol{\theta})\right) - \log Z(\boldsymbol{x};\boldsymbol{\theta}) \\ &= H_{\alpha,\beta}(P(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) - \log Z(\boldsymbol{x};\boldsymbol{\theta}) \\ &= H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) \quad \text{[via Lemma 1]}\end{aligned}$$

Therefore, solving the regularized optimization problem in Eq. (3) is equivalent to solving:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) \tag{6}$$

which ends the proof. ∎

B. Inference

Given an observation $\boldsymbol{x}$, the output label y is inferred by selecting the class that minimizes the new objective function in Eq. (3):

$$y^* = \underset{y\in\mathcal{Y}}{\operatorname{argmin}}\, H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) \tag{7}$$

which leads to not only a high likelihood of the human activity label, but also a low uncertainty in the latent variables.

When the latent variables form a linear chain to model a sequential activity, as illustrated in Figure 1, and the potentials described in Section III-B are used, we can efficiently compute the new objective function in Eq. (7) and solve the inference problem, which is formally proved in Theorem 2. First, we provide the following lemmas:

Lemma 2: The objective function in Eq. (7) can be decomposed as a difference of unnormalized measures (defined in Eq. (1)).

Proof: Under the Kapur entropy constraints $\alpha \neq 1$, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta - 1 > 0$, we obtain the following:

$$\begin{aligned}
&H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) \\
&= \frac{1}{1-\alpha}\log\frac{\sum_{\boldsymbol{h}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\alpha+\beta-1}}{\sum_{\boldsymbol{h}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\beta}} \\
&= \frac{1}{1-\alpha}\left(\log\sum_{\boldsymbol{h}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\alpha+\beta-1} - \log\sum_{\boldsymbol{h}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})^{\beta}\right)
\end{aligned}$$

Since $\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\theta) = \prod_i \psi_i(\boldsymbol{D}_i;\theta_i)$ (discussed in Section III-A), we define the following quantities:

$$\begin{aligned}
\tilde{P}_a(y,\boldsymbol{h}|\boldsymbol{x};\theta) &= \textstyle\prod_i \psi_i^a(\boldsymbol{D}_i;\theta_i) \\
\tilde{P}_b(y,\boldsymbol{h}|\boldsymbol{x};\theta) &= \textstyle\prod_i \psi_i^b(\boldsymbol{D}_i;\theta_i)
\end{aligned}$$

where each potential has the same scope but new values:

$$\begin{aligned}
\psi_i^a(\boldsymbol{D}_i;\theta_i) &= \psi_i(\boldsymbol{D}_i;\theta_i)^{\alpha+\beta-1}, \forall i \\
\psi_i^b(\boldsymbol{D}_i;\theta_i) &= \psi_i(\boldsymbol{D}_i;\theta_i)^{\beta}, \forall i
\end{aligned}$$

As a result, we obtain:

$$\begin{aligned}
&H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) \\
&= \frac{1}{1-\alpha}\left(\log\sum_{\boldsymbol{h}}\tilde{P}_a(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}) - \log\sum_{\boldsymbol{h}}\tilde{P}_b(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})\right) \\
&= \frac{1}{1-\alpha}\left(\log\tilde{P}_a(y|\boldsymbol{x};\boldsymbol{\theta}) - \log\tilde{P}_b(y|\boldsymbol{x};\boldsymbol{\theta})\right)
\end{aligned} \tag{8}$$

Therefore, $H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}))$ is expressed as a difference of unnormalized measures. ∎

Unnormalized measures in Eq. (8) can be very efficiently computed using belief propagation based on the data structure of a clique tree, which is commonly applied to perform inference in conventional HCRFs [23], [24], [25], [2], [26], [20]. If HCRF's latent variables $\boldsymbol{h}$ form an undirected tree $\mathcal{T}_h$, we can always construct a clique tree $\mathcal{T}_c$ using the following steps. First, we construct an undirected tree $\mathcal{T}_c$ that has the same topology as $\mathcal{T}_h$, and assign the singleton potentials $\psi(h_i,y)$ and $\psi(h_i,x_i)$ to the clique $\boldsymbol{C}_i$ with the scope $\{h_i,y,x_i\}$. Second, for each pair of the directly connected cliques $\boldsymbol{C}_i - \boldsymbol{C}_j$, we remove the edge between the cliques, add a new clique $\boldsymbol{C}_{ij}$ with the scope $\{h_i,h_j,y\}$ to form a chain $\boldsymbol{C}_i - \boldsymbol{C}_{ij} - \boldsymbol{C}_j$, and assign the pairwise potential $\psi(h_i,h_j,y)$ to the new clique $\boldsymbol{C}_{ij}$. It can be easily verified that the constructed tree $\mathcal{T}_c$ satisfies the family preservation property and the running intersection property [24], and thus is a clique tree. Because a linear chain structure is a special case of a tree, we can also construct a clique tree for our method. Given such a clique tree, our implementation of the belief propagation algorithm is present in Algorithm 1, with the computational complexity satisfies the following lemma (similar to previous works [2], [20]):

Algorithm 1: Sum-Product Belief Propagation

Input : HCRF's graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, graph potentials $\boldsymbol{\psi}(\boldsymbol{D})$
Output : $\tilde{P}(y|\boldsymbol{x};\boldsymbol{\theta})$

1: Construct clique tree $\mathcal{T}_c = \{\mathcal{V}_c,\mathcal{E}_c\}$ from $\mathcal{G}=(\mathcal{V},\mathcal{E})$;
2: **foreach** node $i \in \mathcal{V}_c$ **do**
3: Initialize clique potentials: $\varphi_i(\boldsymbol{C}_i) = \prod_{\psi_j:\alpha(\psi_j)=i}\psi_j(\boldsymbol{D}_j)$;
4: **while** $\exists i,j$: $\boldsymbol{C}_i$ is ready to send $\delta_{i\to j}(\boldsymbol{S}_{i,j})$ **do**
5: Compute and send the message: $\delta_{i\to j}(\boldsymbol{S}_{i,j}) = \sum_{\boldsymbol{C}_i-\boldsymbol{S}_{i,j}}\left(\varphi_i(\boldsymbol{C}_i)\cdot\prod_{k\in(\mathrm{Nb}_i-\{j\})}\delta_{k\to i}(\boldsymbol{S}_{k,i})\right)$
6: **end**
7: **foreach** node $i \in \mathcal{V}_c$ **do** Compute clique belief: $\beta_i(\boldsymbol{C}_i) = \varphi_i(\boldsymbol{C}_i)\cdot\prod_{k\in\mathrm{Nb}_i}\delta_{k\to i}(\boldsymbol{S}_{k,i})$;
8: **foreach** edge $i-j \in \mathcal{E}_c$ **do** Compute sepset belief: $\mu_{i,j}(\boldsymbol{S}_{i,j}) = \sum_{\boldsymbol{C}_i-\boldsymbol{S}_{i,j}}\beta_i(\boldsymbol{C}_i)$;
9: Compute $\tilde{P}(y|\boldsymbol{x};\boldsymbol{\theta}) = \frac{\prod_{i\in\mathcal{V}_\mathcal{T}}\beta_i(\boldsymbol{C}_i)}{\prod_{(i-j)\in\mathcal{E}_\mathcal{T}}\mu_{i,j}(\boldsymbol{S}_{i,j})}$;
10: **return** $\tilde{P}(y|\boldsymbol{x};\boldsymbol{\theta})$

Lemma 3: Algorithm 1 requires $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$ to compute the quantity $\tilde{P}(y|\boldsymbol{x};\boldsymbol{\theta}) = \sum_{\boldsymbol{h}}\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})$.

Proof: The clique tree $\mathcal{T}_c$ is constructed in $O(|\mathcal{V}|)$ time (line 1). Each pairwise potential is assigned to its corresponding clique in $O(1)$ time. Each singleton potential requires $|\mathcal{H}|$ multiplication to be assigned, and the upper bound for the number of such cliques is $|\mathcal{V}|$. Therefore, the clique potentials are initialized (line 3) in an $O(|\mathcal{V}||\mathcal{H}|)$ runtime. Given a fixed value of $y \in \mathcal{Y}$, there are $2|\mathcal{E}|$ messages that are passed over $\mathcal{T}_c$, each of which requires $O(|\mathcal{H}|^2)$ time to compute, resulting in an $O(|\mathcal{E}||\mathcal{H}|^2)$ runtime. Accordingly, $\forall y$, the total runtime is $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$ (line 5). Finally, $\forall y$, clique beliefs (line 7) and sepset beliefs (line 8) are computed in $O(|\mathcal{V}||\mathcal{Y}||\mathcal{H}|^2)$ and $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$ time, respectively. Thus, Algorithm 1 is performed at $O((|\mathcal{E}|+|\mathcal{V}|)|\mathcal{Y}||\mathcal{H}|^2)$. Since $|\mathcal{E}| = |\mathcal{V}|-1$ in a tree-structured graph[1], the overall time complexity of Algorithm 1 is $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$. ∎

Then, the inference tractability is presented in the following theorem:

[1]Loopy graphs satisfy $|\mathcal{E}| \geq |\mathcal{V}|$.

Theorem 2: Solving the inference problem in Eq. (7) is tractable, with the time complexity of $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$.

Proof: According to Lemma 2, we obtain

$$H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta})) = \frac{1}{1-\alpha}\left(\log \tilde{P}_a(y|\boldsymbol{x};\boldsymbol{\theta}) - \log \tilde{P}_b(y|\boldsymbol{x};\boldsymbol{\theta})\right)$$

Lemma 3 shows the unnomalized measures $\tilde{P}_b(y|\boldsymbol{x};\boldsymbol{\theta})$ and $\tilde{P}_b(y|\boldsymbol{x};\boldsymbol{\theta})$ can be computed in $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$ time. Thus $H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}))$ can be computed with the same runtime. Since $\operatorname{argmin}(\cdot)$ is performed in $O(|\mathcal{Y}|)$ time. Solving the problem in Eq. (7) takes $O(|\mathcal{E}||\mathcal{Y}||\mathcal{H}|^2)$ time. ■

C. Learning

Given i.i.d. training data $\mathcal{D} = \{(\boldsymbol{x}^i, y^i), i = 1,\ldots,N\}$. The method's parameters $\boldsymbol{\theta}$ are estimated by minimizing the following l_2-regularized loss $l(\boldsymbol{\theta}) = \lambda l_r(\boldsymbol{\theta}) + l_{emp}(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* \operatorname*{argmin}_{\boldsymbol{\theta}} \quad \lambda\left(\frac{\|\boldsymbol{\theta}\|^2}{2}\right) + \left(\frac{1}{N}\sum_{i=1}^{N}\max_{y'^i\in\mathcal{Y}}\left(\Delta(y^i,y'^i)+e(y^i,\boldsymbol{x}^i;\boldsymbol{\theta})-e(y'^i,\boldsymbol{x}^i;\boldsymbol{\theta})\right)\right) \tag{9}$$

where $e(y,\boldsymbol{x};\boldsymbol{\theta}) = H_{\alpha,\beta}(\tilde{P}(y,\boldsymbol{h}|\boldsymbol{x};\boldsymbol{\theta}))$. This problem can be solved using the BCP algorithm [18] that is based on the cutting plane technique [27]. A cutting plane of $l_{emp}(\boldsymbol{\theta}) = l_{emp}(\boldsymbol{x},y;\boldsymbol{\theta})$ at $\boldsymbol{\theta}'$ is defined as:

$$\begin{array}{rcl} c_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) & = & a_{\boldsymbol{\theta}'}^{\mathsf{T}}\boldsymbol{\theta} + b_{\boldsymbol{\theta}'} \\ \text{subject to} & & c_{\boldsymbol{\theta}'}(\boldsymbol{\theta}') = l_{emp}(\boldsymbol{\theta}') \\ & & \partial_{\boldsymbol{\theta}} c_{\boldsymbol{\theta}'}(\boldsymbol{\theta}') \in \partial_{\boldsymbol{\theta}} l_{emp}(\boldsymbol{\theta}') \end{array} \tag{10}$$

where $a_{\boldsymbol{\theta}'} = \partial_{\boldsymbol{\theta}} l_{emp}(\boldsymbol{\theta}')$, and $b_{\boldsymbol{\theta}'} = l_{emp}(\boldsymbol{\theta}') - a_{\boldsymbol{\theta}'}^{\mathsf{T}}\boldsymbol{\theta}'$. The cutting plane $c_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$ is a linear lower bound of $l_{emp}(\boldsymbol{\theta})$. The BCP method iteratively builds an increasingly accurate piecewise quadratic lower bound of $l(\boldsymbol{\theta})$. Given an initial value, $\boldsymbol{\theta}$ is iteratively updated by:

$$\begin{array}{l} \boldsymbol{\theta}_{t+1} = \operatorname*{argmin}_{\boldsymbol{\theta}} g_t(\boldsymbol{\theta}) \quad \text{and} \quad v_t = \min_{\boldsymbol{\theta}} g_t(\boldsymbol{\theta}) \\ \text{with} \quad g_t(\boldsymbol{\theta}) = \lambda \cdot l_r(\boldsymbol{\theta}) + \max_{j=1,\ldots,t}(c_j(\boldsymbol{\theta})) \end{array} \tag{11}$$

If $l_{emp}(\boldsymbol{\theta})$ is convex, $c_j(\boldsymbol{\theta}) \equiv c_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$ as defined in Eq. (10).

However, the objective function for learning of our model's parameters is not convex in general, and the commonly used convex solvers, such as BCP, cannot solve Eq. (9). To solve this non-convex optimization problem, we adopt the NRBM algorithm [19] that is described in Algorithm 2. Since a cutting plane of $l_{emp}(\boldsymbol{\theta})$ is not necessarily a lower bound, a *conflict* occurs if and only if the cutting plane does *not* satisfy the following constraint:

$$c_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_t^*) = a_{\boldsymbol{\theta}_t}^{\mathsf{T}}\boldsymbol{\theta}_t^* + b_{\boldsymbol{\theta}_t} \le l_{emp}(\boldsymbol{\theta}_t^*) \tag{12}$$

where $\boldsymbol{\theta}_t^*$ are the best observed parameters up to now (line 4 in Algorithm 2), in which case $l_{emp}(\boldsymbol{\theta})$ is overestimated at $\boldsymbol{\theta}_t^*$. The conflict is solved by tuning the parameters a_t and b_t to form an alternative cutting plane, $c_t(\boldsymbol{\theta}_t) = a_t^{\mathsf{T}}\boldsymbol{\theta}_t + b_t$, which satisfies Eq. (12) and the following condition:

$$\lambda l_r(\boldsymbol{\theta}_t) + c_t(\boldsymbol{\theta}_t) \ge l(\boldsymbol{\theta}_t^*) \tag{13}$$

Algorithm 2: NRBM for Learning MU-HCRFs

Input : $\mathcal{T}_c$, $\boldsymbol{\psi}(\boldsymbol{D})$, $\boldsymbol{\theta}_0$, λ, ϵ, $\mathcal{D}=\{(\boldsymbol{x}^i,y^i), i=1,\ldots,N\}$
Output : $\boldsymbol{\theta}^*$

1: **for** $t \leftarrow 0$ **to** ∞ **do**
2: Compute $l_{emp}(\boldsymbol{\theta})$ over $\mathcal{D}$ acc. to Eq. (9);
3: Define $c_{\boldsymbol{\theta}_t}$ with parameters $(a_{\boldsymbol{\theta}_t}, b_{\boldsymbol{\theta}_t})$ acc. to Eq. (10);
4: Compute $\boldsymbol{\theta}_t^* = \operatorname{argmin}_{\boldsymbol{\theta}_j\in\{\boldsymbol{\theta}_0,\ldots,\boldsymbol{\theta}_t\}} l(\boldsymbol{\theta}_j)$;
5: **if** $c_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_t^*) = a_{\boldsymbol{\theta}_t}^{\mathsf{T}}\boldsymbol{\theta}_t^* + b_{\boldsymbol{\theta}_t} > l_{emp}(\boldsymbol{\theta}_t^*)$ **then**
/*There is a conflict*/
6: Compute upper bound U of b_t acc. to Eq. (12): $U = l_{emp}(\boldsymbol{\theta}_t^*) - a_{\boldsymbol{\theta}_t}^{\mathsf{T}}\boldsymbol{\theta}_t^* \ge b_t$;
7: Compute lower bound L of b_t acc. to Eq. (13): $L = l(\boldsymbol{\theta}_t^*) - \lambda l_r(\boldsymbol{\theta}_t) - a_{\boldsymbol{\theta}_t}^{\mathsf{T}}\boldsymbol{\theta}_t \le b_t$;
8: **if** $L \le U$ **then** Set $a_t = a_{\boldsymbol{\theta}_t}$ and $b_t = L$;
9: **else** Assign $a_t = -\lambda \cdot \partial_{\boldsymbol{\theta}} l_r(\boldsymbol{\theta}_t^*)$ and $b_t = l(\boldsymbol{\theta}_t^*) - \lambda l_r(\boldsymbol{\theta}_t) - a_t^{\mathsf{T}}\boldsymbol{\theta}_t$;
10: Define alternative cutting plane: $c_t(\boldsymbol{\theta}) = a_t^{\mathsf{T}}\boldsymbol{\theta} + b_t$;
11: **else** Set $c_t(\boldsymbol{\theta}) = c_{\boldsymbol{\theta}_t}(\boldsymbol{\theta})$;
12: Update $\boldsymbol{\theta}_{t+1}$ and compute v_t acc. to Eq. (11);
13: Compute gap: $G_t = l(\boldsymbol{\theta}_t^*) - v_t$;
14: **if** $G_t \le \epsilon$ **then return** $\boldsymbol{\theta}_t^*$;
15: **end**

The conflict resolution procedure is described between line 5 and line 11. Using the l_2-regularizer, the NRBM algorithm is guaranteed to produce an approximation gap smaller than ϵ after T iterations and to converge with a convergence rate $O(1/(\lambda\epsilon))$ [19], where $T \le T_0 + 8C^2/(\lambda\epsilon) - 2$ with $T_0 = 2\log(\lambda\|\boldsymbol{\theta}_0 + a_0/\lambda\|/C) - 2$, and C is an upper bound on the norm of the cutting plane direction parameters.

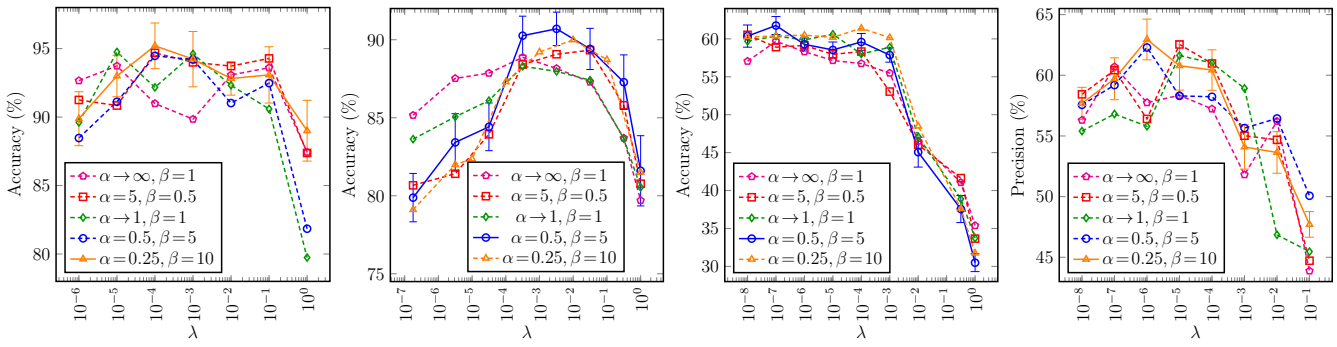
V. Experiments

Extensive empirical study is performed to evaluate the performance of our μHCRF methods on classifying sequential activities, which are represented by BoW or skeleton motion sequences (SMS). We split each dataset into disjoint training and testing sets. Fivefold cross-validation is employed over the training set to estimate model hyper-parameters.

A. Cornell Activity Dataset

The CAD-60 dataset [28] was collected as a benchmark to evaluate personal robots' capability to reason about human behaviors, which provides skeleton motion sequences in 3D space along with color-depth videos captured by a Microsoft Kinect camera. The CAD dataset contains twelve activities of daily living, which are performed by four human subjects in five different environments. We utilize the SMS features that are provided by the dataset, which represent human activities using 15 skeleton joints in 3D space. Following [28], we adopt the "have seen" experimental setting, and randomly select 70% of each subject's available data for hyper-parameter selection and training. As in [28], the performance is reported using precision and recall. In addition, we use accuracy as the performance metric for hyper-parameter selection and model evaluation.

Figure 2(a) and Figure 3(a) illustrate our μHCRF model's accuracy variations on the training sets using different hyper-parameter settings. Given $\lambda = 10^{-4}$, $\alpha = 0.25$ and $\beta =$



(a) CAD-60 (b) MSR Action3D (c) HMDB51 (d) Hollywood-2

Fig. 2. Performance variations of Our μHCRF models on training sets using different hyper-parameter settings. For a clear presentation, standard deviations are depicted only on the curves that contain the best results (depicted with solid lines).

10, our approach achieves the best average accuracy of $94.8 \pm 1.77\%$ over the training set. Employing the same set of hyper-parameter values, our μHCRF model achieves a 92.8% accuracy on the testing set, with a precision/recall of 89.7%/88.6%, which outperforms previous approaches, as demonstrated in Table I. The experiment highlights our μHCRF method's capability of dealing with traditional global skeleton features that have become more accessible with the emergence of affordable color-depth cameras.

TABLE I

PERFORMANCE COMPARISON OF OUR μHCRF MODEL WITH PREVIOUS APPROACHES ON THE CAD-60 DATASET.

Approach	Accuracy (%)	Precision (%)	Recall (%)
SVM [28]	—	66.4	56.0
Piyathilaka et al. [17]	—	84.0	73.0
Sung *et al.* [28]	—	84.7	83.2
Ni *et al.* [29]	65.3	—	—
Wang *et al.* [30]	74.7	—	—
MM-HCRF [30]	88.7	86.5	84.9
μHCRF	**92.8**	**89.7**	**88.6**

B. MSR Action3D Dataset

The MSR Action3D dataset [30] was collected to benchmark the interaction capability of Kinect-enhanced gaming consoles with human subjects, which contains 567 sequences of skeleton motions and depth images, which are grouped into 20 activity classes. Each activity is performed by ten subjects two or three times. In our experiment, we use five subjects for training and the rest for testing as in [30], [12].

Following [12], the HON4D features [12] are applied to represent human activities in depth videos, which generate a 120-dimensional histogram. We extract HON4D features from each frame in a depth video to construct a temporal sequence of such histograms, which serves as the input to our model. Figure 2(b) and Figure 3(b) demonstrate our μHCRF model's accuracy over the training data using different hyper-parameter settings; our model obtains the best cross-validation accuracy of $92.98 \pm 1.01\%$ when $\lambda = 10^{-2}$, $\alpha = 0.5$ and $\beta = 5$. Using these hyper-parameters, our μHCRF model achieves an accuracy of 92.17% on the testing dataset. Comparisons with previous approaches in Table II indicate that our approach achieves the state-of-the-art result.

To show our μHCRF method's capability of modeling sequential activities that are represented by SMS features, we conduct an additional experiment using the skeleton features provided with the dataset. Each skeleton pose contains 20 joint positions with four values per joint. After selecting the hyper-parameter values using the training set, we evaluate our model over the testing set and obtain an accuracy of 90.77%. As compared in Table II, using SMS features, our method still achieves good accuracy that is comparable to the state-of-the-art, although it does not perform as well as our μHCRF method using HON4D features.

C. HMDB51 Dataset

To evaluate our μHCRF's performance in a more realistic daily living scenario, we perform empirical studies based on the HMDB51 dataset [31], which was collected from public video resources including Google Videos and YouTube. The dataset contains 6766 videos in 51 activity categories, each of which has at least 101 instances. We adopt the three training-testing splits provided by the authors for evaluation [31], [32]. Each split contains 70 training and 30 testing clips from every class. Average classification accuracy is applied as our evaluation measure. The same HOG/HOF/MBH features (using improved dense trajectories) in the previous Hollywood-2 experiment are also applied on this dataset.

The average accuracy of our μHCRF model using a variety of hyper-parameter values on the training set is presented in Figure 2(c). Given the hyper-parameters $\alpha = 0.5$, $\beta = 5$ and $\lambda = 10^{-7}$, our approach obtains the best cross-validation accuracy of $61.6 \pm 1.08\%$. Our model's robustness to α and β given $\lambda = 10^{-6}$ is also analyzed, as shown in Figure 3(c). Similar to what we observe in the experiment using the Hollywood-2 dataset, carefully selecting entropy hyper-parameters increases the μHCRF model's activity recognition accuracy. Using these hyper-parameter values, our μHCRF model achieves a classification accuracy of 58.1% over the testing dataset, as demonstrated in Table II. We compare our max-certainty model's performance with the results reported

in previous works in Table II, which highlights our μHCRF model's superior accuracy for human activity recognition.

TABLE II
COMPARISON OF AVERAGE CLASSIFICATION ACCURACY (%) OVER THE COLOR HMDB51 AND DEPTH MSR ACTION3D DATASETS.

HMDB51	Acc. (%)	MSR Action3D	Acc. (%)
Wang *et al.* [10]	48.3	Yang et al. [33]	85.52
Jain *et al.* [34]	52.1	Wang et al. [13]	86.50
Wang *et al.* [35]	57.2	Wang et al. [30]	88.20
MM-HCRF [3]	53.8	Oreifej et al. [12]	88.89
HCRF	50.6	μ**HCRF** (SMS)	90.77
μ**HCRF**	**58.1**	μ**HCRF** (HON4D)	**92.17**

Furthermore, we compare our μHCRF approach, which explicitly models the uncertainty of the latent variables, with other HCRFs. As presented in Table II, the μHCRF model obtains much better results than other previous HCRFs, which demonstrates that explicitly modeling uncertainty in the latent temporal pattern can improve recognition accuracy. At last, compared with conventional MM-HCRF approaches [3], our μHCRF method obtains better performance. To summarize, the comparison in Table II highlights the benefit of modeling the uncertainty in underlying temporal patterns, which results in the state-of-the-art activity recognition accuracy. Similar conclusions are also observed in our other experiments.

D. Hollywood-2 Dataset

In order to explicitly evaluate our methods over sequential activities, we conduct experiments over the Hollywood-2 dataset [36], which contains 12 categories including sequential activities. This dataset contains unconstrained activities from realistic daily living scenes; instances of each activity are typically viewed from different camera angles. Following the standard experimental settings [11], [37], [38], [36], [39], [10], the dataset is divided into 823 training and 884 testing instances; performance is evaluated using precision.

We use the standard BoW representation to evaluate our model. After applying cuboid detectors [9], following [35], we construct a codebook for the HOG, HOF, and MBH descriptors via the k-means quantization. We fix the number of visual words for each descriptor to be 4000, which has empirically shown good results for a wide range of datasets. Then, a total number of 300 words are selected via a feature selection method [40] to reduce the complexity. The resulting histogram of visual word occurrences is computed from each frame in a video and used as our activity representation.

Figure 2(d) depicts our μHCRF model's precision over the training set across different hyper-parameter values. The best cross-validation precision, $62.95 \pm 1.67\%$, is obtained when $\alpha = 0.25$, $\beta = 10$, and $\lambda = 10^{-6}$. Our μHCRF approach's robustness to the entropy hyper-parameters α and β given $\lambda = 10^{-6}$ is shown in Figure 3(d). We observe that, given a fixed regularization hyper-parameter $\lambda = 10^{-7}$, a careful selection of the hyper-parameters α and β is able to improve human activity recognition performance. Using these hyper-parameters, our μHCRF approach obtains a 59.84% overall performance over the testing set.

As compared in Table III, the μHCRF model performs better than previous state-of-the-art approaches, on average. Most importantly, our μHCRF approach significantly improves classification precision over sequential activities, including SitDown, SitUp, StandUp, etc. The great precision improvements demonstrate the importance of modeling latent temporal patterns of sequential activities, and highlight our μHCRF model's superiority on recognizing sequential activities, such as StandUp and SitDown, as demonstrated by the blue font in Table III.

VI. CONCLUSION AND FUTURE WORK

We propose the new μHCRF method to identify sequential human activities, which is critical in many human-centered robotics applications but not well studied in previous work. Besides using the traditional negative log-likelihood of classes as the loss function, we propose a new regularizer that is able to model the uncertainty of latent variables to deal with the gradual transition between continuous human motions in a sequential activity. In addition, we prove that the inference problem in our μHCRF method is tractable, and the learning problem can be performed efficiently. Extensive experiments using public benchmark datasets validate that our μHCRFs approach achieves promising performance for human activity recognition, especially for identifying sequential behaviors. Future work will include implementing and optimizing our μHCRFs in physical robotic systems to enable human-robot interaction in the applications such as playing "Simon Says" games with kids [14].

REFERENCES

[1] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[2] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *PAMI*, vol. 29, no. 10, pp. 1848–1852, 2007.

[3] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," *PAMI*, vol. 33, no. 7, pp. 1310–1323, 2011.

[4] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *NIPS*, 2004.

[5] M. P. Kumar, B. Packer, and D. Koller, "Modeling latent variable uncertainty for loss-based learning," in *ICML*, 2012.

[6] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[7] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *TCSVT*, vol. 18, no. 11, pp. 1473–1488, 2008.

[8] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, pp. 107–123, 2005.

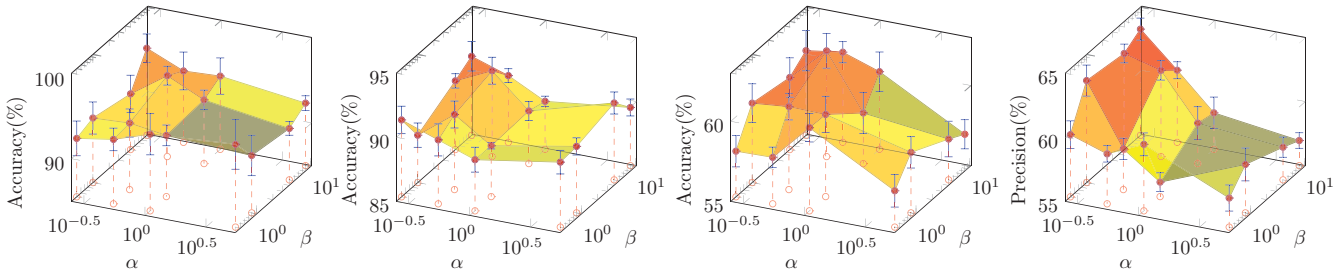
[9] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VSPETS*, 2005.

[10] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[11] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzílez, "Selective spatio-temporal interest points," *CVIU*, vol. 116, no. 3, pp. 396–410, Mar. 2012.

[12] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *CVPR*, 2013.

[13] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *ECCV*, 2012.



(a) CAD-60, $\lambda = 10^{-4}$ (b) MSR Action3D, $\lambda = 10^{-2}$ (c) HMDB51, $\lambda = 10^{-7}$ (d) Hollywood-2, $\lambda = 10^{-6}$

Fig. 3. Classification performance across hyper-parameters α and β, given a fixed λ. The solid circles represent the cases we test in our experiments, with error bars indicating classification errors obtained from cross-validation. The hollow circles are projections of our test cases onto the hyper-parameter plane to illustrate the hyper-parameter values that are employed in corresponding test cases. The surface connecting the points is generated using interpolation for a clear representation.

TABLE III

AVERAGE PRECISION (%) OVER THE HOLLYWOOD-2 DATASET DEMONSTRATES OUR MODEL'S STATE-OF-THE-ART PERFORMANCE, ESPECIALLY ON CLASSIFYING **SEQUENTIAL** HUMAN ACTIVITIES, INCLUDING SITDOWN, SITUP, AND STANDUP AS EMPHASIZED BY THE BLUE FONT.

Activity	Marszalek [36]	Han [41]	Derpanis [37]	Gilbert [38]	Ullah [39]	Wang [10]	Chakraborty [11]	μ**HCRF**
AnswerPhone	13.10	15.57	22.00	40.20	26.30	32.60	**41.60**	34.38
DriveCar	81.00	87.01	83.00	75.00	86.50	88.00	**88.49**	86.65
Eat	30.60	50.93	54.00	51.50	59.20	**65.20**	56.50	58.72
FightPerson	62.50	73.08	72.00	77.10	76.20	81.40	78.20	**82.12**
GetOutCar	8.6	27.19	32.00	45.60	45.70	52.70	47.37	**53.14**
HandShake	19.10	17.17	16.00	28.90	49.70	29.60	**52.50**	50.25
HugPerson	17.00	27.22	37.00	49.40	45.40	54.20	50.30	**54.34**
Kiss	57.60	42.91	59.00	56.60	59.00	**65.80**	57.35	57.20
Run	55.50	66.94	76.00	47.50	72.00	**82.10**	76.73	75.25
SitDown	30.00	41.61	56.00	62.00	62.40	62.50	62.50	**64.77**
SitUp	17.80	7.19	18.00	26.80	27.50	20.00	30.00	**33.65**
StandUp	33.50	48.61	56.00	50.70	58.80	65.20	60.00	**67.67**
Overall	35.50	42.12	48.00	50.90	55.70	58.30	58.46	**59.84**

[14] H. Zhang, C. Reardon, C. Zhang, and L. E. Parker, "Adaptive human-centered representation for activity recognition of multiple individuals from 3d point cloud sequences," in *ICRA*, 2015.

[15] Z. Zeng and Q. Ji, "Knowledge based activity recognition with dynamic bayesian network," in *ECCV*, 2010.

[16] M. Brand, "Structure and parameter learning via entropy minimization, with applications to mixture and hidden Markov models," in *ICASSP*, 1999.

[17] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features," in*ICIEA*, 2013.

[18] C. H. Teo, A. Smola, S. V. Vishwanathan, and Q. V. Le, "Bundle methods for regularized risk minimization," *JMLR*, vol. 11, pp. 311–365, 2010.

[19] T.-M.-T. Do and T. Artières, "Regularized bundle methods for convex and non-convex risks," *JMLR*, vol. 13, no. 1, pp. 3539–3583, Dec. 2012.

[20] Y. Song, L.-P. Morency, and R. Davis, "Action recognition by hierarchical sequence summarization," in *CVPR*, 2013.

[21] J. N. Kapur, "Generalized entropy of order α and type β," *Maths Seminar*, vol. 4, 1967.

[22] G. J. Klir, *Uncertainty & Information: Foundations of Generalized Information Theory*. Wiley-IEEE Press, 2005.

[23] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "Learning partially-observed hidden conditional random fields for facial expression recognition," in *CVPR*, 2009.

[24] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[25] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *CVPR*, 2007.

[26] Y. Song, D. Demirdjian, and R. Davis, "Multi-signal gesture recognition using temporal smoothing hidden conditional random fields," in *FG*, 2011.

[27] V. Franc and S. Sonnenburg, "Optimized cutting plane algorithm for large-scale risk minimization," *JMLR*, vol. 10, pp. 2157–2192, 2009.

[28] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *ICRA*, 2012.

[29] B. Ni, P. Moulin, and S. Yan, "Order-preserving sparse coding for sequence classification," in *ECCV*, 2012.

[30] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *PAMI*, vol. 36, no. 5, pp. 914–927, May 2014.

[31] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *ICCV*, 2011.

[32] S. Narayan and K. Ramakrishnan, "A cause and effect analysis of motion trajectories for modeling actions," in *CVPR*, 2014.

[33] X. Yang and Y. Tian, "EigenJoints-based action recognition using naive-Bayes-nearest-neighbor," *CVPRW*, 2012.

[34] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *CVPR*, 2013.

[35] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[36] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*, 2009.

[37] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," *PAMI*, vol. 35, no. 3, p. 527540, 2013.

[38] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features." in *ICCV*, 2009.

[39] M. M. Ullah, S. N. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues," in *BMVC*, 2010.

[40] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *JMLR*, vol. 13, no. 1, pp. 27–66, Jan. 2012.

[41] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *ICCV*, 2009.