Int J Comput Vis
DOI 10.1007/s11263-016-0897-2

A Hierarchical Video Description for Complex Activity Understanding

Cuiwei Liu[1] · **Xinxiao Wu**[1] · **Yunde Jia**[1]

Received: 18 October 2014 / Accepted: 19 February 2016

Abstract This paper addresses the challenging problem of complex human activity understanding from long videos. Towards this goal, we propose a hierarchical description of an activity video, referring to the "which" of activities, "what" of atomic actions, and "when" of atomic actions happening in the video. In our work, each complex activity is characterized as a composition of simple motion units (called *atomic actions*), and different atomic actions are explained by different video segments. We develop a latent discriminative structural model to detect the complex activity and atomic actions, while learning the temporal structure of atomic actions simultaneously. A segment-annotation mapping matrix is introduced for relating video segments to their associational atomic actions, allowing different video segments to explain different atomic actions. The segment-annotation mapping matrix is treated as latent information in the model, since its ground-truth is not available during both training and testing. Moreover, we present a semi-supervised learning method to automatically predict the atomic action labels of unlabeled training videos when the labeled training data is limited, which could greatly alleviate the laborious and time-consuming annotations of atomic actions for training data. Experiments on three activity datasets demonstrate that our method is able to achieve promising activity recognition results and obtain rich and hierarchical descriptions of activity videos.

Keywords Activity understanding · Hierarchical video description · Atomic action · Latent structural model

Communicated by Junsong Yuan, Wanqing Li, Zhengyou Zhang, David Fleet, and Jamie Shotton.

✉ Xinxiao Wu
wuxinxiao@bit.edu.cn

Cuiwei Liu
liucuiwei@bit.edu.cn

Yunde Jia
jiayunde@bit.edu.cn

[1] Beijing Lab. of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, People's Republic of China

1 Introduction

Understanding human activities in videos has been extensively studied for wide applications such as intelligent surveillance, human computer interaction, content-based video annotation, and video retrieval. Many previous methods (Laptev 2005; Li et al. 2008; Liu et al. 2011; Sadanand and Corso 2012; Yu et al. 2012) focus on the recognition of pre-segmented short videos containing simple and well-defined actions such as running, boxing, and walking. In practice, real-world videos are often of longer length, and contain multiple motions happening at different specific moments and places. Some recent literatures (Niebles et al. 2010; Gaidon et al. 2011; Tang et al. 2012; Izadinia and Shah 2012; Wang et al. 2013b) go beyond the single-label action recognition and deal with understanding more complex activities from long videos.

We focus on understanding and describing complex activities in long videos. The activities are composed of a set of key motion units with simple semantic information, that are referred to as *atomic actions*. For example, the atomic actions of the "triple-jump" activity include: "run-up", "hop and bound", and "jump into sand pit". Due to the fact that different activities have different sequences of different atomic actions, it is beneficial to explore the relationship between activities and atomic actions as well as the temporal

Published online: 22 March 2016

structure of atomic actions to automatically analyze complex activities.

In this paper, we propose to interpret an activity video into a hierarchical description referring to the "which" of activities, "what" of atomic actions, and "when" of atomic actions happening in the video. A novel latent discriminative structural model is developed to capture the intrinsic relationship among high-level activity class, mid-level atomic actions, and low-level video segments. The relationship between video segments and atomic action annotations is represented by a segment-annotation mapping matrix which assigns video segments to the atomic actions. The segment-annotation mapping matrix is treated as latent information in our model, since its ground-truth is not available during both training and testing. The training data to learn the discriminative model is a set of videos with their associated activity labels and atomic action annotations. In the training stage, the model jointly learns the detectors of atomic actions, the overall templates of activities, and the correlation matrix between activities and annotations. For the purpose of learning accurate temporal structure of atomic actions, we enforce the consistency between the segment-annotation mapping matrix and the prior temporal distribution of atomic actions. In the testing stage, the model simultaneously predicts the activity label and the atomic action annotations for a video, and also localizes the predicted atomic actions in time direction via the inferred segment-annotation mapping matrix. Furthermore, in order to alleviate the laborious and time-consuming annotations of atomic actions for training videos, we introduce a semi-supervised method which allows part of the training videos only annotated with activity labels and automatically annotates these training videos with atomic action labels. The overview of our method is illustrated in Fig. 1.

The main contributions of this work include: (1) a hierarchical description of complex activities based on "which" activities, "what" atomic actions, and "when" of atomic actions happening in the video, (2) a novel latent discriminative structural model which jointly performs the detection of atomic actions, the temporal localization of atomic actions, and the classification of overall activity, and (3) a semi-supervised learning method to automatically annotate the training videos with atomic action labels when the training data are partly atomic-action labeled.

2 Related Work

Most existing literatures of human activity analysis focus on the problem of classifying a short video containing a simple action. A lot of approaches (Efros et al. 2003; Yilmaz and Shah 2005; Gorelick et al. 2007; Rodriguez et al. 2008; Li et al. 2008) employ global space-time representations extracted from image sequences. Recently, many standard action recognition methods (Laptev 2005; Dollar et al. 2005; Wang et al. 2013a; Le et al. 2011; Wu et al. 2011; Yu et al. 2012) adopt local spatio-temporal feature representations and achieve impressive results. All the above methods directly establish the mapping from low-level visual features to high-level label information and consequently have innate limitations in semantic description of complex activities in terms of multiple atomic actions.

Some recent literatures have investigated detecting and modeling low-level atomic actions to recognize complex activities. In (Laxton et al. 2007), the hierarchical complex activity model takes input from low-level actions detectors, and manual annotations of actions for each frame are required to learn these detectors. In (Izadinia and Shah 2012), the detectors of low-level events are learned from training video clips, and the co-occurrence relationship between different low-level events is modeled in a latent discriminative framework. Each video is divided into several short clips and each clip is manually annotated using one low-level event label for training the low-level detectors. Sun and Nevatia (2013) defined a set of activity concepts to encode the activity concept transitions over time for describing a video event. But their method requires manual annotations of activity concepts for each clip. Felzenszwalb et al. (2010) presented a representation capturing the temporal dynamics of windowed mid-level concept detectors for improving complex event recognition. And they employed human annotators to mark the approximate beginning and ending frames of the concepts in videos to pre-train the concept detectors. Lillo et al. (2014) developed a hierarchical model based on pose evaluation for recognizing complex activities from RGBD data, and their work need to manually annotate action classes for each frame. In contrast, our method does not need the manual annotation of video segments to learn the atomic action detectors. We integrate the detection of atomic actions, temporal localization of atomic actions, and classification of overall activity into a unified framework. Gaidon et al. (2011) represented the temporal structure of activities as a sequence of histograms of atomic action-anchored visual features, and detected atomic actions in an input activity video based on the prior distributions of atomic actions. Hoai et al. (2011) presented a multi-class SVM based discriminative model for simultaneously performing temporal segmentation and event recognition, and they found the optimal segmentations and atomic action labels of a long video by using dynamic programming. Different from (Gaidon et al. 2011) which annotates time stamps of atomic action clips for training and (Hoai et al. 2011) which assigns a particular class label for each frame, our method does not require the temporal annotations of atomic actions in training. And what we need is only the semantic concept annotations of atomic actions for training.

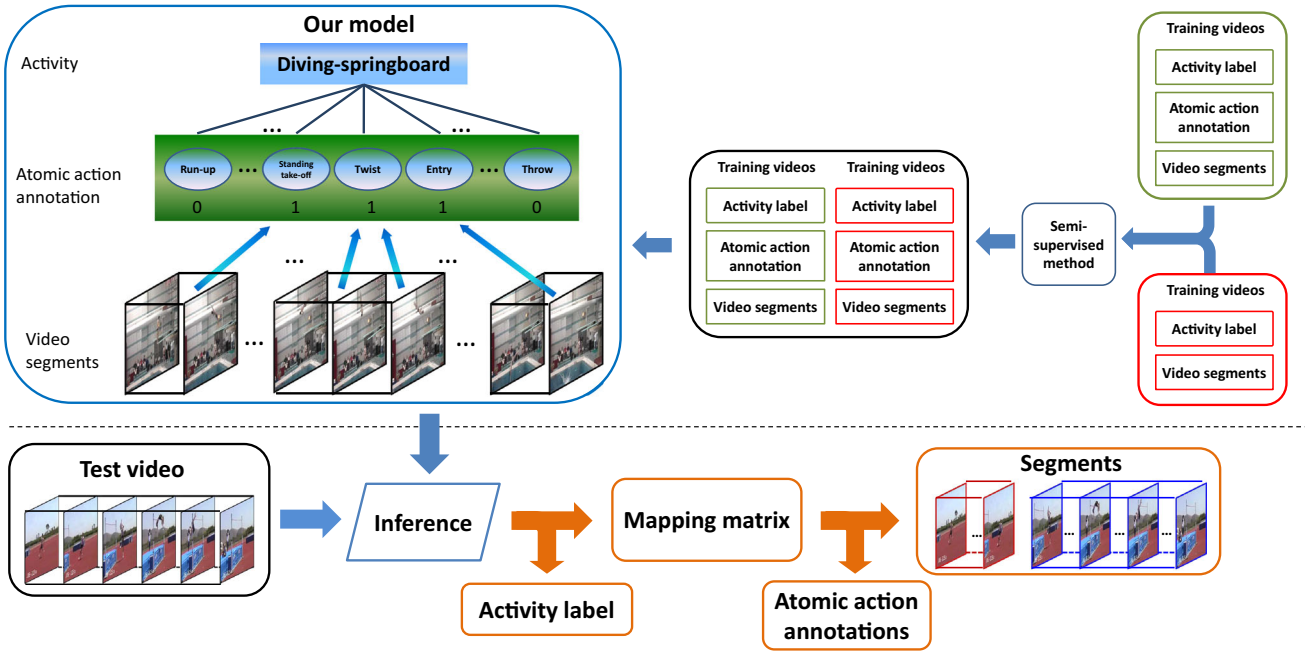

Fig. 1 Overview of the proposed method. During training, our model takes a set of videos with activity labels and atomic action annotations as input. The proposed model jointly learns the relationship among high-level activities, mid-level atomic actions, and low-level video segments. For a new video, the model is able to predict the activity label, annotate the atomic actions, and build the mapping between atomic actions and video segments. We introduce a semi-supervised method which allows part of the training videos only annotated with activity labels and automatically annotates these training videos with atomic action labels (Color figure online)

Niebles et al. (2010) utilized a set of motion segment classifiers to model a complex activity. The classification is based on the quality of matching between motion segment classifiers and temporal segments in the input video by optimizing the temporal positions of motion segment classifiers. The motion segments are not enough for semantic representation, and our method does well in conceptual description of activities by using atomic actions. Moreover, we describe the temporal structure of activities by the mapping matrix between video segments and atomic actions.

Tang et al. (2012) modeled an event by a set of latent state variables and duration variables, and introduced a max-margin based discriminative model to learn the temporal structure of complex events. In their work, the latent states are actually the cluster centers of video clips from training samples. Wang et al. (2013b) decomposed a complex action into a series of motion atoms which are discovered through clustering. A motion atom captures a simple motion in a short temporal scale, and can be considered as an atomic action. Pirsiavash and Ramanan (2014) parsed long videos of actions with segmental grammars to model the hierarchical temporal structure of sub-actions. Wang et al. (2014) developed a latent hierarchical model to decompose an complex activity into sub-activities in a hierarchical way. Hu et al. (2014) presented a hidden CRF model for predicting the underlying sub-activity labels of action videos. The above methods (Tang et al. 2012; Wang et al. 2013b; Pirsiavash and Ramanan 2014; Wang et al. 2014; Hu et al. 2014) model the hierarchical structure of complex activities by introducing the data-driven generated states, atoms, and sub-activities. Different from these work, our method can achieve a more reasonable and logical description of complex activities owing to the available concept annotations of atomic actions in training.

3 Our Model

3.1 Model Formulation

We develop a discriminative structural model with latent variables for jointly capturing the relationship among video segments, atomic action annotations, and overall activity labels. The training data for learning the model are a set of triples $\{(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n)|n = 1, 2, \ldots, N\}$, where $\mathbf{x}^n$ represents the n-th training video which is initially partitioned into several segments, $\mathbf{y}^n$ is the activity label of $\mathbf{x}^n$, and $\mathbf{h}^n = [h_1^n, h_2^n, \ldots, h_V^n]$ indicates the atomic action annotations ($h_i^n = 1$ if the i-th atomic action is present in the video $\mathbf{x}^n$, and $h_i^n = 0$ otherwise). The mapping matrix between video segments and atomic action annotations is treated as latent information in the model. With a set of unobserved latent variables $\{\mathbf{g}^n\}$, $n = 1, 2, \ldots, N$, where $\mathbf{g}^n$ is the segment-annotation mapping matrix of $\mathbf{x}^n$, our learning goal is to learn a prediction rule of the following form:

$$f_{\mathbf{w}}(\mathbf{x}) = \max_{\mathbf{y},\mathbf{h},\mathbf{g}} F(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{g}) = \max_{\mathbf{y},\mathbf{h},\mathbf{g}} \mathbf{w}^\top \Phi(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{g}), \quad (1)$$

where $\Phi(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{g})$ is a joint feature vector that describes the relationship among the activity video $\mathbf{x}$, the activity label $\mathbf{y}$, the atomic action annotations $\mathbf{h}$, and the latent segment-annotation mapping matrix $\mathbf{g}$. The optimization problem in Eq. 1 is typically referred to as the "inference" or "recognition" problem. Detailed explanation of $\mathbf{g}$ and $\mathbf{w}^\top\Phi(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{g})$ will be illustrated in Sects.3.1.1 and 3.1.2, respectively.

3.1.1 Mapping Video Segments to Atomic Action Annotations

Given an activity video $\mathbf{x}$ and its atomic action annotation $\mathbf{h}$, we first partition $\mathbf{x}$ into R segments in time series: $\mathbf{x} = [x_1, x_2, \ldots, x_R]$. Then we interpret the interaction between video segments and atomic action annotations by assuming an unobserved many-to-one mapping which relates R video segments to each of the present atomic actions. Specifically, we introduce a matrix $\mathbf{g} = \{g_{ij}, i = 1, 2, \ldots, R, j = 1, 2, \ldots, V\}$ to translate R segments to V atomic actions, where $g_{ij} = 1$ if the i-th video segment is mapped to the j-th atomic action label, and 0 otherwise.

We do not access to the ground-truth of $\mathbf{g}$, but infer it during both training and testing stages. In order to reduce the searching space of $\mathbf{g}$, we make the following constraints on the relation between $\mathbf{g}$ and $\mathbf{h}$: (1) each video segment is related to at most one annotation term to guarantee that a video segment can not be used to relate more than one annotations; (2) each present atomic action is related to one or more video segments to ensure that for each atomic action assigned to a video, there is at least one video segment explaining it ; (3) each absent atomic action is not related to any video segments to make sure that if an atomic action is not assigned to a video, there are no video segments mapped to it. These constraints are formalized by

$$\sum_j g_{ij} \leq 1, \forall i; \quad \max_i g_{ij} = h_j, \forall j; \quad g_{ij} \in \{0,1\}, \forall i, \forall j. \quad (2)$$

3.1.2 Relationship Among Video Segments, Atomic Action Annotations, and Activity Label

We formulate the relationship among video segments, atomic action annotations, and activity label as

$$\mathbf{w}^\top \Phi(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{g}) = \alpha^\top \phi(\mathbf{x},\mathbf{g}) + \beta^\top \psi(\mathbf{x},\mathbf{y}) + \gamma^\top \varphi(\mathbf{h},\mathbf{y}).$$

The model parameter $\mathbf{w}$ is given by $\mathbf{w} = \{\alpha; \beta; \gamma\}$. The details of each term are described below.

Interaction Function Between Video Segments and Atomic Action Annotations $\alpha^\top\phi(\mathbf{x},\mathbf{g})$. This interaction function captures the compatibility of mapping video segments to the atomic actions, parameterized by

$$\alpha^\top \phi(\mathbf{x},\mathbf{g}) = \sum_{i=1}^{R}\sum_{j=1}^{V} \alpha_j^\top \cdot x_i \cdot g_{ij},$$

where α_j represents a template for predicting a video segment to take the j-th atomic action label, and x_i is the visual feature vector extracted from the i-th video segment. Due to the constraint $\max_i g_{ij} = h_j$ defined in Eq. 2, the definition of this function involves $\mathbf{g}$ instead of $\mathbf{h}$.

Compatibility Function Between Video and Activity Label $\beta^\top\psi(\mathbf{x},\mathbf{y})$. This compatibility function is a standard linear function learned to predict the overall activity label $\mathbf{y}$ for video $\mathbf{x}$, without considering multiple atomic actions, parameterized by

$$\beta^\top \psi(\mathbf{x},\mathbf{y}) = \beta_y^\top \cdot x,$$

where x is the feature extracted from the entire video and β_y represents a template for activity class y.

Correlation function between atomic action annotations and activity label $\gamma^\top\varphi(\mathbf{h},\mathbf{y})$. There is a meaningful relationship between multiple atomic actions and the overall activity. For example, a certain number of atomic actions such as "run-up" and "throw" often happen in a particular activity "javelin-throw", but may not occur in other activities such as "snatch" and "tennis-serve". Therefore, it is beneficial to incorporate a correlation function between an activity label $\mathbf{y}$ and atomic action annotations $\mathbf{h}$, which is defined by

$$\gamma^\top \varphi(\mathbf{h},\mathbf{y}) = \sum_{j=1}^{V} [\gamma_{j,1}^y \cdot h_j + \gamma_{j,0}^y \cdot (1 - h_j)],$$

where $\gamma_{j,1}^y$ represents a template for the j-th atomic action to be present if the activity class is y, while $\gamma_{j,0}^y$ denotes a template for the j-th atomic action to be absent if the activity class is y.

3.2 Model Training

Given a set of N training samples $\{(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n)\}$, $n = 1, 2, \ldots, N$, our goal is to learn the model parameter $\mathbf{w}$. Since the segment-annotation mapping matrix $\mathbf{g}$ is unobserved and hence treated as a latent variable during training, we adopt the Latent Structural SVM framework (Felzenszwalb et al. 2010; Yu and Joachims 2009) by formulating the following optimization problem:

$$\min_{\mathbf{w},\xi^n} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi^n \tag{3}$$

$$\begin{aligned}\text{s.t.} \quad & \max_{\mathbf{g}^n}[\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n) \\ & -C'\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i,j,k,l)] \\ & -\max_{\mathbf{g}} \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g}) \\ & \geq \Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) - \xi^n, \forall n, \forall \mathbf{y}, \forall \mathbf{h}, \end{aligned} \tag{4}$$

where $\rho(i, j, k, l)$ is a penalty function for the co-occurrence of g_{ij}^n and g_{kl}^n, and $\Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h}))$ is a loss function measuring the cost incurred by predicting the ground-truth $(\mathbf{y}^n, \mathbf{h}^n)$ as $(\mathbf{y}, \mathbf{h})$. Here we utilize a simple Hamming loss: $\Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) = \ell(\mathbf{y}^n, \mathbf{y}) + \sum_{j=1}^{V}\ell(h_j^n, h_j)$, where $\ell(a, b)$ is 1 if $a \neq b$ and 0 otherwise.

The constraints in Eq. 4 can be explained as follows: for the n-th training sample, the score $\max_{\mathbf{g}^n}[\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n) - C'\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i, j, k, l)]$ that is associated with the ground-truth activity label $\mathbf{y}^n$ and atomic action annotations $\mathbf{h}^n$ should be no less than the score $\max_{\mathbf{g}} \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g})$ that is associated with any hypothesized activity class label $\mathbf{y}$ and atomic action annotations $\mathbf{h}$. Here, $\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i, j, k, l)$ enforces the consistency between the mapping $\mathbf{g}^n$ and the temporal distribution of atomic action annotations by using the penalty function $\rho(i, j, k, l)$. For example, in the activity "triple-jump", the atomic action "run-up" appears before the atomic action "jump into sand pit", thus the video segments interpreting "run-up" should be ahead of that interpreting "jump into sand pit". The penalty function $\rho(i, j, k, l)$ is defined as

$$\rho(i,j,k,l) = \begin{cases} \eta_{jl} \cdot sgn(i-k), & if \ \ \eta_{jl} \cdot sgn(i-k) > 0 \\ 0, & if \ \ \eta_{jl} \cdot sgn(i-k) \leq 0 \end{cases}$$

where $\eta_{jl} \in [-1, 1]$ represents the possibility of atomic action j happening before atomic action l, and $sgn(i - k)$ indicates the temporal relation of segment i and segment k. In our implementation, η_{jl} is set to be -1, 0, or 1 according to prior knowledge. Particularly, $\eta_{jl} = 1$ represents that atomic action j appears before atomic action l, $\eta_{jl} = -1$ indicates that atomic action j appears after atomic action l, and $\eta_{jl} = 0$ means that the temporal order of atomic action j and atomic action l is indefinite. If $\eta_{jl} \cdot sgn(i-k) > 0$, the temporal relation of segment i and segment k is inconsistent with the prior temporal distribution of atomic action j and atomic action l indicated in η_{jl}, therefore, the co-occurrence of g_{ij}^n and g_{kl}^n is punished with $\rho(i, j, k, l)$.

Since ξ^n is equivalent to

$$\begin{aligned} & \left\{\max_{\mathbf{y},\mathbf{h},\mathbf{g}}(\Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) + \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g})) \right. \\ & -\max_{\mathbf{g}^n}[\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n) \\ & \left. -C'\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i,j,k,l)] \right\} \\ & = \max_{\mathbf{y},\mathbf{h},\mathbf{g}}(\Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) + \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g})) \\ & + \min_{\mathbf{g}^n}\left[C'\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i,j,k,l) \right. \\ & \left. -\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n)\right], \end{aligned} \tag{5}$$

we can rewrite the constrained optimization problem in Eq. 3 as an unconstrained problem:

$$\begin{aligned} & \min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}(L^n + R^n), \\ & L^n = \max_{\mathbf{y},\mathbf{h},\mathbf{g}}(\Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) + \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g})), \\ & R^n = \min_{\mathbf{g}^n}\left[C'\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i,j,k,l) \right. \\ & \left. -\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n)\right], \end{aligned} \tag{6}$$

where $L^n + R^n = \xi^n$.

We employ the non-convex bundle optimization method (NRBM) in Do and Artieres (2009) to solve Eq. 6. Specifically, this optimization algorithm iteratively builds an increasingly accurate piecewise quadratic approximation of Eq. 6 and converges to an optimal solution of $\mathbf{w}$, which requires the calculation of the subgradient of $L^n + R^n$. Suppose $(\mathbf{y}^\star, \mathbf{h}^\star, \mathbf{g}^\star)$ be the solution to the optimization problem

$$\max_{\mathbf{y},\mathbf{h},\mathbf{g}} \Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) + \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g}) \tag{7}$$

and $\mathbf{g}^\dagger$ be the solution to the optimization problem

$$\min_{\mathbf{g}^n}\left[C'\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i,j,k,l) -\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n)\right], \tag{8}$$

then $\partial_{\mathbf{w}}(L^n + R^n)$ can be calculated by $\Phi(\mathbf{x}^n, \mathbf{y}^\star, \mathbf{h}^\star, \mathbf{g}^\star) - \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^\dagger)$. To solve the optimization problem in Eq. 7, we first enumerate all the possible values of $\mathbf{y}$ and then deal with the inner maximization over $\mathbf{h}$ and $\mathbf{g}$ for a fixed $\mathbf{y}$:

$$\begin{aligned}
&\max_{\mathbf{h},\mathbf{g}} \Delta((\mathbf{y}^n,\mathbf{h}^n),(\mathbf{y},\mathbf{h})) + \mathbf{w}^\top \Phi(\mathbf{x}^n,\mathbf{y},\mathbf{h},\mathbf{g}) \\
\Leftrightarrow &\max_{\mathbf{h},\mathbf{g}} \ell(\mathbf{y}^n,\mathbf{y}) + \sum_{j=1}^{V} \ell(h_j^n,h_j) + \sum_{i=1}^{R}\sum_{j=1}^{V} \alpha_j^\top x_i g_{ij} \\
&\quad + \beta_y^\top x + \sum_{j=1}^{V}\left[\gamma_{j,1}^y h_j + \gamma_{j,0}^y (1-h_j)\right] \\
\Leftrightarrow &\max_{\mathbf{h},\mathbf{g}} \ell(\mathbf{y}^n,\mathbf{y}) + \beta_y^\top x + \sum_{i=1}^{R}\sum_{j=1}^{V} \alpha_j^\top x_i g_{ij} \\
&\quad + \sum_{j=1}^{V}\left[\ell(h_j^n,h_j) + \gamma_{j,1}^y h_j + \gamma_{j,0}^y (1-h_j)\right]
\end{aligned} \tag{9}$$

where $\ell(\mathbf{y}^n,\mathbf{y}) + \beta_y^\top x$ is constant for a fixed $\mathbf{y}$, and $\ell(h_j^n,h_j)$ can be re-formulated as

$$\ell(h_j^n,h_j) = \begin{cases} 1-h_j, & if \;\; h_j^n = 1 \\ h_j. & if \;\; h_j^n = 0 \end{cases}$$

By adding the constraints in Eq. 2, the optimization problem in Eq. 9 can be rewritten as

$$\begin{aligned}
\max_{\mathbf{h},\mathbf{g}} \;& \sum_{i=1}^{R}\sum_{j=1}^{V} a_{ij} g_{ij} + \sum_{j=1}^{V} b_j h_j \\
\text{s.t.} \;& \sum_j g_{ij} \le 1, \;\; \max_i g_{ij} = h_j, \;\; g_{ij} \in \{0,1\}, \;\; \forall i, \forall j,
\end{aligned} \tag{10}$$

where a_{ij} is formalized by $a_{ij} = \alpha_j^\top \cdot x_i$, and b_j is defined as

$$b_j = \begin{cases} \gamma_{j,1}^{\mathbf{y}} - \gamma_{j,0}^{\mathbf{y}} - 1, & \textbf{if} \;\; h_j^n = 1, \\ \gamma_{j,1}^{\mathbf{y}} - \gamma_{j,0}^{\mathbf{y}} + 1, & \textbf{if} \;\; h_j^n = 0. \end{cases}$$

Since the optimization problem in Eq. 10 is non-convex, we can relax its constraints and obtain an integer linear problem:

$$\begin{aligned}
\max_{\mathbf{h},\mathbf{g}} \;& \sum_{i=1}^{R}\sum_{j=1}^{V} a_{ij} g_{ij} + \sum_{j=1}^{V} b_j h_j \\
\text{s.t.} \;& \sum_j g_{ij} \le 1, \;\; g_{ij} \le h_j \le \sum_i g_{ij}, \\
& g_{ij} \in \{0,1\}, \;\; h_j \in \{0,1\}, \;\; \forall i, \forall j.
\end{aligned} \tag{11}$$

Since the integral constraints $g_{ij} \in \{0,1\}$ and $h_j \in \{0,1\}$ make the optimization problem NP-hard, we further relax the integer values of h_j and g_{ij} to a real value in the range of [0,1]. Finally, the integer linear problem in Eq. 11 can be relaxed to a linear problem:

$$\begin{aligned}
\max_{\mathbf{h},\mathbf{g}} \;& \sum_{i=1}^{R}\sum_{j=1}^{V} a_{ij} g_{ij} + \sum_{j=1}^{V} b_j h_j \\
\text{s.t.} \;& \sum_j g_{ij} \le 1, \;\; g_{ij} \le h_j \le \sum_i g_{ij} \\
& 0 \le g_{ij} \le 1, \;\; 0 \le h_j \le 1, \forall i, \forall j.
\end{aligned} \tag{12}$$

After solving this problem, we round g_{ij} to the closest integer and obtain h_j by $h_j = \max_i g_{ij}$.

Similarly, the optimization problem in Eq. 8 can be solved by the following quadratic program:

$$\begin{aligned}
\min_{\mathbf{g}^n} \;& C' \sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^n \cdot g_{kl}^n \cdot \rho(i,j,k,l) \\
& - \sum_{i=1}^{R}\sum_{j=1}^{V} a_{ij} g_{ij}^n \\
\text{s.t.} \;& \sum_j g_{ij}^n \le 1, \;\; g_{ij}^n \le h_j^n \le \sum_i g_{ij}^n, \;\; 0 \le g_{ij}^n \le 1, \forall i, \forall j.
\end{aligned}$$

The complexity of computing L^n in Eq. 6 is $O(R^2V^2A)$ and the complexity of computing R^n is about $O(R^4V^4)$, where R is the number of video segments, V is the number of atomic actions, and A is the number of activities. Therefore, the computation complexity of model training is $O((R^4V^4A + R^4V^4)N_t)$, where N_t indicates the number of iterations in NRBM.

Latent Structural SVM has been successfully applied in region-based image annotation (Wang and Mori 2010) which models the mapping between image regions and object tags as latent information to understand the total scene from an image. Different from this method, our method models the temporal relationship between atomic actions and encourages the latent mapping to be consistent with it.

3.3 Model Inference

Given the model parameter $\mathbf{w} = \{\alpha;\beta;\gamma\}$, the inference problem is to simultaneously find the optimal activity label $\mathbf{y}$ and atomic action annotations $\mathbf{h}$ for an input activity video. The inference can be solved by the following optimization problem:

$$(\mathbf{y}^*,\mathbf{h}^*) = \arg\max_{\mathbf{y},\mathbf{h},\mathbf{g}} \mathbf{w}^\top \Phi(\mathbf{x},\mathbf{y},\mathbf{h},\mathbf{g}).$$

We enumerate all the possible activity labels $\mathbf{y}$ to find the optimal $\mathbf{y}^*$. For a fixed $\mathbf{y}$, we can modify the optimization problem in Eq. 12 by re-defining b_j as $b_j = \gamma_{j,1}^{\mathbf{y}} - \gamma_{j,0}^{\mathbf{y}}$, and adopt the similar linear program to find the final solution of $\mathbf{h}$. We can also obtain the optimized latent mapping matrix $\mathbf{g}$ as a by-product. The computational complexity of model inference is $O(R^2V^2A)$.

4 Automatically Learning of Atomic Action Annotations

4.1 Model Formulation

Our model is trained on videos annotated with activity labels and atomic action annotations. However, it is laborious and time-consuming to manually annotate atomic actions for each video. Besides, it is hard to distinguish atomic actions in some videos, and gathering atomic action annotations from humans would be prone to noise. To deal with these problems, we resort to semi-supervised learning.

In our setting, both the activity and the atomic action annotations are available for a portion of training videos, and the rest training videos are only annotated with activity labels. We introduce a semi-supervised method to automatically predict atomic action labels of the rest of the training videos only annotated with activity labels. Our method aims to learn a discriminative compatibility function:

$$G(\mathbf{x}, \mathbf{y}, \mathbf{h}) = \mathbf{w}^\top \Psi(\mathbf{x}, \mathbf{y}, \mathbf{h}),$$

where $\Psi(\mathbf{x}, \mathbf{y}, \mathbf{h})$ is a joint feature vector which describes the relationship among the video $\mathbf{x}$, the activity label $\mathbf{y}$, and the atomic action annotation $\mathbf{h}$. The model parameter includes two parts $\mathbf{w} = \{\lambda; \mu\}$. The relationship among a video $\mathbf{x}$, an activity label $\mathbf{y}$, and the atomic action annotations $\mathbf{h}$ is formulated as

$$\begin{aligned}
\mathbf{w}^\top \Psi(\mathbf{x}, \mathbf{y}, \mathbf{h}) &= \lambda^\top \psi_1(\mathbf{x}, \mathbf{y}) + \mu^\top \psi_2(\mathbf{h}, \mathbf{y}),\\
\lambda^\top \psi_1(\mathbf{x}, \mathbf{y}) &= \sum_{t=1}^{A} \lambda_t^\top \cdot x \cdot \mathbf{I}_t(\mathbf{y}),\\
\mu^\top \psi_2(\mathbf{h}, \mathbf{y}) &= \sum_{j=1}^{V} \sum_{t=1}^{A} \mu_{j,1}^t \cdot \mathbf{I}_1(h_j) \cdot \mathbf{I}_t(\mathbf{y})\\
&\quad + \mu_{j,0}^t \cdot \mathbf{I}_0(h_j) \cdot \mathbf{I}_t(\mathbf{y}),
\end{aligned}$$

where $\lambda^\top \psi_1(\mathbf{x}, \mathbf{y})$ measures the compatibility of a video $\mathbf{x}$ and certain activity class $\mathbf{y}$, and $\mu^\top \psi_2(\mathbf{h}, \mathbf{y})$ models the relationship between atomic action annotations $\mathbf{h}$ and the overall activity $\mathbf{y}$. Here, $\mathbf{I}_a(b)$ is an indicator function, namely, $\mathbf{I}_a(b) = 1$ if $a = b$, and 0 otherwise.

4.2 Learning Procedure

The training data of our model include a set of triples $X^L = \{(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n) | n = 1, 2, \ldots, N\}$ which are annotated with both activity labels $\mathbf{y}^n$ and atomic action annotations $\mathbf{h}^n$, and a set of two-tuples $X^U = \{(\mathbf{x}^m, \mathbf{y}^m) | m = 1, 2, \ldots, M\}$ which are only annotated with activity labels $\mathbf{y}^m$. Since the atomic action annotations $\mathbf{h}^m$ of videos in X^U are not available, they are regarded as latent variables. The model is formulated in a latent structural SVM framework for learning:

$$\begin{aligned}
\min_{\mathbf{w}, \xi_L^n, \xi_U^m, \delta^m} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \left(\sum_{n=1}^{N} \xi_L^n + \sum_{m=1}^{M} \xi_U^m \right) + C_2 \sum_{m=1}^{M} \delta^m \qquad (13)\\
\text{s.t.} \quad & \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n) - \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h})\\
& \geq \Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) - \xi_L^n, \forall \mathbf{x}^n \in X^L, \forall \mathbf{y}, \forall \mathbf{h} \qquad (14)\\
& \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m) - \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}, \mathbf{h})\\
& \geq \Delta(\mathbf{y}^m, \mathbf{y}) - \xi_U^m, \forall \mathbf{x}^m \in X^U, \forall \mathbf{y}, \forall \mathbf{h} \qquad (15)\\
& \Theta(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m, X^L, X^U) \leq \delta^m, \forall \mathbf{x}^m \in X^U. \qquad (16)
\end{aligned}$$

Equations 14 and 15 denote the max margin constraints for data in X^L and X^U, respectively. These constraints optimize the parameters by classifying training data correctly. Note that, due to the non-available ground-truth of atomic action annotations in X^U, the loss function $\Delta(\mathbf{y}^m, \mathbf{y})$ only involves the overall activity labels.

The constraints in Eq. 16 model the relationship among atomic action annotations of different training videos, in which $\Theta(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m, X^L, X^U)$ is a loss function of assessing the dissimilarity between atomic action annotations of $\mathbf{x}^m$ and the rest training videos. For simplicity, we only consider the relationship among videos with the same activity label, and define Θ as

$$\begin{aligned}
\Theta(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m, X^L, X^U) &= \sum_{\mathbf{x}^k \in X^L; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m)\\
&\quad + \sum_{\mathbf{x}^k \in X^U; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m),\\
\theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m) &= \langle \mathbf{x}^k, \mathbf{x}^m \rangle \cdot \mid \mathbf{h}^k - \mathbf{h}^m \mid,
\end{aligned} \qquad (17)$$

where $\theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m)$ is a pairwise cost function of annotating video $\mathbf{x}^k$ with $\mathbf{h}^k$ and annotating video $\mathbf{x}^m$ with $\mathbf{h}^m$. $\langle \mathbf{x}^k, \mathbf{x}^m \rangle$ measures the similarity between appearances of $\mathbf{x}^k$ and $\mathbf{x}^m$, and $\mid \mathbf{h}^k - \mathbf{h}^m \mid$ indicates the Hamming distance between $\mathbf{h}^k$ and $\mathbf{h}^m$. Although our method allows videos within the same activity class to be annotated with different atomic actions, it is reasonable that videos with similar appearances should also have similar atomic action annotations.

We utilize the non-convex bundle optimization method (NRBM) (Do and Artieres 2009) to solve Eq. 13. If we substitute Eqs. 14–16 into Eq. 13, the objective function can be rewritten as

$$\begin{aligned}
Q(\mathbf{w}) = {} & \tfrac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{\mathbf{x}^m \in X^U} \max_{\mathbf{y}, \mathbf{h}} \left[\mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}, \mathbf{h}) + \Delta(\mathbf{y}^m, \mathbf{y}) \right]\\
& + C_1 \sum_{\mathbf{x}^n \in X^L} \left\{ \max_{\mathbf{y}, \mathbf{h}} \left[\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}) + \Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) \right] - \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n) \right\}\\
& - \max_{\mathbf{h}^1 : \mathbf{h}^M} \left[C_1 \sum_{\mathbf{x}^m \in X^U} \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m) - C_2 \sum_{\mathbf{x}^m \in X^U} \Theta(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m, X^L, X^U) \right].
\end{aligned}$$

The subgradient of $Q(\mathbf{w})$ required in each iteration of the non-convex bundle optimization algorithm is given by

$$\begin{aligned}\frac{\partial Q(\mathbf{w})}{\partial \mathbf{w}} &= \mathbf{w} + C_1 \sum_{\mathbf{x}^n \in X^L} \Phi(\mathbf{x}^n, \mathbf{y}^n_*, \mathbf{h}^n_*) - \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n) \\ &\quad + C_1 \sum_{\mathbf{x}^m \in X^U} \Phi(\mathbf{x}^m, \mathbf{y}^m_*, \mathbf{h}^m_*) - \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m), \\ (\mathbf{y}^n_*, \mathbf{h}^n_*) &= \arg\max_{\mathbf{y},\mathbf{h}} [\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}) \\ &\quad + \Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h}))], \forall \mathbf{x}^n \in X^L, \qquad (18) \\ (\mathbf{y}^m_*, \mathbf{h}^m_*) &= \arg\max_{\mathbf{y},\mathbf{h}} [\mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}, \mathbf{h}) \\ &\quad + \Delta(\mathbf{y}^m, \mathbf{y})], \forall \mathbf{x}^m \in X^U, \qquad (19)\end{aligned}$$

$$\begin{aligned}\{\mathbf{h}^1 : \mathbf{h}^M\} = \arg\max_{\mathbf{h}^1:\mathbf{h}^M} \Bigg[& C_1 \sum_{\mathbf{x}^m \in X^U} \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m) \\ & - C_2 \sum_{\mathbf{x}^m \in X^U} \Theta(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m, X^L, X^U) \Bigg]. \qquad (20)\end{aligned}$$

In each iteration of the non-convex bundle optimization algorithm, we need to solve the inference problems defined in Eqs. 18 and 19, and the latent variables $\{\mathbf{h}^1 : \mathbf{h}^M\}$ of all the videos in X^U are jointly inferred according to Eq. 20.

For a fixed activity label $\mathbf{y}$, the inference problem of a video $\mathbf{x}^n \in X^L$ defined in Eq. 18 can be relaxed to a linear problem

$$\begin{aligned}&\arg\max_{\mathbf{h}} \sum_{j=1}^{V} b_j h_j, \\ &\text{s.t.} \quad 0 \le h_j \le 1, \forall j. \\ &b_j = \begin{cases} \mu^{\mathbf{y}}_{j,1} - \mu^{\mathbf{y}}_{j,0} - 1, & \textbf{if} \quad h^n_j = 1, \\ \mu^{\mathbf{y}}_{j,1} - \mu^{\mathbf{y}}_{j,0} + 1, & \textbf{if} \quad h^n_j = 0. \end{cases}\end{aligned}$$

Similarly, for a fixed activity label $\mathbf{y}$, the inference problem of a video $\mathbf{x}^m \in X^U$ defined in Eq. 19 can be relaxed to a linear problem:

$$\begin{aligned}&\arg\max_{\mathbf{h}} \sum_{j=1}^{V} (\mu^{\mathbf{y}}_{j,1} - \mu^{\mathbf{y}}_{j,0}) h_j, \\ &\text{s.t.} \quad 0 \le h_j \le 1, \forall j.\end{aligned}$$

We employ dual decomposition in (Sontag et al. 2011) to solve the challenging problem defined in Eq. 20. Substituting Eq. 17 into Eq. 20, we get

$$\begin{aligned}&\{\mathbf{h}^1 : \mathbf{h}^M\} \\ &= \arg\max_{\mathbf{h}^1:\mathbf{h}^M} \Bigg\{ C_1 \sum_{\mathbf{x}^m \in X^U} \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m) \\ &\quad - C_2 \sum_{\mathbf{x}^m \in X^U} \Bigg[\sum_{\mathbf{x}^k \in X^L; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m) \\ &\quad + \sum_{\mathbf{x}^k \in X^U; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m) \Bigg] \Bigg\} \qquad (21) \\ &= \arg\max_{\mathbf{h}^1:\mathbf{h}^M} \Bigg\{ \sum_{\mathbf{x}^m \in X^U} [C_1 \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m) \\ &\quad - C_2 \sum_{\mathbf{x}^k \in X^L; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m)] \\ &\quad - C_2 \sum_{\mathbf{x}^m \in X^U} \sum_{\mathbf{x}^k \in X^U; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m) \Bigg\}.\end{aligned}$$

We define a graphical model $G = \{Ns, Es\}$, where $Ns = \{\mathbf{h}^m | \mathbf{x}^m \in X^U\}$ and $Es = \{(\mathbf{h}^{m1}, \mathbf{h}^{m2}) | \mathbf{y}^{m1} = \mathbf{y}^{m2}\}$ denote nodes and edges, respectively. Then Eq. 22 can be re-formulated as a standard problem of dual decomposition:

$$\begin{aligned}\{\mathbf{h}^1 : \mathbf{h}^M\} = \arg\max_{\mathbf{h}^1:\mathbf{h}^M} \Bigg[& \sum_{\mathbf{h}^m \in Ns} \tau_n(\mathbf{h}^m) \\ & + \sum_{(\mathbf{h}^{m1}, \mathbf{h}^{m2}) \in Es} \tau_e(\mathbf{h}^{m1}, \mathbf{h}^{m2}) \Bigg], \qquad (22)\end{aligned}$$

where $\tau_n(\cdot)$ is the energy function of a node, and $\tau_e(\cdot)$ represents the energy function of an edge:

$$\begin{aligned}\tau_n(\mathbf{h}^m) &= C_1 \mathbf{w}^\top \Phi(\mathbf{x}^m, \mathbf{y}^m, \mathbf{h}^m) \\ &\quad - C_2 \sum_{\mathbf{x}^k \in X^L; \mathbf{y}^k = \mathbf{y}^m} \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m), \\ \tau_e(\mathbf{h}^{m1}, \mathbf{h}^{m2}) &= \theta(\mathbf{x}^k, \mathbf{h}^k, \mathbf{x}^m, \mathbf{h}^m).\end{aligned}$$

In order to solve the problem in Eq. 22 by dual decomposition, we must reduce the state space of a latent variable $\mathbf{h}^m$. If there are V atomic actions, the number of possible atomic action annotations is 2^V. However, most of them are unreasonable annotations since some atomic actions (such as "basketball dribble" and "jump into sand pit") cannot appear simultaneously. In the implementation, we reduce the state space of latent variables by finding all the reasonable atomic action annotations from dataset X^L.

5 Experiments

5.1 Human Activity Datasets

To evaluate the effectiveness of our model, we conduct experiments on the three human activity datasets.

Synthesized Multi-view IXMAS Activity Dataset. We construct a synthesized set of complex activities by concatenating simple actions from the multi-view IXMAS dataset

Table 1 Synthesized complex activity classes from the multi-view IXMAS dataset. The column of "Activity" indicates the synthesized activity classes: #1–#8 and the column of "Atomic actions" represents the concatenated simple action classes from the IXMAS dataset

Activity	Atomic actions
#1	Check watch, cross arms, scratch head, sit down, get up
#2	Cross arms, scratch head, sit down, get up, turn around
#3	Scratch head, sit down, get up, turn around, walk
#4	Sit down, get up, turn around, walk, wave
#5	Get up, turn around, walk, wave, punch
#6	Turn around, walk, wave, punch, kick
#7	Walk, wave, punch, kick, point
#8	Wave, punch, kick, point, pick up

(Weinland et al. 2007), which contains 12 simple action classes. Each activity video is constructed by concatenating five different simple actions selected from the 12 classes. For each view, we synthesize eight complex activity classes and different activity classes have different five atomic actions. The detailed definitions of activity classes for each view are illustrated in Table 1. Accordingly, the number of total atomic action annotation terms is 12 and the number of present annotation terms is 5. Each activity class is conducted by 12 subjects. We adopt the leave-one-subject-out cross validation setting, in which videos of 11 subjects are used as training data and videos of the remaining one subject are used for testing.

Olympic Sports Dataset. The Olympic sports dataset (Niebles et al. 2010) contains 16 different Olympic sports activity classes: High-jump, Long-jump, Triple-jump, Pole-vault, Gymnastics-vault, Shot-put, Snatch, Clean-jerk, Javelin-throw, Hammer-throw, Discus-throw, Diving-platform, Diving-springboard, Basketball-layup, Bowling, and Tennis-serve. We define 24 atomic actions and manually annotate each activity video by assigning it to 24 annotation terms. The whole dataset is split into 649 videos for training and 134 videos for testing.

UCF101 Dataset. The UCF101 dataset consists of realistic user-uploaded videos from 101 action classes, and video clips of each action are divided into 25 groups. We conduct experiments on the following 13 actions which can be decomposed into a series of meaningful atomic actions: Balance Beam, Basketball, Bowling, Cliff Diving, Diving, Hammer-throw, High-jump, Javelin-throw, Long-jump, Pole-vault, Shot-put, Throw-discus, and Uneven Bars. For each action class, we use the video clips from 12 of the 25 groups as testing samples, leaving the rest for training.

5.2 Experimental Setting

In this paper, we adopt a video description (Wang et al. 2013a) based on dense trajectories and motion boundary histogram descriptors. Dense trajectories are obtained by tracking dense patches in videos, and two local descriptors (i.e., HOG and MBH) are extracted from the spatiotemporal region around each trajectory to describe the appearance and motion information. The parameters of trajectory length and dense sample step are set to 15 and 5, respectively. For the more complex and challenging Olympic and UCF101 datasets , the Motion Interchange Pattern (MIP) (Kliper et al. 2012) feature is extracted to further improve the recognition performance. The MIP descriptor encodes local motion patterns by matching patches across successive video frames, and captures local changes in motion directions. The standard bag-of-words approach is utilized to construct a codebook for each visual descriptor separately. The number of visual words per descriptor is fixed to 400. To reduce the computational complexity, a subset of 100,000 features are randomly selected from the training data and clustered to generate the codebook using k-means algorithm.

In our experiments, we split each video into several segments with equal length, and the length of each segment is set to 30 frames. We generate a feature vector for a video with all the local descriptors extracted from the entire video, and create a representation for each segment using descriptors within it.

The initialization of latent variables is important in practice for our method because NRBM can only guarantee a local optimum solution. Since atomic action annotations of the whole video are known for training samples, we initiate the latent variables according to the atomic action annotations under the constraints defined in Eq. 2.

5.3 Activity Recognition Results

We compare our method with several baseline methods on the three datasets. The first baseline method is a linear SVM model based on the bag-of-words visual feature extracted from the entire activity video, without considering atomic actions. The second baseline method is from our framework, without considering the mapping model between video segments and atomic actions. For the synthesized multi-view IXMAS activity dataset, we evaluate performances of different methods with the recognition accuracy. For the Olympic

Table 2 Comparison of action recognition performance between our method and the baseline methods on the three datasets

Method	IXMAS					Olympic	UCF101
	View 1	View 2	View 3	View 4	View 5		
Linear SVM	0.729	0.844	0.771	0.844	0.656	0.737	0.682
No segments	0.781	0.854	0.833	0.844	0.760	0.794	0.708
Our method	0.959	0.959	0.969	0.969	0.938	0.820	0.743

Table 3 Mean average precision values for activity recognition between our method and other state-of-the-art methods on the Olympic sports dataset

Method	Mean AP
Niebles et al. (2010)	0.625
Tang et al. (2012)	0.668
Liu et al. (2011)	0.743
Zhou and Wang (2012)	0.710
Li and Vasconcelos (2012)	0.765
Jiang et al. (2012)	0.806
Wang et al. (2013a)	0.772
Li et al. (2013)	0.782
Zhou et al. (2013)	0.783
Gaidon et al. (2014)	0.850
Our method	**0.820**

Table 4 Recognition accuracies of different methods from five views on the synthesized multi-view IXMAS activity dataset

Method	View 1	View 2	View 3	View 4	View 5
Tang et al. (2012)	0.854	0.833	0.823	0.875	0.667
Zhou et al. (2013)	0.865	0.896	0.813	0.875	0.813
Our method	0.959	0.959	0.969	0.969	0.938

sports dataset and the UCF101 dataset, we compute the Average Precision (AP) for each activity class and report the mean AP over all the activity classes.

We summarize the comparisons between our method and the baseline methods on the synthesized multi-view IXMAS activity dataset, Olympic sports dataset, and UCF101 dataset in Table 2. The methods of "Linear SVM" and "No segments" indicate the first baseline and second baseline methods, respectively. From Table 2, we have the interesting observations as follows: (1) our method outperforms the linear SVM on the three datasets, which obviously demonstrates that it is beneficial to exploit a set of atomic actions for distinguishing different complex activities; (2) by incorporating the segment-annotation mapping matrix into model, our method achieves better results than the second baseline method which does not consider the relationship between video segments and atomic actions. Clearly, exploring temporal structure of atomic actions further improves the activity classification performance.

Table 3 compares our method with state-of-the-art methods on the Olympic sports dataset by evaluating the mean average precision over all activity classes. As shown in Table 3, our method performs better than (Niebles et al. 2010; Tang et al. 2012; Liu et al. 2011; Zhou and Wang 2012; Li and Vasconcelos 2012; Jiang et al. 2012; Wang et al. 2013a; Li et al. 2013; Zhou et al. 2013) and achieves comparable perfor-

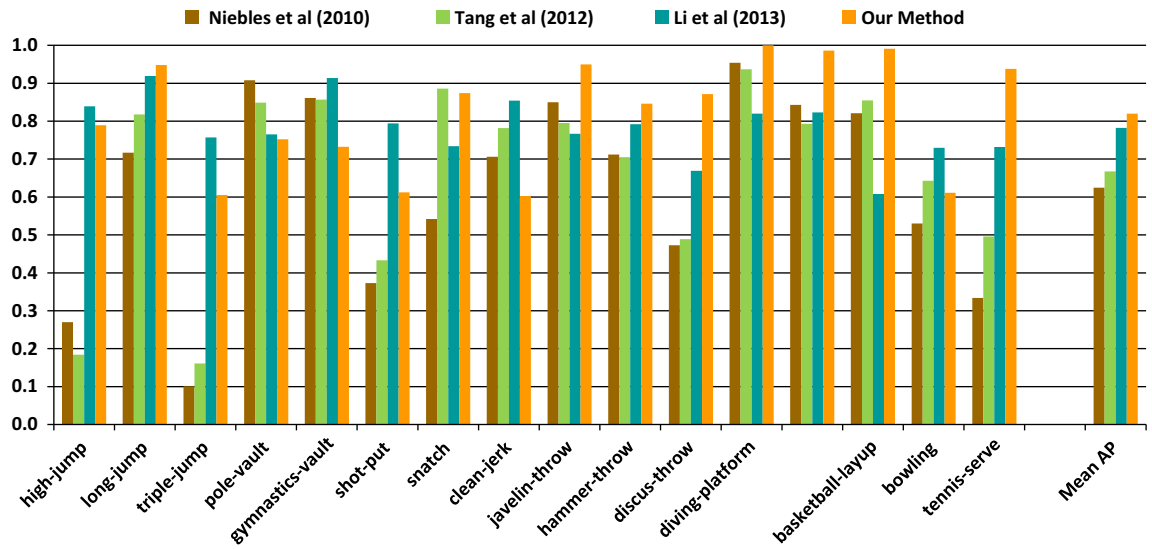

Fig. 2 Average precision values for activity recognition of different methods on the Olympic sports dataset (Color figure online)

Table 5 Average Precision values for activity recognition of three methods on the UCF101 dataset

Activity	Tang et al. (2012)	Zhou et al. (2013)	Our method
Balance-beam	0.724	0.824	**0.858**
Basketball	0.678	0.671	**0.756**
Bowling	0.977	0.965	**0.967**
Cliff-diving	**0.973**	0.952	0.947
Diving	0.967	0.970	**0.977**
Hammer-throw	0.705	0.637	**0.719**
High-jump	0.555	0.605	**0.636**
Javelin-throw	**0.509**	0.500	0.466
Long-jump	0.643	0.629	**0.655**
Pole-vault	0.619	0.772	**0.848**
Shot-put	0.292	0.324	**0.330**
Discus-throw	0.406	0.454	**0.543**
Uneven-bars	0.934	0.924	**0.948**
MAP	0.691	0.710	**0.742**

Bold values indicate the best performance

mance with (Gaidon et al. 2014) which preprocesses videos by using camera motion compensation method. Although the Olympic sports dataset is difficult with large appearance and background variability, our result is still encouraging which shows that our model is able to capture the semantic concept and temporal structure of atomic actions for distinguishing different activity classes. Figure 2 compares the per-class Average Precision values for activity recognition between our method and three methods. From Fig. 2, we can observe that our method achieves better performance for most activities.

To show the promising performance of our method, we also compare it with two relevant methods (Tang et al. 2012; Zhou et al. 2013) on the synthesized multi-view IXMAS activity dataset and UCF101 dataset. Table 4 summarizes the recognition accuracies of three methods for different views on the synthesized multi-view IXMAS activity dataset. For all views, our method achieves the highest recognition accuracy with the same feature and the same evaluation strategy. Table 5 depicts the Average Precision (AP) values of each activity on the UCF101 dataset. As is shown in Table 5, for most of the activities from the UCF101 dataset, our method yields higher AP than both (Tang et al. 2012) and (Zhou et al. 2013) using the same evaluation setting and the same feature. We note that the result on "shot-put" is relatively poor in comparison with other activities. One possible reason is the large variations within this class due to various factors such as motion style.

5.4 Evaluation of the Semi-supervised Method

We evaluate the performance of our semi-supervised method for learning atomic action annotations of training videos on the Olympic sports dataset and the UCF101 dataset. In this experiment, all the training videos are annotated with the overall activity labels, and we randomly annotate 25, 50, and 75 % of the training videos with atomic actions. For each setting, we compute the accuracy of atomic action annotations for the rest training videos, and summarize the results in Fig. 3. A video is considered to be correctly annotated only if all the atomic action annotations are consistent with the ground-truth. From Fig. 3, we can see that performance of atomic action annotation increases with the amount of training videos annotated with atomic action labels.

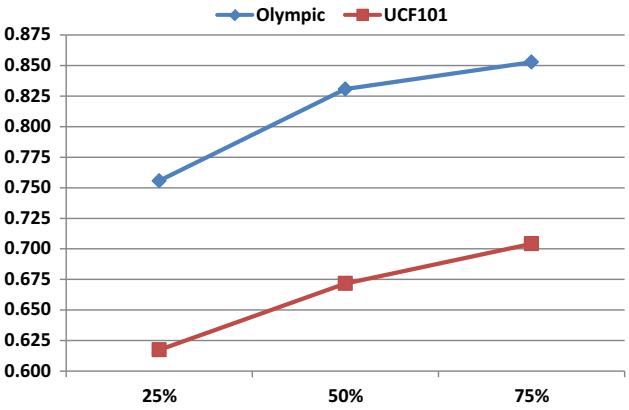

Fig. 3 Accuracies of atomic action annotations with a varying number of training videos annotated with both activity and atomic action labels (Color figure online)

Figure 4 compares our model using the semi-supervised method to learn the atomic action annotations of training videos with the Linear SVM baseline method. Notice that, when the proportion of training videos annotated with atomic actions are increased to 100 %, all the training videos

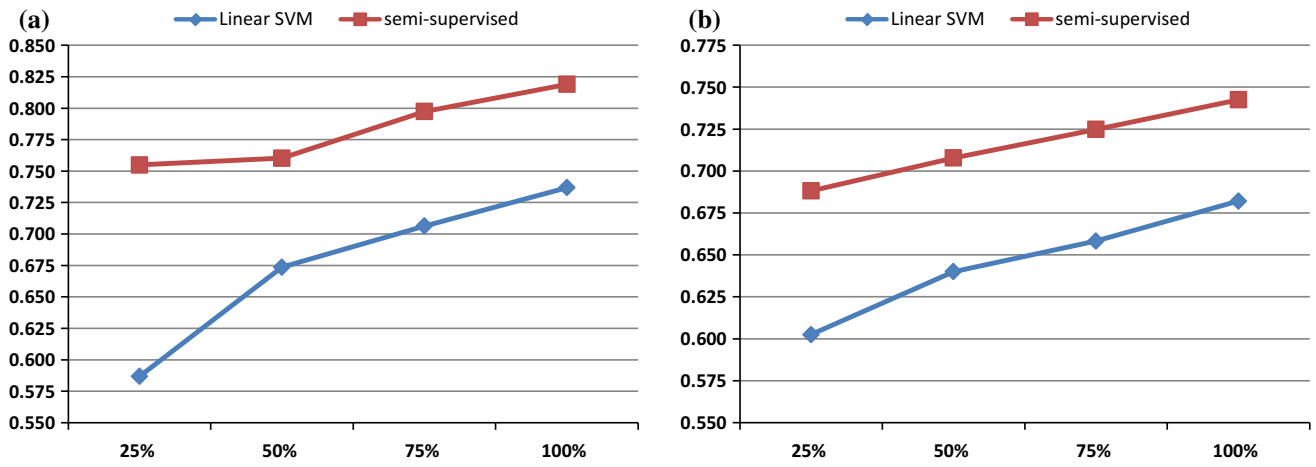

Fig. 4 Mean average precision values for activity recognition of two methods. **a** The Olympic sports dataset, **b** the UCF101 dataset (Color figure online)

are annotated with atomic actions and our method is fully supervised. Figure 4 shows that our semi-supervised method outperforms the Linear SVM baseline method by making using of the videos not annotated with atomic action labels. Furthermore, when only 25 % of training videos are annotated with atomic actions, our model with semi-supervised learning can perform better than the Linear SVM baseline method using all training videos. If the proportion of training videos annotated with atomic actions is increased to 75 %, our model with semi-supervised learning achieves comparable performance to the supervised model using 100 % annotated training videos, which demonstrates the effectiveness of the semi-supervised method.

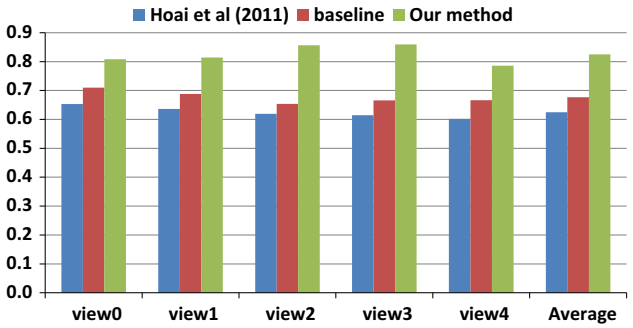

Fig. 5 Accuracies of segment annotations for three methods on the synthesized multi-view IXMAS activity dataset. This figure is best seen in color (Color figure online)

5.5 Video Description Results

In addition to performing competitive recognition results on difficult datasets, our method is also capable of obtaining a rich description of a long activity video by capturing both semantic concept and temporal structure of atomic actions. Specifically, our method can interpret a new complex activity via finding "what" simple atomic actions happening in the video based on the predicted atomic action annotations **h**, as well as detecting "when" these atomic actions occurring in the temporal direction based on the predicted segment-annotation mapping matrix **g**.

5.5.1 Quantitative Evaluation

We conduct experiments on the synthesized multi-view IXMAS activity dataset and quantitatively evaluate the performance of our method for video description. According to the predicted segment-annotation mapping matrix, each segment is annotated with at most one atomic action, and the performance is evaluated by computing the accuracy of segment annotations.

We compare our method with the method in (Hoai et al. 2011), which jointly performs action segmentation and action recognition in videos. In our experiments, the method in (Hoai et al. 2011) takes segments of a synthesized activity video as input, and annotates the video segments with different atomic actions. This method focuses on the annotation of atomic actions for video segments, without considering the complex activity label of the whole video. Different from it, our method utilizes a unified framework to capture the relationship among video segments, atomic actions and complex activities, and obtains a hierarchical description based on "which" activity, "what" atomic actions, and "when" of atomic actions happening in a video. Our method is also compared with a baseline method, which is from our framework without considering the temporal distribution of atomic actions. Particularly, the baseline method neglects $\sum_{i=1}^{R}\sum_{j=1}^{V}\sum_{k=1}^{R}\sum_{l=1}^{V} g_{ij}^{n} \cdot g_{kl}^{n} \cdot \rho(i,j,k,l)$ that forces the mapping matrix to be consistent with the tem-

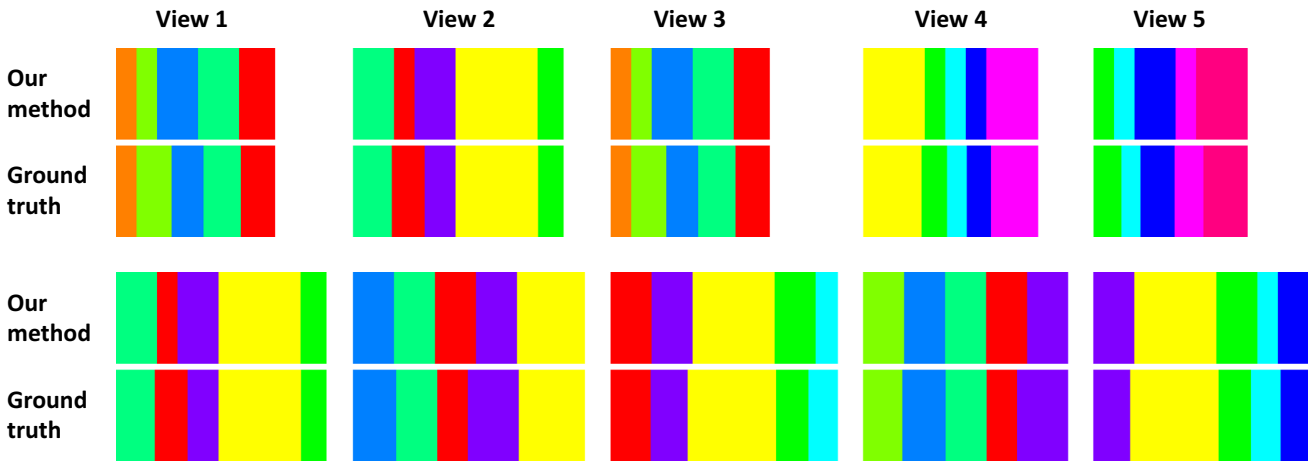

Fig. 6 Comparison between our prediction results and the ground truth on detecting semantic concept and temporal structure of atomic actions for the synthesized multi-view IXMAS activity dataset. In each view, we show two examples (i.e., test activities) depicted by *colorbars*. *Different colors* represent different atomic action classes (i.e., annotation terms). The width of color bin indicates the temporal duration of the corresponding atomic action class. This figure is best seen in color (Color figure online)

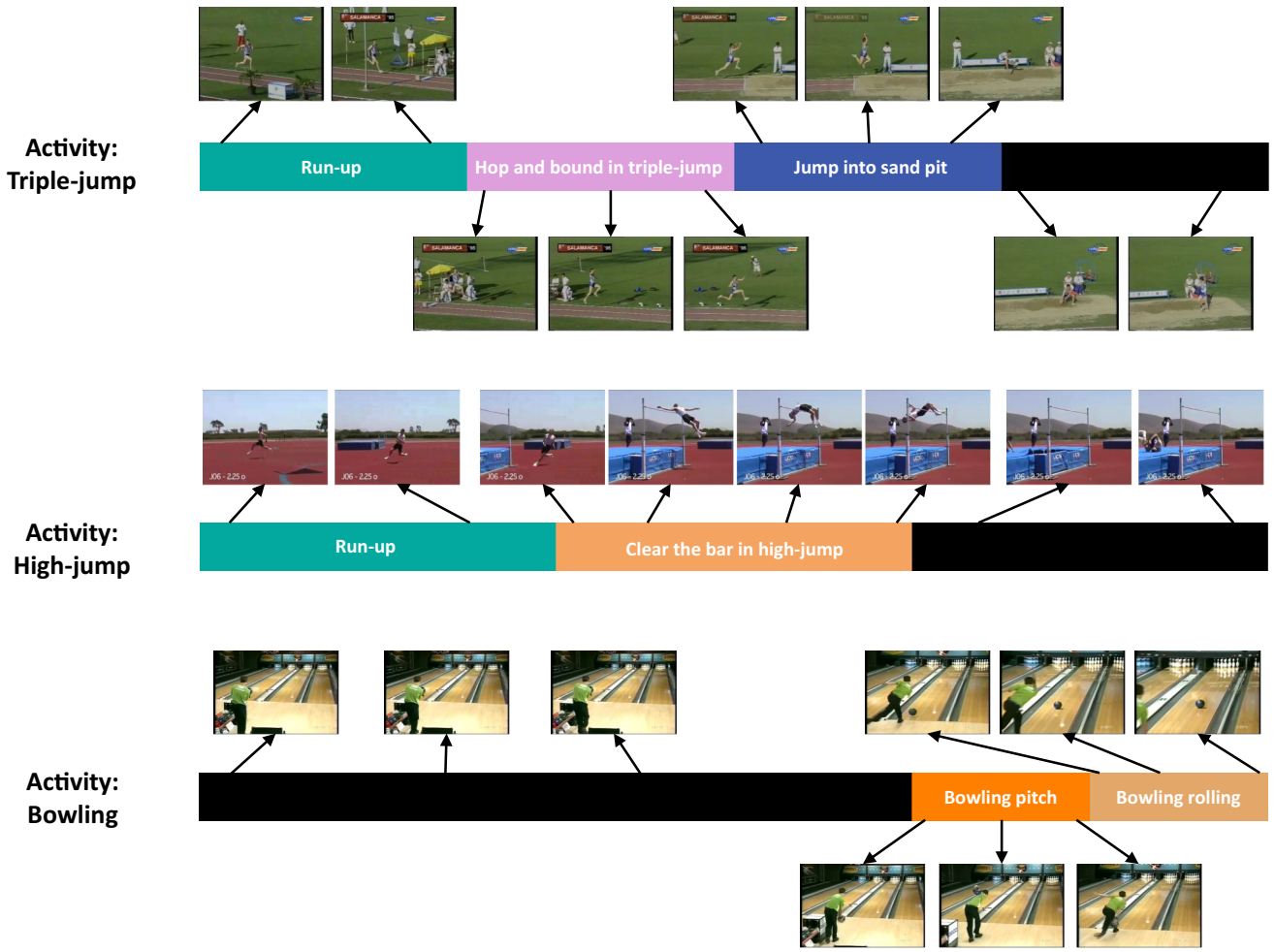

Fig. 7 Examples of activity video description results on the Olympic sports dataset. The timeline shows the detected semantic concepts of each video, where different colors indicate different atomic actions and "*black*" segments are not associated with any atomic action. This figure is best seen in color (Color figure online)

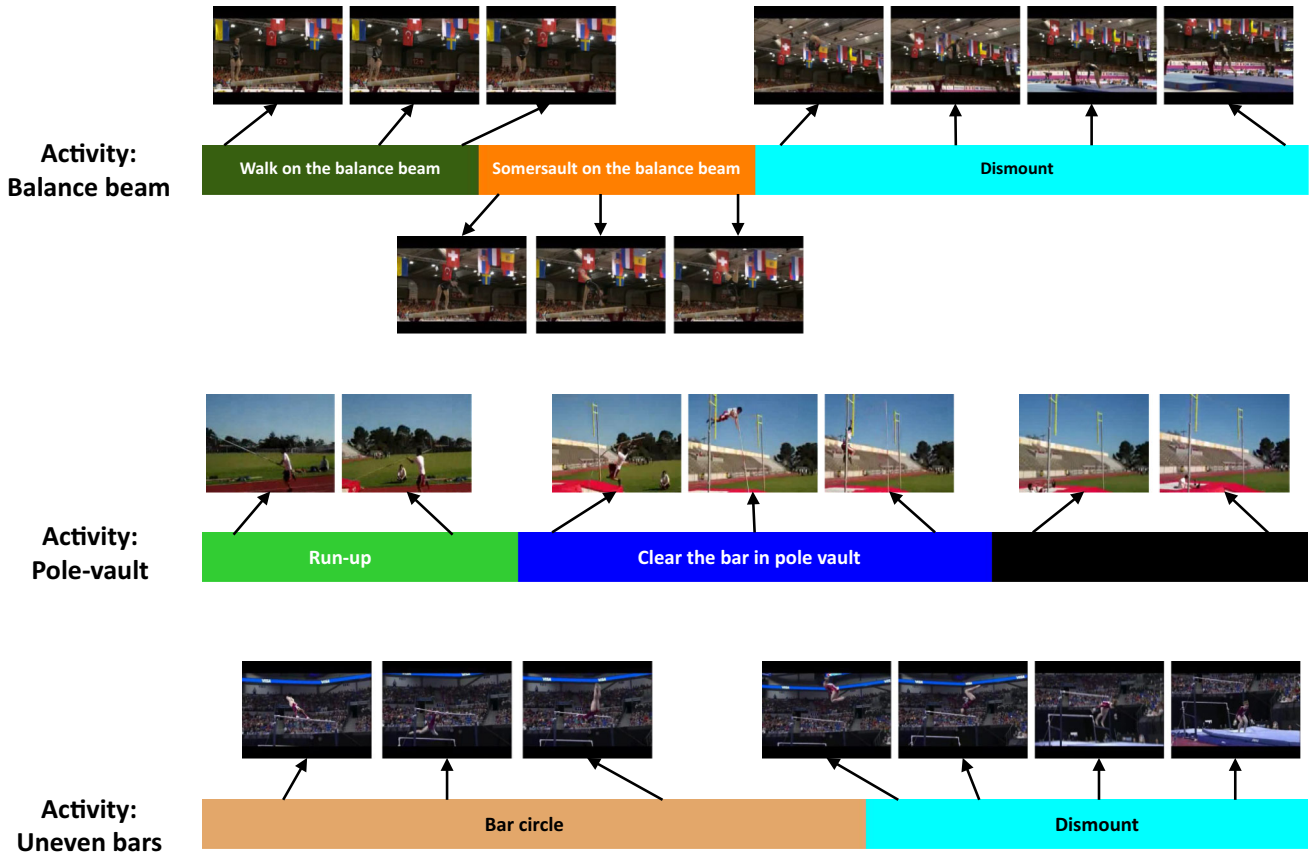

Fig. 8 Examples of activity video description results on the UCF101 dataset. The timeline shows the detected semantic concepts of each video, where different colors indicate different atomic actions. This figure is best seen in color (Color figure online)

poral distribution of atomic actions, and Eq. 4 is rewritten as $\max_{\mathbf{g}^n}[\mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}^n, \mathbf{h}^n, \mathbf{g}^n) - \max_{\mathbf{g}} \mathbf{w}^\top \Phi(\mathbf{x}^n, \mathbf{y}, \mathbf{h}, \mathbf{g}) \geq \Delta((\mathbf{y}^n, \mathbf{h}^n), (\mathbf{y}, \mathbf{h})) - \xi^n, \forall n, \forall \mathbf{y}, \forall \mathbf{h}]$.

Comparison of our method, the baseline method, and the method in (Hoai et al. 2011) is shown in Fig.5. From Fig.5, we have two observations as follows: (1) our method is able to achieve much better performance than the method in Hoai et al. (2011) by jointly modeling the relationship among video segments, atomic actions, and complex activities, which demonstrates the ability of our method for video description; (2) for each view, our method performs significantly better than the baseline method, which demonstrates that it is reasonable and effective to enforce the consistency between the segment-annotation mapping matrix and the temporal distribution of atomic actions.

Examples of comparisons between our prediction results and the ground-truth on the synthesized multi-view IXMAS activity dataset are demonstrated in Fig.6. In most cases, our method succeeds in accurately predicting both the semantic concept and temporal localization of atomic actions. The reason for error occurring at the boundaries between atomic actions may be attributed to the initial segmented video clips, in which some clips may cross two different atomic actions.

5.5.2 Qualitative Evaluation

Since the ground-truth of segment-annotation mapping matrix is not available for the Olympic sports and UCF101 datasets, we show qualitative results in this section. Examples of descriptions for activities by the semantic annotation and temporal localization of atomic actions for the Olympic sports dataset and the UCF101 dataset are illustrated in Figs. 7 and 8, respectively. Taking "triple-jump" in Fig. 7 for example, it is associated with atomic actions "run-up", "hop and bound in triple-jump", and "jump into sand pit". It is also interesting to observe that these atomic actions are roughly localized in the video time. Beside, our method is able to roughly temporal localize atomic actions that only appear in one activity, such as "bowling pitch" and "bowling rolling" in activity "bowling", by enforcing the consistency between the segmentation-annotation mapping matrix and the prior temporal distribution of atomic action annotations. From Figs. 7 and 8, we can see that the "black" segments not related to any atomic action are motionless video segments (See "bowling" activity) or irrelevant actions (See "high-jump", "triple-jump" and "pole-vault" activities). Furthermore, atomic actions are shared among different

activities, for example, the atomic action "run-up" appears in both "triple-jump" and "high-jump" activities in Fig. 7, and the atomic action "dismount" appears in both "balance beam" and "uneven bars" activities in Fig. 8.

6 Conclusions

We have presented a hierarchical and complete description of an activity video by automatically inferring the "which" of activities, "what" of atomic actions, and "when" of present atomic actions. A novel latent discriminative structural model is developed to model the relationship among video segments, atomic action annotations, and overall activities. Competitive activity recognition results have been shown on difficult datasets. Meanwhile, quantitative and qualitative experiments have demonstrated the capability of our model for video description.

Acknowledgments This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant Nos. 61203274, 61375044 and 61472038.

References

Bhattacharya, S., Kalayeh, M. M., Sukthankar, R. & Shah, M. (2014). Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Do, T. M. T. & Artieres, T. (2009). Large margin training for hidden markov models with partially observed states. In *IEEE international conference on machine learning (ICML)*.

Dollar, P., Rabaud, V., Cottrell, G. & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *VS PETS*.

Efros, A.A., Berg, A.C., Mori, G. & Malik, J. (2003). Recognizing action at a distance. In *IEEE international conference on computer vision (ICCV)*.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *32*(9), 1627–1645.

Gaidon, A., Harchaoui, Z. & Schmid, C. (2011). Actom sequence models for efficient action detection. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Gaidon, A., Harchaoui, Z., & Schmid, C. (2014). Activity representation with motion hierarchies. *International Journal of Computer Vision (IJCV)*, *107*(3), 219–238.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, *29*(12), 2247–2253.

Hoai, M., Lan. Z. & Torre, F. (2011). Joint segmentation and classification of human actions in video. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Hu, N., Englebienne, G., Lou. Z. & Krose, B. (2014). Learning latent structure for activity recognition. In *IEEE international conference on robotics and automation (ICRA)*.

Izadinia, H. & Shah, M. (2012). Recognizing complex events using large margin joint low-level event model. In *European conference on computer vision (ECCV)*.

Jiang, Y., Dai, Q., Xue, X., Liu, W. & Ngo, C. W. (2012). Trajectory-based modeling of human actions with motion reference points. In *European conference on computer vision (ECCV)*.

Kliper, O., Gurovich, Y., Hassner, T., Wolf, L. (2012). Motion interchange patterns for action recognition in unconstrained videos. In *European conference on computer vision (ECCV)*.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision (IJCV)*, *64*, 117–123.

Laxton, B., Lim, J. & Kriegman, D. (2007). Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Le, Q., Zou, W., Yeung, S. & Ng, A. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Li, W. & Vasconcelos, N. (2012). Recognizing activities by attribute dynamics. In *Neural information processing systems conference (NIPS)*.

Li, W., Zhang, Z., & Liu, Z. (2008). Expandable data-driven graphical modeling of human actions based on salient postures. *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, *18*(11), 1499–1510.

Li, W., Yu, Q., Sawhney, H. & Vasconcelos, N. (2013). Recognizing activities via bag of words for attribute dynamics. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Lillo, I., Soto, A., Niebles, J.C. (2014). Discriminative hierarchical modeling of spatio-temporally composable human activities. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Liu, J., Kuipers, B., Savarese, S. (2011). Recognizing human actions by attributes. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Niebles, J., Chen, C., Li, F. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision (ECCV)*.

Pirsiavash, H., Ramanan, D. (2014). Parsing videos of actions with segmental grammars. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Rodriguez, M.D., Ahmed, J. & Shah, M. (2008). Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Sadanand, S. & Corso, J.J. (2012). Action bank: a high-level representation of activity in video. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Sontag, D., Globerson, A., & Jaakkola, T. (2011). Introduction to dual decomposition for inference. *Optimization for Machine Learning*, *1*, 219–254.

Sun, C. & Nevatia, R. (2013). Active: Activity concept transitions in video event classification. In *IEEE international conference on computer vision (ICCV)*.

Tang, K., Li, F.F., Koller, D. (2012). Learning latent temporal structure for complex event detection. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Wang, H., Klaser, A., Schmid, C., & Liu, C. L. (2013a). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)*, *103*(1), 60–79.

Wang, L., Qiao, Y., Tang, X., et al. (2013b). Mining motion atoms and phrases for complex action recognition. In *IEEE international conference on computer vision (ICCV)*.

Wang, L., Qiao, Y., & Tang, X. (2014). Latent hierarchical model of temporal structure for complex activity classification. *IEEE Transactions on Image Processing (T-IP)*, *23*(2), 810–822.

Wang, Y. & Mori, G. (2010). A discriminative latent model of image region and object tag correspondence. In *Neural information processing systems conference (NIPS)*.

Weinland, D., Boyer, E. & Ronfard, R. (2007) Action recognition from arbitrary views using 3d exemplars. In *IEEE international conference on computer vision (ICCV)*.

Wu, X., Xu, D., Duan, L. & Luo, J. (2011). Action recognition using context and appearance distribution features. In *IEEE international conference on computer vision and pattern recognition (CVPR)*

Yilmaz, A. & Shah, M. (2005). Action sketch: a novel action representation. In *IEEE international conference on computer vision and pattern recognition (CVPR)*.

Yu, C. N. J. & Joachims, T. (2009). Learning structural svms with latent variables. In *IEEE international conference on machine learning (ICML)*.

Yu, G., Yuan, J. & Liu, Z. (2012). Propagative hough voting for human activity recognition. In *European conference on computer vision (ECCV)*.

Zhou, Q. & Wang, G. (2012). Atomic action features: A new feature for action recognition. In *European conference on computer vision (ECCV)*.

Zhou, Q., Wang, G., Jia, K. & Zhao, Q. (2013). Learning to share latent tasks for action recognition. In: *IEEE international conference on computer vision (ICCV)*.