# HIERARCHICAL LATENT DIRICHLET ALLOCATION MODELS FOR REALISTIC ACTION RECOGNITION

*Heping Li, Jie Liu, Shuwu Zhang*

Hi-tech Innovation Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China

## ABSTRACT

It has always been very difficult to recognize realistic actions from unconstrained videos because there are tremendous variations from camera motion, background clutter, object appearance and so on. In this paper, a Single-Feature Hierarchical Latent Dirichlet Allocation model called SF-HLDA by extending Latent Dirichlet Allocation to the hierarchical one is first proposed for realistic action recognition. And then, by extending SF-HLDA, we present another model called Multi-Feature Hierarchical Latent Dirichlet Allocation model MF-HLDA which can effectively fuse several different features into one model for recognizing the realistic actions. Experiments demonstrate the effectiveness of our proposed models.

***Index Terms***— action recognition, hierarchical Latent Dirichlet Allocation, multi-feature model

## 1. INTRODUCTION

Action recognition is driven by a wide range of applications, such as visual surveillance, advanced user interface, video meeting, behavior based video index and retrieval and so on, and it is an active research topic in video processing and multimedia signal processing. The aim of action recognition is to get semantic descriptions and understanding of dynamic scene by analyzing low-level features. To this end, one key step is to model and recognize actions.

Most of the early existing works on action modeling and recognition are in the limited conditions like those in [1-8]. For example, Bobick and Davis [1] used temporal template called motion-history image to model the behaviors without camera motion. Similar to Bobick and Davis's work, the methods [2-3] used spatial-temporal features to recognize the indoor behaviors. In the works [4-5], 3D features were adopted to recognize the behaviors in the simple scene. And space-time points were used in [6]. Different from the above methods, Mikolajczyk and Uemura [7] used the motion-appearance vocabulary forest to model the actions. And Dhillon et al. [8] combined appearance and motion features for human action classification and tested their algorithm in videos with simple background. Recently, some works [9-11] have been reported on action recognition from unconstrained videos. For example, Laptev et al. [9] modeled realistic human actions movies by extending several recent ideas including local space-time features, space-time pyramids and multi-channel non-linear SVMs. Reliable and informative features including motion and static features were extracted from the unconstrained videos in [10]. And Hu et al. [11] detected actions in complex scenes with spatial and temporal ambiguities.

Different from the above-mentioned methods, Niebles et al. [12] used Latent Dirichlet Allocation (LDA) to model human actions in real scene. LDA, a very simple model, was first described by Blei et al. [13] in detail. And it can handle noisy feature points arisen from dynamic background and moving cameras, and has been applied to challenging computer vision tasks [12]. Based on this model, Hospedale et al. [14] constructed a Markov clustering topic model to mine behaviors in videos with complex background.

In this paper, by extending LDA, we propose two novel Hierarchical Latent Dirichlet Allocation models for recognizing realistic actions from unconstrained videos. The first model is a Single-Feature Hierarchical Latent Dirichlet Allocation model (SF-HLDA) by extending Latent Dirichlet Allocation to the hierarchical one. This model assumes every action video as one document which is a random mixture of document topics and one document topic is a mixture of one type of feature topics. With the hierarchical structure including two topic layers, SF-HLDA can further reduce the effect of the noise led by the tremendous variations from camera motion, background clutter, object appearance and so on, and improve the correct rate of action recognition. By extending SF-HLDA to combine two different types of features, we propose the second model called Multi-Feature Hierarchical Latent Dirichlet Allocation model (MF-HLDA). Different from SF-HLDA, MF-HLDA is not only a hierarchical model but also a multi-feature based model. In this paper, two types of features including motion and static features are used in MF-HLDA. Experiments show that MF-HLDA model can improve the recognition performance even if one type of the features is badly extracted.

The remainder of the paper is organized as follows: Section 2 is a detailed description of the proposed models. Experiments and results are reported in section 3, and followed by some conclusions in section 4.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

## 2. THE PROPOSED MODELS

Two novel hierarchical generative models are proposed for recognizing realistic actions from unconstrained videos. The models shown in Fig. 1(b) and (c) follow the bag-of-word framework [12-14]. In this section, we will give a detailed description of the proposed models about their generative process, Bayesian decision and parameter learning.

### 2.1. Single-feature hierarchical latent dirichlet allocation model

By extending LDA model shown in Fig.1 (a), we get the first model called Single-Feature Hierarchical Latent Dirichlet (SF-HLDA) model shown in Fig. 1 (b). Different from LDA, SF-HLDA has two topic layers. One layer is called document topic *z*, and another is called feature topic *m*. Given one video $W_d$ like LDA, the generative process of SF-HLDA can be formalized as follows:

(1) Choose $\theta \sim Dir(\alpha)$ ,

(2) Choose $z \sim Mult(\theta)$ , where $z$ is a *Z*-dim vector,

(3) Choose $\omega \sim Dir(\kappa, z)$ , where $\omega$ is mixture parameter over topic $z$ and $\kappa$ is the Dirichlet prior of $\omega$, a matrix of size $Z \times M$ , $M$ is the total number of latent topic *m*,

(4) Choose $H \sim Poisson(\xi')$ ,

(5) For *h=1* to *H*,

  i) Choose $m_h$~ $Mult(\omega)$ , where $m_h$ is a *M*-dim vector,

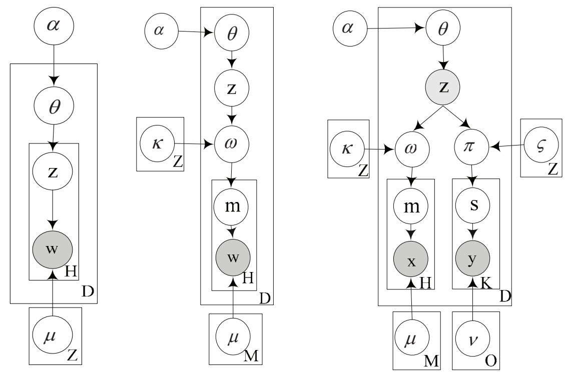
  ii) Choose $w_h$ from $w_h \sim p(w_h \mid m_h, \mu)$ , a multinomial probability conditioned on $m_h$.

In the above process of SF-HLDA model, $Poisson(\xi)$ is a Poisson distribution over parameter $\xi$ , the parameter $\theta$ is the mixture parameter of action topics, $Dir(\alpha)$ represents the Dirichlet distribution with prior parameter $\alpha$ , $Mult(\theta)$ is a multinomial distribution over parameter $\theta$ , and $\mu$ is a matrix of size $M \times U$ , where $\mu_{zu}$ indicates the probability of the word $w^u$ within the topic *z*, and $U$ is the number of vocabularies. Given the parameters, the joint distribution for $\theta$ , *z*, $\omega$ , *m* and $W_d$ is

$$p(W_d, \theta, z, \omega, m \mid \alpha, \kappa, \mu) = p(\theta \mid \alpha) \ p(z \mid \theta) p(\omega \mid z, \kappa) \cdot \prod_{h=1}^{H} p(m_h \mid \omega) p(w_h \mid m_h, \mu) \qquad (1)$$

### 2.2. Multi-feature hierarchical latent dirichlet allocation model

By extending SF-HLDA to combine two different features including motion and static features, we get the Multi-Feature Hierarchical Latent Dirichlet Allocation (MF-LDA) model shown in Fig. 1(c). This model assumes the following generative process for each action document:



(a) (b) (c)

Fig 1. Three generative models:
(a) LDA, (b) SF-HLDA, and (c) MF-HLDA

(1) Choose a mixture proportion $\theta \sim Dir(\alpha)$ for each action document, where $\theta$ is a *Z*-dim vector,

(2) Choose a document topic $z \sim Mult(\theta)$ , where $z$ is a *Z*-dim vector,

(3) Choose mixture proportion $\omega \sim Dir(\kappa, z)$ , $\pi \sim Dir(\varsigma, z)$ . $\kappa$ is the Dirichlet prior of $\omega$ , a matrix of size $Z \times M$ , where $M$ is the total number of the latent topic about motion features, and $\varsigma$ is Dirichlet prior of $\pi$ , a matrix of size $Z \times S$ , where *S* is the total number of the latent topics about static features,

(4) Choose $H \sim Poisson(\xi)$ , and $K \sim Poisson(\xi')$ . $H$ is the number of the words about motion features, and $K$ is the number of the words about static features,

(5) For each word about motion features: *h=1* to *H*,

  i) Choose a topic about motion feature $m_h \sim Mult(\omega)$ , where $m_h$ is a *M*-dim vector,

  ii) Choose a word about motion feature $x_h$ from $x_h \sim p(x_h \mid m_h, \mu)$ , where $\mu$ is a matrix of size $M \times U$ . $U$ is the total number of the vocabularies in the motion codebook for *x*. $\mu$ is the multinomial parameter for *x*,

(6) For each word about static feature: *k=1* to *K*,

  i) Choose a topic about static feature $s_k \sim Mult(\pi)$ , where $s_k$ is a *S*-dim vector,

  ii) Choose a word about static feature $y_k$ from $y_k \sim p(y_k \mid s_k, \nu)$ , where $\nu$ is a matrix of size $S \times P$ , and *P* is the total number of the vocabularies in the static codebook for $y$, $\nu$ is the multinomial parameter for *y*.

Given the parameters $\Theta = \{\alpha, \kappa, \varsigma, \mu, \nu\}$ and an observed action video, we get the joint distribution of the latent topics, mixture proportions, the words about motion and static features as follows:

$$p(x, y, m, s, \omega, \pi, z, \theta \mid \Theta) = p(\theta \mid \alpha) \ p(z \mid \theta) \ p(\omega \mid z, \kappa) \ p(\pi \mid z, \varsigma)$$

$$\cdot \prod_{h=1}^{H} p(m_h \mid \omega) p(x_h \mid m_h, \mu) \cdot \prod_{k=1}^{K} p(s_k \mid \pi) p(y_k \mid s_k, \nu)\,, \quad (2)$$

where $x$, y represent the motion properties of the spatial-temporal interest points and the static properties of spatial interest points from the action video respectively.

### 2.3. Bayesian decision

Before giving the algorithm of learning the models, we first discuss the Bayesian decision about action recognition. Here we only give the Bayesian decision about MF-HLDA model for saving the space of this paper.

An unknown action video sequence is first represented by two different collections of codewords including motion feature words $x$ and static feature words $y$. Assuming there are $C$ classes in action dataset and the parameters $\Theta_c$ ($1 \le c \le C$) are given, the probability of the action class $c$ is computed as follows:

$$p(c \mid x, y, \Theta_c) \propto p(x, y \mid \Theta_c) p(c) \propto p(x, y \mid \Theta_c)\,, \quad (3)$$

where $\Theta_c = \{\alpha, \kappa, \varsigma, \mu, \nu\}$ are parameters gotten by the following learning algorithm, and $p(c) = 1/C$ for convenience. If $c = \arg\max_c p(x, y \mid \Theta_c)$, then the action video belongs to the $c^{th}$ class of actions. The term $p(x, y \mid \Theta_c)$ can be computed as follows:

$$\begin{aligned} p(x, y \mid \Theta_c) = \int \sum_{z_i} p(z_i \mid \theta) p(\omega \mid z_i, \kappa) p(\pi \mid z_i, \varsigma) \\ \cdot \prod_{h=1}^{H} \sum_{m_h} p(m_h \mid \omega) p(x_h \mid m_h, \mu) \\ \cdot \prod_{k=1}^{K} \sum_{s_k} p(s_k \mid \pi) p(y_k \mid s_k, \upsilon)\ d\omega d\pi \end{aligned} \quad (4)$$

The distribution shown in the equation (4) is intractable for exact inference. Similar to the algorithm in [12], we adopt the variational approximation in this paper.

### 2.4. Learning the models

Similar to the section 2.3, this section will only give the description about the learning method of MF-HLDA model too. In this section, we introduce a variational distribution $q(m, s, \omega, \pi, z, \theta \mid \beta, \delta, \vartheta, \tau, \phi, \gamma)$ as an approximation of true posterior distribution $p(m, s, \omega, \pi, z, \theta \mid \Theta)$ over the latent variables with the following formalization,

$$q(m, s, \omega, \pi, z, \theta \mid \beta, \delta, \vartheta, \tau, \phi, \gamma) =$$

$$q(\theta \mid \gamma)\ q(\omega_t \mid \vartheta)\ q(\pi \mid \tau)\ q(z \mid \phi) \prod_{h=1}^{H} q(m_h \mid \beta_h) \prod_{k=1}^{K} q(s_k \mid \delta_k)\,, \quad (5)$$

where $\phi$, $\beta_h$, $\delta_k$ represent multinomial parameter over $Z$, $M$, $S$ topics respectively, and $\gamma$, $\vartheta$, $\tau$ are the Dirichlet parameter.

Similar to the LDA model in paper [12-13], these parameters are the free variational parameters. By using the Jensen's inequality directly [13], we have

$$\log\ p(x, y \mid \Theta) =$$

$$\log \int \frac{p(x, y, m, s, \omega, \pi, z, \theta \mid \Theta) q(m, s, \omega, \pi, z, \theta)}{q(m, s, \omega, \pi, z, \theta)} dm ds d\omega d\pi dz d\theta$$

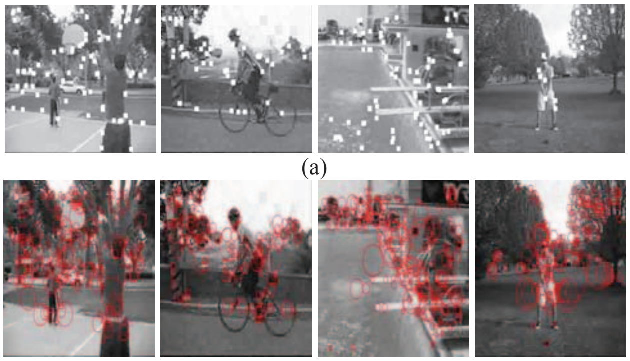


Fig 2. Some examples: (a) Motion features, and (b) Static features.

$$\ge E_q[\log p(x, y, m, s, \omega, \pi, z, \theta \mid \Theta)] - E_q[\log q(m, s, \omega, \pi, z, \theta)]\,. \quad (6)$$

We use $L$ to represent the right term in equation (6). Following [13], the lower bound $L$ is

$$\begin{aligned} L = E_q[\log p(\theta \mid \alpha)] + E_q[\log p(\mathbf{z} \mid \theta)] + E_q[\log p(\omega \mid z, \kappa)] \\ + E_q[\log p(\pi \mid z, \varsigma)] + E_q[\log p(m \mid \omega)] + E_q[\log p(x \mid m, \mu)] \\ + E_q[\log p(s \mid \pi)] + E_q[\log p(y \mid s, \upsilon)] - E_q[\log q(m)] \\ - E_q[\log q(s)] - E_q[\log q(\omega)] - E_q[\log q(\pi)] - E_q[\log q(z)] \\ - E_q[\log q(\theta)]\,. \end{aligned} \quad (7)$$

By maximizing the lower bound $L$, we can infer the variational parameters $\phi$, $\beta_h$, $\delta_k$, $\gamma$, $\vartheta$, $\tau$, and then by using variational EM algorithm, we can estimate the parameter $\Theta = \{\alpha, \kappa, \varsigma, \mu, \nu\}$.

## 3. EXPERIMENTS AND RESULTS

We employ the YouTube action dataset [10]. The videos in this dataset including about 1600 action videos with 11 categories are challenging for recognizing actions. There are 25 groups in each category and there are 4~23 action videos in each group. We choose 80 videos from each category video at random for training our models and the rest for testing the performance of the proposed models.

### 3.1. Action video representation

We first transfer them from the color space to grey space. And then the action videos are represented by two different collections of action features including motion and static features. We use the method proposed by Dollar et al. [15] to extract motion features shown in Fig. 2(a) and every motion feature is described by a 96-dim vector, and then k-means is adopted to cluster these features into $U$ classes as the vocabulary codebook of motion features. The method SURF proposed by Bay et al. [16] is used to get the static features shown in Fig. 2(b). Each static feature is described by a 128-dim vector. All the static features are clustered into $P$ classes as the vocabulary codebook of static features.

Table 1. Recognition Results

| U=P | LDA1 | SF-HLDA1 | LDA2 | SF-HLDA2 | MF-HLDA |
|---|---|---|---|---|---|
| **600** | 35.56 | 35.41 | 73.89 | 75.56 | **76.25** |
| **1000** | 31.53 | 33.33 | 73.33 | 77.64 | **78.89** |
| **1500** | 34.03 | 35.28 | 76.11 | 75.97 | **78.06** |
| **2000** | 32.36 | 35.56 | 75.69 | 77.22 | **79.17** |
| **2500** | 30.56 | 33.61 | 75.69 | 78.33 | **79.03** |
| **3000** | 31.53 | 36.53 | 76.39 | 77.08 | **80.41** |

### 3.2. Results

Table 1 shows the comparison results of action recognition assuming *U=P* for 600, 1000, 1500, 2000, 2500, 3000 and Z=11, M=S=15. In this table, the number 1 indicates motion feature is only used to learn the models, and 2 represents static feature is only used to learn the models.

From this table, we can see that motion features are badly extracted by the effect of the noise from the tremendous variations, so both LDA1 and SF-HLDA1 which are learnt by only motion features have bad performance. But compared with the results of LDA1, SF-HLDA1 can improve the average recognition rate from 32.60% to 34.95%. And if the static features are only adopted, compared with LDA2, SF-HLDA2 can improve the average recognition rate from 75.18% to 76.97%. Although the motion features are badly extracted, compared with LDA2, MF-HLDA by combining motion and static features can improve the average recognition rate from 75.18% to 78.64%. These results show that our proposed models are effective.

## 4. CONCLUSIONS

By extending LDA to hierarchical one, we get the Single-Feature Hierarchical Latent Allocation model which can further reduce the effect of the noise led by the tremendous variations from camera motion, background clutter, object appearance and so on, and improve the correct rate of action recognition. By extending SF-HLDA to combine two different features, we get Multi-Feature Hierarchical Latent Dirichlet Allocation model which can improve the average recognition performance even if one type of the features is badly extracted.

## ACKNOWLEDGEMENTS

This work has been supported by the National Key Technology R&D Program of China under Grant No. 2009BAH48B 02, 2009BAH43B04, 2011BAH16B01 and 2011BAH16B 02.

## REFERENCES

[1] A. Bobick and J. Davis. “The recognition of human movement using temporal templates”. PAMI, 23(3): 257 - 267, 2001.

[2] H. Li and M. Greenspan, “Multi-scale Gesture Recognition from Time-Varying Contours”, ICCV, pp.236-243, 2005.

[3] H. P. Li, Z. Y. Hu, Y. H. Wu, and F. C. Wu, “Behavior modeling and recognition based on space-time image features”, ICPR, pp. 243-246, 2006.

[4] Y. Ke, R. Sukthankar, and M. Hebert, “Efficient Visual Event Detection Using Volumetric Features”, ICCV, pp.166-173, 2005.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R.Basri, “Actions as Space-Time Shapes”, PAMI, 29(12), pp.2247–2253, 2007.

[6] I. Laptev and T. Linderberg, “Space-Time Interest Points”, ICCV, pp.432-439, 2003.

[7] K. Mikolajczyk and H. Uemura, “Action recognition with motion-appearance vocabulary forest”, CVPR, 2008.

[8] P. S. Dhillon, S. Nowozin, and C. H. Lampert, “Combining Appearance and Motion for Human Action Classification in Videos”, CVPR, pp. 22–29, 2009.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “ Learning Realistic human actions from movies”, CVPR, 2008.

[10] J. G. Liu, J.B. Luo, and M. Shah, “Recognizing Realistic Actions from Videos in the Wild”, CVPR , 2009.

[11] Y. X. Hu, L. L. Cao, F. J. Lv ,etc. “Action Detection in Complex Scenes with Spatial and Temporal Ambiguities”, ICCV, 2009

[12] J. C. Niebles, H. C. Wang and F. F. Li, “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words”. IJCV, 79, pp. 299-318, 2008.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocaion”, JMLR, 3, pp. 993-1022, 2003.

[14] T. Hospedale, S. G. Gong, and T. Xiang, “A Markov Clustering Topic Model for Mining Behavior in Video”, ICCV, 2009.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior Recognition Via Sparse Spatio-Temporal Features”, In Proceeding 2nd Joint IEEE Int’l Workshop on VS-PETS, Beijing, 2005.

[16] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “ SURF: Speeded Up Robust Features”,CVIU, 110(3), pp. 346-359, 2008.