Noname manuscript No.
(will be inserted by the editor)

Human Action Segmentation and Recognition using Discriminative Semi-Markov Models

Qinfeng Shi · Li Wang · Li Cheng · Alex Smola

Received: date / Accepted: date

Abstract A challenging problem in human action understanding is to jointly segment and recognize human actions from an unseen video sequence, where one person performs a sequence of continuous actions.

In this paper, we propose a discriminative semi-Markov model approach, and define a set of features over boundary frames, segments as well as neighboring segments. This enable us to conveniently capture a combination of local and global features that best represent a specific action type. To efficiently solve the inference problem of simultaneously segmentation and recognition, we devise a Viterbi-like dynamic programming algorithm, which is able to process 20 frames per second in practice. Moreover, the model is discriminatively learned from large margin principle, and is formulated as an optimization problem with exponentially many constraints. To solve it efficiently, we present two different optimization algorithms, namely cutting plane method and bundle method, and demonstrate that each can be alternatively deployed in a "plug and play" fashion. From its theoretical aspect, we also analyze the generalization error of the proposed approach and provide a PAC-Bayes bound.

The proposed approach is evaluated on a variety of datasets, and is shown to perform competitively to the state-of-the-art methods. For example, on KTH dataset, it achieves $95\% \pm 0.01$ recognition accuracy, where the best known result on this dataset is $92\% \pm 0.03$ [8].

Keywords Action Segmentation and Recognition · Large-margin Method · Semi-Markov Model

A preliminary version has been published at [28].

Qingfeng Shi
NICTA and ANU, Canberra, Australia
E-mail: qinfeng.shi@rsise.anu.edu.au

Li Wang
Southeast University, Nanjing, China
E-mail: wang.li.seu.nj@gmail.com

Li Cheng
TTI-Chicago, USA
E-mail: licheng@tti-c.org

Alex Smola
Yahoo! Research, USA
E-mail: alex@smola.org

1 Introduction

A challenging problem in human action understanding is to recognize a sequence of continuous actions, that is, to segment and recognize elementary actions such as running, walking and drawing on board, from a video sequence where one person performs a sequence of such actions. This has a wide range of applications in *e.g.* surveillance, video retrieval and intelligent interface. It is nevertheless challenging due to the high variability of appearances, shapes and possible occlusions. Things are further complicated for continuous action recognition since it is also necessary to segment the sequence of actions.

This problem could however be addressed by considering the proper temporal context of each elementary action. Motivated by this observation, in this paper, we consider a discriminative learning approach that is capable of incorporating both local and long-range information. To better motivate our proposed model, we will describe in turn three categories of statistical models that can be used to represent human actions (illustrated from top to bottom panels in Figure 1).

Figure 1 (a) depicts the first category of models: By simply ignoring the temporal dependencies among video frames, each frame is assumed to be independent of the rest. Models such as support vector machines

(SVMs), naive Bayes classifier, nearest neighbor classifier (KNNs) fall into this category. There is however a significant limitation in their prediction abilities when analyzing action sequences. this limitation is partially circumvented in [21, 27, 36], where their feature descriptors utilize the spatial-temporal characteristics of each type of actions, and a variant of the first model is applied in this feature space to decide to which category a new action sequence belongs. They nevertheless require a pre-segmentation of the continuous action sequence into elementary segments, a tedious manual operation.

The second category of models is the Markov chain models delineated in Figure 1 (b) (include *e.g.* hidden Markov models (HMMs) [2, 16], conditional random fields (CRFs) [30] or SVM-HMMs [34]) that consider statistical dependencies over adjacent frames and show good performance on pre-segmented datasets. We argue that these models are not well suited to the problem considered in this paper. First, continuous action recognition inherently has a segmentation problem, where each action starts, lasts for a varying period of frames and then transits to another action. This is however difficult to be dealt with by Markov chain models. Second, although Markov chain models utilize local interaction between adjacent frames, it does not have access to long-range or global characteristics, such as the duration of one action segment, or interactions between adjacent segments.

1.1 Our Model

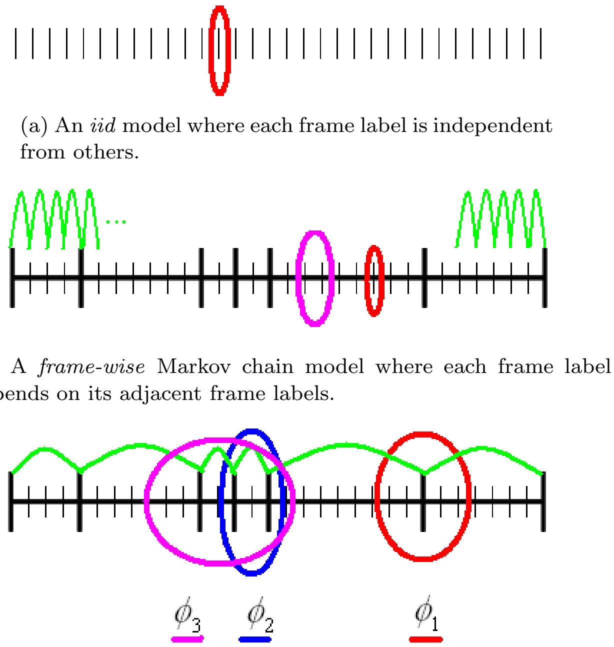
The third and the model we consider is a semi-Markov model (SMM) [5, 23], shown in Figure 1 (c). Essentially, it is an extension of HMM by allowing the underlying process to be a semi-Markov chain with a variable duration for each state. In particular, this enables the exploitation of the segmentation nature of our problem, where the modeling emphasis now shifts more towards segment-wise properties involving individual segments of variable length as well as adjacent segments.

Inspired by [5], we propose a *discriminative* SMM model, and define a set of distinct features at our disposal, which includes (a) the boundary frames of each segment, (b) the content characteristics of segments, and (c) the interactions between neighboring segments. This allows us to conveniently capture a combination of local as well as longer-range features that best represent a specific action type. It turns out that our discriminative SMM approach recovers the first two categories of models (i.e. multiclass SVM, SVM-HMM) as special cases, by properly setting the maximum segment length M and the feature function Φ that can be decomposed into (ϕ_1, ϕ_2, ϕ_3). They are depicted in Figure 1 (c), using (red, blue, purple) color, respectively. To efficiently solve the inference problem involving simultaneously segmentation and recognition, we devise a Viterbi-like dynamic programming algorithm that is able to process 20 frames per second in practice. This model is discriminatively learned from large margin principle, and is formulated as an optimization problem with exponentially many constraints. To solve the learning problem efficiently, we present two optimization algorithms, namely cutting plane method and bundle method, and demonstrate that each can be alternatively deployed in a "plug and play" fashion. From its theoretical aspect, we also analyze the generalization error of the proposed approach and provide a PAC-Bayes bound. Empirical simulations, presented in section 6, support that the proposed discriminative SMM approach is indeed well-suited to the problem of segmenting and recognizing human action sequences.

1.2 Related Literature

There exists a wealth of literature on topics related to human action recognition. As it is beyond the scope of this paper to review these existing literature, interested readers may refer to e.g. Gavrila [6] or Moeslund et al. [20] for a survey of the field. Here we instead focus our discussions only on closely related work.

Traditionally generative statistical approaches, especially the Markov models [2, 9, 16, 38] have been in wide use to model and analyze human actions, *e.g.* HMMs and its variants such as coupled HMMs [2, 38]. Recently, large margin based discriminative learning schemes [35] are extended to cases where there are structured dependencies among the outputs [25, 32, 34] (*e.g.* SVM-HMM where the output could be time series sequences), and encouraging results are obtained in bioinformatics and natural language processing related applications [25]. As far as we are aware, there is not much work along this line conducted in the field of video action analysis. The most relevant work is [25] where a SMM approach is proposed for gene structure prediction applications. They propose a two-stage learning algorithm where binary SVM classifiers are firstly used to identify segment boundaries and the *content* of each segment is recognized separately in the second stage. This sequential procedure is quite different from the unified approach we propose in section 3 where action segmentation and recognition can be dealt with simultaneously. We note in the passing that conditional random field (CRF), as a discriminative model that deals with structured outputs, has recently been applied to human action understanding where the underlining model is a Markov Chain model [30, 36]. Learning in CRF is approached by solving an induced optimization problem, which is closely related to that of SVM-HMM, but with a different loss function.



(a) An *iid* model where each frame label is independent from others.

(b) A *frame-wise* Markov chain model where each frame label depends on its adjacent frame labels.

(c) The proposed *semi-Markov model* where frames in one segment share one label, and this label depends on its adjacent segment labels.

Fig. 1: We compare three categories of statistical models on the continuous action sequence prediction problem. The Markovian dependency in each model is illustrated as green arcs. It turns out that our discriminative SMM approach recovers the first two models (i.e. multiclass SVM, SVM-HMM) as special cases, by properly setting the maximum segment length M and the feature function Φ that can be decomposed into (ϕ_1, ϕ_2, ϕ_3) and depicted in (red, blue, purple) color, respectively.

1.3 Paper Outline

The remaining sections are organized as follows: In section 2 we give a probabilistic account of the proposed discriminative approach. To solve the induced optimization problem, in section 3 we introduce efficient learning and inference algorithms. We proceed to provide details of the feature representation scheme in section 4, and analyze from theoretical viewpoint the generalization property of our approach in section 5. In section 6, extensive experiments are conducted on standard testbeds, where our approach is compared against state-of-the-art methods. This is followed by a summary as well as future directions in section 7.

2 Our Semi-Markov Model

Define the set of action labels as $\mathcal{C} = \{1, \cdots, C\}$, and the set of persons $\mathcal{I} = \{1, \cdots, I\}$. Without loss of generality, we assume that there is exactly one person P in a given video sequence X performing actions Y. In this paper, we formulate the human action analysis problem as solving a convex optimization problem over a probabilistic semi-Markov model.

Semi-Markov Model (SMM): *Consider a graph defined on the action sequence label Y for person $P \in \mathcal{I}$. More precisely we consider a semi-Markov model, where each node in this graph corresponds to a segment of video frames having the same action label, and each edge captures the statistical dependency between adjacent segments. Given a video sequence of length m as $X = \{x_k\}_{k=0}^{m-1}$, we attach a dummy node x_m to this sequence to denoate the end of the sequeunce. Let l denote the number of segments, and define a set of segment boundaries $\{n_k\}_{k=0}^{l-1}$ with $n_{k-1} < n_k < n_{k+1}, \forall k$. Fix $n_0 = 0$, $n_l = m$ to satisfy boundary conditions. As a consequence, the first segment is $[0, n_1)$, and the last segment is $[n_{l-1}, m)$. Its action sequence label can be equivalently represented as $Y = \{(n_k, c_k)\}_{k=0}^{l-1}$, where each pair (n_k, c_k) denotes the starting position and the corresponding action label for the kth segment $[n_k, n_{k+1})$.*

During training we have access to a set of T video sequences $\mathcal{X} = \{X_t\}_{t=1}^{T}$, as well as corresponding labels $\mathcal{Y} = \{Y_t\}_{t=1}^{T}$, accordingly. We further assume that the *conditional* distribution over the labels Y given the observation sequence $X = X_t$ can be written as a log-linear model,

$$\log p(Y|X, W) = \langle W, \Phi(X, Y)\rangle - A_W(X). \quad (1)$$

Therefore, the joint *conditional* probability over training sequences is $p(\mathcal{Y}|\mathcal{X}, W) = \prod_t p(Y_t|X_t, W)$, since all action sequences are statistically independent. Here $A_W(X)$ is the normalization constant to ensure $p(Y|X, W)$ respects a valid probability distribution, with $W \in \mathcal{W}$ being the parameter vector. $\Phi(X, Y)$ is a feature map over the joint input-output space, which can be decomposed with respect to the SMM graph structure of Figure 1 (c) as

$$\Phi(X, Y) = \Bigg(\sum_{i=0}^{l-1} \phi_1(X, n_i, c_i), \sum_{i=0}^{l-1} \phi_2(X, n_i, n_{i+1}, c_i), \sum_{i=0}^{l-1} \phi_3(X, n_i, n_{i+1}, c_i, c_{i+1}) \Bigg). \quad (2)$$

As indicated in the beginning of this paper, ϕ_1 and ϕ_2 capture the observation-label dependency in the current action segment: ϕ_1 concentrates on a segment's boundary frame, and ϕ_2 takes care of global characteristics of the segment. The interaction between two neighboring segments is encoded in ϕ_3. In the same manner, W can as well be decomposed.

For an unseen video sequence X, its action sequence is optimally labeled as maximum likelihood decoding of the *conditional* probability

$$Y^* = \arg\max_Y \log p(Y|X, W) = \arg\max_Y F(X, Y; W), \tag{3}$$

where $F(X, Y; W) \triangleq \langle W, \Phi(X, Y)\rangle$ is the discriminant function. Therefore the optimal sequence label Y^* amounts to the one attaining the maximum vale of the discriminant function.

Learning in our discriminative SMM approach is accomplished, similar to that of [32, 34], by solving a regularized optimization problem with respect to the parameter W: We would like W to be bounded to avoid over-fitting, meanwhile maximize the minimum log ratio of the *conditional* probabilities

$$\min_W \frac{\|W\|^2}{2} \quad s.t. \quad \log \frac{p(Y_t|X_t, W)}{p(Y|X_t, W)} \geq \Delta(Y_t, Y) \ \ \forall t, Y \tag{4}$$

for the set of video sequences $\{t : t \in 1, \cdots, T\}$. Here the margin is $\Delta(Y_t, Y)$, the label loss between the two feasible label assignments Y_t and Y. Now, we invoke (1), and add the non-negative slack variables ξ to account for the non-separable case. As the normalization terms cancel out, the optimization problem becomes

$$\min_{W,\xi} \quad \frac{\|W\|^2}{2} + \frac{\eta}{T}\sum_t \xi_t \tag{5}$$
$$s.t. \quad \langle W, \triangle\Phi(X_t, Y)\rangle \geq \Delta(Y_t, Y) - \xi_t \quad \forall t, Y,$$

where $\triangle\Phi(X_t, Y) := \Phi(X_t, Y_t) - \Phi(X_t, Y)$. This optimization problem is highly intuitive: The margin $\Delta(Y_t, Y)$ reflects the magnitude of mis-predicted assignment Y w.r.t. the truth Y_t. We would like to safeguard ourselves mostly against those mis-predictions Y which incur a large label loss. The non-negative ξ_t in the constraints relaxes the hard inequality by allowing few violations while penalizing these violations in the objective function by the extra cost $\frac{1}{T}\sum_t \xi_t$.

For the sake of completeness, we also present here the corresponding dual program

$$\max_\alpha \quad \sum_{t,Y} \alpha_{t,Y}\Delta(Y_t, Y) - \frac{\eta}{2}\left\|\sum_{t,Y} \alpha_{t,Y}\triangle\Phi(X_t, Y)\right\|^2 \tag{6}$$
$$s.t. \quad \alpha_{t,Y} \in \mathcal{M} \quad \forall t,$$

where $\mathcal{M}$ denotes the probability simplex constraints. Applying the Representer theorem [11] directly yields a dual representation of the discriminant function,

$$F(X, Y; W) = \sum_{t,Y'} \alpha_{tY'} \langle \triangle\Phi(X_t, Y'), \Phi(X, Y)\rangle .$$

Following those of W and Φ, F can also be decomposed into three components $f_i(X, Y) = \langle w_i, \phi_i(X, Y)\rangle, \forall i = \{1, 2, 3\}$ as

$$\sum_{i=0}^{l-1}\Big(f_1(X, n_i, c_i) + f_2(X, n_i, n_{i+1}, c_i) + f_3(X, n_i, n_{i+1}, c_i, c_{i+1})\Big).$$

An important aspect of the proposed discriminative SMM model is its generality, where the other two categories of models can be recovered as its special cases: Let $M \geq 1$ upper-bound the maximum number of frames a segment would last. By fixing $M = 1$ (which implies $\phi_1 = \phi_2$) and using only features ϕ_1 and ϕ_2 (*i.e.,* setting $\phi_3 = 0$), we recover the multi-class SVM model as displayed in Figure 1 (a). By fixing $M = 1$ and utilizing all three features, we obtain the discriminative HMM model (includes e.g. SVM-HMM [34]) as illustrated in Figure 1 (b).

3 Efficient Algorithms For Learning and Inference

One standing issue is that both the primal (5) and the dual problem (6) are practically intractable: Since the configuration space of $\mathcal{Y}$ is in the order of $T \times C^m$, the number of constraints grows exponentially as the length of training sequences increases. Even for videos of moderate length, its optimization problem would come with an astronomical amount of constraints. Nevertheless, as we show next, this problem can be solved approximately up to precision ϵ by optimization techniques such as the cutting plane [34] or the bundle method [33] in a "plug and play" manner.

3.1 Learning: the Cutting Plane vs. the Bundle Method

The main procedure of the cutting plane method is to find the most violated constraint using current solution of (5), then iteratively add these constraints to the optimization problem. This is guaranteed to converge to the the optimal solution [34], while it approximates the optimal solution to precision ϵ in a polynomial number

Algorithm 1 Cutting Plane Method

Input: sequence X_t and true label Y_t for example t, sample size T, precision $\epsilon > 0$
Initialize the constraint set $R_t = \emptyset$ for every t.
repeat
  for $t = 1$ **to** T **do**
    $Y^* = \mathrm{argmax}_Y\ \Delta(Y_t, Y) + F(X_t, Y; W)$
    $\xi_t = \max\{0, \max_{Y \in R_t} \Delta(Y_t, Y) + F(X_t, Y; W) - F(X_t, Y_t; W)\}$
    if $\Delta(Y_t, Y^*) + F(X_t, Y^*; W) - F(X_t, Y_t; W) > \xi_t + \epsilon$ **then**
      Add this constraint into $R_t \leftarrow R_t \cup \{Y^*\}$
      Optimize (6) using only α_{tY} where $Y \in R_t$.
    end if
  end for
until $R = \{R_1, \cdots, R_T\}$ has not changed in this iteration

of iterations. By adapting to our context, the cutting plane method is presented in Algorithm 1.

The bundle methods can be viewed as a quadratic counterpart of the cutting plane algorithm using line search. Both of they attempt to decrease the true objective function at every iteration. While the cutting plane algorithm relies on the monotonicity of the approximating function to guarantee convergence to an optimal solution, the bundle method directly attempts to decrease the true objective function. Recently, a bundle method solver BMRM is proposed in [31, 33] for solving general non-smooth convex optimization problems. Similar to the cutting plane method, we need to compute Y^* which can be efficiently obtained by the inference procedure. In addition, it requires as input two other quantities: the empirical loss

$$R_{\mathrm{emp}}(W) := \frac{1}{T} \sum_t \Delta(Y_t, Y_t^*) - \langle W, \triangle\Phi(X_t, Y_t^*) \rangle \,, \quad (7)$$

as well as its gradient with respect to W that yields

$$-\frac{1}{T} \sum_t \triangle\Phi(X_t, Y_t^*). \quad (8)$$

Algorithm 2 Bundle Method

Input: sequence X_t and true label Y_t for example t, sample size T, precision $\epsilon > 0$
Initialize $W = 0$
repeat
  Obtain current W from BMRM
  for $t = 1$ **to** T **do**
    $Y_t^* = \mathrm{argmax}_Y\ \Delta(Y_t, Y) + F(X_t, Y; W)$
    Compute the empirical loss $\Delta(Y_t, Y_t^*) - \langle W, \triangle\Phi(X_t, Y_t^*) \rangle$
    Compute the gradient $-\triangle\Phi(X_t, Y_t^*)$
  end for
  Report (7) and (8) to BMRM
until $R_{\mathrm{emp}}(W) \le \epsilon$

Empirical studies in Section 6 show that the bundle method often delivers superior results to those of the cutting plane method. This observation aligns with those that have been reported in [31, 33].

3.2 Viterbi-Like Inference

For both learning algorithms, we need to solve in our context an assignment problem

$$Y^* = \underset{Y \in \mathcal{Y}}{\mathrm{argmax}}\ \Delta(Y_t, Y) + F(X_t, Y; W). \quad (9)$$

It is easy to see that the result Y^* corresponds to the *most violated* constraint in (5) as long as $Y^* \neq Y_t$. For this purpose, we devise a Viterbi-like dynamic programming procedure, which is presented in Algorithm 3. Besides, we use the Hamming distance to measure the label loss $\Delta(Y, Y')$ between alternative action sequence labels as

$$\sum_{k=0}^{m-1} (1 - \delta(y_k = y'_k)),$$

where $\delta(x)$ is the indicator function.

Algorithm 3 Viterbi-Like Inference

Input: sequence X_t of length m, its true label Y_t, and maximum length of a segment M
Output: score s, optimal label Y^*
Initialize matrices $S \in \mathbb{R}^m \times C$, $J \in \mathbb{Z}^m$, and $L \in \mathbb{Z}^m$ to 0, $Y^* = \emptyset$
for $i = 1$ **to** m **do**
  for $c_i = 1$ **to** C **do**
    $(J_i, L_i) = \underset{j, c_j}{\mathrm{argmax}}\, S(j, c_j) + g(j, i, c_j, c_i)$
    $S(i, c_i) = S(j^*, c^*_{j^*}) + g(j^*, i, c^*_{j^*}, c_i)$
  end for
end for
$c^*_m = \underset{c_m}{\mathrm{argmax}}\, S(m, c_m)$
$s = S(m, c^*_m)$
$Y^* \leftarrow \{(m, c^*_m)\}$
$i \leftarrow m$
repeat
  $Y^* \leftarrow \{(J_i, L_i), Y^*\}$
  $i \leftarrow J_i$
until $i = 0$

For any segment i, we denote its related boundaries as $n_- \triangleq n_{i-1}$ and $n \triangleq n_i$. Similarly the related labels are $c_- \triangleq c_{i-1}$ and $c \triangleq c_i$. Now, we maintain a partial score $S(X, n, c)$ that sums up to segment i (*i.e.,* starts at position 0 and ends with the segment $[n_-, n)$ with labels c_- (for n_-) and c (for n), respectively), and it is defined as

$$\max_{c_-, \max\{0, n-M\} \leq n_- < n} \{ S(X, n_-, c_-) + g(X, n_-, n, c_-, c) \}. \tag{10}$$

Here the increment $g(X, n_-, n, c_-, c)$ equals to

$$f_1(X, n_-, c_-) + f_2(X, n_-, n, c_-) + f_3(X, n_-, n, c_-, c) + 1 - \sum_{k=n_-}^{n-1} \delta(y_k = c_-).$$

It is easy to verify that in the end, the sum of two terms in the RHS of (9) amounts to $S(m, c_m)$. This algorithm, after minor modification, is also used to solve the Maximum Likelihood assignment problem of (3) in the prediction phase.

This inference algorithm is very efficient: Its time complexity is $O(mMC^2)$, linear with respect to the sequence length m; Its memory complexity is $O(m(C+2))$. Our C++ implementation[1] processes the video sequences at 20 frames per second (FPS), an average speed obtained throughout our experiments. In terms of hardware, the desktop we use comes with an Intel Pentium 4 3.0GHz processor and 512M memory.

4 Feature Representation

Neuro-psychological findings such as [24] suggest that the visual and motor cortices of human perception system are more responsible than the semantic ones for retrieval and recognition of visual action patterns. This motivates us to represent action features Φ of (2) by a set of local features that capture the salient aspect of spatial and temporal video gradients.

The foreground object in each image is obtained using an efficient background subtraction method [3]. By applying the SIFT [15] key points detector, the object is represented as a set of key feature points extracted from the foreground with each point having a 128-dimensional feature vector. Importantly, SIFT features bear these properties that are critical in our context, as being relatively invariant to illumination and view-angle changes, as well as being insensitive to the objects' color appearance by capturing local image textures in the gradient domain. In addition, from each feature point, we construct an additional 60-dimensional shape context [1] features that roughly encode how each point "sees" the remaining points. The two sets of features are then concatenated with proper scaling to form a 188-dimensional vector. This point-set object representation are further transformed into a 50-dimensional codebook using K-means, similar to the visual vocabulary approach of [29]. Therefore, once a new frame is presented, each of the key points is projected into this codebook space with a cluster assignment. Thus the object is now represented as a 50-dimensional histogram vector h. Typical results of this codebook representation are illustrated in Figure 5 bottom row, where we randomly choose four codebook clusters and impose the assigned feature point locations on individual images. This convincingly shows that each cluster is able to pick up reasonably similar patches over time and across people.

Equipped with this codebook representation, we construct feature functions ϕ_1, ϕ_2 and ϕ_3 as follows.

Boundary Frame Features $\phi_1(X, n_i, c_i) = \psi_1(X, n_i) \otimes c_i$, where $\otimes$ denotes a tensor product. ψ_1 is a concatenation of two features. The first is a constant 1 which acts as the bias term. The second part is obtained from a sliding window of size w_s centered on the boundary frame. When $w_s = 1$ it becomes the single histogram vector h_{n_i}.

Node Features on Segment Node features are devised to capture the characteristics of the segment. ϕ_2 is defined as $\phi_2(X, n_i, n_{i+1}, c_i) = \psi_2(X, n_i, n_{i+1}) \otimes c_i$. $\psi_2(X, n_i, n_{i+1})$ contains three components: the length of this segment, the mean and the variance of the histogram vector of the segment (*i.e.*, over frames from n_i to $n_{i+1} - 1$).

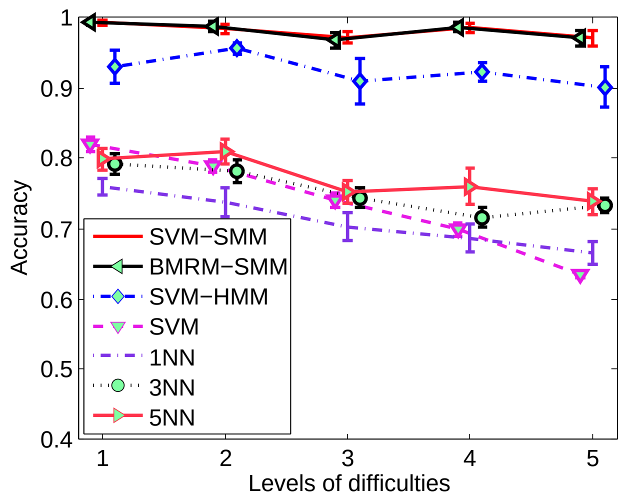


Fig. 2: Comparing seven methods for action recognition on the synthetic dataset. See text for details.

[1] Source code can be downloaded from http://users.rsise.anu.edu.au/~qshi/code/smm_release.tgz.

Edge Features on Neighboring Segments As in practice we do have prior knowledge about how long a segment would at least last, we define the minimum duration of a segment as d. Similarly $\phi_3(X, n_i, n_{i+1}, c_i, c_{i+1}) = \psi_3(X, n_i, n_{i+1}) \otimes c_i \otimes c_{i+1}$, and it is a concatenation of the following components: (1) the mean of the histogram vector from frames n_i to $n_{i+1} - 1$, and (2) from frames n_{i+1} to $n_{i+1} + d$, as well as (3) the variance of the histogram vector from n_i to $n_{i+1} - 1$, and (4) from n_{i+1} to $n_{i+1} + d$.

Before carrying on to conduct simulations, we would like to pause for a moment, and investigate from theoretical viewpoint to understand how the proposed approach would generalize on unseen test action sequences.

5 Generalization Error

Our generalization analysis of the proposed approach is based on the PAC-Bayes theory introduced by McAllester and co-workers [12, 17–19]. Germain et al. [7] recently show a simplified PAC-Bayes generalization proof technique for linear classifiers in a more general setting.

We start by adopting the PAC setting where an instance-label pair (X, Y) is drawn from a fixed but unknown distribution D over the input-output space. For any discriminant function $F(X, Y; W)$, let $Y^* = \max_{Y' \neq Y} F(X, Y'; W)$, and define its difference

$$M(X, Y; W) := F(X, Y; W) - F(X, Y^*; W). \tag{11}$$

Assume for now Y is the true label of X, then we would enforce the margin constraint $M(X, Y; W) \geq \gamma$, where the margin is $\gamma \geq 0$ to ensure the separability of an input-output pair by applying the discriminant functions. A soft constraint $M(X, Y; W) \geq \gamma - \xi$ is then adopted to allow the existence of outliers. Here $\xi \geq 0$. This can be further extended when W is sampled from a posterior Q over W [7],

$$M(X, Y; Q) = \max_{Y' \neq Y} \mathbb{E}_Q \left[F(X, Y; W) - F(X, Y'; W) \right]. \tag{12}$$

In addition, we define the true risk

$$R(D) = P_{(X,Y) \sim D} \left(\operatorname*{argmax}_{Y' \in \mathcal{Y}} \{F(X, Y'; W)\} \neq Y \right),$$

and the γ-empirical risk over the training set S w.r.t. Q as

$$R_Q(S, \gamma) = P_{(X,Y) \sim D} \left(M(X, Y; Q) \leq \gamma \right).$$

With the above setup, the generalization ability of our proposed approach can be upper-bounded by the following theorem:

Theorem 1 (PAC-Bayes Risk Bound) *Let $\delta \in (0, 1]$, assume $F(X, Y; W) \in \mathcal{F}$ is bounded, and the parameter $W \in \mathcal{W}$ with $\mathcal{W}$ being a measurable parameter space. Then, with probability at least $1 - \delta$, for a sample S with m instance-label pairs drawn from data distribution D, for any prior P and posterior Q over W, and for any margin $\gamma > 0$, we have*

$$R(D) \leq R_Q(S, \gamma) + O\left(\sqrt{\frac{\frac{1}{2}(\gamma)^{-2}(||W||^2) \ln(m|\mathcal{Y}|) + \ln m + \ln \frac{1}{\delta} + 2}{m}} \right).$$

We omitted the detailed proof as it essentially follows Theorem 5 of [13], as well as Lemma 4.2 of [14], to deal with structured output. Notice that this generalization error depends not on the dimensionality of the feature space, rather it depends on the size of the observation sample S and the margin γ: As we increase the sample size m and margin γ, the risk bound becomes tighter. In particular, with high probability, the empirical risk deviates from the true risk with an additive term that diminishes quickly as m goes to infinity.

6 Experiments

During the follow-up experiments, the proposed discriminative SMM approach is compared against three other algorithms: KNN (where K=1, 3, 5), SVM multiclass and SVM-HMM [34]. In particular, two variants of discriminative SMM are considered, namely the one with cutting plane method (SVM-SMM) and the one with bundle method (BMRM-SMM).

By default, we fix $\epsilon = 1e - 4$, $M = 3$, and $w_s = 3$. The trade-off parameter η of each method (SVM multiclass, SVM-HMM, SVM-SMM and BMRM-SMM) is tuned separately using cross-validation. Moreover, we evaluate the action recognition and segmentation performance separately: A frame-wise recognition rate is utilized to benchmark the recognition performance for each of the comparison algorithms. To measure segmentation performance, we adopted the F_1-score, which is often used in information retrieval tasks, and is given by $(2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$.

6.1 Synthetic dataset

We start with a controlled setting where we are able to quantitatively measure the performance of comparison algorithms by varying the difficulty level of problems from easy to difficult. We do this by constructing a two-person two-action synthetic dataset consisting of five trials, where each trial has a set of ten sequences and

Fig. 3: Sample frames of one person engaging in six types of actions in the KTH dataset.

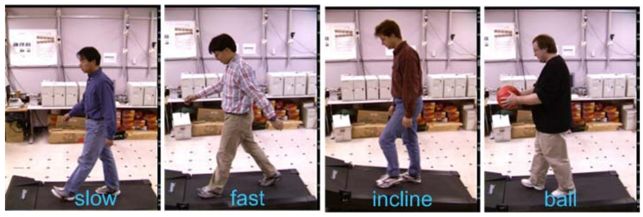


Fig. 4: Sample frames of subjects each performs one of the four actions: slow walk, fast walk, incline walk and walk with a ball, in an action sequence of the CMU MoBo dataset.

corresponds to a certain level of difficulty [2]. Here each person P equals to one semi-Markov model containing its own Gaussian emission probabilities $\mathcal{N}(\mu_{c,P}, \sigma_{c,P})$ and duration parameters $\lambda_{c,P}$ for the two actions $c = 1, 2$, respectively. Each sequence of length 150 frames is generated by sampling from a SMM model, and as a result contains continuous actions. Now, we build five trials as follows: For each trial, five sequences are generated from each person's model, and in the end we have ten sequences. Across trials, we vary the level of difficulty by moving μ_2 toward μ_1 and fixing other parameters of the models.

Figure 2 displays the action recognition results on this dataset, where 5-fold cross-validation are utilized. Here both discriminative SMM variants consistently outperform others: In fact, both SVM-SMM and BMRM-SMM gives almost the same recognition accuracy regardless the levels of difficulty. They are followed by SVM-HMM while the rest methods (namely SVM and KNNs) have inferior performance. This clearly shows that as we further exploiting the contextual information from neighboring nodes up to neighboring segments, the gains in term of recognition rate become more significant.

6.2 KTH dataset

The KTH dataset [27] contains 25 individuals performing six activities: *running*, *walking*, *jogging*, *boxing*, *handclapping* and *handwaving*, where each sequence contains single action with multiple action cycles. Figure 3 displays exemplar frames of one person taking each of the six activities.

To make direct comparisons to existing methods in literature presented in Table 1, in this experiment we adopt a "cuboid" [4] feature (instead of SIFT) that captures the local spatio-temporal characteristics using Gabor filters. More specifically, this detector is tuned to fire whenever variations in local image intensities contain distinguishing spatio-temporal characteristics. At each detected interest point location, a 3D cuboid is then extracted and represented as a flattened vector that contains the spatio-temporal windowed information including normalized pixel values, brightness gradient and windowed optical flow.

Table 1 shows the results of our methods averaged over 5-fold cross-validation. Our SVM baseline is comparable to similar methods (e.g. SVM of [4, 22, 37]) reported in literature, while our BMRM-SMM performs favorably comparing to these state-of-the-art methods. We attribute it to the contextual information that we are able to exploit through the usage of ϕ_2 features in our SMM framework. Tables 2 displays the confusion matrix of the BMRM-SMM method, where *handwaving* action can be perfectly identified from the rest actions. On the other hand, there are a few mistakes among the three easy-to-be-confused categories: *walking*, *jogging*, and *running*.

6.3 CMU MoBo dataset

This dataset [26] contains 24 individuals[3] walking on a treadmill. As illustrated in Figure 4, each subject performs in a video clip one of the four different actions: *slow* walk, *fast* walk, *incline* walk and slow walk *with a ball*. Each sequence has been pre-processed to contain several cycles of a *single* action and we additionally manually label the boundary positions of these cycles. The task on this dataset is to automatically partition a sequence into atomic action cycles, as well as predict the action label of this sequence.

Table 3 presents the results averaged over 6-fold cross-validation. The results of 3NN and 5NN are omit-

[2] This dataset can be downloaded at http://users.rsise.anu.edu.au/~qshi/code/smm_release.tgz.

[3] The dataset originally consists of 25 subjects. We drop the last person since we experienced technical problems obtaining the sequences of this individual walking with balls.

Method		Brief Description	Accuracy
Ke et al. [10]	ICCV'05	cascade classifier, spatio-temporal volumetric features, feature selection	0.63
Schuldt et al. [27]	ICPR'04	SVM, local space time features	0.72
Dollar et al. [4]	VSPETS'05	SVM, "cuboid" features	0.88 ± 0.02
Wong et al. [37]	CVPR'07	pLSA-ISM, "cuboid" features	0.84
Jhuang et al. [8]	ICCV'07	SVM, GrC2 sparse features, feature selection	0.92 ± 0.03
Nowozin et al. [22]	ICCV'07	baseline SVM linear kernel, "cuboid" features	0.83
		subsequence boosting, "cuboid" features	0.85
Our SVM		baseline SVM, "cuboid" features	0.85 ± 0.04
Our BMRM-SVM		discriminative SMM, "cuboid" features	0.95 ± 0.01

Table 1: Comparisons of action recognition rates on KTH dataset.

truth vs. predict	boxing	handclapping	handwaving	jogging	running	walking
boxing	0.91	0.09	0.00	0.00	0.00	0.00
handclapping	0.00	0.96	0.00	0.00	0.04	0.00
handwaving	0.00	0.00	1.00	0.00	0.00	0.00
jogging	0.00	0.00	0.00	0.89	0.00	0.11
running	0.00	0.00	0.00	0.08	0.92	0.00
walking	0.00	0.00	0.00	0.12	0.00	0.88

Table 2: Confusion matrix of BMRM-SMM on the KTH dataset for action recognition.

	1NN	SVM	HMM	SVM-HMM	SVM-SMM	BMRM-SMM
Act.	0.65 ± 0.02	0.67 ± 0.03	0.68 ± 0.08	0.75 ± 0.06	0.75 ± 0.03	0.78 ± 0.07
Seg.	0.16 ± 0.05	0.15 ± 0.03	n/a $\pm$ n/a	0.43 ± 0.01	0.59 ± 0.03	0.59 ± 0.03

Table 3: Comparison on CMU MoBo dataset. The first row presents action recognition rate,while the second row gives F_1-score for segmentation measurement. See text for details.

ted here as they are very similar to 1NN. We also experiment with generative HMM on the task of solely action recognition (predicting action label for each sequence), where one HMM is trained for each action type using BaumWelch algorithm. It performs slightly better than the baseline methods including KNN (K=1,3,5) and SVM, but is still inferior to SVM-HMM [34], its discriminative counterpart. Note that both SMM variants (SVM-SMM and BMRM-SMM) significantly outperforms the other methods including SVM-HMM on action label prediction as well as on segmentation of action cycles.

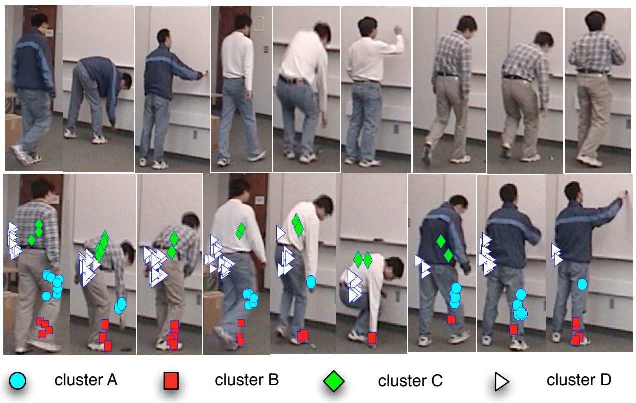


Fig. 5: A Walk-Bend-Draw (WBD) dataset. **Top** shows some sample frames of the dataset . **Bottom** displays the assignments of image feature points on four randomly chosen codebook clusters over time and across person.

6.4 WBD: A Dataset of Continuous Actions

In addition to the existing datasets (such as the MoBo and the KTH datasets), where each sequence contains exactly one type of action, we construct a Walk-Bend-Draw (WBD) dataset of continuous actions. Some exemplar frames are displayed in Figure 5. This is an indoor video dataset contains three subjects, each performs six action sequences at 30 FPS at a resolution of 720×480, and each sequence consists of three continuous actions: slow *walk*, *bend* body and *draw* on board, and on average each action lasts about 2.5 seconds. We subsample each sequence to obtain 30 key frames, and manually label the ground truth actions.

The comparison results, obtained using 6-fold cross-validation, are summarized in table 4. Both discriminative SMM variants consistently deliver the best results, while here BMRM-SMM slightly outperforms SVM-SMM. Similar to the results presented for the synthetic dataset,

	1NN	3NN	5NN	SVM	SVM-HMM	SVM-SMM	BMRM-SMM
Action Recognition	0.82 ± 0.02	0.80 ± 0.03	0.77 ± 0.03	0.84 ± 0.03	0.87 ± 0.02	0.91± 0.02	0.94± 0.01

Table 4: A summary of the action recognition methods performed on the WBD dataset.

truth vs. predict	walk	bend	draw
walk	0.93	0.07	0.00
bend	0.05	0.93	0.01
draw	0.02	0.09	0.89

Table 5: Confusion matrix of SVM-SMM applied on the WBD dataset for action recognition.

truth vs. predict	walk	bend	draw
walk	0.91	0.09	0.00
bend	0.02	0.93	0.04
draw	0.00	0.04	0.96

Table 6: Confusion matrix of BMRM-SMM applied on the WBD dataset for action recognition.

they are then followed by SVM-HMM, SVM, and the KNN methods. Furthermore, Tables 5 and 6 display the confusion matrices of the two SMM variants: SVM-SMM vs. BMRM-SMM. where the two actions – *walk* and *draw* – seem to be rarely confused with each other, nevertheless both sometimes are mis-interpreted as bend. This is to be expected, as although *walk* and *draw* appear to be more similar to human observer in isolated images, it nevertheless can be learned by discriminative SMM methods that *walk*, *bend* and *draw* are usually conducted in order.

7 Outlook and Future Work

We present a novel discriminative semi-Markov approach to human action analysis, where we intend to simultaneously segment and recognize continuous action sequences. We then devise a Viterbi-like dynamic programming algorithm that is able to efficiently solve the inference problem, and show the induced learning problem can be casted as a convex optimization problem with many constraints, that can be subsequently solved and we present two such solvers. We also analyze the generalization error of the proposed approach and provide a PAC-Bayes bound. Empirical simulations demonstrate that our approach is competitive to and often outperforms the state-of-the-art methods.

Our approach can be extended in several directions. It is promising to explore the dual representation in order to incorporate matching cost between point sets. On future work we also plan to apply this approach to closely related problems, such as detecting unusual actions from a large video dataset.

Acknowledgements We thank Baochun Bai and Cheng Lei for their help in creating the WBD dataset, Piotr Dollar for generously providing the "cuboid" feature implementation, the authors of [26, 27] for sharing the CMU MoBo and KTH datasets, respectively.

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Analysis and Machine Intell. **24**(4), 509–522 (2002)
2. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, p. 994. IEEE Computer Society, Washington, DC, USA (1997)
3. Cheng, L., Wang, S., Schuurmans, D., Caelli, T., Vishwanathan, S.: An online discriminative approach to background subtraction. In: IEEE international conference on advanced video and signal based surveillance (AVSS) (2006)
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS workshop (2005)
5. Ferguson, J.: Variable duration models for speech. In: Symposium on the Application of Hidden Markov Models to Text and Speech, pp. 143–179 (1980)
6. Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding **73**(1), 82–98 (1999)
7. Germain, P., Lacasse, A., Laviolette, F., Marchand, M.: PAC-Bayesian learning of linear classifiers. In: ICML, pp. 353–360. ACM (2009)
8. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV (2007)
9. Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N., RoyChowdhury, A., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Trans. on Image Processing pp. 1163–1173 (2004)
10. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: ICCV, vol. 1, pp. 166 – 173 (2005)

11. Kimeldorf, G., Wahba, G.: Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33**, 82–95 (1971)
12. Langford, J., McAllester, D.: Computable shell decomposition bounds. J. Mach. Learn. Res. **5**, 529–547 (2004)
13. Langford, J., Seeger, M., Megiddo, N.: An improved predictive accuracy bound for averaging classifiers. In: ICML, pp. 290–297 (2001)
14. Langford, J., Shawe-taylor, J.: PAC-Bayes and Margins. In: NIPS, pp. 439–446. MIT Press (2002)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
16. Lv, F., Nevatia, R.: Recognition and segmentation of 3-d human action using HMM and multi-class adaboost. In: European Conference on Computer Vision, pp. IV: 359–372 (2006)
17. McAllester, D.: Some PAC-Bayesian theorems. In: COLT, pp. 230–234. ACM (1998)
18. McAllester, D.: PAC-Bayesian stochastic model selection. Mach. Learn. **51**(1), 5–21 (2003)
19. McAllester, D.: Simplified PAC-Bayesian margin bounds. In: COLT, pp. 203–215. ACM (2003)
20. Moeslund, T., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2), 90–126 (2006)
21. Niebles, J., Fei, L.F.: A hierarchical model of shape and appearance for human action classification. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2007)
22. Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: ICCV (2007)
23. Ostendorf, M., Digalakis, V., Kimball, O.: From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. on Speech and Audio Processingn **4**(5), 360–378 (1996)
24. Phillips, J., Humphreys, G., Noppeney, U., Price, C.: The neural substrates of action retrieval: An examination of semantic and visual routes to action. Visual Cognition **9**(4-5), 662–685 (2002)
25. Ratsch, G., Sonnenburg, S.: Large scale hidden semi-markov svms. In: B. Schölkopf, J. Platt, T. Hoffman (eds.) NIPS, pp. 1161–1168. MIT Press (2006)
26. R.Gross, Shi, J.: The CMU motion of body (MoBo) database. Tech. Rep. Tech. Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University (2001)
27. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proc. Intl. Conf. Pattern Recognition, pp. 32–36. IEEE Computer Society, Washington, DC, USA (2004)
28. Shi, Q., Wang, L., Cheng, L., Smola, A.: Discriminative human action segmentation and recognition using semi-markov model. In: CVPR (2008)
29. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1470–1477 (2003)
30. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: IEEE International Conference on Computer Vision, pp. 1808–1815 (2005)
31. Smola, A., Vishwanathan, S., Le, Q.: Bundle methods for machine learning. In: NIPS (2007)
32. Taskar, B., Guestrin, C., Koller, D.: Max-margin Markov networks. In: S. Thrun, L. Saul, B. Schölkopf (eds.) NIPS, pp. 25–32. MIT Press, Cambridge, MA (2004)
33. Teo, C., Le, Q., Smola, A., Vishwanathan, S.: A scalable modular convex solver for regularized risk minimization. In: KDD (2007)
34. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6**, 1453–1484 (2005)
35. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
36. Wang, L., Suter, D.: Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1–8 (2007)
37. Wong, S., Kim, T., Cipolla, R.: Learning motion categories using both semantic and structural information. In: CVPR, pp. 1–6 (2007)
38. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden markov model. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 379–385 (1992)