MINING ACTIVITIES USING STICKY MULTIMODAL DUAL HIERARCHICAL DIRICHLET PROCESS HIDDEN MARKOV MODEL

Guodong Tian[1], *Chunfeng Yuan*[1], *Weiming Hu*[1], *Zhaoquan Cai*[2]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]Huizhou University, Huizhou, China

ABSTRACT

In this paper, a new nonparametric Bayesian model called Sticky Multimodal Dual Hierarchical Dirichlet Process Hidden Markov Model (SMD-HDP-HMM) is proposed for mining activities from a collection of time series. An activity is modeled as an HMM where each state corresponds to an atomic activity. By extensively using Dirichlet Process (DP), multiple HMMs sharing a common set of states are learned and the numbers of HMMs and states are both automatically determined. Each time series is modeled to be generated by one of the HMMs such that all time series are clustered into activities. Simultaneously state sequences for time series are learned and each of them is decomposed into a sequence of atomic activities. Experimental results on KTH activity dataset demonstrate the advantage of our method.

Index Terms— Dirichlet process, HMM, HDP, activity mining, time series

1. INTRODUCTION

In this paper, we address the problem of mining activities from time series. This general problem is frequently encountered in the field of computer vision, for activities is naturally represented by time series of visual features extracted from video sequences. Given a dataset of complex time series that may have multiple multimodal observations per time step, our goal is to 1) cluster them into different activities without the true number of categories known a priori, and 2) simultaneously learn a hierarchical probabilistic explanation in which an activity is composed by atomic activities and different categories have different rules of time dependencies between them. This probabilistic model has many potential applications such as activity classification, abnormality detection, video segmentation and video annotation.

We propose a novel nonparametric Bayesian model, Sticky Multimodal Dual Hierarchical Dirichlet Process Hidden Markov Model (SMD-HDP-HMM), based on Dirichlet Process (DP) [1, 2, 3] and its extension Hierarchical Dirichlet Process (HDP) [4, 5]. Through our model, arbitrary number of HMMs each with arbitrary number of states are learned. Each HMM corresponds to a category of activity and each state to an atomic activity. All the HMMs share a common set of states. Different HMMs have different subsets of states and transition matrices. A time series is generated by one of the HMMs so clustering can be realized. The transition matrices of the HMMs are regularized by a stickiness prior, which makes the learned sequences of states vary smoothly and the model more robust to the variation among frames of the same atomic activity. For each state, a multimodal emission with arbitrary number of modes can be learned. In this way, complex time series with multiple multimodal observations per time step are allowed. Gibbs sampler is developed to learn SMD-HDP-HMM. Our method is evaluated on KTH activity dataset [6] and achieves convincing performance.

Related work. There exists plenty of research work for activity mining. They can be roughly divided into two categories: similarity-based models and Bayesian models.

Lots of similarities for time series are proposed, such as Euclidean distance [7] and Dynamic Time Warping (DTW) [8]. These similarities are employed by clustering methods such as KMeans and spectral clustering. The performance of different similarities and clustering methods are experimentally compared by [9] and [10]. Similarity-based methods are simple and achieve good performance in many applications, however, their limitations are obvious: 1) they cannot determine the number of clusters automatically; 2) they can do nothing more than clustering/classification.

Among large number of Bayesian models, topic models have achieved great success for activity mining mainly because topics are naturally related with activities or atomic activities. Latent Dirichlet Allocation (LDA) [11] is a classical topic model. It is applied to unsupervised learning of human actions by [12]. This work achieves good classification performance, however, it has three limitations: 1) the number of categories has to be known a priori; 2) activities are modeled as "bag of words" (BoW) where no time dependencies is modeled; and 3) no atomic activity is modeled. [13] proposes Dual Hierarchical Dirichlet Process (Dual-HDP) extending

This work is partly supported by NSFC (Grant No. 60935002 and 61100099), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

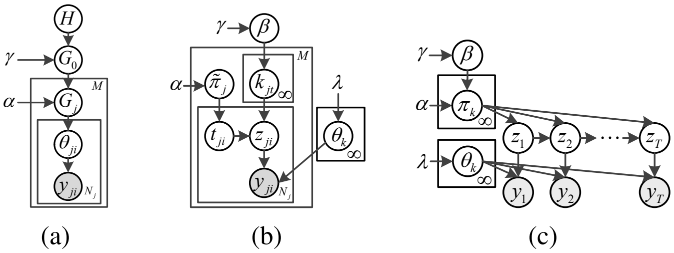

Fig. 1. Graphical model of (a) HDP, (b) another form of HDP and (c) HDP-HMM.

HDP so that the activities are modeled in a hierarchy and the number of clusters in each layer is automatically determined, however, it is also a BoW model. In order to learn temporal scene rules, [14] proposes Dependent Dirichlet Process Hidden Markov Model (DDP-HMM) built on HDP-HMM which is a dynamic variant of HDP. The DDP-HMM models temporal dependencies with arbitrary number of HMMs and allows for multiple observations per time step, but it lacks the ability of clustering high level activities and sharing atomic activities between different activities. [15] proposes a sticky and multimodal extension of HDP-HMM that is more robust to noise and learns state sequences more precisely.

Fusing the ideas of [13], [14] and [15], we develop a novel model that has the ability of learning different levels of activities without cluster numbers predefined, modeling temporal rules and handling complex noisy time series.

2. BACKGROUND

We denote a DP as DP(γ, H), where $\gamma > 0$ is the concentration parameter and H the base probability measure. A draw from DP(γ, H) is an infinite discrete distribution $G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta\left(\theta - \theta_k\right)$, which can be obtained by stick-breaking construction [16]: $\beta'_k \sim \text{Beta}(1, \gamma), \beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l), \theta_k \sim H(\theta|\lambda)$, where the construction of β is commonly denoted by $\beta \sim \text{GEM}(\gamma)$. The HDP is an extension of DP to model multiple mixtures that share components. It contains two levels of DPs, as shown in Fig. 1(a). At the first level, a global distribution G_0 is drawn from DP(γ, H). At the second level, DP(α, G_0) uses G_0 as the base distribution and generates multiple distributions $G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta - \theta_k), j = 1, \ldots, M$, which have the same support $\{\theta_k\}$ and each is used to generate a group of data $Y_j = \{y_{ji}\}$. The HDP can be considered as a topic model where topics $\{\theta_k\}$ are shared among documents $\{Y_j\}$. The distribution G_j also can be obtained by stick-breaking construction. This gives another form of HDP, as shown in Fig. 1(b). It is described by:

$$\begin{array}{lll} \beta \sim \text{GEM}(\gamma) & k_{jt} \sim \beta & \tilde{\pi}_j \sim \text{GEM}(\alpha) \\ t_{ji} \sim \tilde{\pi}_j, z_{ji} = k_{jt_{ji}} & \theta_k \sim H(\theta|\lambda) & y_{ji} \sim F(y|\theta_{z_{ji}}). \end{array} \quad (1)$$

Note that t_{ji} for different i and k_{jt} for different j and t may be identical so that clusters in two levels are formed. The posterior distributions of t_{ji} and k_{jt} on clusters is described as Chinese Restaurant Franchise (CRF), where a group of data Y_j corresponds to a restaurant and an observation y_{ji} to a customer that sits at a table t_{ji} with dish k_{jt}. The CRF plays an important role in the inference of HDP.

The HDP-HMM is an infinite state HMM with an HDP prior, as shown in Fig. 1(c). It is described by

$$\begin{array}{lll} \beta \sim \text{GEM}(\gamma) & \pi_k \sim \text{DP}(\alpha, \beta) & z_t | z_{t-1} \sim \pi_{z_{t-1}} \\ \theta_k \sim H(\theta|\lambda) & y_t \sim F(y|\theta_{z_t}), & \end{array} \quad (2)$$

where each state z_t corresponds to a group in HDP and each group-specific distribution $\pi_{z_{t-1}}$ contains transition probabilities from z_{t-1} to z_t. Due to the properties of HDP, the number of states is infinite.

For more details of DP, HDP and HDP-HMM, please refer to [1, 2, 3, 4, 5].

3. SMD-HDP-HMM

3.1. Proposed Model

Our goal is to cluster time series into activities and simultaneously segment each of them into atomic activities. HDP-HMM is suitable to model an activity as a sequence of atomic activities, because each state can be considered as an atomic activity and HDP-HMM solves the problem of complexity selection in traditional HMM. However, when modeling multiple activities, HDP-HMM is incompetent. To solve this problem, we construct an infinite mixture model of HDP-HMMs where each component corresponds to an activity. HDP-HMM still has other drawbacks: 1) the state sequences learned through HDP-HMM tend to have redundant fast switching states, so precise segmentation of a time series is hard to be obtained; 2) multimodal data are not well modeled. The sticky multimodal HDP-HMM proposed by [15] is a good solution to these two problems and it's approaches are adopted by our model. In addition, multiple observations are allowed to be generated at each time step in the proposed model.

For the sake of efficient computation and easy handling, the features in the raw time series are quantized into words and a codebook is obtained. An encoded time series $Y_j = \{Y_{jt}\}_{t=1}^{T_j}$ can be considered as a sequence of documents each containing words, i.e. $Y_{jt} = \{y_{jti}\}_{i=1}^{N_{jt}}$.

The proposed model is shown in Fig. 2. Its associated equations are:

$$\begin{array}{ll} \beta_0 \sim \text{GEM}(\gamma_0) & \beta_c \sim \text{DP}(\gamma, \beta_0) \\ \pi_{ck} \sim \text{DP}(\alpha + \kappa, \dfrac{\alpha\beta_c + \kappa\delta_k}{\alpha + \kappa}) & \\ \psi_k \sim \text{GEM}(\sigma) & \omega \sim \text{GEM}(\xi) \\ c_j \sim \omega & z_{jt} | z_{j,t-1} \sim \pi_{c_j, z_{t-1}} \\ s_{jti} \sim \psi_{z_{jt}} & \theta_{ks} \sim \text{Dir}(\lambda) \\ y_{jti} \sim \text{Discrete}(\theta_{z_{jt}, s_{jti}}). & \end{array} \quad (3)$$

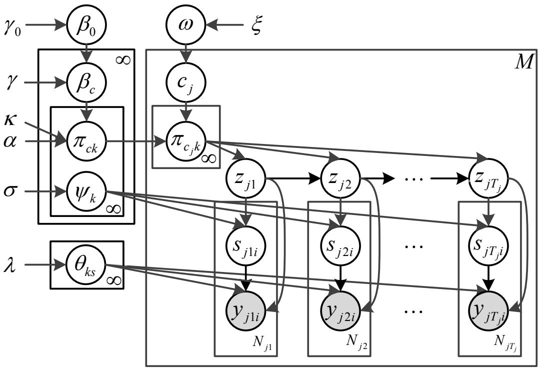

Fig. 2. Graphical model of SMD-HDP-HMM.

In our model, each β_c corresponds to an activity class and is an infinite discrete distribution on the labels of atomic activities. Infinite number of β_cs are drawn from DP(γ, β_0) in order to model arbitrary number of activities that share atomic activities. For each activity c, a transition matrix $\{\pi_{ck}\}_{k=1}^{\infty}$ with infinite dimension is obtained by drawing each row vector π_{ck} from DP($\alpha+\kappa$, $\frac{\alpha\beta_c+\kappa\delta_k}{\alpha+\kappa}$)), where $\kappa > 0$ is used to increase the self-transition prior probability, so that the states in a sequence switch more smoothly and the model is more robust to the variation among frames belonging to the same atomic activity [15]. Each atomic activity k is a multimodal distribution for words and modeled by a mixture of multinomials with weight vector ψ_k. A component of atomic activity k corresponds to a discrete distribution θ_{ks} on the codebook, which is drawn from the Dirichlet distribution Dir(λ). The distribution ω is drawn as the prior on labels of activities. Generating a time series Y_j follows the following steps: 1) an activity label c_j is drawn from ω; 2) the c_jth transition matrix $\{\pi_{c_jk}\}_{k=1}^{\infty}$ is chosen to produce a sequence of atomic activities $\{z_{jt}\}_{t=1}^{T_j}$; 3) the component label s_{jti} is drawn from $\psi_{z_{jt}}$ for each word i at time step t in sequence j; 4) the word y_{jti} is finally drawn from the discrete distribution on the codebook parameterized by $\theta_{z_{jt},s_{jti}}$.

3.2. Inference

We develop a Gibbs sampler to do inference for SMD-HDP-HMM. It alternatively samples: $\{c_j\}$, ω, $\{z_{jt}\}$, $\{s_{jti}\}$, $\{\psi_k\}$, β_0, $\{\beta_c\}$, $\{\pi_{ck}\}$ and $\{\theta_{ks}\}$. We use the efficient sampling strategy in [15] that employs a truncated approximation of DP. The numbers of activities, atomic activities and components of each atomic activity are limited by large numbers K_c, K_z and K_s, respectively.

Sampling $\{c\}$**.** The posterior distribution of c_j is given by:

$$p\left(c_j=c|\omega,\{\pi_{ck}\},\{z_{jt}\}\right)=\omega_c\prod_k\prod_l\pi_{ckl}^{n_{jkl}}, \qquad (4)$$

where n_{jkl} is the number of state transitions from k to l in instance j and can be easily computed from $\{z_{jt}\}$.

Sampling $\{\omega\}$**.** The discrete distribution ω is sampled by:

$$\omega|\{c_j\}\sim\text{Dir}\left(n'_1+\xi/K_c,\dots,n'_{K_c}+\xi/K_c\right), \qquad (5)$$

where n'_c is the number of instances that are assigned to activity c.

Sampling $\{z_{jt}\}, \{s_{jti}\}$**.** First the backward messages are computed recursively by:

$$m_{j,t,t-1}(z_{j,t-1})\propto\sum_{z_{jt}}p(z_{jt}|\pi_{c_j,z_{j,t-1}})m_{j,t+1,t}(z_{jt}) \\ \prod_i\sum_{s_{jti}}p(s_{jti}|\psi_{z_{jt}})p(y_{jti}|\theta_{z_{jt},s_{jti}}), \qquad (6)$$

then z_{jt} and s_{jti} are recursively sampled by:

$$p(z_{jt}|z_{j,t-1},\{y_{jti}\},\{\pi_{ck}\},\{\psi_k\},\{\theta_{ks}\},c_j)\propto m_{j,t+1,t}(z_{jt}) \\ p(z_{jt}|\pi_{c_j,z_{j,t-1}})\prod_i\sum_{s_{jti}}p(s_{jti}|\psi_{z_{jt}})p(y_{jti}|\theta_{z_{jt},s_{jti}}), \qquad (7)$$

$$p(s_{jti}|z_{jt},y_{jti},\{\psi_k\},\{\theta_{ks}\})=p(s_{jti}|\psi_{z_{jt}})p(y_{jti}|\theta_{z_{jt},s_{jti}}). \qquad (8)$$

Sampling $\{\psi_k\}$**,** β_0**,** $\{\beta_c\}$ **and** $\{\pi_{ck}\}$**.** The weight vector ψ_k is sampled by:

$$\psi_k|\{z_{jt}\},\{s_{jti}\}\sim\text{Dir}\left(\sigma/K_s+\tilde{n}_{k1},\dots,\sigma/K_s+\tilde{n}_{kK_s}\right), \qquad (9)$$

where $\tilde{n}_{ks}$ is the number of observations assigned to component s of state k. For sampling $\{\beta_c\}$ and β, first the number of tables $\{m_{ckl}\}$, $\{\bar{m}_{ckl}\}$ and $\{m'_{ck}\}$ in CRF with loyalty customers are sampled as auxiliary variables [15], then we have:

$$\beta_0|\{m'_{ck}\}\sim\text{Dir}\left(\gamma_0/K_z+m'_{\cdot 1},\dots,\gamma_0/K_z+m'_{\cdot K_z}\right), \qquad (10)$$

$$\beta_c|\{\bar{m}_{ckl}\},\beta_0\sim\text{Dir}\left(\gamma\beta_{01}+\bar{m}_{c\cdot 1},\dots,\gamma\beta_{0K_z}+\bar{m}_{c\cdot K_z}\right), \qquad (11)$$

where $\bar{m}_{c\cdot l}=\sum_k\bar{m}_{ckl}$ and $m'_{\cdot k}=\sum_c m'_{ck}$. The transition distribution $\{\pi_{ck}\}$ is sampled by:

$$\pi_{ck}|\{z_{jt}\},\{c_j\},\beta_c\sim\text{Dir}(\alpha\beta_{c1}+\sum_{j|c_j=c}n_{jk1},\dots, \\ \alpha\beta_{ck}+\kappa+\sum_{j|c_j=c}n_{jkk},\dots,\alpha\beta_{cK_z}+\sum_{j|c_j=c}n_{jkK_z}) \qquad (12)$$

Sampling $\{\theta_{ks}\}$**.** The parameter θ_{ks} is sampled by:

$$\theta_{ks}|\{z_{jt}\},\{s_{jti}\},\{y_{jti}\}\sim\text{Dir}(\lambda_1+n''_{ks1},\dots,\lambda_{K_v}+n''_{ksK_v}), \qquad (13)$$

where n''_{ksw} is the number of word w assigned to the component s of state k.

The hyper-parameters γ_0, γ_c, α, κ, σ and ξ are given non-informative priors and also sampled by the Gibbs sampler. For more details, please refer to [17].

4. EXPERIMENTS

Experiments are carried out on KTH activity dataset [6] widely used in the field of activity recognition. In our experiments, we do unsupervised learning on it. It contains 599 instances involving 6 classes, 25 persons and 4 scenarios. The subset containing 150 instances in scenario 1 is used in our experiments.

The features based on space time interest points (STIP) used by [18] are extracted. A video is represented by a time series of features and the numbers of features in different frames are not equal. Four frames containing features from the same video are shown in Fig. 3, where each circle corresponds to a feature. A codebook of size 1000 is obtained by KMeans and each feature is encoded by a word. The encoded time series are used to learn an SMD-HDP-HMM model by our Gibbs sampler.

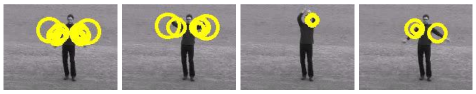

Fig. 3. STIPs detected in four frames of the same video.

The clustering performance of our model are evaluated and compared with KMeans (KM), Spectral Clustering (SC), LDA and Dual-HDP. Because the labels of the learned clusters are arbitrary, in order to evaluate the clustering performance, we have to find an mapping between the estimated labels and the true labels. This is done by maximizing the number of instances with matched labels using Munkres algorithm [19]. The ratio of the instances with matched labels to the whole instances, which is called the correct clustering rate (CCR) [9, 10], is then used to be the evaluation criteria. For KM and SC, a video is represented by the histogram of its visual words. The distance matrix for SC is computed by χ^2 distance. True number of activities is used for KM, SC and LDA. LDA is also used for unsupervised learning on KTH dataset in [12], but no clustering result is reported by [12]. The comparison of different methods' clustering performance is shown in Table 1. It's clear that our model has the highest CCR. This result demonstrates our model's advantage of modeling time dependencies between different frames, because the other methods are all BoW-based. There are 8 clusters produced by our model. The numbers of instances in the two redundant clusters are 8 and 1. This bias has been punished by the CCR. The relatively high CCR of our model also confirms its ability of finding the number of clusters.

Another advantage of our model lies in the ability of learning atomic activities and the rules governing activities. We find that the states of HMMs learned do correspond to some semantically meaningful atomic activities. Fig. 4(a)(b) show a part of the transition matrix and its corresponding state transition diagram containing four atomic activities that occur most often in running. Some sample frames assigned to these states are also shown. State A and B of running correspond to the actions of dropping a leg and lifting a leg, respectively. State C and D have the same semantic meanings but contrary running directions as A and B. There are considerable reciprocal transition probabilities between A and B, as well as between C and D. This fits the fact that lifting a leg and dropping a leg alternatively occur in the process of running. The transition relations for the seven most frequent happening atomic activities in hand waving are shown in Fig. 4(c)(d). A cycle of transition is formed by these states, which reflects the repetition of waving arms up and down.

Table 1. Clustering performance of different methods.

Method	KM	SC	LDA	Dual-HDP	Our model
CCR	25.93%	55.04%	42.67%	50.67%	60.67%
Finding the number of clusters	No	No	No	Yes	Yes

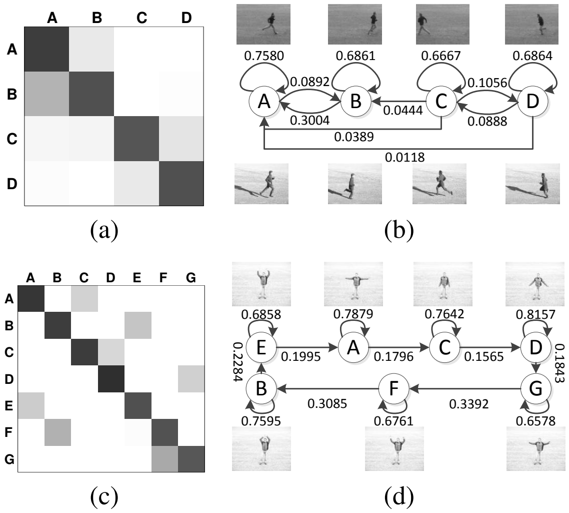

Fig. 4. Patial transition matrices and transition diagrams for running and hand waving.

5. CONCLUSION

A novel nonparametric Bayesian model, SMD-HDP-HMM, has been proposed. Complex time series containing multiple observations with multimodal distributions per time step can be clustered into activities by SMD-HDP-HMM. The numbers of clusters and states of HMMs are both automatically determined. With the stickiness prior on self-transitions, large variation among frames belonging to the same atomic activity is well handled, so that meaningful atomic activities and rules governing activities are also obtained. Experiments demonstrate that SMD-HDP-HMM has superb performance on activity clustering and semantic learning.

6. REFERENCES

[1] T. S. Ferguson, “A bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[2] Y. W. Teh, “Dirichlet processes,” in *Encyclopedia of Machine Learning*. Springer, 2010.

[3] R. M. Neal, “Markov chain sampling methods for dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierar-chical dirichlet process,” *Journal of the American StatisticalAssociation*, vol. 101, no. 476, pp. 1566–1581, 2006.

[5] Y. W. Teh and M. I. Jordan, “Hierarchical Bayesian nonparametric models with applications,” in *Bayesian Nonparametrics: Principles and Practice*, N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge University Press, 2010.

[6] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local svm approach,” in *International Conference on Pattern Recogntion (ICPR)*, 2004.

[7] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank, “A system for learning statistical motion patterns,” *IEEE Trans. on Pattern Analysis and Machine Intelligent*, vol. 28, no. 9, pp. 1450–1464, 2006.

[8] E. Keogh and M. Pazzani, “Scaling up dynamic time warping for datamining application,” in *International Conference on Knowledge discovery and data mining (KDD)*, 2004.

[9] Z. Zhang, K. Huang, and T. Tian, “Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes,” in *International Conference on Pattern Recogntion (ICPR)*, 2006.

[10] B. Morris and M. Trivedi, “Learning trajectory patterns by clustering: Experimental studies and comparative evaluation,” in *International Conference on Computer Vision and Pattern Recogntion (CVPR)*, 2009.

[11] D. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] J. C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[13] X. Wang, K. T. Ma, G. Ng, and E. Grimson, “Trajectory analysis and semantic region modeling using a nonparameteric hierarchical bayesian model,” in *International Conference on Computer Vision and Pattern Recogntion (CVPR)*, 2008.

[14] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari, “What’s going on? discovering spatio-temporal dependencies in dynamic scenes,” in *International Conference on Computer Vision and Pattern Recogntion (CVPR)*, 2010.

[15] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “An HDP-HMM for systems with state persistence,” in *International Conference on Machine Learning (ICML)*, 2008.

[16] J. Sethuraman, “A constructive definition of dirichlet prior,” *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[17] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “A Sticky HDP-HMM with Application to Speaker Diarization,” *Annals of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[18] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in *International Conference on Computer Vision and Pattern Recogntion (CVPR)*, 2008.

[19] J. Munkres, “Algorithms for the assignment and transportation problems,” *Journal of the Society of Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.