Multi-Feature Max-Margin Hierarchical Bayesian Model for Action Recognition

Shuang Yang, Chunfeng Yuan, Baoxin Wu, Weiming Hu
NLPR, Institution of Automation, CAS
Beijing, China
{syang,cfyuan,bxwu,wmhu}@nlpr.ia.ac.cn

Fangshi Wang
Beijing Jiaotong University
Beijing, China
fshwang@bjtu.edu.cn

Abstract

In this paper, a multi-feature max-margin hierarchical Bayesian model (M^3HBM) is proposed for action recognition. Different from existing methods which separate representation and classification into two steps, M^3HBM jointly learns a high-level representation by combining a hierarchical generative model (HGM) and discriminative max-margin classifiers in a unified Bayesian framework. Specifically, HGM is proposed to represent actions by distributions over latent spatial temporal patterns (STPs) which are learned from multiple feature modalities and shared among different classes. For recognition, we employ Gibbs classifiers to minimize the expected loss function based on the max-margin principle and use the classifiers as regularization terms of M^3HBM to perform Bayeisan estimation for classifier parameters together with the learning of STPs. In addition, multi-task learning is applied to learn the model from multiple feature modalities for different classes. For test videos, we obtain the representations by the inference process and perform action recognition by the learned Gibbs classifiers. For the learning and inference process, we derive an efficient Gibbs sampling algorithm to solve the proposed M^3HBM. Extensive experiments on several datasets demonstrate both the representation power and the classification capability of our approach for action recognition.

1. Introduction

Human action recognition in video is an active area for its potential in a number of real-world applications. Various methods have been proposed to achieve automatic action recognition. The traditional procedure for action recognition [25, 2, 32] includes two separate steps: designing descriptors to represent actions and then training classifiers to predict the action class for test video.

Recently, many work [15, 13, 27] propose to perform representation and classification together in a single model with the aid of probabilistic graphical models and present successful results. A common idea of these methods is to construct a graphical model for each class and describe actions by combining the learned class-specific parameters. Then a prediction score for each action class is obtained by a class-specific inference process, and the final action class is obtained through a maximum-voting process based on the class-specific scores. One challenge of these methods is that the powerful capability of discriminative classifiers, like max-margin classifiers, is excluded in the learning process, which makes the learned representations may be appropriate for description, but may be not optimal for classification.

In this paper, we propose a multi-feature max-margin hierarchical Bayesian model (M^3HBM), which jointly learns a multi-feature hierarchical generative model (HGM) as the representation part together with max-margin classifiers in a unified Bayesian framework for action recognition. Inspired by the recent success of hierarchical structures for representation, we model actions by a hierarchical generative model (HGM) including three layers: point-level visual observations, region-level local STPs which are distributed in many different small neighbourhoods, and top-level global STPs which are shared among all different classes without position limitation.

Through HGM, we learn from multiple feature modalities to jointly characterize different aspects of actions and represent actions by the probabilistic distributions over the top-level global STPs. For the classification part in M^3HBM, Gibbs classifiers are applied to minimize the expected loss based on the max-margin principle, and Gaussian priors are introduced to the classifiers to perform Bayesian estimation for classifier parameters. We employ the classifiers as regularization terms of M^3HBM to jointly learn the parameters in HGM and Gibbs classifiers in a united process. By learning the representations together with the classifiers in a unified framework, the learned latent STPs are both descriptive and discriminative for action recognition.

Additionally, we integrate the multi-task learning into our model to learn shared latent STPs from multiple fea-

ture modalities for different classes.In the end, an efficient collapsed Gibbs sampling algorithm is derived to learn the proposed M^{3}HBM. We validate the proposed model on both the representation and recognition capabilities, and the experimental results clearly show the effectiveness of our model for action recognition.

The reminder of this paper is organized as follows. Section 2 briefly reviews the related work. Section 3 presents the details of the proposed model and derives the Gibbs sampling algorithm for the learning and inference process. Section 4 illustrates the experimental settings and presents the empirical results. Finally, Section 5 concludes the paper.

2. Related Work

During the past few years, significant progress has been made to develop some appropriate descriptors for action recognition. However, many of them are carefully engineered and/or hand-crafted which are designed for specific tasks and so have limited generality.

Recently, some work propose to automatically learn the mid-level or high-level representations by introducing hierarchical structures for recognition [24, 24, 9] . One line of work is based on topic models [4, 1] which have received increasing attention in recent years for their appealing motivation and great success to discover the latent semantic topics in documents. To make the model structure more suitable for computer vision tasks, various types of discriminative information [7, 16, 27] and relations between topics and words have been gradually introduced into the models [11, 5, 30, 14]. For example, Niu *et al*. [17] propose to cast the spatial context information in a discriminative LDA model for scene recognition. They include the class label in the model and perform recognition by maximum a posteriori (MAP) estimation of the image class. In [12], Lin and Xiao integrate the spatial relationships between image regions into a spatial topic process for representation and employ a linear SVM for scene classification. For recognition, many methods perform classification by training extra independent classifiers in a next separate stage. Some others learn the model parameters for each class respectively, give the class-specific prediction scores by maximizing the likelihood or posterior of the model for each class, and then perform maximum voting to predict the final class of test samples, which considers little the advantages of discriminative classifiers for recognition.

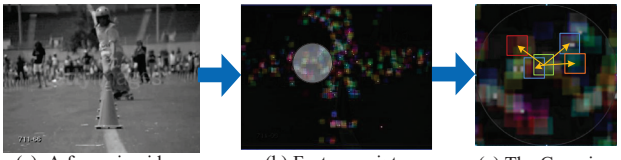
Some recent approaches have taken several attempts to combine the learning process of high-level representations together with discriminative classifiers in a single graphical model for recognition. Two most related methods to our work are the relevance topic model in [29] and the max-margin latent Dirichlet allocation in [28]. In the relevance topic model [29], sparse Bayesian learning is incorporated into the Replicated Softmax model [21] to discover the topics for representation and recognition. Although they learn the representations and classifiers together like us, they optimize the parameters in a principle of automatic relevance determination which is totally different from our max-margin principle. In [28], Wang and Mori develop a max-margin latent Dirichlet allocation(MMLDA) model by combining the max-margin learning and latent Dirichlet allocation (LDA) together to learn the representations for image classification. In spite of the good performance, the model suffers from computational challenges in the following two main aspects: the model is learned by variational approximations which involve complex computations of the derivatives of the probabilities, and the classifiers are optimized by solving an optimization problem similar to a multi-class SVM problem, both of which are time-consuming especially when the number of classes is large.

Recently, Zhu *et al*. [31] propose a Gibbs max-margin topic model to combine Gibbs classifiers with LDA and show successful results for document analysis. Inspired by [31] which has proved the computational efficiency of minimizing the expected loss function (for the linearity of expectation operation), we employ Gibbs classifiers in our M^{3}HBM model to form the classification part and minimize the expected loss function based on the max-margin principle. For action recognition, M^{3}HBM learns from multiple feature modalities by multi-task learning, instead of only the word-frequency features in [31]. Furthermore, rather than assuming text words to be independent, we jointly models high-level relations including both context of feature words and relations between latent STPs. In addition, M^{3}HBM performs inference similar to [4] and directly employs learned statistics of training data together with test data for inference without estimation. Finally, we also compare the two models in our experimental results which shows an advantage of M^{3}HBM for action recognition.

3. Multi-Feature Max-Margin Hierarchical Bayesian Model (M^{3}HBM)

Given a collection of videos, the proposed M^{3}HBM jointly learns a probabilistic distribution over latent STPs for representation together with max-margin classifiers for recognition.

In the following, we first give the low-level representations to form the observations for the proposed model. Then a multi-feature hierarchical generative model (HGM) is presented to learn latent STPs as high-level representations based on the observations and then the simplified generative process with max-margin classifiers as known is illustrated. Afterwards we give the details of the max-margin classifiers and the multi-task learning method to jointly deal with multiple feature modalities and multiple action classes in a unified framework. In the end, we derive an efficient Gibbs sampling algorithm for the proposed M^{3}HBM.



(a). A frame in video (b).Feature points (c).The Gaussian Context

Figure 1. The low level context information in our model.

3.1. Point-Level Visual Observations

In this paper, we employ two feature modalities, but it is easy to extend the following process and the proposed model to three or more types of feature modalities.

Specifically, we extract two complementary types of features to generate the point-level observations. The first type is employed to describe the appearance characteristics of actions by drawing 3D SIFT descriptors [23] based on sparse spatial-temporal interest points [8]. Compared with the traditional 2D SIFT descriptors, the 3D SIFT descriptors are able to capture the appearance variations in both the spatial and temporal domains. However, interest points are always too sparse to capture the details of motions. Therefore, for the second type, we sample densely in each frame on multiple scales and employ the motion boundary histogram (MBH) descriptors [25] to capture the motion information. By using the relative difference between optical flows, MBH descriptors are effective to handle background movements and scale changes.

Then we quantize the sparse appearance features to generate a V_s-word codebook and the dense motion features to generate a V_t-word codebook respectively. The feature points are described by the nearest cluster index to form visual words $\{\boldsymbol{w}\}$. In addition, we keep the continuous 3D position coordinates $\{\boldsymbol{x}\}$ of each word and use them to restrict the mid-level STPs in different small neighborhoods. As shown in Figure 1, by introducing the position information $\{\boldsymbol{x}\}$, the mid-level STP corresponding to each word in M^{3}HBM is not only decided by the appearance and motion of the current word, but also influenced by the neighbour words, which will be reflected by the Gaussian distributions and the learning process as shown in the following subsections. The class label y of each clip is also introduced to the training process to performs parameter estimation.

3.2. The Multi-Feature Hierarchical Generative Model

The goal of HGM is to learn latent STPs from multiple feature modalities and represent actions by probabilistic distributions over the learned STPs. The latent STPs are shared among different action classes, but different actions have different priorities on each latent STP. Our HGM is based on latent Dirichlet allocation (LDA) [1] which is proposed to discover latent topics in documents. In the following, we will use the similar terminologies as that in document analysis and give the generative process of HGM in the following with the assumption that the parameters of max-margin classifiers, denoted by $\{\boldsymbol{\eta}, \boldsymbol{\lambda}\}$, are fixed to be known values in this subsection.

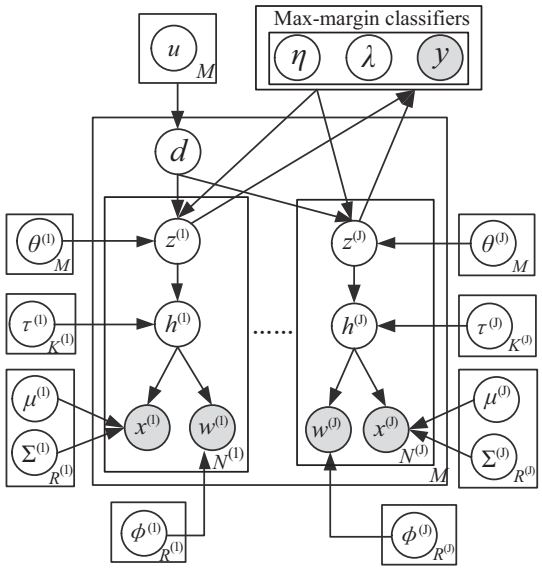


Figure 2. Graphical model representation of the proposed M^{3}HBM. We have omitted the hyper-parameters $\alpha^{(j)}$, $\beta^{(j)}$, $\gamma^{(j)}$, $\xi^{(j)}$, $\mu^{(\eta)}$ and $\Sigma^{(\eta)}$ to simplify the graphical model.

We divide each video into several overlapped temporal clips and treat each clip as an individual document d. As shown in Figure 2, we model the actions in a clip with three layers: point-level visual observations $\{\boldsymbol{w} = \{\boldsymbol{w}^{(j)}\}, \boldsymbol{x} = \{\boldsymbol{x}^{(j)}\}\}$, mid-level local STPs $\boldsymbol{h} = \{\boldsymbol{h}^{(j)}\}$ which are centered at several certain locations and top-level global STPs $\boldsymbol{z} = \{\boldsymbol{z}^{(j)}\}$ which are shared among different actions, where $j = 1, ..., J$ and J is the number of different feature modalities. In the same type of feature modality, the local STPs $\boldsymbol{h}$ are modeled as the probabilistic distributions over visual observations $\{\boldsymbol{w}, \boldsymbol{x}\}$, and the global STPs $\boldsymbol{z}$ are modeled as mixture distributions over local STPs .

Given the parameters $\{\boldsymbol{\eta}, \boldsymbol{\lambda}\}$ of the max-margin classifiers, the generative process to produce observations $\{\boldsymbol{w}^{(j)}, \boldsymbol{x}^{(j)}\}$ in the j-th feature modality for a collection of clips $\boldsymbol{D} = \{d, d = 1, ..., M\}$ is defined as follows. Suppose a clip d with class label y_d has $N_d^{(j)}$ visual words $\{w_{d,n}^{(j)}, x_{d,n}^{(j)}\}$ $(n = 1, ..., N_d^{(j)})$ in the j-th feature modality. Each clip d is assumed to be drawn from an Uniform distribution $Uniform(M)$ so that each clip is selected without priorities but with equal chance; Within each clip, $N_d^{(j)}$ global STPs $z_{d,n}^{(j)}$ $(1 \le n \le N_d^{(j)})$ are drawn independently from a multinomial distribution $p(z_{d,n}^{(j)} = k | d, \boldsymbol{\eta}, y_d, \theta^{(j)})$ $(1 \le k \le K^{(j)})$ conditioned on clip d, the class label y_d, the classifier parameters $\boldsymbol{\eta}$ and the distribution parameter $\theta^{(j)}$, where $K^{(j)}$ is the number of global STPs; Then

$N_d^{(j)}$ local STPs $h_{d,n}^{(j)}$ are drawn from a multinomial distribution $p(h_{d,n}^{(j)} = r|z_{d,n}^{(j)}, \tau^{(j)})$ $(1 \leq r \leq R^{(j)})$, conditioned on the n-th global STP $z_{d,n}^{(j)}$ and the parameter $\tau^{(j)}$, where $R^{(j)}$ is the number of local STPs; Afterwards, $N_d^{(j)}$ visual words $w_{d,n}^{(j)}$ are drawn from a multinomial distribution $p(w_{d,n}^{(j)}|h_{d,n}^{(j)}, \phi^{(j)})$ conditioned on the n-th local STP $h_{d,n}^{(j)}$ and the parameter $\phi^{(j)}$; Finally, the position $x_{d,n}^{(j)}$ is drawn from a Gaussian distribution $p(x_{d,n}^{(j)}|h_{d,n}^{(j)}, \mu^{(j)}, \Sigma^{(j)})$ conditioned on the n-th local STP $h_{d,n}^{(j)}$ and parameters $\{\mu^{(j)}, \Sigma^{(j)}\}$.

For computational efficiency, we assume Dirichlet distributions with hyper-parameters $\{\alpha^{(j)}, \gamma^{(j)}, \beta^{(j)}\}$ as the conjugate priors over $\{\theta^{(j)}, \tau^{(j)}, \phi^{(j)}\}$. In addition, we assume Normal-inverse-Wishart distribution $\mathcal{NIW}(\mu^{(j)}, \Sigma^{(j)}|m_0^{(j)}, \kappa_0^{(j)}, \nu_0^{(j)}, S_0^{(j)})$ as the conjugate prior over the Gaussian distribution parameters $\{\mu^{(j)}, \Sigma^{(j)}\}$ and use symbol $\xi^{(j)}$ to denote $\{m_0^{(j)}, \kappa_0^{(j)}, \nu_0^{(j)}, S_0^{(j)}\}$ for short. For the j-th feature modality, we summarize the complete generative model as:

$$p(\theta_d^{(j)}|\alpha^{(j)}) = Dir(\theta_d^{(j)}; \alpha^{(j)}), \quad d = 1, 2, ..., M; \tag{1}$$

$$p(\tau_k^{(j)}|\gamma^{(j)}) = Dir(\tau_k^{(j)}; \gamma^{(j)}), \quad k = 1, 2, ..., K^{(j)}; \tag{2}$$

$$p(\phi_r^{(j)}|\beta^{(j)}) = Dir(\phi_r^{(j)}; \beta^{(j)}), \quad r = 1, 2, ..., R^{(j)}; \tag{3}$$

$$\begin{aligned} p(\mu_r^{(j)}, \Sigma_r^{(j)}|\xi^{(j)}) &= \mathcal{NIW}(\mu_r^{(j)}, \Sigma_r^{(j)}|\xi^{(j)}) \\ &= \mathcal{N}(\mu_r^{(j)}|\nu_0^{(j)}, \Sigma_r^{(j)})\mathcal{IW}(\Sigma_r^{(j)}|\kappa_0^{(j)}, S_0^{(j)}); \end{aligned} \tag{4}$$

$$p(z_{d,n}^{(j)}|\theta^{(j)}, \boldsymbol{D}, \boldsymbol{\eta}, \boldsymbol{y}) = p(z_{d,n}^{(j)}|\boldsymbol{\eta}, \boldsymbol{y}) \cdot Mult(\theta_d^{(j)}); \tag{5}$$

$$p(h_{d,n}^{(j)}|\tau^{(j)}, z_{d,n}^{(j)} = k) = Mult(\tau_k^{(j)}); \tag{6}$$

$$p(w_{d,n}^{(j)}|h_{d,n}^{(j)} = r, \phi^{(j)}) = Mult(\phi_r^{(j)}); \tag{7}$$

$$p(x_{d,n}^{(j)}|h_{d,n}^{(j)} = r, \mu^{(j)}, \Sigma^{(j)}) = \mathcal{N}(\mu_r^{(j)}, \Sigma_r^{(j)}). \tag{8}$$

Given the generative process above, the joint distribution of HGM with fixed classifiers parameters $\boldsymbol{\eta}$ and class labels $\boldsymbol{y}$ is

$$\begin{aligned} &p(\boldsymbol{z}, \boldsymbol{h}, \boldsymbol{w}, \boldsymbol{x}, d, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{\eta}, \boldsymbol{y}, \alpha, \beta, \gamma, \xi) \\ &= \prod_{j=1}^{J} \Bigg[\Bigg(\prod_{d=1}^{M} p(\theta_d^{(j)}|\alpha^{(j)}) \prod_{k=1}^{K^{(j)}} p(\tau_k^{(j)}|\gamma^{(j)}) \\ &\cdot \prod_{r=1}^{R^{(j)}} p(\phi_r^{(j)}|\beta^{(j)}) p(\mu_r^{(j)}, \Sigma_r^{(j)}|\xi^{(j)}) \Bigg) \\ &\cdot \Bigg(\prod_{d=1}^{M} \prod_{n=1}^{N_d^{(j)}} p(z_{d,n}^{(j)}|\theta_d^{(j)}, \boldsymbol{\eta}, y_d) p(h_{d,n}^{(j)}|\tau^{(j)}, z_{d,n}^{(j)}) \\ &\cdot p(w_{d,n}^{(j)}|h_{d,n}^{(j)}, \phi^{(j)}) p(x_{d,n}^{(j)}|h_{d,n}^{(j)}, \mu^{(j)}, \Sigma^{(j)}) \Bigg) \Bigg]. \end{aligned} \tag{9}$$

3.3. Multi-Feature Max-Margin Hierarchical Bayesian Model

To make the learned STPs more discriminative for classification, we introduce max-margin classifiers to learn the representations. In particular, we employ Gibbs classifiers [31] to minimize the expected margin-based classification loss and assume Gaussian priors over the classifiers to perform parameter estimation together with the learning process of HGM. In addition, we apply the multi-task learning method to learn the shared latent STPs from multiple feature modalities and estimate the parameters for different classes. The classifiers and HGM are coupled via the latent global STPs $\boldsymbol{z}$, which makes the representations in our model both descriptive and predictive for action recognition.

As shown in Figure 3, learning the classifier parameters from each feature modality j within each action class l is viewed as a single task and there are $I = L \cdot J$ tasks in total where L is the number of action classes.

We define the linear discriminant function of task i $(1 \leq i \leq I)$ as

$$F_i(\boldsymbol{\eta}_i, \boldsymbol{z}_i; \boldsymbol{w}_i, \boldsymbol{x}_i) = \boldsymbol{\eta}_i^T \overline{\boldsymbol{z}}_i, \tag{10}$$

where $\boldsymbol{\eta}_i$ are parameters, $\{\boldsymbol{w}_i, \boldsymbol{x}_i\}$ are observations, and $\overline{\boldsymbol{z}}_i$ are probabilistic representations over the global STPs for task i which can be defined as the learned parameter θ or the average frequency of each global STP in the clip.

The prediction rule of task i for the training process is defined as

$$\hat{y}^i(\eta_i, z_i) = \text{sign } F(\boldsymbol{\eta}_i, \boldsymbol{z}_i; \boldsymbol{w}_i, \boldsymbol{x}_i), \tag{11}$$

where sign$(\cdot)$ is the sign function.

The hinge loss of task i is defined as

$$\mathcal{R}(\boldsymbol{\eta}_i, \boldsymbol{z}_i) = \sum_{d=1}^{M} \max(0, T - y_d^i \boldsymbol{\eta}_i^T \overline{z}_{d,i}), \tag{12}$$

where T is the threshold for margin values and

$$y_d^i = \begin{cases} 1, & if \;\; y_d = l_i, \text{ where } l_i \text{ is the label of task } i; \\ -1, & otherwise. \end{cases} \tag{13}$$

To transform the classifiers into probabilistic form, we consider $\boldsymbol{\eta}$ as random variables and minimize the expected hinge loss over the joint distribution $p(\boldsymbol{\eta}, \boldsymbol{z})$ as

$$\mathcal{R}'(\boldsymbol{\eta}_i, \boldsymbol{z}_i) = E_{p(\boldsymbol{\eta}, \boldsymbol{z})}[\mathcal{R}(\boldsymbol{\eta}_i, \boldsymbol{z}_i)]. \tag{14}$$

To perform Bayesian estimation for $\boldsymbol{\eta}$, we introduce augmented variables $\boldsymbol{\lambda}$ [19] to express the max function in Eq.12 as

$$\begin{aligned} \varphi_i(y_d^i|\boldsymbol{z}_d, \boldsymbol{\eta}) &= \exp\left(-2c \max(0, T - y_d^i \boldsymbol{\eta}_i^T \overline{z}_d^i)\right) \\ &= \int_0^\infty \mathcal{N}(c\zeta_d^i| - \lambda_d^i, \lambda_d^i) d\lambda_d^i, \end{aligned} \tag{15}$$

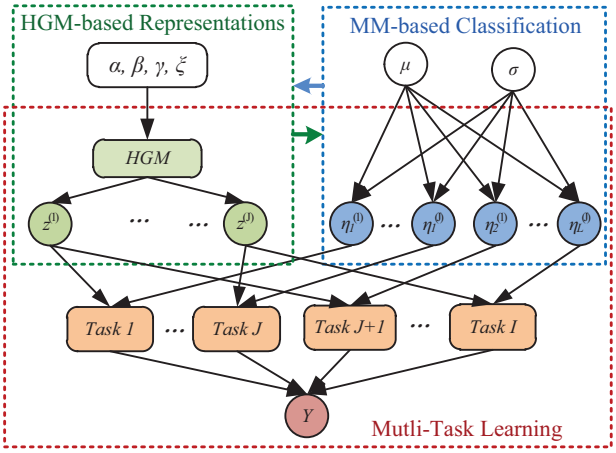


Figure 3. The learning framework of M^{3}HBM in which a unified Bayesian framework is employed to combine the proposed HGM with the max-margin classifier together. The green box shows the HGM for action representation and the blue box illustrates the max-margin component for classification. We employ multi-task learning to integrate the two parts to simultaneously learn the representations and classifiers from multiple feature modalities for multiple classes, as shown by the red box. The green and blue arrows in the figure are used to denote that the representations and classifiers are interacted with each other in our model. (Best viewed in color.)

where $\zeta_d^i = T - y_d^i \eta_i^T \overline{z}_{d,i}$ and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. Considering the conjugacy property, we assume Gaussian priors over $\boldsymbol{\eta}$ as

$$p_0(\boldsymbol{\eta}) = \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{\eta}_i; \boldsymbol{\mu}_0^{(\eta)}, \boldsymbol{\sigma}_0^{(\eta)}), \tag{16}$$

where $\boldsymbol{\mu}_0^{(\eta)}$ and $\boldsymbol{\sigma}_0^{(\eta)}$ are the mean and variance of the prior respectively.

By combining the generative process of HGM and the distributions over classifiers, we can get the posterior of M^{3}HBM as

$$p(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{h}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}) = \frac{p_0(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{h}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x} | \boldsymbol{z}, \boldsymbol{h}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathcal{Z}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x})}, \tag{17}$$

where $\mathcal{Z}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x})$ is the normalized constant and independent of the parameters and variables.

3.4. Model Learning

By taking advantage of conjugacy, we integrate out the multinomial parameters $\{\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\phi}\}$ and the Gaussian parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$, and then obtain the collapsed posterior distribution as

$$p(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{h} | \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}) \propto p_0(\boldsymbol{\eta}) \varphi(\boldsymbol{y}, \boldsymbol{\lambda} | \boldsymbol{z}, \boldsymbol{\eta}) p(\boldsymbol{z}, \boldsymbol{h}, \boldsymbol{w}, \boldsymbol{x}). \tag{18}$$

For computational convenience, we assume $\boldsymbol{\mu}_0^{(\eta)} = 0$ and $\boldsymbol{\sigma}_0^{(\eta)} = \sigma_i^2$ for $p_0(\boldsymbol{\eta}_i)$ and derive the Gibbs sampling process for learning M^{3}HBM as follows.

Sampling $h_{d,n}^{(j)}$:

$$p(h_{d,n}^{(j)} = r | \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{w}, \gamma, \xi) \propto \frac{\gamma + n_{k,r}^{(j)-}}{R\gamma + n_{k,\cdot}^{(j)-}} \cdot \frac{\beta + n_{r,\mathrm{w}}^{(j)-}}{V^{(j)}\beta + n_{r,\cdot}^{(j)-}} \cdot G_{\boldsymbol{x}}^{(j)}. \tag{19}$$

where $n_{k,r}^{(j)}$ is the number of words with global STP k assigned to local STP r in the j-th modality, $n_{r,\mathrm{w}}^{(j)}$ is the number of words with word value w assigned to local STP r in the j-th modality, the symbol $\cdot$ in $n_{k,\cdot}^{(j)}$ and $n_{r,\cdot}^{(j)}$ means summation over the local STPs and the visual words respectively. The superscript $^-$ denotes the count excluding the n-th word. $G_{\boldsymbol{x}}^{(j)}$ is decided by context $\boldsymbol{x}^{(j)}$ and can be simplified by using Stirling's approximations to obtain

$$G_{\boldsymbol{x}}^{(j)} \propto \frac{(\kappa_{N-}^{(j)} + 1)^{\frac{-\rho_x^{(j)}}{2}}}{(\kappa_{N-}^{(j)})^{-\rho_x^{(j)}/2}} \cdot \frac{\|S_{N-+1}^{(j)}\|^{-\frac{\nu_{N-}^{(j)}+1}{2}}}{\|S_{N-}^{(j)}\|^{-\frac{\nu_{N-}^{(j)}}{2}}} \cdot \prod_{a=1}^{\rho_x^{(j)}} \sqrt{\nu_{N-}^{(j)} + 1 - a}, \tag{20}$$

where $\rho_x^{(j)}$ is the dimensionality of $\boldsymbol{x}^{(j)}$ and

$$\begin{aligned} \kappa_N &= \kappa_0 + n_{r,\cdot}^{(j)-}; \\ \nu_N &= \nu_0 + n_{r,\cdot}^{(j)-}; \\ m_N &= \frac{\kappa_0 m_0 + n_{r,\cdot}^{(j)-} \overline{x}}{\kappa_N}; \\ S_N &= S_0 + \kappa_0 m_0 m_0^T + \sum_{n=1}^{n_{r,\cdot}^{(j)-}} x_n x_n^T - \kappa_N m_N m_N^T. \end{aligned} \tag{21}$$

Sampling $z_{d,n}^{(j)}$:

$$p(z_{d,n}^{(j)} = k | \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{w},) \propto \frac{\alpha + n_{d,k}^{(j)-}}{K\alpha + n_{d,\cdot}^{(j)}} \cdot \frac{\gamma + n_{k,r}^{(j)-}}{R\gamma + n_{k,\cdot}^{(j)}} \cdot \prod_{i=1}^{L} \exp\left(\frac{(y_d^i c(cT + \lambda_d^i) - \frac{c^2}{2N_d^{(j)}}\eta_{ik} - \frac{2N_d^{(j)}-2}{N_d^{(j)}}\Lambda_{dn}^i)\eta_{ik}}{N_d^{(j)}\lambda_d^i} \right), \tag{22}$$

where $\Lambda_{dn}^i = \frac{1}{N_d - 1} \sum_{k=1}^{K} \eta_{ik} n_{d,k}^{(j)-}$ and $n_{d,k}^{(j)-}$ is the number of words assigned to global STP k in the j-th feature modality in document d. From the above we can see that, the value of each STP $z_{d,n}^{(j)}$ is influenced by each task i, which means that all the feature modalities are influenced by each other.

Sampling λ_d^i:

$$p(\lambda_d^i|\boldsymbol{z}_{d,i},\boldsymbol{\eta}) \propto \frac{1}{\sqrt{2\pi\lambda_d^i}}\exp\left(-\frac{(\lambda_d^i+c\zeta_d^i)^2}{2\lambda_d^i}\right) = \mathcal{GIG}(\lambda_d^i;\frac{1}{2},1,c^2(\zeta_d^i)^2), \tag{23}$$

Where $\mathcal{GIG}(x;p,a,b)$ is a generalized inverse Gaussian distribution [3].

Sampling $\boldsymbol{\eta}_i$:

$$p(\boldsymbol{\eta}|\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{h}) = \prod_{i=1}^{I} p(\boldsymbol{\eta}_i|\boldsymbol{z}_i,\boldsymbol{\lambda}_i) = \prod_{i=1}^{I}\mathcal{N}(\boldsymbol{\eta}_i;\boldsymbol{\mu}_i^{(\eta)},\boldsymbol{\Sigma}_i^{(\eta)}), \tag{24}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_i^{(\eta)-1} &= \frac{1}{\sigma_i^2}I + c^2\sum_{d=1}^{N}\frac{\overline{\boldsymbol{z}}_{d,i}\overline{\boldsymbol{z}}_{d,i}^T}{\lambda_d^i},\\ \boldsymbol{\mu}_i^{(\eta)} &= \boldsymbol{\Sigma}_i^{(\eta)}(c\sum_{d=1}^{D} y_d^i\frac{\lambda_d^i+cT}{\lambda_d^i}\overline{\boldsymbol{z}}_{d,i}).\end{aligned} \tag{25}$$

3.5. Inference and Recognition

The goal of this step is to infer the latent STPs $\boldsymbol{z}^{(j)}$ and $\boldsymbol{h}^{(j)}$ in each feature modality for test clip d with observations $\{\boldsymbol{w},\boldsymbol{x}\}$ and then classify the actions with the learned classifier parameters $\boldsymbol{\eta}$. We perform inference by

$$p(\tilde{\boldsymbol{z}}_{dn}^{(j)}=k|\tilde{\boldsymbol{w}},\mathcal{D}) \propto \frac{\alpha_k+\tilde{n}_{d,k}}{K\alpha_k+\tilde{n}_d}\frac{\gamma+\tilde{n}_{k,r}+n_{k,r}}{R\gamma+\tilde{n}_k+n_k}, \tag{26}$$

$$p(\tilde{h}_{dn}^{(j)}=r|\tilde{\boldsymbol{w}},\mathcal{D}) \propto \frac{\gamma+\tilde{n}_{k,r}+n_{k,r}}{R\gamma+\tilde{n}_k+n_k}\frac{\beta+\tilde{n}_{r,w}+n_{r,w}}{V\beta+\tilde{n}_r+n_r}G_{\tilde{x}}^{(j)}, \tag{27}$$

where the symbol $\tilde{}$ is used to denote the variables in test clip and $G_{\tilde{x}}^{(j)}$ is computed as the gaussian probability $p(\tilde{x}|\mu_r^{(j)},\Sigma_r^{(j)})$ of $\tilde{x}$ in new document with trained parameters $\mu_r^{(j)}=m_N$ and $\Sigma_r^{(j)}=S_N$ for simplicity.

Having obtained the global STPs $\boldsymbol{z}$, we compute the predict score of each clip for each class by Eq.10 and then perform maximum majority voting on these prediction scores to obtain the final class label of the video which is composed by several clips.

4. Experiments

In this section, we evaluate our approach on two public datasets, one benchmark dataset and one more challenging dataset. Firstly, we test on the benchmark KTH action datasets [22]: There are six human action classes performed by 25 subjects in four different scenarios. We perform five-fold cross-validation on this dataset where sequences of 20 persons are used for training and the other 5 persons are for test each time. Secondly, we test on a much more challenging UCF Sports action dataset [20]: There are 150 sports action videos from 10 classes in total and most of the videos have severe camera motions and viewpoints variations. We extend the dataset by mirror-flipping the videos as [6, 26] and perform five-fold cross validation on this dataset. We test several aspects of our model and compare with several related methods on both the two datasets to show the effectiveness of our model.

Figure 4. Some sample frames in each datasset.

4.1. Experimental Settings

For the sparse interest points based 3D SIFT features, the spatial and temporal scale parameters are empirically fixed to 2 respectively. The size of the cuboid is set to be $6\times6\times3$ for all the datasets. For the dense sampling based MBH descriptors, we sample on a grid spaced by 5 pixels and use 8 spatial scales spaced by a factor of $1/\sqrt{2}$ as in [25]. For all the datasets, we set $K^{(j)}$ and $R^{(j)}$ of each feature modality j as 100 and 400 respectively.

Experiments were conducted in two groups to verify the representation power and the recognition capability of our approach respectively. In the first group, we evaluate the representation performance of M^{3}HBM by combining HGM with linear SVM classifiers for action recognition and compare with related LDA based methods using exactly the same features. In the second group, we demonstrate the classification capability of the proposed M^{3}HBM for action recognition through two groups of comparison: (1). we compare M^{3}HBM with the HGM+SVM method which shares the same generative part and similar classifiers to M^{3}HBM but performs representation and classification in separate steps. (2). we compare with other related methods which use similar hierarchical structures and similar features for action recognition.

4.2. Evaluation of HGM for Action Representation

In this part, we test the representation capability of HGM by combining it with linear SVM classifiers for action recognition. We compare the representations obtained in several settings, denoted by *LDA+S*, *LDA+M*, *LDA+F*, *HGM+S*, *HGM+M*, and *HGM+F* respectively. We use symbols "S", "M" and "F" to denote the representations obtained from only SIFT descriptors, only MBH descriptors and the

Methods	Fold1	Fold2	Fold3	Fold4	Fold5	Average
LDA+S	83.19	88.33	**94.12**	96.67	93.33	91.14
LDA+M	87.39	95.00	90.76	91.67	91.67	91.30
LDA+F	**90.76**	95.83	90.76	96.67	94.17	93.65
HGM+S	**90.76**	95.83	93.28	96.67	95.00	94.31
HGM+M	**90.76**	94.17	92.44	96.67	96.67	94.15
HGM+F	89.92	**96.67**	**94.12**	**99.17**	**97.50**	**95.48**

Table 1. Comparison results by using HGM and LDA for representation and SVM for classification on the KTH dataset (%).

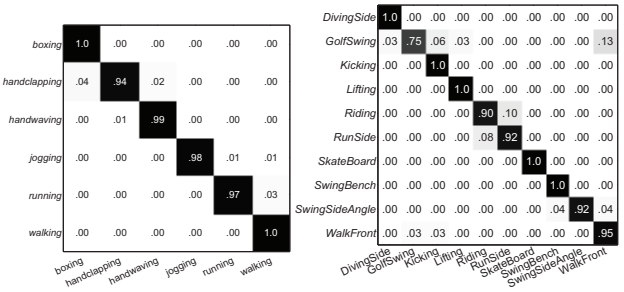
Method	Fold1	Fold2	Fold3	Fold4	Fold5	Average
LDA+S	66.67	70.00	83.33	80.00	78.33	75.67
LDA+M	93.33	81.67	83.33	85.00	76.67	84.00
LDA+F	81.67	81.67	86.67	88.33	91.67	86.00
HGM+S	81.67	78.33	86.67	85.00	88.33	84.00
HGM+M	91.67	91.67	90.00	95.00	90.00	91.67
HGM+F	**95.00**	**93.33**	**91.67**	**95.00**	**93.33**	**93.67**

Table 2. Comparison results by using HGM and LDA for representation and SVM for classification on the UCF Sports action dataset (%).

concatenation of SIFT and MBH descriptors respectively. Here HGM is implemented by excluding the influence of the class labels y and the max-margin classifiers $\boldsymbol{\eta}$ in the sampling process, which makes it an unsupervised model similar to LDA. We implement with the same parameters for LDA and HGM, including the hyper-parameters, the convergence principle, the generation method to get documents, and so on. We use the average frequency of each global STP z in each clip as the representation generated by HGM and the average frequency of topics generated by LDA in each clip as the representation to input to the following SVM classifiers.

The comparison results on the two action datasets are shown in Table 1 and Table 2 respectively. We can see that HGM outperforms LDA on both datasets, which clearly shows the representation power of our HGM for action recognition. In particular, on the KTH dataset, HGM outperforms LDA with an average accuracy of 95.48% with the peak accuracy of 99.17% in Fold 3 when using both the SIFT features and the MBH descriptors.

Furthermore, HGM also shows a stable performance on each fold with high accuracy. For example, when using only the SIFT features, the divergence between the peak and lowest accuracy in the five folds of HGM on the KTH dataset is no more than 6% which is lower than a half of the divergence of 13.48% obtained by LDA. On the other hand, the divergence between the five folds of HGM on the UCF Sports action dataset is only 5% when using only MBH descriptors and it is only about one third of the divergence of 16.66% obtained by LDA.



(a) KTH dataset: 97.99% (b) UCF Sports dataset: 94.07%

Figure 5. Confusion tables on the KTH dataset and the UCF Sports action dataset.

In short, the average recognition accuracy of our HGM is much higher than LDA on both datasets which verifies the advantages of our HGM for describing actions. In addition, the comparison results also proves the stability of our HGM for action representation.

4.3. Evaluatation of M^{3}HBM for Action Recognition

For action recognition, we compare M^{3}HBM with other related methods in two groups. In the first group, we compare the recognition performance of M^{3}HBM with HGM+SVM on the two action datasets and demonstrate the benefits of jointly learning the representations and classifiers in a unified process. In the second group, we compare M^{3}HBM with other similar methods which also employ hierarchical structures, including deep structures in [10], to build representations and use similar features for action recognition.

Figure 5 shows the recognition confusion tables of M^{3}HBM on the KTH dataset and the UCF Sports action dataset. In general, we can see that M^{3}HBM achieves an accuracy of 100% on 2 classes in the total 6 classes on KTH dataset and the lowest accuracy in these classes is as high as 94.00%. On the UCF Sports action dataset, we can see that M^{3}HBM achieves an accuracy of 100% on as many as 5 classes in the total 10 classes. Moreover, we achieve an accuracy higher than 90% in most of the total 10 classes, which proves the stability of our approach for recognition.

In the following, Table 3 compares M^{3}HBM with HGM+SVM using exactly the same features and representation model HGM, and the difference only lies on that the representations are learned separately or jointly with the classifier part. It is clear to see that M^{3}HBM outperforms HGM+SVM on both datasets with a significant improvement in general, which proves the effectiveness of the joint learning framework of M^{3}HBM for action recognition. In particular, M^{3}HBM achieves an accuracy of 97.99% on the KTH

	KTH dataset		UCF Sports dataset	
	HGM+SVM	M^{3}HBM	HGM+SVM	M^{3}HBM
Fold1	89.92	96.64	95.00	95.00
Fold2	96.67	98.33	93.33	93.33
Fold3	94.12	97.48	91.67	93.33
Fold4	99.17	99.17	95.00	93.33
Fold5	97.50	98.33	93.33	96.67
Average	95.48	**97.99**	93.67	**94.07**

Table 3. Recognition results on the KTH and UCF Sports action dataset (%).

dataset, which is higher than HGM+SVM by more than 2%, and 94.07% on the UCF Sports action dataset respectively. Moreover, the divergence of the recognition accuracy obtained by HGM+SVM on the two datasets achieves 9.25% and 3.33% respectively, while the divergence of M^{3}HBM are less than 3% on both datasets, which clearly show the advantages of the proposed M^{3}HBM for action recognition. Because the comparisons are performed on exactly the same features and mainly differently on the learning framework, the results demonstrate that M^{3}HBM is benefit from the joint learning framework of representation and classifiers. Taken together, Table 1, 2 and Table 3 show that the proposed M^{3}HBM outperform the baseline LDA+SVM on both datasets by a large margin and present a clear improvement for action recognition.

In the next part, we compare with other related methods in Table 4, including both discriminative methods and graphical models based methods which employ hierarchical structures and similar features for action recognition. In [25], Wang combine the MBH descriptors and other appearance features to represent actions, which is similar to our point-level representation, but use a non-linear SVM with a χ^2 kernel for classification. As a popular technique, deep architectures are receiving more and more attention. Among the method using deep architectures for action recognition, [10] is an outstanding work which uses a high-level representation obtained from spatial temporal features for recognition. As the results in Table 4 shown, our M^{3}HBM outperforms most related methods and achieves a relatively higher performance on both datasets, especially on the UCF Sports action dataset which is much more challenging than KTH dataset. We also compare with other similar graphical model based methods in Table 4, like [13, 16] and [27], and the results show the advantages of our method on both datasets.

5. Conclusions

This paper has proposed a multi-feature max-margin hierarchical Bayesian model (M^{3}HBM) for action recognition. A three-layer hierarchical generative model (HGM) is constructed to learn a high-level representation based on latent spatial temporal patterns (STPs) which are learned from multiple feature modalities and shared among all classes. We introduce Gibbs classifiers into M^{3}HBM and employ Gaussian priors to learn the classifiers in a Bayesian framework together with the learning process of STPs. In addition, multi-task learning is introduced into our model to learn the latent STPs and the classifiers from multiple feature modalities for different classes. An efficient Gibbs sampling algorithm is derived for both the learning and inference process of M^{3}HBM. Experiments have shown the advantages of our methods for representation and have also demonstrated the effectiveness of M^{3}HBM for action recognition. Several future work may be developed in view of the following appealing properties of our model. Firstly, it is easy to extend our model to three or more feature modalities to enrich the representations. In addition, our model is also suitable for many other applications, such as image classification. Last but not least, it is also interesting to integrate many powerful classifiers, other than max-margin classifiers, into the proposed M^{3}HBM to improve the recognition performance.

	KTH dataset	UCF Sports dataset
Wang *et al*. [25]	94.2	88.2
O'Hara *et al*. [18]	97.9	91.3
Le *et al*. [10]	93.9	86.5
Kovashka *et al*. [6]	94.53	85.49
Malgireddy *et al*. [13]	94.6	-
Niebles *et al*. [16]	83.33	-
Wang and Mori [27]	91.20	-
M^{3}HBM	**97.99**	**94.07**

Table 4. Comparison of recognition accuracy on the KTH dataset and the UCF Sports action dataset (%).

6. Acknowledgements

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), NSFC (Grant No. 61472420, 61303086, 61472227, 61202327),and the project supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage. Action recognition by learning discriminative key poses. In *ICCV Workshops*, pages 1302–1309, 2011.

[3] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[4] T. Hofmann. Probabilistic latent semantic indexing. SIGIR '99, pages 50–57, 1999.

[5] T. M. Hospedales, S. Gong, and T. Xiang. Video behaviour mining using a dynamic topic model. *IJCV*, 98(3):303–323, 2012.

[6] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, pages 2046–2053, 2010.

[7] S. Lacoste-julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, pages 897–904. 2008.

[8] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[9] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, June 2011.

[10] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, 2011.

[11] L. Li, R. Socher, and F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, pages 2036–2043, 2009.

[12] D. Lin and J. Xiao. Characterizing layouts of outdoor scenes using spatial topic processes. In *ICCV*, pages 841–848, 2013.

[13] M. Malgireddy, I. Inwogu, and V. Govindaraju. A temporal bayesian model for classifying, detecting and localizing activities in video sequences. In *CVPR Workshp*, pages 43–48, 2012.

[14] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou. Multi-modal image annotation with multi-instance multi-label lda. In *IJCAI*, pages 1558–1564, 2013.

[15] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, pages 1–8, 2007.

[16] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.

[17] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *CVPR*, pages 2743–2750, 2012.

[18] S. O'Hara and B. Draper. Scalable action recognition with a subspace forest. In *CVPR*, pages 1210–1217, 2012.

[19] N. G. Polson and S. L. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

[20] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008.

[21] R. Salakhutdinov and G. E. Hinton. Replicated softmax: an undirected topic model. In *NIPS*, pages 1607–1614, 2009.

[22] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, pages 32–36, 2004.

[23] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. ACM Multimedia*, pages 357–360, 2007.

[24] Y. Song, L.-P. Morency, and R. Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, pages 3562–3569, 2013.

[25] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[26] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, pages 124.1–124.11, 2009.

[27] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *PAMI*, 31(10):1762–1774, 2009.

[28] Y. Wang and G. Mori. Max-margin latent dirichlet allocation for image classification and annotation. In *BMVC*, pages 1–11, 2011.

[29] F. Zhao, Y. Huang, L. Wang, and T. Tan. Relevance topic model for unstructured social group activity recognition. In *NIPS*, pages 2580–2588. 2013.

[30] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *CVPR*, pages 3441–3448, 2011.

[31] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with data augmentation. *JMLR*, 15(1):1073–1110, 2014.

[32] Y. Zhu, N. Nayak, and A. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *CVPR*, pages 2491–2498, 2013.