2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

Robust Action Recognition and Segmentation with Multi-Task Conditional Random Fields

Masamichi Shimosaka, Taketoshi Mori and Tomomasa Sato

Abstract— **In this paper, we propose a robust recognition and segmentation method for daily actions with a novel Multi-Task sequence labeling algorithm called Multi-Task conditional random field (MT-CRF). Multi-Task sequence labeling is a task of assigning input sequence to sequence of multi-labels that consist of one or multiple symbols in single frame. Multi-Task sequence labeling is essential for action recognition, since motions can be often classified into multi-labels, e.g. he is folding arms while sitting. The MT-CRFs: extensions of conditional random fields (CRFs), incorporate jointly interaction between action labels as well as Markov property of actions, to improve the performance of the joint accuracy: the accuracy for whole labels at specific time. The MT-CRFs offer several advantages over the generative dynamic Bayesian networks (DBNs), which are often utilized as Multi-Task sequence labelers. First, the MT-CRFs allow relaxing the strong assumption of conditional independence of observed motion, which is used in DBNs. Second, the MT-CRFs exploit the power of non-Markovian discriminative classification frameworks instead of generative models in DBNs. With deep insight of the problem Multi-Task sequence labeling, the inference process of the classifier gains more efficiency than the previous Markov random fields that tackle Multi-Task sequence labeling. The experimental results show that classifiers with MT-CRFs have better performance than cascaded classifiers with a couple of CRFs.**

I. Introduction

Recognizing human action is one of essential foundations to achieve smooth communication between intelligent robotics systems and human. It is also a key technical element in achieving analysis and surveillance of human activity by intelligent systems. In action recognition, input is time-series human motion. Thus, it is interesting to formulate action recognition as a statistical sequence labeling problem where the output is a sequence of labels rather than a single label, as well as POS tagging in computational linguistics, function analysis in bioinformatics, and speech recognition. In mobile robotics, a sequence of range scan data and state of robots can be an input and an output of this problem.

There exist common factors for realizing robust sequence labeling in various domains. One is to leverage Markov assumption. In action recognition, this is related to the time-dependency problem or segmentation, which specifies the start and end points of action, because human requires a certain time interval to behave that action. This is also known as chunking in computational linguistics. Another important factor is to design good label-observation mapping: a mapping problem. For example of this for speech recognition, specific frequency of sound serves as a cue for estimating the specific phoneme.

Another factor to realize robust sequence labeling in practical problems is to incorporate multi-label problems [14]. Multi-label is a tuple of labels where the number of the symbols is variable where it is important to consider the pair of labels interact with each other. This is an essential for daily action recognition, since motions can be often classified into multi-labels, e.g. he is *folding arms* while *sitting*. In other words, it's not always true that all the labels to be annotated are exclusive such as pair of *standing* and *sitting*. Instead, it often occurs that there are non-exclusive pair of labels: *sitting on chair*, and *sitting* or *showing hand* and *standing*. In this paper, we call the sequence labeling problem where the output in a single frame is a multi-label as Multi-Task sequence labeling.

In order to incorporate the properties mentioned above, statistical approaches are proposed in many research works. Popular approach in this framework is to use dynamic Bayesian networks (DBNs) [8], such as hidden Markov models (HMMs) and their extensions [3]. Because of their systematic formulation, they have achieved privilege in action recognition domain [16]. However, there is a critical restrictions in the generative approach: strong assumption of conditional independence of observed motion. This restriction is related to the mapping problem. In action recognition, relevant motion features vary widely with the target actions. In DBNs, it is common to use a single or mixture of Gaussians to compute likelihood of labels from the observed motion. In case we want to classify actions regardless of actions ranging from dynamic action to static postures of human in systematic manner, DBNS limits a designer of labeling algorithms to utilize flexible motion cues.

As a resolution for the inflexibility of mapping design of DBNs, some researchers recently proposed flexible Markov-based models that allow observations to be represented as arbitrary overlapping features, e.g. conditional random fields (CRFs) [5]. They are not generative but their inference is in discriminative manner. This approach seems to be very good for us, and drive researchers to make a novel action recognition methodology [12], because they allow us to exploit motion cues or several non-Markovian discriminative method [1] as a mapping from motions to action.

In this paper, we propose an extention of CRFs to tackle the Multi-Task sequence labeling. It is possible to cascade CRFs for the Multi-Task sequence labeling. However, errors on early processing influence through the chain and cause errors in the final output. To attain higher joint accuracy: the accuracy for whole labels in single time, it is natural to couple CRFs systematically. In natural language processing,

Authors are with Dept. of Mechano-Informatics, The University of Tokyo, Japan simosaka@ics.t.u-tokyo.ac.jp

1-4244-0602-1/07/$20.00 ©2007 IEEE.

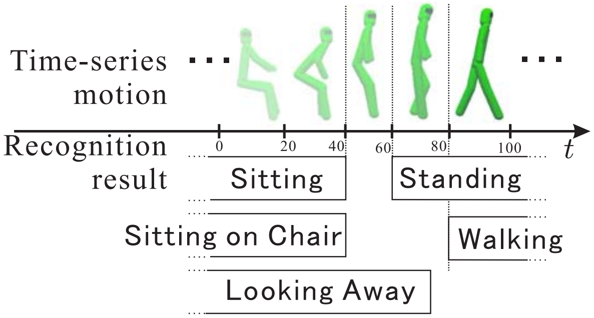

Fig. 1. Input and output of daily action recognition is shown. Input: time-series of human motion. Output: chunked recognition results in synchronization with input motion. It often occurs that multi-labels are annotated, like he is *sitting* (specifically *sitting on chair*) and *looking away* at $t = 30$.

an extetion of CRFs called a factorial CRFs [13]: a systematic Multi-Task sequence labeler, is already proposed and achieves higher performance than the traditional cascaded CRFs. But their inference process based on a loopy belief propagation [9] lacks the efficiency in action recognition. Hence, we propose an efficient alternative inference based on variational approach to focusing on the influence related to the interaction in multi-labels would be smaller than the interaction in Markov and the mapping property,

The rest of the paper proceeds as follows. Section II outlines action recognition framework with complicated semantics and the formulation of it as a Multi-Task sequence labeling problem. Section III introduces definition, labeling procedure and learning process of our new labeling model, MT-CRFs. Section IV presents results of several experiments about multiple-task sequence labeling. We conclude in section V with some directions for future research.

II. Time-Series daily action recognition as Multi-Task sequence labeling

The input of action recognition is time-series data of motion features and output of the recognition is a sequence of multi-labels that consist of one or multiple action symbols (see Fig. 1).

A. Graphical representation of Multi-Task sequence labeling

In this subsection, we model the time-series action recognition problem with graphical model representation, which is suitable for the Multi-Task sequence labeling. The structure used in this paper is shown in Fig. 2. This modeling can incorporate all the properties for robust action recognition: mapping property, Markov property, symbol interaction within a single frame. The variables for this problem is as follows: input data at time t is depicted by $\boldsymbol{x}_t \in \mathcal{X}$, where $\mathcal{X}$ means arbitrary motion data structure. Let $\boldsymbol{y}_t \in \boldsymbol{\mathcal{Y}} = \{\mathcal{Y}_1 \oplus \cdots \oplus \mathcal{Y}_K\}$ be a tuple of labels at time t, where $\mathcal{Y}_k$ represents set of symbols for k-th tasks. $y_{t,k}$ corresponds to the label of k-th task at that time. $X = \boldsymbol{x}_{1:T} \in \mathbb{S}_{\mathcal{X}}$ denotes a input sequence with length T and $Y = \boldsymbol{y}_{1:T} \in \mathbb{S}_{\boldsymbol{\mathcal{Y}}}$ indicates the corresponded label sequence. Collection $\mathbb{S}_A$ represents a set of sequences of set A. Let $Y_k = y_{1:T,k} \in \mathbb{S}_{\mathcal{Y}_k}$ be a sequence of labels for k-th task.

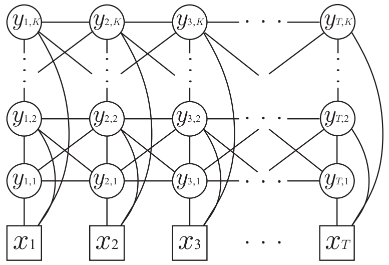

Fig. 2. Graphical model for Multi-Task Sequence Labeling. Circles represent hidden probabilistic variables and Squares denote observed non-probabilistic variables. In Multi-Task sequence labeling, a label is influenced by input data $\boldsymbol{x}$ and interacts with the other labels.

B. Semi-Hierarchical Representation of Actions

The above models and setting seems to provide us substantial information to implement recognizers, however, it remains an important issue to be solved before implementation. The issue is how to set $\mathcal{Y}_k$.

The most primitive approach in this setting defines each task as a binary classification e.g. "sitting" vs "non-sitting". This means the number of the symbols in each task is 2: $|\mathcal{Y}_k| = 2$. This means that a sequence labeler integrates the outputs of non-Markovian binary classifiers of single action symbol. However, this approach is too naive because the number of the tasks K grows linearly when the number of the target action increases. In addition, the complexity of the label interaction in single frame drastically increases. Hence another designing approach $\mathcal{Y}_k$ must be proposed.

To take deep insight of semantics of action symbols, there are some obvious relations of actions: 1) hierarchical representation, e.g. "sitting on chair" is a kind of "sitting", 2) exclusive relation, e.g. "standing" never occurs when human is "lying," 3) some relation that can be depicted by rule but has influence label assignment, e.g. "standing" does not influenced by "folding arms," however, "folding arms" never occurs when he is lying.

To incorporate the insights of action semantics for designing the set $\boldsymbol{\mathcal{Y}}$ and $\mathcal{Y}_k$, we adopt a semi-hierarchical structure of actions. An example of the structure of semantics of action is illustrated in Fig. 3. In this framework, there are several *groups* of action categories. For example, a group of action categorized by gazing full-body posture, what we call *root* group, contains "standing", "lying", "sitting", and "on four limbs". Another group called *sitting* group contains "sitting on chair" or "sitting on floor", *lying* group that treats lying actions and the rest group treats actions determined by arms posture. This structure provides us information of hierarchy, exclusiveness, the other relations between actions. For example this structure tells that *sitting on floor* is a kind of *sitting*, "lying" never occurs when "sitting" and "folding arms" may occur when "sitting" occurs. The reason why we adopt this categorization scheme is that this can make the Multi-Task sequence labeling with small number of tasks K and relatively compact size of $\mathcal{Y}_k$. Instead of using the semi-hierarchical structure, we can handle hierarchical structures with "flat" symbol space. However, the symbol space grows

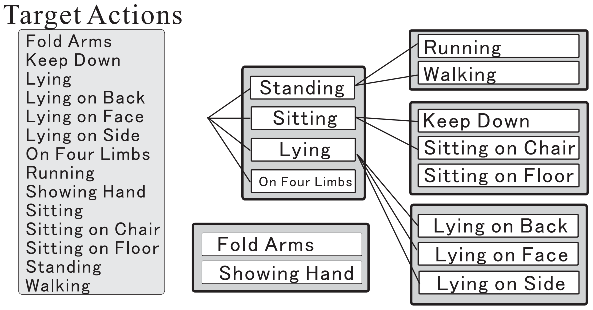

Fig. 3. Relation between Actions. It contains semi-hierarchical structures of actions.

very large when the layer of hierarchy grows. Put it all together, inference for action recognition in this paper can be formulated as integration of the results of couple of interdependent multi-class classifiers.

III. Multi-Task conditional random fields

In this section, we introduce a probabilistic model to annotate Multi-Task sequence labeling. At first, we introduce conditional random fields as a basis of our model, then we propose and define the Multi-Task Conditional Random Fields, and illustrate the process of their inference and learning from the data.

A. Conditional random fields: CRFs

Conditional Random Fields (CRFs) [5] are undirected graphical models that encode a conditional probability using a set of given feature templates. Originally, standard CRFs are developed as alternatives of hidden Markov models. In this paper, we call original CRFs as standard CRFs. In standard CRFs, a first-order Markov assumption is made on the label variables, and the number of the tasks is 1.

CRFs are defined as follows. The output of CRFs for $X = \boldsymbol{x}_{1:T}$ is $Y = y_{1:T}$. In order to incorporate first-order Markov assumption, local feature templates should be defined as $\boldsymbol{f}(X, y_{t-1}, y_t, t)$. For example, i-th feature template of $\boldsymbol{f}(X, y_{t-1}, y_t, t)$ is $\{\boldsymbol{f}(X, y_{t-1}, y_t, t)\}_i = [\![y_t = \text{“walking''}]\!][\![y_{t-1} = \text{“walking''}]\!]$, where $[\![b]\!]$ returns binary result of boolean value b. Furthermore, local feature templates can be freely designed with X. For example, i'-th feature template of $\boldsymbol{f}(X, y_{t-1}, y_t, t)$ is $\{\boldsymbol{f}(X, y_{t-1}, y_t, t)\}_{i'} = [\![y_t = \text{“walking''}]\!][\![v_t > \theta_{i'}]\!]$, where v_t denotes some motion information such as, forward velocity of hips. Informal interpretation of this template is “*walking makes human move forward*”. A parameter θ_i is a kind of adjustable threshold. Then a standard CRF can be defined with a probability distribution as

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\boldsymbol{w}^\top \boldsymbol{F}(X, Y)\right) \tag{1}$$

where $\boldsymbol{F}(X, Y) = \sum_t \boldsymbol{f}(X, y_{t-1}, y_t, t)$ is global feature vector of the sequence. The parameter $\boldsymbol{w}$ denotes a set of real weights and $Z(X)$ is a normalization factor of the distribution that satisfies $Z(X) = \sum_Y \exp\left(\boldsymbol{w}^\top \boldsymbol{F}(X, Y)\right)$ .In standard CRFs, probability of label sequence $p(Y|X)$ and $Z(X)$ can be analytically solved via generalized forward and backward algorithm similar to that of HMMs [5] once input sequence X is given.

B. Model representation

Following from the definition of the Multi-Task sequence labeling problem in the previous sections and borrowing the sense of standard CRFs, we formulate Multi-Task CRFs as

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\breve{\boldsymbol{w}}^\top \breve{\boldsymbol{F}}(X, Y) + \sum_k \boldsymbol{w}_k^\top \boldsymbol{F}_k(X, Y_k)\right), \tag{2}$$

where $Z(X) = \sum_Y \exp\left(\breve{\boldsymbol{w}}^\top \breve{\boldsymbol{F}}(X, Y) + \sum_k \boldsymbol{w}_k^\top \boldsymbol{F}_k(X, Y_k)\right)$. A “feature” template $\breve{\boldsymbol{F}}(X, Y)$ provides cues of mapping from input X to Y where each task is interdependent to the other tasks, and can be defined as $\breve{\boldsymbol{F}}(X, Y) = \sum_t \breve{\boldsymbol{f}}(X, \boldsymbol{y}_{t-1}, \boldsymbol{y}_t, t)$. For example of this feature template, i''-th feature template $\{\breve{\boldsymbol{f}}(X, \boldsymbol{y}_{t-1}, \boldsymbol{y}_t, t)\}_{i''} = [\![y_{t,k} = \text{“lying on side''}]\!] \cdot [\![y_{t,k'} = \text{“lying''}]\!]$. Another feature template $\boldsymbol{F}_k(X, Y_k)$ provides cues of mapping from input X to labels of k-th task Y_k, and can be denoted as $\boldsymbol{F}_k(X, Y_k) = \sum_t \boldsymbol{f}_k(X, y_{t-1,k}, y_{t,k}, t)$. This factorized representation means $\boldsymbol{F}_k$ depends only on the labels of the k-th task. The model parameters are a set of real weights $\boldsymbol{w} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K, \breve{\boldsymbol{w}}\}$. The weight parameter indicates how correct the feature template is for sequence labeling.

C. Inference in a MT-CRF

Unlike inference of standard CRFs, the inference cannot be solved analytically in a MT-CRF, because the Multi-Task sequence problem is a inference problem of the graphs with loops (see Fig. 2). Gibbs sampling for sequence labeling is very useful and easy to be implemented, however, the computational cost for the sampling is very expensive. Another major approach for this problem is to use loopy belief propagation (Loopy BP) [9] algorithms. This can be viewed as a general form of forward and backward algorithms of HMMs, and approximately estimates the posterior distribution of the label sequence. This method is known to be empirically successful for the inference in the graph with loops, however, naive implementation of Loopy BP makes the inference very slowly. Dynamic CRFs [13] utilized an efficient version of Loopy BP [15], however, some heuristics are inevitable to run this. Thanks to the result that the interaction in a multi-label would be smaller than the interaction in Markov and the mapping property, an alternative efficient inference method should be investigated.

In this research, we adopt structured variational approximation [2] as an alternative. The procedure of the inference is as follows. First, we approximate $p(Y|X)$ by some simple distribution $Q(\cdot)$ for inference. We factorize $Q(\cdot) = Q_X(Y; \boldsymbol{\nu}_{1:K})$ parameterized with auxiliary functions $\boldsymbol{\nu}_1(Y_1), \ldots, \boldsymbol{\nu}_K(Y_K)$ as $Q_X(Y; \boldsymbol{\nu}_{1:K}) = \prod_{k=1}^K q^{(k)}(Y_k; \boldsymbol{\nu}_k)$, to divide the Multi-Task labeling problem with a MT-CRF into a couple of Single-Task problems. Kullback-Leibler (KL) divergence is leveraged as the measure of the similarity between $p(Y|X)$ and $Q_X(Y;\boldsymbol{\nu}_{1:K})$ so as to get appropriate $Q_X(Y;\boldsymbol{\nu}_{1:K})$. KL divergence can be written as

$$\begin{aligned}&\mathrm{KL}(Q(Y;\boldsymbol{\nu}_{1:K})||p(Y|X))\\ &\quad= \langle \ln Q(Y;\boldsymbol{\nu}_{1:K}) - \ln p(Y|X)\rangle_{Q(Y;\boldsymbol{\nu}_{1:K})} \qquad (3)\end{aligned}$$

where $\langle \boldsymbol{f}(A)\rangle_{p(A)}$ represents the expected value of $\boldsymbol{f}(A)$ over a distribution $p(A)$. Then we minimize KL divergence between $Q_X(Y)$ and $p(Y|X)$ with respect to $\boldsymbol{\nu}_{1:K}$. We make an alias for approximated posterior distribution as $q^{(k)}(Y_k) := q(Y_k;\boldsymbol{\nu}_k)$ so as to keep the notation simple. In this research, we formulate the factorized approximated posterior distribution for k-th task, $q^{(k)}(Y_k)$, as

$$q^{(k)}(Y_k) = \frac{1}{Z_k(X)} \exp\left(\breve{\boldsymbol{w}}^{\mathsf{T}} \boldsymbol{\nu}_k(Y_k) + \boldsymbol{w}_k \boldsymbol{F}_k(X,Y_k)\right), \quad (4)$$

where $Z_k(Y_k) = \sum_{Y_k} \exp\left(\breve{\boldsymbol{w}}^{\mathsf{T}} \boldsymbol{\nu}_k(Y_k) + \boldsymbol{w}_k \boldsymbol{F}_k(X,Y_k)\right)$. The benefit with such a model factorization is that we can acquire exact $q^{(k)}(Y_k)$ and $\ln Z_k(X)$ efficiently, once $\boldsymbol{\nu}^{(k)}(Y_k)$ is given and if $\boldsymbol{\nu}^{(k)}(Y_k)$ can be written with first-order Markov assumption. This is because $q^{(k)}(Y_k)$ is equivalent to standard CRFs. With the formulation of $q^{(k)}$ and a result of the stationary point of KL divergence, we can acquire the optimal auxiliary function $\boldsymbol{\nu}_k^*$ as

$$\boldsymbol{\nu}_k^*(Y_k) \;=\; \sum_{k'\neq k}\sum_{Y_{k'}} q^{(k')}(Y_{k'})\breve{\boldsymbol{F}}(X,Y) \qquad (5)$$

This result leads to the intuitive interpretation: $\boldsymbol{\nu}_k^*$ is the expected function of the feature templates $\breve{\boldsymbol{F}}(X,Y)$ by all the approximated distributions $q^{(k')}(Y_{k'})$ except k-th task.

If we can assume the KL divergence is close to 0, then the log of the normalize factor of the MT-CRF can be approximated as $\ln Z(X) \approx \ln \tilde{Z}(X)$

$$\begin{aligned}\ln \tilde{Z}(X) \;\equiv\;& \breve{\boldsymbol{w}}^{\mathsf{T}} \left\langle \breve{\boldsymbol{F}}(X,Y)\right\rangle_{Q_X(Y)}\\ &+\sum_{k=1}^{K}\left(\ln Z_k(X) - \breve{\boldsymbol{w}}^{\mathsf{T}}\boldsymbol{\lambda}_k\right), \qquad (6)\end{aligned}$$

where $\boldsymbol{\lambda}_k = \langle \boldsymbol{\nu}_k^*(Y_k)\rangle_{q^{(k)}(Y_k)}$. In this inference, we must initialize the distribution $q^{(k)}(Y_k)$ and iteratively optimize the distribution $q^{(k)}(Y_k)$ with fixed-point iteration method until $\ln \tilde{Z}(X)$ converges. This is because the optimal auxiliary function $\boldsymbol{\nu}_k$ depends on the approximated distribution for the other tasks. In this paper, we initialize the distribution $q^{(k)}(Y_k) \propto 1$. Thus our inference algorithm with variational approximation is summarized in Table I.

D. Parameter estimation in a MT-CRF

The parameter estimation problem is to find a set of parameter vectors $\boldsymbol{w} = \{\breve{\boldsymbol{w}}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_K\}$ given the training dataset $D_{X,Y} = \{X^{(n)}, Y^{(n)}\}_{n=1}^N$. More specifically, we find the optimal parameter $\boldsymbol{w}$ by MAP estimation. From Bayes' theorem, the following relation satisfies

$$p(\boldsymbol{w}|D_{X,Y}) \propto p(\boldsymbol{w})\prod_{n=1}^{N} p(Y^{(n)}|X^{(n)}), \qquad (7)$$

TABLE I

INFERENCE ON MT-CRFS

0	Setting the approximated distributions $q^{(k)}(Y_k) \propto 1$ for $\forall k$, given input motion sequence X
1	Iterating k to update the approximated distribution $q^{(k)}(Y_k)$ and calculate $\ln Z_k(X)$ by using forward and backward procedure of standard CRFs. Before updating the distribution, the auxiliary function $\boldsymbol{\nu}_k(Y_k)$ can be calculated by the other approximated distributions as in (5).
2	Computing pseudo log of the normalization factor of MT-CRFs: $\ln \tilde{Z}(X)$ as in (6).
3	When $\ln \tilde{Z}(X)$ does not converge, returning to 1, otherwise, terminating the inference and outputting approximated distribution $p(Y|X) \approx Q_X(Y;\boldsymbol{\nu}_{1:K}) = \prod_{k=1}^{K} q^{(k)}(Y_k)$

hence, the optimal parameter can be defined as $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w})\prod_{n=1}^{N} p(Y^{(n)}|X^{(n)})$. In this research, we use Gaussian (normal) distribution as the prior distribution of $\boldsymbol{w}$: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, I/C)$ with $C > 0$ for simplicity. $\mathcal{N}(\cdot)$ represents Gaussian distribution as $\mathcal{N}(\boldsymbol{a}|\boldsymbol{\mu},\Sigma) \propto \exp\left(-\frac{1}{2}(\boldsymbol{a}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{a}-\boldsymbol{\mu})\right)$. MAP estimation under the above condition is equal to the following numerical optimization problem: $\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} J(\boldsymbol{w})$, where the target function $J(\boldsymbol{w})$ satisfies $J(\boldsymbol{w}) = \sum_{n=1}^{N}\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{F}(X^{(n)}, Y^{(n)}) - \ln Z(X^{(n)})\right) - \frac{C}{2}|\boldsymbol{w}|^2$. This optimization problem can be simply solved by several gradient-based methodsod because $J(\boldsymbol{w})$ is convex. In the implementation of the MT-CRFs, we use a limited memory version of BFGS update in quasi-Newton optimization algorithm [6]. The gradient of the MAP function w.r.t. $\boldsymbol{w}$ is $\nabla J(\boldsymbol{w}) = \sum_{n=1}^{N}\boldsymbol{F}(X^{(n)}, Y^{(n)}) - \sum_{n=1}^{N}\left\langle \boldsymbol{F}(X^{(n)}, Y^{(n)})\right\rangle_{p(Y^{(n)}|X^{(n)})} - C\boldsymbol{w}$. $J(\boldsymbol{w})$ requires $\ln Z(X)$ and its gradient requires expectation over the distribution $p(Y|X)$, however, both of them cannot be acquired analytically. Hence, we must replace $\ln Z(X^{(n)})$ by $\ln \tilde{Z}(X^{(n)})$, and $\left\langle \boldsymbol{F}(X^{(n)}, Y^{(n)})\right\rangle_{p(Y^{(n)}|X^{(n)})}$ by $\left\langle \boldsymbol{F}(X^{(n)}, Y^{(n)})\right\rangle_{Q_X(Y^{(n)};\boldsymbol{\nu}_{1:K})}$ from (6).

IV. EXPERIMENTAL RESULT

In this section, we illustrate the performance of the MT-CRFs in sequence labeling problem with synthetic and real time-series motion dataset.

A. Classification task evaluation with synthetic dataset

In this experiment, we evaluate the validity of MT-CRFs with synthetic dataset. The goal of this experiment is to clarify the tractability of the variational inference proposed in this paper, and to clarify the impact of leveraging graphical models represented in Fig. 2. Hence, we compare the performance of several types of CRFs.

a) **Dataset:** In this experiment, the task of inference is to annotate character sequences $\boldsymbol{x}_t \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \ldots, \mathrm{Z}\}$ by multiple labels as $y_{t,1} \in \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$, $y_{t,2} \in \{\mathrm{d}, \mathrm{e}, \mathrm{f}\}$. Here, the number of the tasks K is 2. We use multiple hidden Markov models to generate a synthetic data. To simulate the interaction between tasks, we must carefully design the rule of generation. Specifically, we design a basic Markov process for $y_{t,1}$ and the other three complementary Markov models for $y_{t,2}$. The complementary processes are influenced

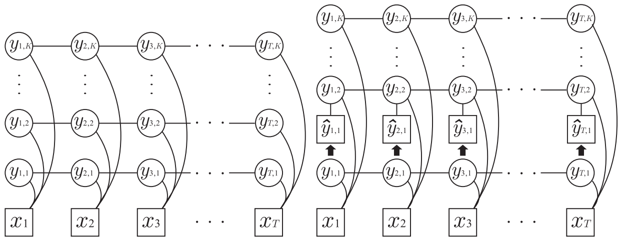

Fig. 4. Graphical model with parallel (left) and cascaded (right) style Multi-Task sequence labeling. In cascaded style of sequence labeling, estimated labels for 1st Single-Task problem $\hat{y}_{t,1}$ is used and equally treated as input $\boldsymbol{x}_t$ for 2nd Single-Task sequence labeling $y_{t,2}$.

by the output of the basic model. The basic Markov process is designed with a ring topology to generate $y_{t,1} = \{\mathrm{a}, \mathrm{b}, \mathrm{c}\}$. We design one of the complementary Markov models generates $y_{t,2} = \{\mathrm{e}, \mathrm{f}\}$ when $y_{t,1} = \mathrm{a}$. The others generates $y_{t,2} = \{\mathrm{d}, \mathrm{f}\}$ when $y_{t,1} = \mathrm{b}$, $y_{t,2} = \{\mathrm{d}, \mathrm{e}\}$ when $y_{t,1} = \mathrm{c}$. Thus the combination such as $y_{t,1} = \mathrm{a}$ and $y_{t,2} = \mathrm{d}$ never happens in this dataset. The parameters of these generators are described in the following sections.

b) **Evaluation method:** In order to clarify the validity of capability to incorporate the interaction between tasks, we evaluate the performance of MT-CRFs where the number of the tasks is 2, and compare the performance of the multiple standard CRFs. Specifically, we prepare three types of CRFs. First model to be compared is called parallel CRFs. In this model, the multiple tasks are factorized and executed in parallel and independently (see Fig. 4). Second model to be compared is called cascaded CRFs. In this model, the multiple tasks are factorized and executed in sequence. This model can tightly incorporate the relation between tasks, however, the model of accuracy will be poor when the performance of the task for the basic Markov process is poor (see Fig. 4). The final one is called flat CRF. This is a standard CRF where the number of the labels is 6, because we can convert the dataset with 2 tasks as a Single-Task sequence labeling problem with 6 states.

The evaluation criteria we used in this experiment are frame-wise accuracy: the count of the matches between the estimated result $\hat{y}_{t,k}$ and $y_{t,k}$ for $k = 1, 2$, and the joint accuracy: the count of the matches $\hat{\boldsymbol{y}}_t$ and $\boldsymbol{y}_t$. The performance is calculated via 20 times of training and testing.

c) **Parameters and condition:** In the basic Markov process, we set the transition probability to the other state is 0.1. We set the start probability distribution as flat. In the complementary Markov processes, the state is initialized when the state in the basic Markov process changes with flat probability. The transition probability to the other state is set 0.2. The emitter functions are attached to the complementary processes. They are designed to output $\boldsymbol{x}_t = \mathrm{A}, \mathrm{B}, \mathrm{C}$ with probability $20/83$ and the others with probability $1/83$ when the combination of labels is as $y_{t,1} = \mathrm{a}, y_{t,1} = \mathrm{e}$. The others are designed in same way. This setting of emitter functions leads to analytical result of the performance of non-Markovian local classification algorithms. The local classifier would achieve the accuracy 81.9 % and the joint accuracy 75.9 % respectively. The length of each sequence is set about 50 to 60. The dataset for training and testing is randomly generated and the number of each dataset is 100.

TABLE II

ACCURACY IN EACH METHOD FOR SYNTHETIC DATASET

	Parallel	Cascaded	Flat	**MT-CRF**
A	87.9 ± 0.5	86.9 ± 0.6	88.1 ± 0.6	$\mathbf{89.7 \pm 0.2}$
JA	79.6 ± 0.7	79.7 ± 0.8	82.2 ± 0.8	$\mathbf{83.9 \pm 0.4}$

As for all the CRFs, we set the parameter of the prior distribution of $\boldsymbol{w}$ as $C = 20.0$. We prepare the following feature templates: simple emitter functions that corresponding to $p(\boldsymbol{x}_t | y_{t,2})$ in HMMs, start features corresponding to $p(y_{1,\cdot})$, edge features $p(y_{t,\cdot} | y_{t-1,\cdot})$. As for MT-CRF, we utilize inter-label emitter functions that corresponding to $p(\boldsymbol{x}_t | y_{t,1}, y_{t,2})$. As for the variational inference in MT-CRFs, we set the terminal condition as the difference of $\ln \tilde{Z}(\cdot)$ through iteration is less than 10^{-3}.

d) **Result:** The performance in each model obtained of this experiment is summarized in Table II. Abbreviation A and JA in Table II represents accuracy and joint accuracy, respectively. From Table II, MT-CRF outperforms the other models. In other words, MT-CRF improve the performance the parallel CRFs model. It can be found that the cascaded model realized few improvement of joint accuracy compared to the parallel CRFs. The flat CRF achieves high performance than the parallel models, however, the performance is not so high as that of MT-CRF. From our qualitative analysis of this result, the result comes from that the flat CRF requires much more state transition parameters than the MT-CRF and the competitive result of the flat CRF relative to the MT-CRF requires larger size of the dataset. From another perspective of validating the proposed model, we calculated the number of iteration for variational inference. In this experiment, the pseudo log likelihood $\ln \tilde{Z}(\cdot)$ in (6) converges through 5.4 ± 1.4 times of iteration.

B. Classification task evaluation with real motion data

In this experiment, we evaluate the performance of action recognition based on MT-CRFs and compare the performance of the parallel and the cascaded CRFs.

At first we design the semantics of action: $\boldsymbol{y}_t$. In this experiment, we utilize the structural semantics as in Fig. 3. Specifically, we set labels as $y_{t,1} \in \{$ "lying", "on four limbs", "sitting", "standing", "other"$\}$, $y_{t,2} \in \{$"lying on back", "lying on face", "lying on side", "other"$\}$, $y_{t,3} \in \{$"keep down", "sitting on chair", "sitting on floor", "other"$\}$, $y_{t,4} \in \{$"stand still", "walking", "running", "other"$\}$, $y_{t,5} \in \{$"fold arms", "showing hand", "other"$\}$. This means the number of annotated actions at one frame is from 0 to 3, and the total number of action labels is 15.

Next we design local features from motion observation. It's important to remember that we target dynamical and posture action simultaneously in this experiment. This situation prevents us from utilizing a naive hidden Markov models because of the variety of relevant motion features. In this experiment, a discriminative classifier optimized support vector learning [11] is utilized as a strong cue for the label-observation mapping. There are several approaches to use binary discriminative classifiers for multi-class or multi-label problems, we adopt *one vs. the other* approach. Specifically, we optimize a binary classifier that discriminates whether some action occurs or not. In this experiment, we transfer the output of the classifier $h_{l_{k(c)}}(\boldsymbol{x}_t)$ for c-th action in k-th task $l_{k(c)}$ so as to interpret probability value

$$p_{\mathsf{local}}(y_{t,k} = l_{k(c)} | \boldsymbol{x}_t, h_{l_{k(c)}}) = \frac{1}{1 + \exp\left(-\sigma h_{l_{k(c)}}(\boldsymbol{x}_t)\right)} \quad (8)$$

$$h_{l_{k(c)}}(\boldsymbol{x}_t) = \sum_{n,t^{(n)}} \alpha^{(n)}_{t^{(n)}} \operatorname{sgn}\left(y^{(n)}_{t^{(n)},k} = l_{k(c)}\right) \mathcal{K}\left(\boldsymbol{x}_t, \boldsymbol{x}^{(n)}_{t^{(n)}}\right) + b, \quad (9)$$

where $\mathcal{K}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ represents a kernel function which computes similarity between motions. Specifically, $\mathcal{K}\left(\boldsymbol{x}_t, \boldsymbol{x}^{(n)}_{t^{(n)}}\right)$ represents similarity between input motion at t frame and the reference motion $\boldsymbol{x}^{(n)}_{t^{(n)}}$ at $t^{(n)}$ frame in n-th sequence. Function $\operatorname{sgn}(\cdot)$ returns $+1$ if an input is true, otherwise, returns -1. α and b represents real values optimized by support vector learning. This is similar idea of logistic regression. The parameter σ is optimized via maximum a posterior estimation with cross validation techniques [10]. Then we build feature templates for the CRFs as $\{\boldsymbol{f}(X, \boldsymbol{y}_{t-1}, \boldsymbol{y}_t, t)\}_i = [\![y_{t,k} = l_{k(c)}]\!] \cdot p_{\mathsf{local}}(y_{t,k} = l_{k(c)} | \boldsymbol{x}_t, h_{l_{k(c)}})$.

e) **Dataset:** In the following sentences, we illustrate the training and testing of motion dataset used in this experiment. We utilize ICS Action database [7]. This is a collection of annotated motion capture data. In the database, the motion sequences are annotated with 25 daily actions, such as walking and showing hand per frame. All the target actions in the experiments are involved in them. The motion data in the database contain human skeletal configuration and its time-series of joints angles acquired by a magnetic motion capture system. Specifically, the format of the motion data is BVH. The specification and the quantum of the motion data used in this experiment is as follows. An actor of this dataset is a 20s male. The number of degree of freedoms of the motion is 36. The Posture and position of the motions are measured by magnetic motion capture systems sampled at 30 Hz. The number of the files is 125. Total time of the files is about 400 sec. (avg. 3.2 sec.). It contains 5 sub-datasets. In each sub-dataset, an actor behaves similar actions, such as getting up, sitting on chair, and lying down. In this experiment, we utilize 5-fold cross validation: an iteration to execute the learning of CRFs and SVMs from 4 sub-datasets and evaluate the performance of the classifiers with the rest sub-dataset.

f) **Evaluation method:** At first, we compute the performance of the non-Markovian local kernel classifiers as the baseline performance of the sequence labeler. Specifically, we compute the joint accuracy of the training dataset as the performance of the classifier. As in the case of the experiment with synthetic dataset, we compare the performance of the several types of Multi-Task sequence labeler based on CRFs with the performance of the MT-CRFs. In this experiment, the number of the tasks of MT-CRFs is 5.

The first compared model is called parallel CRFs. These are 5 standard CRFs running independently. The second model is called cascaded CRFs. We cascade CRFs focusing on the hierarchical structures of actions. Hence, the inference procedure is done in the standard CRF for $y_{t,1}$, then the classifier independently infers the labels of $y_{t,2}$, $y_{t,3}$, $y_{t,4}$ leveraging $\hat{y}_{t,1}$. This type would lead poor performance when the performance of inference for $y_{t,1}$ is poor. The final model to be compared is called flat CRF. Because the hierarchical structure of actions can be transformed into flat symbol space. Specifically speaking, we can set the symbol space for $y_{t,1:4}$ to the flat space with 13 symbols. Then we set parallel type CRFs for the new flat symbol space and for $y_{t,5}$.

g) **Parameters and condition:** In this experiment, we utilize edge features that correspond to the transition probability of HMMs, and feature templates based on the local kernel classifiers. To build local kernel classifiers, we design kernels with two strategies. One is for posture action such as sitting and lying. For these actions, it is a natural idea to utilize motion features to be recovered posture information and to calculate similarity from the features. Thus we select motion features for the input of kernel: posture and position of the specific body parts. Specifically, we selected the orientation of hips and height of hips, and the positions of head, hands, and foots with respect to the frame of hips. In this setting, a input feature $\boldsymbol{x}_t$ can de depicted by a vector of 18 dimension. Next, we set radial basis functions (RBF) as the kernel function $\mathcal{K}(\cdot,\cdot)$. This can be written as $\mathcal{K}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \exp\left(-\frac{|\boldsymbol{x}-\tilde{\boldsymbol{x}}|^2}{\sigma^2}\right)$, where σ is an adjustable positive parameter, and in our setting, $\boldsymbol{x}$ is a 18 dimensional vector that contains the selected motion features noted above. In this experiment, we optimize the local classifier a priori. In this optimization procedure, the parameter σ in RBF kernel is set as $\sigma = 0.80$.

Another strategy to design kernels is for dynamic action such as walking and running. Because the posture of such actions varies from time to time, thus we exploit dynamics property of gait motions. Specifically speaking, we utilize the foot motions w.r.t. the frame of hip as a motion feature and assume these motions are driven by linear dynamics. We compute similarity between the sequence of the motion features from the probability product kernels [4]. As for support vector learning [11], we set the max of α as 500.

As for all the CRFs, we set the parameter of the prior $p(\boldsymbol{w})$ as $C = 100$. The setting of terminal condition in variational inference is the same with the experiment with synthetic dataset. We also design inter-label feature templates. We set this with focusing on hierarchical structures. In this experiment the template is designed as $\{\check{\boldsymbol{f}}(X, \boldsymbol{y}_{t-1}, \boldsymbol{y}_t)\}_i = [\![y_{t,k} = l_{k(c)}]\!][\![y_{t,k'} = l_{k'(c')}]\!] p_{\mathsf{local}}\left(y_{t,k} = l_{k'(c')} | \boldsymbol{x}_t, h_{l_{k(c)}}\right)$, where k' represents child or parental category, and k represents k's parent or child category.

h) **Result:** The performance of the baseline: the result of local kernel classifiers is averaged 87.6 ± 4.9. In this experiment, the goal is to show the superiority of MT-CRFs

TABLE III
RELATIVE PERFORMANCE OF CRFS FOR ACTION RECOGNITION

	Parallel	Cascaded	Flat	MT-CRF
JA Error reduction	-10.8	0.97	-2.97	$\mathbf{+5.32}$

to the others. Thus we calculate the relative performance of the CRFs. Specifically, we calculate the rate of error reduction. The error is Joint Accuracy error. The largest error reduction score is equal to the best performance of the classifier. The average of the relative performance of the CRFs is shown in Table III. This result shows the proposed MT-CRF is superior to the other CRFs. Unlike the result of the experiment with synthetic dataset, the performance of the flat CRF is relatively poor. This would come from the number of the state: 13 is much larger than the other CRFs, and the quantum of the dataset is not enough to acquire the sufficient performance.

Fig. 5 shows a classification result of the sequence labelers built on this experiment for motion from sitting on floor (with keeping down shortly) to standing. Parallel CRFs output conflicted results at around $t = 50$ as "he is not sitting but keeping down." Cascaded version does not detect "keep down" even if the output of the corresponding SVM is high. Relative to these models, MT-CRF can output reasonable results for this motion. It is important to make mention that non-Markovian classifier mistakes to output the $y_{t=100,4} =$ "$walking''$ because the output of the SVM is quite large and the classifier ignores the output in successive frames. On the other hand, all the CRFs do not output mistaken result.

V. CONCLUSION

In this paper, we propose a robust action recognition framework with Multi-Task conditional random fields (MT-CRFs) that can treat multi-labels problem, hierarchical structures of action semantics, label-label interaction, and flexible designing framework of label-observation mapping. This model can be great extention of conditional random fields proposed by Lafferty and factorial dynamic Bayesian networks. To make efficient inference for this model, we made an efficient structured variational inference with smaller heuristics than the previous works. The experimental results using synthetic and real motion capture data show that our model outperforms the Multi-Task labeling frameworks with cascaded or parallel connected standard CRFs. It also clears that the convergence of variational inference of the model is very fast and stable.

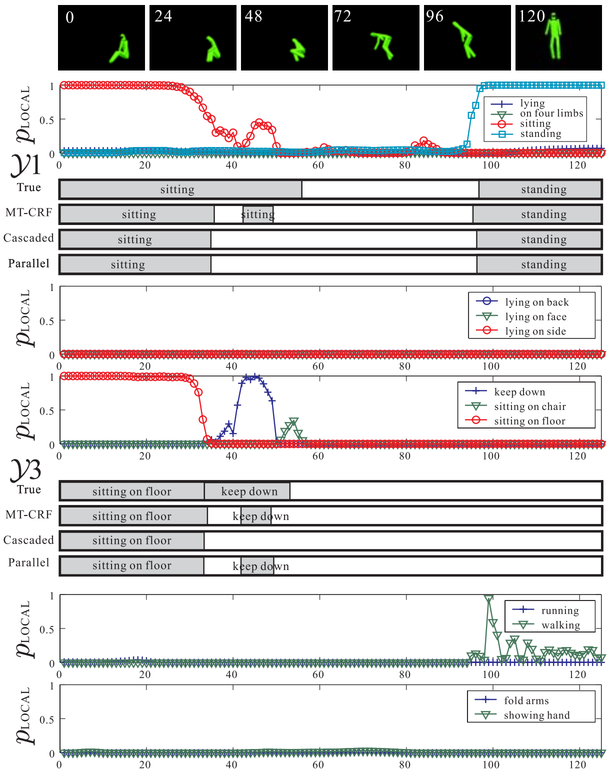

Fig. 5. Thumbnails of input standing up motion, time-series of p_{LOCAL} in each action, and classification results are shown. The 1st time-lined row represents an output of SVMs for $\mathcal{Y}_1$. The 2nd, 4th and 5th time-lined row represent the outputs of the sequence labelers. The 3rd time-lined row denotes the output of SVMs for $\mathcal{Y}_3$ and the next to this depicts the output of labels.

REFERENCES

[1] D. Cao, O. Masoud, D. Boley, and N. Papanikolopoulos. Online motion classification using support vector machines. In *Proc. of the 2004 ICRA*, volume 3, pages 2291–2296, 2004.

[2] Z. Ghahramani. On structured variational approximations . Technical Report CRG-TR-97-1, University of Toronto, 1997.

[3] Z. Ghahramani and M. Jordan. Factorial hidden Markov models. *Machine Learning*, 29(2-3):245–273, 1997.

[4] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th ICML*, pages 282–289, 2001.

[6] D. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(3):503–528, 1989.

[7] T. Mori, M. Shimosaka, and K. Tsujioka. ICS action database. http://www.ics.t.u-tokyo.ac.jp/action/, 2003.

[8] K. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[9] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proc. of the 15th Conference on UAI*, pages 467–475, 1999.

[10] J. Platt. Probabilities for SV machines. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[11] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[12] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional random fields for contextual human motion recognition. In *Proc. of the 10th ICCV*, volume 2, pages 1808–1815, 2005.

[13] C. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data . In *Proc. of the 21st ICML*, number 99, 2004.

[14] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *Advances in NIPS 15*, pages 721–728, 2003.

[15] M. Wainwright, T. Jaakkola, and A. Willsky. Tree-based reparameterization for approximate inference on loopy graphs. In *Advances in NIPS 14*, pages 1001–1008, 2002.

[16] A. Wilson and A. Bobick. Parametric hidden Markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):884–900, September 1999.