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

Shuang Yang[1], Chunfeng Yuan[1], Baoxin Wu[1], Weiming Hu[1], Fangshi Wang[2]
[1] NLPR, Institute of Automation, Chinese Academy of Sciences (CASIA). [2] Beijing Jiaotong University.

Recently, many work propose to unify representation and classification in a single model to make the representations both characteristic and discriminative. Some related outstanding methods include the relevance topic model [5], the max-margin LDA [4] and the Gibbs max-margin topic model [6], which show successful results with the aid of topic models. Inspired by them, this paper proposes a multi-feature max-margin hierarchical Bayesian model (M^{3}HBM) for action recognition and differs from them mainly in the designing principle, the optimization method, the introduction of multiple feature modalities, and the modeling of high-level relations.

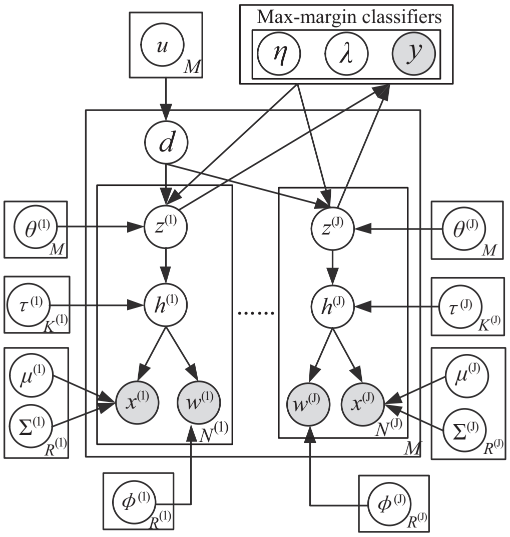

Figure 1: Graphical model representation of the proposed M^{3}HBM.

In this paper, M^{3}HBM learns a high-level representation by combining a hierarchical generative model (HGM) and discriminative max-margin classifiers in a unified Bayesian framework. For the representation part, HGM is proposed to represent actions by distributions over latent spatial temporal patterns (STPs). As shown in Fig.1,HGM includes three layers: point-level visual observations $\{\boldsymbol{w},\boldsymbol{x}\}$, region-level local STPs $\boldsymbol{h}$ scattered in many different small regions, and top-level global STPs $\boldsymbol{z}$ which are shared among all the classes without position limitation. For point-level observations, we extract two complementary types of features ($J=2$): the sparse interest points [1] based 3D SIFT descriptors [2] and the dense sampling based MBH descriptors [3]. For computational efficiency, we assume conjugate priors and summarize the HGM as follows:

$$
\begin{aligned}
p(\theta_d^{(j)}|\alpha^{(j)}) &= Dir(\theta_d^{(j)};\alpha^{(j)}), \quad d=1,2,...,M;\\
p(\tau_k^{(j)}|\gamma^{(j)}) &= Dir(\tau_k^{(j)};\gamma^{(j)}), \quad k=1,2,...,K^{(j)};\\
p(\phi_r^{(j)}|\beta^{(j)}) &= Dir(\phi_r^{(j)};\beta^{(j)}), \quad r=1,2,...,R^{(j)};\\
p(\mu_r^{(j)},\Sigma_r^{(j)}|\xi^{(j)}) &= \mathcal{NIW}(\mu_r^{(j)},\Sigma_r^{(j)}|\xi^{(j)})\\
&= \mathcal{N}(\mu_r^{(j)}|\nu_0^{(j)},\Sigma_r^{(j)})\mathcal{IW}(\Sigma_r^{(j)}|\kappa_0^{(j)},S_0^{(j)}); \qquad (1)\\
p(z_{d,n}^{(j)}|\theta^{(j)},\boldsymbol{D},\boldsymbol{\eta},\boldsymbol{y}) &= p(z_{d,n}^{(j)}|\boldsymbol{\eta},\boldsymbol{y})\cdot Mult(\theta_d^{(j)});\\
p(h_{d,n}^{(j)}|\tau^{(j)},z_{d,n}^{(j)}=k) &= Mult(\tau_k^{(j)});\\
p(w_{d,n}^{(j)}|h_{d,n}^{(j)}=r,\phi^{(j)}) &= Mult(\phi_r^{(j)});\\
p(x_{d,n}^{(j)}|h_{d,n}^{(j)}=r,\mu^{(j)},\Sigma^{(j)}) &= \mathcal{N}(\mu_r^{(j)},\Sigma_r^{(j)}).
\end{aligned}
$$

To make the latent STPs both representative and discriminative, we introduce Gibbs classifiers and employ the multi-task learning to jointly learn the representations and classifiers from multiple feature modalities. As shown in Fig.2, learning classifier parameters from each feature modality j within each action class l is viewed as a single task. For each task i, a linear classifier is defined with a Sign function as the prediction rule. The expected loss function is

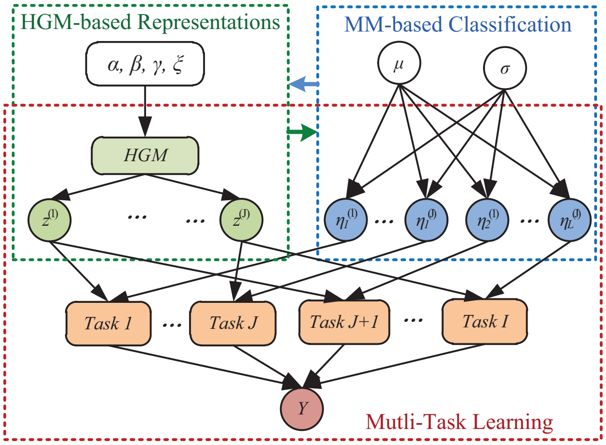

Figure 2: The learning framework of M^{3}HBM. (Best viewed in color.)

$$\mathcal{R}'(\boldsymbol{\eta}_i,\boldsymbol{z}_i) = E_{p(\boldsymbol{\eta},\boldsymbol{z})}[\mathcal{R}(\boldsymbol{\eta}_i,\boldsymbol{z}_i)] = E_{p(\boldsymbol{\eta},\boldsymbol{z})}[\sum_{d=1}^{M}\max(0,T-y_d^i\eta_i^T\bar{z}_{d,i})]. \quad (2)$$

For the loss function, we introduce augmented variables $\boldsymbol{\lambda}$ and get

$$\varphi_i(y_d^i|\boldsymbol{z}_d,\boldsymbol{\eta}) = e^{-2c\max(0,T-y_d^i\boldsymbol{\eta}_i^T\bar{z}_d^i)} = \int_0^\infty \mathcal{N}(c\zeta_d^i|-\lambda_d^i,\lambda_d^i)d\lambda_d^i, \quad (3)$$

where $\zeta_d^i = T-y_d^i\eta_i^T\bar{z}_{d,i}$ and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution.

With Gaussian priors over $\boldsymbol{\eta}$, we get the posterior of M^{3}HBM as

$$p(\boldsymbol{\eta},\boldsymbol{\lambda},z,\boldsymbol{h},\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{y},\boldsymbol{w},\boldsymbol{x}) = \frac{p_0(\boldsymbol{\eta},\boldsymbol{\lambda},z,\boldsymbol{h},\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\Sigma})p(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x}|z,\boldsymbol{h},\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\phi},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathcal{Z}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x})}, \quad (4)$$

where $\mathcal{Z}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x})$ is the partition function.

According to the representations and classifiers given above, we derive the Gibbs sampling algorithm to solve the model. For test videos, we perform inference to obtain the latent STPs $\boldsymbol{z}^{(j)}$ and $\boldsymbol{h}^{(j)}$ in each feature modality with observations $\{\boldsymbol{w},\boldsymbol{x}\}$ and then classify the actions with the learned classifier parameters $\boldsymbol{\eta}$.

The detailed illustrations and implementations are described in the paper, and it is proved to be beneficial to jointly learn the representations and classifiers together. The proposed model is easy to be extended to three or more feature modalities and other applications. It is also valuable to integrate other powerful classifiers to improve the performance.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

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

[2] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proc. ACM Multimedia*, pages 357–360, 2007.

[3] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

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

[5] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. Relevance topic model for unstructured social group activity recognition. In *NIPS*, pages 2580–2588. 2013.

[6] Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *JMLR*, 15(1):1073–1110, 2014.