# AUTOMATIC VIDEO ANNOTATION VIA HIERARCHICAL TOPIC TRAJECTORY MODEL CONSIDERING CROSS-MODAL CORRELATIONS

*Takuho Nakano*[⋆]*, Akisato Kimura*[†]*, Hirokazu Kameoka*[†]*, Shigeki Miyabe*[⋆]*, Shigeki Sagayama*[⋆]*, Nobutaka Ono*[⋆]*, Kunio Kashino*[†]*, Takuya Nishimoto*[⋆]

[⋆]Graduate School of Information Science and Technology, The University of Tokyo
[†]NTT Communication Science Laboratories, NTT Corporation

## ABSTRACT

We propose a new statistical model, named Hierarchical Topic Trajectory Model (HTTM), for acquiring a dynamically changing topic model that represents the relationship between video frames and associated text labels. Model parameter estimation, annotation and retrieval can be executed within a unified framework with a few computation. It is also easy to add new modals such as audio signal and geotags. Preliminary experiments on video annotation task with manually annotated video dataset indicate that our proposed method can improve the annotation accuracy.

***Index Terms*—** Video annotation, generative approach, topic model, canonical correlation analysis, hidden Markov model

## 1. INTRODUCTION

Content-based retrieval has been the subject of a significant amount of research in this decade [1, 2, 3]. Especially, we are focusing on video signals as the subject of research, since video signals often include various types of information such as text, audio and visual information.

The realization of general-purposed image/video annotation retrieval has still been a challenging problem. Previous efforts have been mainly directed to how to construct and combine binary classifiers such as support vector machine (SVM) [4, 5] and supervised multi-class learning (SML) [6] to annotate with respect to the presence or absence of each text label. However, this approach has the following crucial drawbacks. 1) Generally speaking, incorporating *co-occurrences* among text labels into this approach is quite difficult. For example, if a image contains buses, cars and an unknown object, that is hard to recognise, and we have to decide the object is a cow or a bike. If we use binary classifiers for cows and bikes, that may give us no information. But we know that bikes co-occurs with buses and cars more frequently than cows. Although some previous work [7] tried to integrate collocations between text labels with two different attributes, it is impossible to extend it to general types of co-occurrences. 2) In many methods, "one vs the rest" classifiers have been utilized mainly due to computational efficiency. However, this strategy often encounters the so-called *masking problem*[8] because of the disparity in samples, which means that discrimination for each class often fails.

Recently, inference techniques based on *topic models* (see Fig. 1) have been proposed for acquiring topic models. Probabilistic latent semantic analysis (pLSA) [9] and latent Dirichlet allocation (LDA) [10] are widely known and have been exploited for image annotation retrieval [6, 11, 12, 13, 14]. Canonical correlation analysis (CCA) [15, 16, 17], which is a generalized variant of Fisher linear discriminant analysis (FDA) for multi-category classification, is also known as one of them. CCA is easy-to-use and feasible for efficiently acquiring topic models. Its effectiveness on image annotation and retrieval has been presented in some previous researches [18, 19].

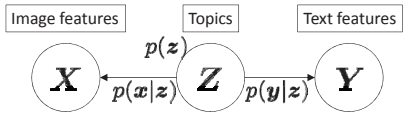



**Fig. 1**. Topic model representing topics of both image features and text features in symmetric translation model.

Another significant issue for video annotation retrieval is a way of representing temporal dynamics of videos. Two typical approaches exist: 1) representing a video as a set of keyframes and replacing the problem into image annotation retrieval, and 2) representing a video as a statistical model that assumes some Markov properties, such as a state space model and a hidden Markov model (HMM). Although the former approach makes the problem simple, any types of temporal information have been removed, which would be quite significant to capture a video concept composed of a sequence of visual scenes. Meanwhile, the latter approach might be redundant since a shot includes so many video frames similar to each other. Keyframe extraction would be important to obtain a concise representation of shots. Previously, layered dynamic mixture model [20] using hierarchical hidden Markov model (HHMM) are presented. However, they needs considerable computational cost for model parameter estimation and inference.

To this end, we propose a new statistical model which incorporates 1) co-occurrences among visual information and text information and 2) temporal dynamics of videos simultaneously. The proposed model shown in Figure 2 is composed of keyframe-wise topic models and a hidden Markov model connecting topic models smoothly. From this viewpoint, we call the proposed model as a hierarchical topic trajectory model (HTTM).

## 2. HIERARCHICAL TOPIC TRAJECTORY MODEL

### 2.1. Framework

Our proposed model consists of four layers: (a) time-series data, such as video frames, text labels and audio signals, (b) features extracted from data, (c) latent variables and (d) state variables .

Figure 2 overviews our proposed model, HTTM. The bottom

⋆ {t-nakano, miyabe, sagayama, onono, nishi}@hil.t.u-tokyo.ac.jp
† akisato@ieee.org, {kameoka, kunio}@eye.brl.ntt.co.jp

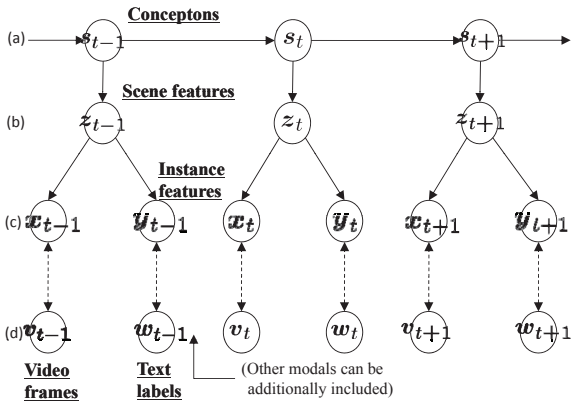



**Fig. 2**. Overview of proposed model (HTTM). Direct dependencies of values are showed with arrow lines.

layer corresponds to observations namely video frames $v_x$ and text labels $w_t$. The second layer from the bottom corresponds to features $\boldsymbol{x}_t, \boldsymbol{y}_t$ extracted from the observations. The third layer from the bottom corresponds to latent variables $\boldsymbol{z}_t$ representing the relationship between video and audio features. The top layer consists of hidden state series $s_t$, which outputs the latent variables.

HTTM can be formulated by the following joint probability density function (PDF):

$$p(X, Y, Z, S) = \prod_{t=1}^{T} p(s_t|s_{t-1})p(z_t|s_t)p(x_t|z_t)p(y_t|z_t)$$

, where $X = \{x_1, x_2, ..., x_T\}$ ($Y$, $Z$, $S$ are all defined similarly), $T$ is the number of keyframes in a given shot, and $p(s_1|s_0) = p(s_1)$. We will describe every component PDF in the following.

The feature vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are assumed to be independently generated given the latent variable $\boldsymbol{z}$ from a normal distribution with a mean vector given by an affine transformation of $\boldsymbol{z}$: $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; W_x\boldsymbol{z} + \bar{\boldsymbol{x}}, \Psi_x)$ and $p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{y}; W_y\boldsymbol{z} + \bar{\boldsymbol{y}}, \Psi_y)$, where $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \Sigma)$ denotes the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$. A latent space provides a compact representation (topic model) reflecting cross-modal correlations. This model can be easily extended to more than two types of feature vectors, which implies that our proposed model can deal with multiple modals such as audios and geotags.

The latent variables $\boldsymbol{z}_t$ (more precisely their conditional expectation) are modeled as the observations of an HMM with hidden states $s_t$ in layer (d). At each time $t$, a state variable $s_t$ takes values from a finite set $\{1, \ldots, K\}$. We define a transition probability $p(s_t = j|s_{t-1} = i)$ from a state $s_{t-1} = i$ at time $t$ to a state $s_t = j$ at time $t$ as $p_{ij}$. The output probability distributions of the states are modeled using a Gaussian mixture model (GMM) as described by the equation below:

$$\begin{aligned} p(\boldsymbol{z}_t|s_t = k) &= \textstyle\sum_{j=1}^{L_k} \pi_{k,j} p(\boldsymbol{z}_t|j, s_t = k), \qquad (1) \\ p(\boldsymbol{z}_t|j, s_t = k) &= \mathcal{N}(\boldsymbol{z}_t; \bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j}), \end{aligned}$$

where $L_k$ is the number of Gaussians and $\bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j}, \pi_{k,j}$ are the mean vector, the covariance matrix and the mixture weight of the $j$-th component of state $k$. For simplicity, we assume the number of Gaussians to be common between states ($L_k = L$). This HMM enables us to control temporal dependency of topic models and variety of feature vectors simultaneously.

### 2.2. Model training

In this framework, the parameter estimation method can be achieved by a combination of those of topic models and HMM. It consists

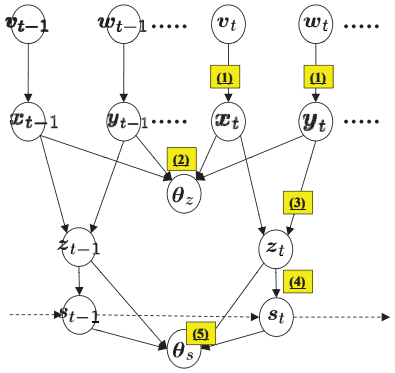



**Fig. 3**. Procedure of parameter estimation for HTTM. $\boldsymbol{\theta}_z$ and $\boldsymbol{\theta}_s$ are a set of parameters of topic models and HMM, respectively.

of five steps, shown in Fig. 3, i.e.: (1) extracting features from data, (2) estimation parameters for model parameters of topic models, (3) extracting latent variables, (4) temporal clustering of scene features via Viterbi search or forward-backward algorithm and (5) estimation parameters for HMM.

Maximum-likelihood estimates for the parameters $\boldsymbol{\theta}_z = \{\Lambda, A, B, \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, C_{yy}\}$ of topic models can be obtained with the help of probabilistic canonical correlation analysis (PCCA)[16]. The mean values $\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}$ are simply calculated as the average of the training data $\boldsymbol{x}_t, \boldsymbol{y}_t$. The estimation of the other parameters reduces to solving a generalized eigenvalue problem. Let $d_x$ and $d_y$ be the number of dimensions of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively and let $d$ ($d \leq \min(d_x, d_y)$ ) be the number of dimensions of the latent variable. By solving the following generalized eigenvalue problems, we can obtain the top $d$ eigenvalues $\lambda_i, (i = 1, 2, \ldots d)$ in descending order and the $d$ associated eigenvectors $(\boldsymbol{a}_i, \boldsymbol{b}_i) \in \mathbf{R}^{d_x+d_y}$:

$$\begin{pmatrix} \mathbf{0} & C_{xy} \\ C_{yx} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} = \lambda \begin{pmatrix} C_{xx} & \mathbf{0} \\ \mathbf{0} & C_{yy} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix},$$

where $\mathbf{0}$ denotes zero matrix and $C_{xx}, C_{yy}, C_{xy}, C_{yx}$ denotes covariance matrices calculated from training data. We can define a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, and matrices $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d)$ and $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)$. The conditional expectation of the latent variables $\boldsymbol{z}_t$ are then obtained, respectively for the case where only the image feature $\boldsymbol{x}$ is given and that where both the image feature $\boldsymbol{x}$ and the label feature $\boldsymbol{y}$ are given, as:

$$\boldsymbol{z}(\boldsymbol{x}) = \Lambda^{\frac{1}{2}} A^{\top} (\boldsymbol{x} - \bar{\boldsymbol{x}}), \qquad (2)$$

$$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{y}) = \Lambda^{\frac{1}{2}} (I + \Lambda) \left( A^{\top} (\boldsymbol{x} - \bar{\boldsymbol{x}}) + B^{\top} (\boldsymbol{y} - \bar{\boldsymbol{y}}) \right). \qquad (3)$$

Parameters $\boldsymbol{\theta}_s = \{\{p_{ij}\}, \{\bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j}, \pi_{k,j}\}\}$ for HMM can be estimated from the latent variables $\boldsymbol{z}_t$. That is achieved by using Baum-Welch algorithm [21], which is an iterative method for estimating parameters in two steps. At first, model parameters are set randomly.

- In the expectation step, the hidden state series $s_t$ is estimated stochastically, with Viterbi search or forward-backward algorithm.
- In the maximization step, model parameters $\boldsymbol{\theta}_s$ are estimated considering state series estimated in the expectation step.

These two steps are repeated until the estimated parameters converge or the iteration count reaches the predefined maximum.

### 2.3. Recognition (Estimation of lacking features)

Recognition can be considered as estimation of label features from video features only. It consists of six steps, shown in Fig. 4, i.e.: (1) extract features from data, (2) estimate latent variables, (3) estimate

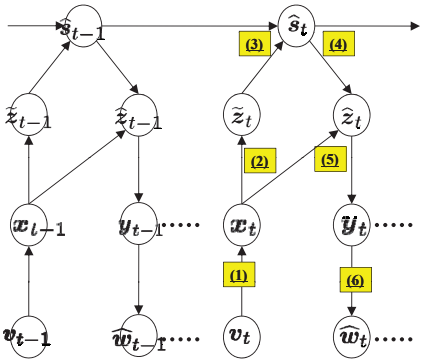
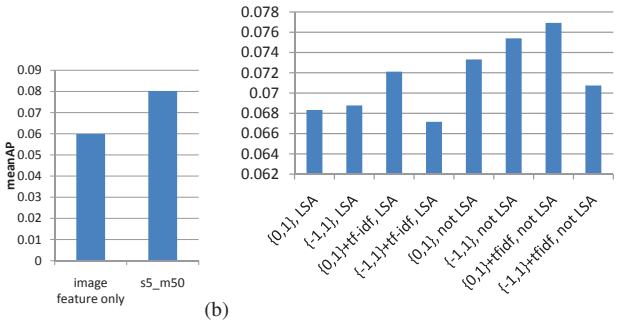



**Fig. 4**. Procedure of recognition using HTTM.




(a) (b)

**Fig. 5**. Recognition results. (a) Proposed model outperforms the model without state estimation. (b) Label feature $\{0, \mathrm{idf}(l)\}$ performs best. ("$sn_s$_m$n_m$" means $K = n_x, L = n_m$)

state variables, (4) re-estimate latent variables considering instance features and state variables, (5) estimate label features, (6) output recognition result considering estimated label features. Our main contribution in this stage is the fourth step, considering not only observations but also temporal dependencies to estimate latent variables. Re-estimation considering state variables, which represents temporal dependencies, would improve estimation accuracy.

Image features $\boldsymbol{x}_t$ are extracted from the images of the given video. Latent variables $\tilde{\boldsymbol{z}}_t$ are estimated with the extracted image features $\boldsymbol{x}_t$ by Eq. (2), considering the model parameter set $\boldsymbol{\theta}_z$ learned in Sec. 2.2. Hidden state series $\hat{\boldsymbol{S}} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_t, \ldots\}$ are estimated from the estimated latent variable series $\tilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_1, \tilde{\boldsymbol{z}}_2, \ldots, \tilde{\boldsymbol{z}}_t, \ldots\}$, considering the model parameter set $\boldsymbol{\theta}_s$ by Viterbi decoding:

$$\hat{\boldsymbol{S}} \approx \underset{S}{\operatorname{argmax}}\, p(\tilde{\boldsymbol{z}}_1|s_1)p(s_1)\textstyle\prod_{t=2}^{T} p(\tilde{\boldsymbol{z}}_t|s_t)p(s_t|s_{t-1}),$$

where $T$ is the size of series (the number of images used) and $p(s_t|s_{t-1})$ is transition probability[1].

Latent variables $\hat{\boldsymbol{z}}_t$ are re-estimated by considering estimated hidden states $\hat{S}$ and instance features $\boldsymbol{x}_t$. At each time $t$, the hidden state $s_t$ gives a distribution dependent of latent variable $\boldsymbol{z}_t$ as described above. Using this information, estimation accuracy for latent variable $\boldsymbol{z}_t$ can be improved. Given a state variable $s_t = k$, the distribution of latent variable $\boldsymbol{z}_t$ can be described by GMM with parameters $\bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j}, \pi_{k,j} (j = 1, 2, \ldots, L)$ for mean vectors, covariance matrices, mixture weights. $\hat{\boldsymbol{z}}_t$ are calculated by below equation:

$$\hat{\boldsymbol{z}}_t = \textstyle\sum_{j=1}^{L} \tilde{\pi}_j \bar{\boldsymbol{z}}_{k,j},$$

where $\tilde{\pi}_i$ are given by:

$$\tilde{\pi}_j = \frac{\pi_{k,j}\mathcal{N}(\tilde{\boldsymbol{z}}_t; \bar{\boldsymbol{z}}_{k,j}, \Sigma_{k,j})}{\sum_{l=1}^{L} \pi_{k,l}\mathcal{N}(\tilde{\boldsymbol{z}}_t; \bar{\boldsymbol{z}}_{k,l}, \Sigma_{k,l})}.$$

A label feature $\hat{\boldsymbol{y}}_t$ can be estimated with re-estimated latent variables $\hat{\boldsymbol{z}}_t$. That is archived in the framework of PCCA:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{y}(\hat{\boldsymbol{z}}_t) = W_y \hat{\boldsymbol{z}}_t + \bar{\boldsymbol{y}} \tag{4}$$

where $W_y$ is given by $W_y = C_{yy} B \Lambda^{\frac{1}{2}}$.

Labels are estimated or annotated from estimated label features $\hat{\boldsymbol{y}}_t$. We use ranked output of image for each label i.e. fix a label and rank images in a high likelihood order.

## 3. EXPERIMENTS

### 3.1. Experimental Conditions

We conducted experiments with TRECVID [3] 2005 data including 127 videos and 56191 shots. We divide them into two data, one is for training, containing 102 videos, 45689 shots, the other is for testing, containing 25 videos, 10502 shots. Bag of Features (BoF) with SIFT local descriptors provided by vireo374 [22] were used as image features. We chose 47 labels[2] from LSCOM-Lite and LSCOM annotation [23, 24] and remove shots without any of 47 labels. We adopted the following 3 alternatives and test $8 = 2^3$ methods to extract label feature:

- Label features with zero and positive values such as $\{0, 1\}$, or negative and positive values$\{-1, 1\}$.
- Weighting label features with tf-idf or not.
- Use latent semantic analysis (LSA) to extract correlations among labels, or not.

The number of dimensions $d$ of latent variables was set to 47. In GMM we used diagonal covariance matrices.

We used mean average precision (meanAP), namely the mean value of average precision values over all the labels. Average precision is defined by follow equation:

$$\mathrm{AP} = \frac{1}{R}\textstyle\sum_{k=1}^{N} r_k p_k,$$

where $r_k$ takes the value $r_k = 1$ if the $k$-th output is true, otherwise $r_k = 0$, $N$ is the number of annotated samples, $R$ takes the value $R = \sum_{k=1}^{N} r_k$, and $p_k$ denotes the precision value when considering the 1-st through $k$-th outputs as all the results. If a system outputs randomly, the precision takes values equal to chance level at each rank and average precision also takes chance level. For the $l$-th label, values of the $l$-th dimension of $\hat{\boldsymbol{y}}_t$s are compared and the output are the indexes in descending order.

### 3.2. Results

We conducted two experiments to verify the effectiveness of the proposed method for automatic video annotation and to analyze relationships between model parameters and annotation accuracy. In the experiment (a), 8 types of methods for label feature extraction were compared on the basis of meanAP. In the experiment (b), we chose the best label features and evaluated the relationship between the annotation accuracy and the number of states $K$ and mixtures $L$ ($KL = 240$, described below).

[1] Hidden state series $\hat{S}$ can be also estimated stochastically using forward-backward algorithm for Viterbi decoding.

[2] Used labels are: Airplane, Airplane_Flying, Animal, Boat_Ship, Building, Bus, Car, Charts, Cityscape, Classroom, Computer_TV-screen, Corporate-Leader, Court, Crowd, Demonstration_Or_Protest, Desert, Entertainment, Explosion_Fire, Face, Flag-US, Government-Leader, Hand, Maps, Meeting, Military, Mountain, Natural-Disaster, Nighttime, Office, Outdoor, People-Marching, Person, Police_Security, Prisoner, Road, Singing, Sky, Snow, Sports, Studio, Telephones, Truck, Urban, Vegetation, Walking_Running, Waterscape_Waterfront, Weather.

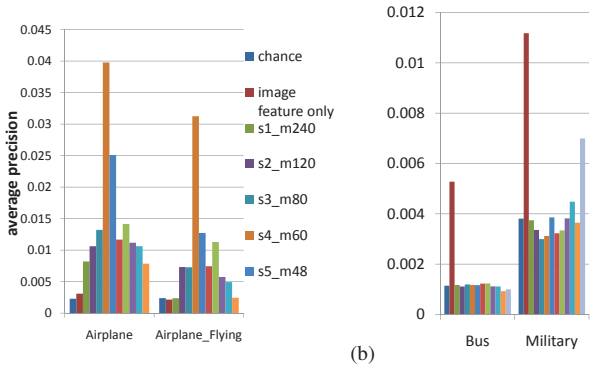



**Fig. 6**. Results of labels: Airplane, Airplane_Flying, Bus and Military.

Figure 5 (a) shows the meanAP obtained from the framewise topic model (left) and the proposed method, HTTM, with the fixed number of states $K = 5$ and mixtures $L = 50$ (right). This figure indicates that our proposed model considering temporal dependency outperformed a set of topic models without considering temporal dependency.

Figure 5 (b) shows the meanAP of 8 different methods for label feature extraction. This figure indicates that the label feature expressed with zero or positive value and tf-idf weighting, performed best.

In the experiment (b), we fixed the number of states $K$ and mixtures $L$ satisfying $KL = 240$ because $K = 5$ and $L = 50$ performed best in experiment (a) and 240 has many divisors around 250. There might be some trade-offs between the number of hidden states and mixtures. Many hidden states and a few mixtures would emphasize temporal structures of videos, while the opposite case would pay attention to the current frame features more.

Figure 6 shows the results of two labels each. Figure 6 (a) shows that both Airplane and Airplane_Flying performed best with $K = 4, L = 60$. This suggests that correlation information was accurately used in model learning with that condition. Figure 6 (b) indicates the results of Bus and Military. This shows that sometimes HMMs performed worse than considering only image features. One possible reason is that GMMs does not match to very small chance levels those may considered as outliers.

## 4. CONCLUDING REMARKS

We proposed a new statistical model, Hierarchical Topic Trajectory Model (HTTM), for acquiring a dynamically changing topic model that represents the relationship between video frames and associated text labels. Label features with zero or tf-idf values performs best. The recognition results revealed some relationships among the number of states, the number of GMM mixtures and annotation accuracy.

Our future work includes some comparison with other video recognition methods and automatic determination extension to estimate the number of states or mixtures.

## 5. ACKNOWLEDGMENTS

The authors thank Dr. Naonori Ueda, Dr. Eisaku Maeda, Dr. Futoshi Naya and Dr. Junji Yamato of NTT Communication Science Laboratories for their help.

## 6. REFERENCES

[1] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[2] K. Kashino, T. Kurozumi, and H. Murase, "A quick search method for audio and video signals based on histogram pruning," *IEEE Trans. MM*, vol. 5, no. 3, pp. 348–357, 2003.

[3] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. MIR*, pp. 321–330, 2006.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, pp. 2169–2178, October 2006.

[5] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *Journal of Machine Learning Research*, vol. 8, pp. 725–760, April 2007.

[6] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. PAMI*, vol. 29, pp. 394–410, March 2007.

[7] G. Wang and D. Forsyth, "Joint learning of visual attributes, object classes and visual saliency," in *Proc. ICCV*, pp. 537 –544, sep 2009.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Springer, February 2009.

[9] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1-2, pp. 177–196, 2001.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, January 2003.

[11] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[12] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. ICCV*, vol. 1, pp. 370–377, 2005.

[13] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. CVPR*, vol. 2, pp. 524 – 531 vol. 2, jun 2005.

[14] F. Monay and D. Gatica-Perez, "Modeling semantic aspects for cross-media image indexing," *IEEE Trans. PAMI*, vol. 29, pp. 1802–1817, August 2007.

[15] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, 1933.

[16] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," *Technical Report 688, Department of Statistics, University of California, Berkeley*, 2005.

[17] C. Wang, "Variational bayesian approach to canonical correlation analysis," *IEEE Trans. NN*, vol. 18, no. 3, pp. 905 –910, 2007.

[18] T. Bailloeul, C. Zhu, and Y. Xu, "Automatic image tagging as a random walk with priors on the canonical correlation subspace," in *Proc. MIR*, pp. 75–82, 2008.

[19] T. Harada, H. Nakayama, and Y. Kuniyoshi, "Image annotation and retrieval based on efficient learning of contextual latent space," in *Proc. ICME*, pp. 858–861, August 2009.

[20] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin, "Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams," in *Proc. ICASSP*, vol. 2, pp. 1053–1056, mar. 2005.

[21] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[22] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," *Proc. CIVR*, 2007.

[23] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for trecvid 2005," *IBM Research Technical Report*, 2005.

[24] M. Naphade, J. R. Smith, J. Tesic, S. F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," in *IEEE Trans. MM*, vol. 13, pp. 86–91, 2006.