[NTT-UT SIN & KIS]
SEMANTIC INDEXING AND KNOWN ITEM SEARCH BASED ON A UNIFIED MODEL WITH TOPIC TRANSITION REPRESENTATION

Takuho Nakano**, Akisato Kimura*†*, Hirokazu Kameoka*†*,*
Shigeki Sagayama**, Nobutaka Ono***, and Kunio Kashino*†

* Graduate School of Information Science and Technology, The University of Tokyo
† NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
Contact address: t-nakano@hil.t.u-tokyo.ac.jp, akisato@ieee.org

ABSTRACT

We applied a generative approach to the TRECVID 2010 Semantic Indexing (SIN) and Known-Item Search (KIS) tasks, using a probabilistic network called Hierarchical Topic Trajectory Model (HTTM). It is our newly-developed model that can integrate multiple sources of potentially associated information such as video frames and texts, as well as dynamically changing high-level pieces of information such as topics. With this model, the semantic indexing and the known-item search tasks were dealt within a single unified framework. We show how it worked, and present some analysis for the SIN task.

Index Terms— Semantic indexing, known-item search, generative approach, topic model, canonical correlation analysis, hidden Markov model

1. INTRODUCTION

In the field of content analysis, indexing, and retrieval, in recent years, inference techniques have been highlighted for acquiring topic models. For example, probabilistic latent semantic analysis (pLSA) [1] and latent Dirichlet allocation (LDA) [2] are widely used for image annotation and retrieval [3, 4, 5]. Canonical correlation analysis (CCA) [6, 7, 8], which is a generalized variant of Fisher linear discriminant analysis (FDA) for multi-category classification, is also known as one of them. Its effectiveness on image annotation and retrieval has been presented in some previous researches [9, 10, 11]. Furthermore, modeling temporal dynamics of videos has been a key for accurate video analysis, indexing, and retrieval. Typical approaches include: 1) representing a video as a set of keyframes to reduce the problem into "image" annotation and retrieval, and 2) representing a video as a statistical model [12].

With this background, we developed a new statistical model called a hierarchical topic trajectory model (HTTM) that can handle the semantic indexing task and the known-item search task within a single framework. The model incorporates (1) co-occurrences among visual information and text information and (2) temporal dynamics of videos simultaneously. As shown in Figure 2, it comprises a series of keyframe-wise topic models and an HMM that connects them.

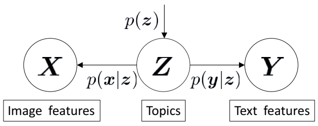


Fig. 1. Topic model representing relationships between image and text features.

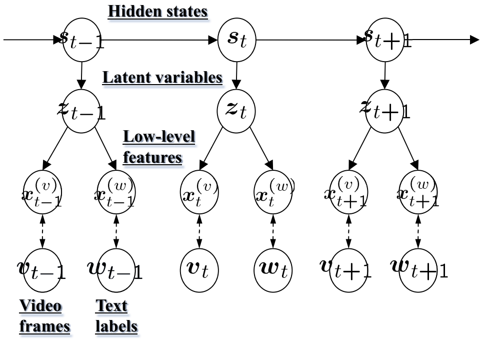


Fig. 2. Overview of hierarchical topic trajectory model (HTTM).

2. HIERARCHICAL TOPIC TRAJECTORY MODEL

2.1. Framework

This section briefly describes the method we adopted for Semantic Indexing and Known-Item Search tasks in TRECVID2010. Details can be seen in [13]. Figure 2 overviews the structure of HTTM, which consists of four layers: (a) raw data such as video frames and text tags, (b) low-level features, (c) latent variables and (d) hidden states.

The bottom layer corresponds to video frames v_t and text tags w_t. The second layer from the bottom represents low-level features $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ extracted from the video frames and text tags, respectively. The third layer stands for latent variables $\boldsymbol{z}_t$ representing the relationship between video and text features. The top layer consists of hidden states s_t, which inherits temporal relationships.

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

$$p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{S}) = \prod_{t=1}^{T} p(s_t|s_{t-1}) p(\boldsymbol{z}_t|s_t) p(\boldsymbol{x}_t|\boldsymbol{z}_t) p(\boldsymbol{y}_t|\boldsymbol{z}_t), \quad (1)$$

where $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ ($\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{S}$ 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 $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ at frame t are assumed to be independently generated given a latent variable $\boldsymbol{z}_t$. The PDF is assumed to be a Gaussian with a mean vector given by an affine transformation of $\boldsymbol{z}_t$:

$$\begin{aligned} p(\boldsymbol{x}_t|\boldsymbol{z}_t) &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{W}_x \boldsymbol{z}_t + \overline{\boldsymbol{x}}, \boldsymbol{\Psi}_x), & (2) \\ p(\boldsymbol{y}_t|\boldsymbol{z}_t) &= \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{W}_y \boldsymbol{z}_t + \overline{\boldsymbol{y}}, \boldsymbol{\Psi}_y), & (3) \end{aligned}$$

where $\mathcal{N}(\boldsymbol{z}; \overline{\boldsymbol{z}}, \boldsymbol{\Psi})$ denotes a Gaussian PDF with mean $\overline{\boldsymbol{z}}$ and covariance matrix $\boldsymbol{\Psi}$. A latent space provides a compact representation reflecting cross-modal correlations. The PDF of latent variables given a hidden state $s_t = k$ can be modeled by using a Gaussian mixture model (GMM) as described by the equation below:

$$\begin{aligned} p(\boldsymbol{z}_t|s_t = k) &= \sum_{j=1}^{L_k} \pi_{k,j} p(\boldsymbol{z}_t|s_t = k, r_{t,k} = j), & (4) \\ p(\boldsymbol{z}_t|s_t = k, r_{t,k} = j) &= \mathcal{N}(\boldsymbol{z}_t; \overline{\boldsymbol{z}}_{k,j}, \boldsymbol{\Psi}_{k,j}), & (5) \end{aligned}$$

where L_k is the number of Gaussians of the k-th GMM, and $\overline{\boldsymbol{z}}_{k,j}$, $\boldsymbol{\Psi}_{k,j}$, $\pi_{k,j}$ are the mean vector, the covariance matrix and the mixture weight of the j-th component of the k-th GMM. simplicity, we assume the number of Gaussians to be common for all the GMMs, namely $L_k = L$.

2.2. Model training

In this framework, the parameter estimation method can be achieved by a combination of canonical correlation analysis (CCA) and Viterbi learning. It consists of 4 steps as shown in Fig. 3: (1) extracting low-level features, (2) estimating topic model parameters with probabilistic CCA [14], (3) extracting latent variables, (4) estimateing HMM parameters via Viterbi learning. See [13] for the details.

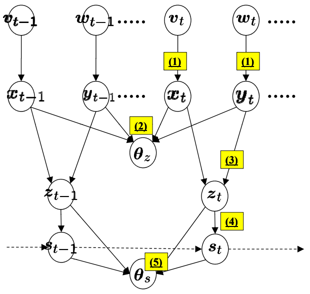


Fig. 3. Procedure for parameter estimation.

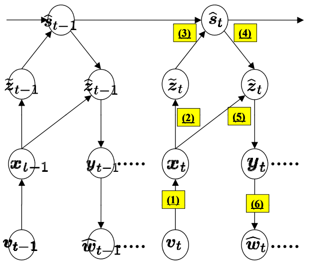

$\widehat{\boldsymbol{s}}_{t-1}$
$\widehat{\boldsymbol{s}}_t$
(3)
(4)
$\widetilde{\boldsymbol{z}}_{t-1}$
$\widehat{\boldsymbol{z}}_{t-1}$
$\widetilde{\boldsymbol{z}}_t$
$\widehat{\boldsymbol{z}}_t$
(2)
(5)
$\boldsymbol{x}_{t-1}$
$\boldsymbol{y}_{t-1}$.....
$\boldsymbol{x}_t$
$\boldsymbol{y}_t$.....
(1)
(6)
$\boldsymbol{v}_{t-1}$
$\widehat{\boldsymbol{w}}_{t-1}$.....
$\boldsymbol{v}_t$
$\widehat{\boldsymbol{w}}_t$.....

Fig. 4. Procedure of semantic indexing.

2.3. Semantic indexing and retrieval

Semantic indexing can be considered as a process of label estimation from video features only. It consists of six steps as shown in Fig. 4. (1) extracting low-level features from video frames, (2) obtaining latent variables only from video features, (3) estimating hidden states by Viterbi decoding, (4) re-estimating the latent variable considering low-level features and hidden states, (5) estimating label features from the re-estimated latent variables, (6) emitting the indexing result. For bravity, we will describe only the 4th and 5th steps.

Latent variables $\widehat{\boldsymbol{z}}_t$ are re-estimated by considering the estimated hidden state s_t obtained in the 3rd step and the low-level feature $\boldsymbol{x}_t$. The PDF of a latent variable $\boldsymbol{z}_t$ is given by the following GMM:

$$\begin{aligned} \widehat{\boldsymbol{z}}_t(\boldsymbol{x}_t, s_t) &= \sum_{j=1}^{L} \widetilde{\pi}_j(\boldsymbol{x}_t, s_t) \overline{\boldsymbol{z}}_{s(t),j}, & (6) \\ \widetilde{\pi}_j &= \frac{\pi_i \mathcal{N}(\boldsymbol{z}(\boldsymbol{x}_t); \overline{\boldsymbol{z}}_i, \boldsymbol{\psi}_i)}{\sum_{j=1}^{L} \pi_j \mathcal{N}(\boldsymbol{z}(\boldsymbol{x}_t); \overline{\boldsymbol{z}}ji, \boldsymbol{\psi}_j)}. & (7) \end{aligned}$$

Label features $\widehat{\boldsymbol{y}}_t$ can be estimated with the re-estimated latent variables $\widehat{\boldsymbol{z}}_t$ with the the framework of PCCA.

$$\widehat{\boldsymbol{y}}_t = \boldsymbol{y}(\widehat{\boldsymbol{z}}_t) = \boldsymbol{W}_y \widehat{\boldsymbol{z}}_t + \overline{\boldsymbol{y}}. \quad (8)$$

Note that HTTM is symmetric among low-level features, which implies that we can utilize almost the same approach also to KIS task.

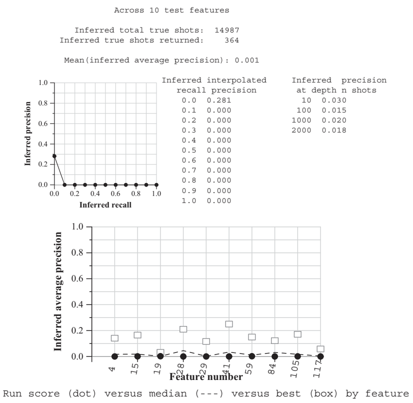
3. TRECVID2010 SUBMISSIONS

The results obtaind for the semantic indexing task and the known-item search task are shown in Figs. 5, 6, and 7.

4. ADDITIONAL EXPERIMENTS

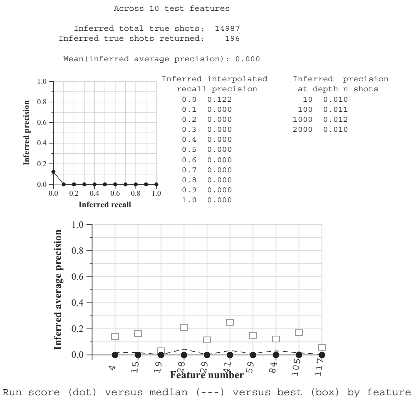
4.1. Experimental Conditions

For further analysis of performances of our method in SIN task, we tested our method with TRECVID 2005 data including 127 videos and 56191 shots. We divide them into two sets,



```
            Across 10 test features

     Inferred total true shots:  14987
   Inferred true shots returned:    364

    Mean(inferred average precision): 0.001

                Inferred interpolated     Inferred  precision
                  recall precision         at depth n shots
                    0.0  0.281              10  0.030
                    0.1  0.000             100  0.015
                    0.2  0.000            1000  0.020
                    0.3  0.000            2000  0.018
                    0.4  0.000
                    0.5  0.000
                    0.6  0.000
                    0.7  0.000
                    0.8  0.000
                    0.9  0.000
                    1.0  0.000
```

Fig. 5. Semantic indexing result ($K = 40, L = 40$)



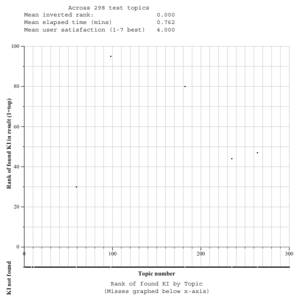
```
            Across 10 test features

     Inferred total true shots:  14987
   Inferred true shots returned:    196

    Mean(inferred average precision): 0.000

                Inferred interpolated     Inferred  precision
                  recall precision         at depth n shots
                    0.0  0.122              10  0.010
                    0.1  0.000             100  0.011
                    0.2  0.000            1000  0.012
                    0.3  0.000            2000  0.010
                    0.4  0.000
                    0.5  0.000
                    0.6  0.000
                    0.7  0.000
                    0.8  0.000
                    0.9  0.000
                    1.0  0.000
```

Fig. 6. Semantic indexing result ($K = 5, L = 50$)

one set (containing 102 videos and 45689 shots) is for model parameter estimation, , the other (containing 25 videos and 10502 shots) is for testing. Bag of Features (BoF) with SIFT local descriptors provided by vireo374 [15, 16] were used as image features. We chose 47 text tags[1] from LSCOM-Lite and LSCOM annotation [17, 18] and remove shots without any of 47 tags. We adopted word occurrence vectors weighted by idf scores as a text feature.

4.2. Results

Figure 8 shows the performance measured by mean average precision. We compared HTTM with the framewise topic model under the constraint that the number of states K=5 and mixtures L=50 were fixed. This figure indicates temporal transition is effective to enhance the performance.



```
             Across 298 test topics
Mean inverted rank:                   0.000
Mean elapsed time (mins)              0.762
Mean user satisfaction (1-7 best)     4.000
```

Fig. 7. Known-item search result ($K = 40, L = 40$)

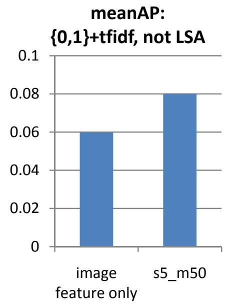


Fig. 8. Recognition results. Proposed model outperforms the model without state estimation.

In the second experiment, we fixed the number of states K and mixures L satisfying $KL = 240$ because $K = 5$ and $L = 50$ performed best in the 1st experiment 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 9 shows the results of two labels each. Figure 9 (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 9 (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.

5. CONCLUDING REMARKS

We applied our new approach based on the Hierarchical Topic Trajectory Model to the SIN and KIS tasks and analyzed some basic behaviors of the proposed method. Although the current performance is still limited, we anticipate much future work; it will include, for example, more detailed discussion regarding the pros and cons when compared with the classifier-based

[1] 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.

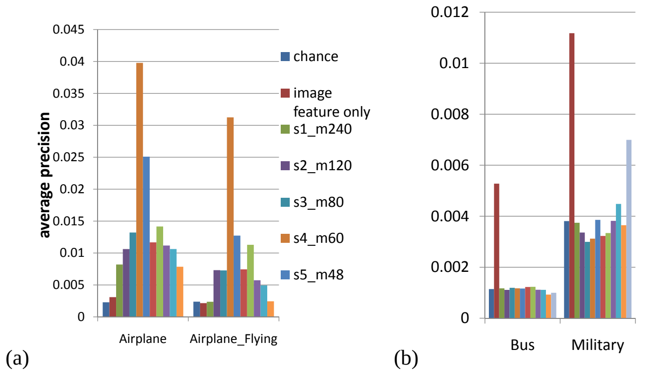


Fig. 9. Results of labels: Airplane, Airplane_Flying, Bus and Military.

methods such as support vector machine (SVM) [19, 20] and supervised multi-class learning (SML) [21] [22].

6. ACKNOWLEDGMENTS

The authors would express thanks to Prof. Takuya Nishimoto and Dr. Shigeki Miyabe at the University of Tokyo for their valuable discussions and comments, and also to Dr. Naonori Ueda, Dr. Eisaku Maeda, Dr. Futoshi Naya and Dr. Junji Yamato at NTT Communication Science Laboratories for their help and encouragement.

7. REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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, 2005.

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

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

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

[9] 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.

[10] 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, 2009.

[11] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, “Semicca: Efficient semi-supervisedllearning of canonical correlations,” in *Proc. International Conference on Pattern Recognition (ICPR)*, 2010.

[12] 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, 2005.

[13] T. Nakano, A. Kimura, H. Kameoka, S. Miyabe, S. Sagayama, N. Ono, K. Kashino, and T. Nishimoto, “Hierarchical topic trajectory model considering cross-modal correlations for video annotation retrieval,” submitted to ICASSP2011.

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

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

[16] Y.G. Jiang, J. Yang, C.W. Ngo, A. G. Hauptmann: “Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study”, *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42-53, 2010.

[17] 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.

[18] 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.

[19] 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, 2006.

[20] 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, 2007.

[21] 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, 2007.

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