Generating Representative Views of Landmarks via Scenic Theme Detection

Yi-Liang Zhao[1], Yan-Tao Zheng[2], Xiangdong Zhou[3], and Tat-Seng Chua[1]

[1] Department of Computer Science, National University of Singapore, Singapore
[2] Institute for Infocomm Research, Singapore
[3] Fudan University, China
{zhaoyl,chuats}@comp.nus.edu.sg, yantaozheng@gmail.com,
xdzhou@fudan.edu.cn

Abstract. Visual summarization of landmarks is an interesting and non-trivial task with the availability of gigantic community-contributed resources. In this work, we investigate ways to generate representative and distinctive views of landmarks by automatically discovering the underlying *Scenic Themes* (e.g. sunny, night view, snow, foggy views, etc.) via a content-based analysis. The challenge is that the task suffers from the subjectivity of the scenic theme understanding, and there is lack of prior knowledge of scenic themes understanding. In addition, the visual variations of scenic themes are results of joint effects of factors including weather, time, season, etc. To tackle the aforementioned issues, we exploit the Dirichlet Process Gaussian Mixture Model (DPGMM). The major advantages in using DPGMM is that it is fully unsupervised and do not require the number of components to be fixed beforehand, which avoids the difficulty in adjusting model complexity to avoid over-fitting. This work makes the first attempt towards generation of representative views of landmarks via scenic theme mining. Testing on seven famous world landmarks show promising results.

Keywords: Dirichlet Process, Dirichlet Process Gaussian Mixture Model, Scenic Theme Detection.

1 Introduction

The fast development and proliferation of digital photo-capture devices and the growing practice of online photo-sharing have resulted in huge online photo collections, which cover virtually everywhere on earth. The fast-growing of this huge image collection opens up many opportunities to research communities to work on more effective and efficient searching, viewing, archiving and interaction with such collections. Recently, much work aims to organize this huge collection or mine important landmarks worldwide based on the context information: geography, user, tag, etc. [5][9][19]. However, less efforts have been put into generating representative views of landmarks via recognizing the underlying scenic themes. As shown in Figure 1, a landmark may present different scenery views under different time, season and weather circumstances. Here, we define a distinct landmark scenery view as a scenic theme of the landmark. A scenic theme affords distinct vistas of the landmark with notably different aesthetic and visual

K.-T. Lee et al. (Eds.): MMM 2011, Part I, LNCS 6523, pp. 392–402, 2011.

Fig. 1. Different scenic themes of Golden Gate Bridge

perceptions. Effective scenic theme detection helps to better organize, browse and index image collections of a particular landmark. In particular, we show in this work that the summary of distinctive and representative views is generated with satisfaction on top of the detected scenic themes.

However, mining representative scenic themes from community-contributed image collections itself is a difficult task. Variations of scenic themes are the results of joint effects of seasons, weather conditions and time. Although we have the corresponding context information associated with the images most of the time, it is very difficult to model the underlying scenic themes using a combination of these meta information. In addition, scenic theme understanding requires subjective judgements. People from different background, culture and with different personal experience may have different perceptions on the same scenes. Figure 2 shows the variations of scenic views of The Eiffel Tower in a cloudy or overcast day. People may judge that it is either cloudy or overcast based on their own experience and background. In addition, we find that landmarks in different parts of the world often have different sets of distinctive scenic themes. For example, in tropical area, there is probably no snowy scenes but lots of beautiful night views whereas in subpolar zones, we can find nice snowscapes. After all, we do not have a fixed list of scenic themes, which makes automatic detection and generation of scenic themes summary a necessity. To tackle this problem, we adopt a generative probabilistic approach to visually model the scenic themes of a landmark. The rational is that scenic theme of a landmark is a cause-effect process and the generative model is expected to capture the underlying rules of producing different landmark sceneries.

Fig. 2. Scene-variation views of The Eiffel Tower in a cloudy or overcast day

The task of mining distinct scenic themes is formulated as a clustering problem in this paper. Traditional probabilistic models are often used throughout machine learning to model the distributions over observed data, but they tend to suffer from either under-fitting or over-fitting. The choice of using bayesian nonparametric approach resolves both under-fitting by using a model with an unbounded complexity and over-fitting by approximating the full posterior over parameters. Dirichlet Process (DP) is currently one of the most popular nonparametric bayesian model [7][2][6][12][14] due to its simplicity and efficiency. Here we show that the Gaussian mixtures whose parameters are generated by DP is able to create satisfactory view summaries with distinctive scenic themes.

In summary, the contribution of this paper is: we demonstrate a novel attempt in generating view summaries via underlying scenic themes detection on community-contributed images by using DPGMM. Testing on the seven landmarks shows that the proposed approach delivers promising results. The remaining parts of the paper is organized as follows. In Section 2, we review the related work. Formulation of the problem together with the proposed model are presented in Section 3. We present the experiments in Section 4 followed by the conclusion in Section 5.

2 Related Work

We review two areas, which are most related to our work: (1) Visual scene understanding; (2) Visual summarization through Landmark mining from community-contributed collections.

The use of global features for scene classification dates back in 1998, when Martin, et al. [18] showed how global features can be used to model each scene as an individual object for classification. Their approaches, however, are normally only used to classify a small number of scene categories. Recently, probabilistic models are used extensively in visual scenes categorization in the computer vision literature [10][17][4][8]. Fei-Fei, et al. [10] proposed a bayesian hierarchical model for learning natural scene categories based on Latent Dirichlet Allocation (LDA). However, a fixed list of categories are readily available in their work while we need to tackle the problem of variable number of scenic themes for different landmarks.

Visual summarization through landmark mining from online community-contributions is a recent trend [3][9][1][19][16]. Lyndon, et al. [9] proposed a way to generate representative views of important landmarks based on both context and content information. The statistical approaches adopted by them showed effectiveness in aggregating the representative views of landmarks in San Francisco area. However, statistical approach needs a highly accurate and sufficiently large dataset. Simon, et al. [16] worked on finding a set of canonical views to summarize a visual scene. Their work aimed at constructing a guidebook which contains a summary on representative views of large landmarks. Comparing to their work, ours focuses on generating a representative view summarization distinguished by the underlying scenic themes, which is more difficult due to its subjectivity and uncertainty. Zheng, et al. [19] built a landmark recognition engine in modeling and recognizing landmarks at world-scale level. The graph clustering result however only shows strong visual correlations between the traditional representative views of landmarks. While these work did not look into ways to organize the images based on scenic themes, the approaches adopted in mining representative landmarks, however, can be used in the preprocessing steps of our work in generating clearer subsets of the collections.

3 Our Approach

In this section, we formally define the problem and elaborate on details of the proposed approach.

3.1 Problem Formulation

Let $\boldsymbol{y}$ denote the scenery or visual appearance of a landmark. $\boldsymbol{y}$ is the consequence of joint effects of factors $\boldsymbol{q}$, like weather, season, lighting, etc, with markov condition $\boldsymbol{q} \rightarrow \boldsymbol{y}$. To model this cause-effect process is, however, a challenge, as the causality relationship is nondeterministic and not tractable with a finite number of rules and parameters. To simplify the modeling, we introduce an intermediate variable $\boldsymbol{t}$, i.e., scenic theme. In the new cause-effect process, scenic theme is a result of joint effects of factors like weather, season, etc, while landmark scenery (visual appearance) is an observation conditioned on scenic theme only. The markov condition now becomes $\boldsymbol{q} \rightarrow \boldsymbol{t} \rightarrow \boldsymbol{y}$. Scenic theme

here corresponds to a distribution of random variable $\boldsymbol{y}$ of landmark visual appearances. Intuitively, a scenic theme captures the characteristics of landmark scenery appearances from certain aspect. For example, day view and night view could be examples of scenic theme, while day view can be subdivided further. As our goal is to generate visual summarization of landmarks, we focus only on the second half of the cause-effect process, i.e., $\boldsymbol{t} \rightarrow \boldsymbol{y}$ and model it using a generative probabilistic model.To avoid suffering from the problem of over-fitting or under-fitting associated with traditional parametric models when there is a mismatch between the complexity of the model and the amount of the data available, we propose to use the bayesian nonparametric approach for the clustering problem. According to Rasmussen et al. [15], reasonable and proper bayesian methods do not have over-fitting problem as the number of latent variables does not grow with the number of mixture components in the model inference. Formally, given observations: $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}$ with $\boldsymbol{y}_i \sim G$ drawn from some unknown distribution G, we first place a prior over G then compute the posterior over G given the observations. In this case, we use Dirichlet Process Gaussian Mixture Model (DPGMM) as the prior distribution of the model since DP has a tractable posterior distribution [7][14]. The nonparametric nature of the DP translates to mixture models with a countably infinite number of components. Use of countably infinite gaussian mixtures bypasses the need to determine the "correct" number of components in a finite mixture model, which is technically much more difficult.

In generating view summaries based on the detected scenic themes, we seek to find the most distinct and representative scenic theme clusters. The detected scenic themes are each modeled as a probability distribution; in this case, each gaussian component represents a distinct detected scenic themes. We then seek to find the differences between each detected scenic themes by calculating the Kullback-Leibler (KL) Divergence between each corresponding probability distributions. The KL Divergence between two Gaussian distributions: $\mathcal{N}_i(\boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}_j(\boldsymbol{\mu_j}, \boldsymbol{\Sigma}_j)$ is:

$$D_{KL}(\mathcal{N}_i||\mathcal{N}_j) = \frac{1}{2}(\ln(\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|}) + tr(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) - N) \quad (1)$$

For each landmark, we rank the detected scenic themes according to the sum of KL Divergence values with respect to other theme clusters and select the top ten themes to form the visual summary of the landmark. The overall framework is presented in Figure 3.

3.2 Dirichlet Process Gaussian Mixture Model (DPGMM)

As one of the bayesian non-parametric model, DPGMM does not require the number of Gaussian components to be fixed in advance. Instead, the number of components is determined by the model and data in the subsequent inferences. The infinite DPGMM for scenic themes detection is defined as follows: we model a set of observations: $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}$ using a set of latent parameters $\{(\boldsymbol{\mu}_1, \mathbf{S}_1), ..., (\boldsymbol{\mu}_n, \mathbf{S}_n)\}$, where $\boldsymbol{\mu}_i$ are the means, $\mathbf{S}_i$ are the precisions (inverse

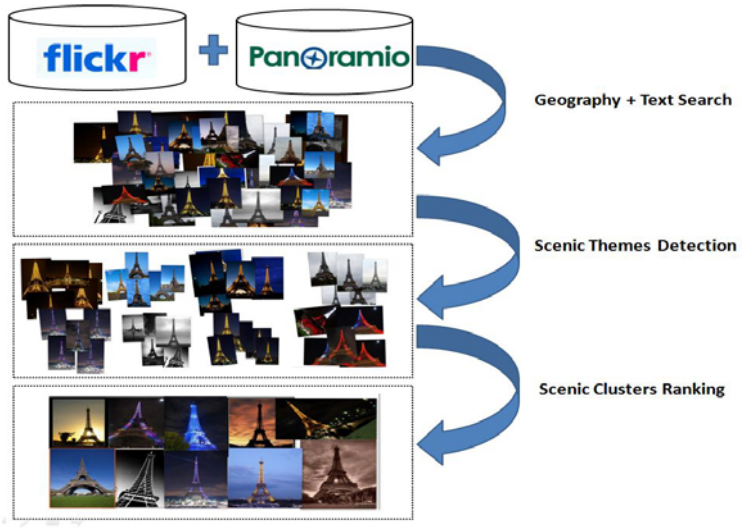

Fig. 3. Overall Framework

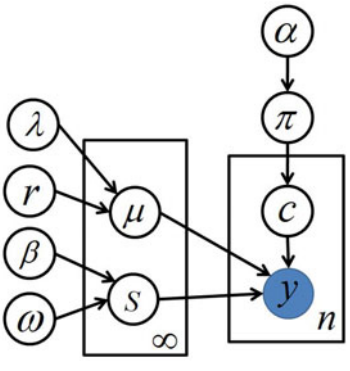

Fig. 4. Dirichlet Process Gaussian Mixture Model: y is the observation and c is the class label

variances), both of which are generated by a DP which can be seen as a distribution over the parameters of other distributions. In this case, each $\boldsymbol{\mu}_i$ is drawn independently and identically from another Gaussian distribution parameterized by the hyperparameters: $\boldsymbol{\lambda}$ and $\boldsymbol{r}$ and each $\mathbf{S}_i$ is drawn independently and identically from a Gamma distribution parameterized by $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. Together, $\boldsymbol{S}$ and $\boldsymbol{\mu}$ form the scenic theme intermediate variable $\boldsymbol{t}$ as introduced in Section 3.1. Let $\boldsymbol{c}_i$ be a cluster assignment variable, which takes on value k with probability π_k. Thus each $\boldsymbol{y_i}$ has distribution $p(\boldsymbol{y}|\boldsymbol{\mu}, \mathbf{S}, \boldsymbol{c})$ parameterized by $\boldsymbol{\mu}, \mathbf{S}$ and $\boldsymbol{c}$. The graphical representation of this model is presented in Figure 4 and is defined as follows:

$$\begin{aligned}
&\boldsymbol{\pi}|\alpha \sim Dirichlet(\tfrac{\alpha}{k}, ..., \tfrac{\alpha}{k}) = \tfrac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{k})^k} \textstyle\prod_{j=1}^{k} \pi_j^{\frac{\alpha}{k}-1} \\
&\boldsymbol{c}|\boldsymbol{\pi} \sim Multinomial(\pi) = \textstyle\prod_{j=1}^{k} \pi_j^{n_j} \\
&\boldsymbol{\mu}|\boldsymbol{\lambda}, \boldsymbol{r} \sim \mathcal{N}(\boldsymbol{\lambda}, \boldsymbol{r}^{-1}) \\
&\mathbf{S}|\boldsymbol{\beta}, \boldsymbol{\omega} \sim Gamma(\boldsymbol{\beta}, \boldsymbol{\omega}^{-1}) \\
&p(\boldsymbol{y}|\boldsymbol{\mu}, \mathbf{S}, \boldsymbol{c}) = \textstyle\sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{S}_j^{-1})
\end{aligned} \tag{2}$$

In equation (2), n_j is the number of images belonging to the jth scenery cluster. The indicator variables $\boldsymbol{c}$ is introduced to encode which class has generated the observation so that the inference is possible using finite amounts of computation with the maximum number of components not exceeding the number of observations. Observations: $\boldsymbol{y}_i$ are color histogram feature vectors in the current setting. The mixing proportions $\boldsymbol{\pi}$ are positive and sum to one. From equation (2), we can write the prior directly in terms of the indicators by integrating out the mixing proportions:

$$\begin{aligned}
p(c_1, ..., c_k|\alpha) &= \int p(c_1, ..., c_k|\pi_1, ..., \pi_k) d\pi_1 \cdots d\pi_k \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_j + \alpha/k)}{\Gamma(\alpha/k)}
\end{aligned} \tag{3}$$

With the analytical tractability of equation (3), we can now work directly with the finite number of indicator variables, rather than the infinite number of mixing proportions. We use Markov Chain which relies on Gibbs Sampling for inference in this work [13]. Each variable is updated by sampling from its posterior distribution conditional on all other variables. We repeatedly sample the parameters, hyperparameters and indicator variables from their posterior distributions conditioned on all other variables: (1) sample parameters conditioned on the indicators and hyperparameters; (2) sample hyperparameters conditioned on the parameters; (3) update each indicator, conditioned on other indicators and hyperparameters; and (4) update Dirichlet process concentration parameter α. The process will repeat until termination condition is met. In this paper, we set the maximum number of iteration to be $5,000$ for each landmark.

4 Experiment

The experiments are performed on seven worldwide famous landmarks: The Eiffel Tower, Golden Gate Bridge, The Great Sphinx, Notre Dame, Leaning Tower Pisa, Statue of Liberty and Basil Cathedral. To make the dataset more complete, we crawl data from two online collections: Flickr[1] and Paronamia[2]. Comparing with Panoramio, which is a geolocation-oriented photo sharing website, Flickr has much more user contribution while Panoramio images have more accurate geographical locations. Using public APIs provided by both web services, we

[1] `http://www.flickr.com`
[2] `http://www.panoramio.com/`

specify a restricted bounding box on geographical locations as well as key words related to corresponding landmarks. After cleaning the data, we have an average of 806 images of each landmark. In the scenic theme detection stage, we exploit global features in the work due to its capability in producing compact representations of images. We did a comparison test with different global features related to the color distributions: Color Histogram, Color Moment and Color Correlogram. We found that color histogram gives the best results in generating the most coherent scenic theme classes. After feature extraction, we normalize the feature vectors such that the sum of each element equals to a constant: $\sum_k y_{ik} = \sum_k y_{jk} = a$, for every $i \neq j$. We then adopt Principle Component Analysis (PCA) approach to reduce the dimensionality of the feature vectors such that: $d_{new} = \lceil b \times \sqrt{n} \rceil$, where d_{new} is the reduced dimension and n is the number of images in the dataset for that particular landmark. Empirically, we choose $b = 2.5$ in the current setting because it produces better results compared to other values. After sampling from the posterior distribution of the DPGMM according to the Markov Chain inference procedure described in Section 3, we obtained an average of 15 scenic themes for each landmarks. Some detected scenic themes of The Eiffel Tower are presented in Figure 5. Finally, to give a view summarization with the most interesting and representative scenic themes for each landmark, we exploit the calculated probability measures of each mixture model and measure the pair-wise KL Divergences according to equation (1). We rank the detected scenic themes according to the sum of KL Divergence values with respect to other theme clusters and select the top ten themes to form the visual summary of each landmark. We then calculate the probability

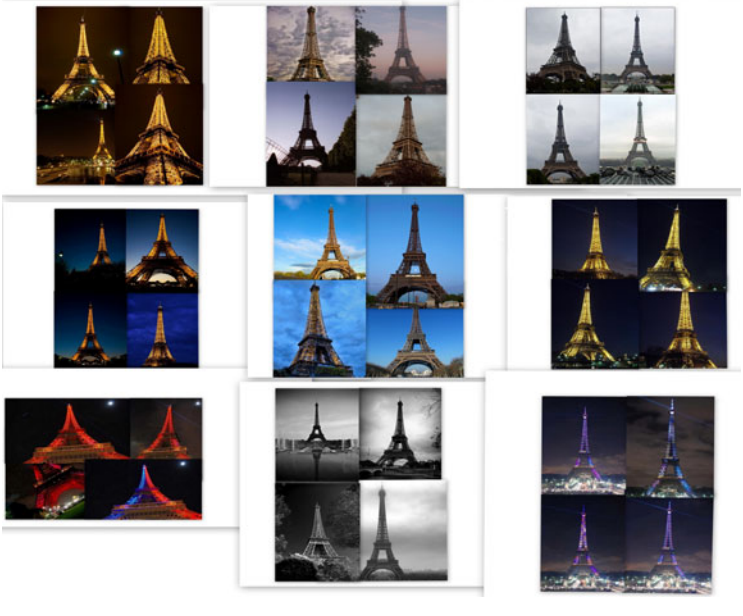

Fig. 5. Scenic themes mined for The Eiffel Tower

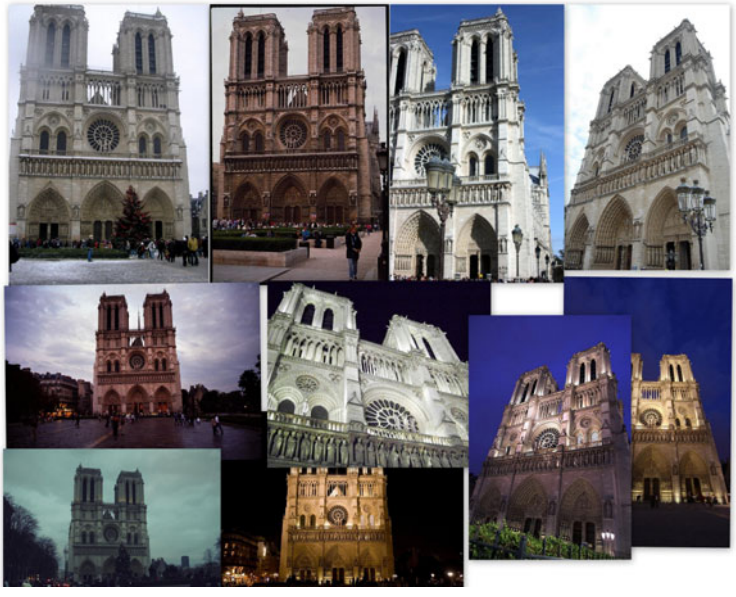

Fig. 6. View summarization with representative scenes detected for Notre Dame

of images in each selected theme clusters and choose the one with the highest probability to represent the theme. Figure 6 shows the view summary of Notre Dame generated by our approach.

To evaluate the correctness and representativeness of the mined scenic themes, we have conducted a user study, where a group of users are selected to judge the mined scenic themes for each landmarks. The judgments are based on four criteria: (1) the level of consistency of the scenic themes mined for each landmark; (2) the level of completeness of the generated scenic themes summary; (3) the level of redundancy of the summary and (4) how satisfying are the summaries? We randomly selected twenty users: seven from Singapore; others from Shanghai, China. The evaluation result is depicted in Figure 7. The results for each question are averaged over all users for each landmark. The average satisfaction score is 7.97. We observe that Statue of Liberty and Basil Cathedral obtained the best scores while The Great Sphinx does not perform well in terms of the consistency level. The reasons could be: (1) there are much fewer distinct scenic themes The Great Sphinx has as compared to that of The Statue of Liberty and Basil Cathedral; and (2) the color distribution of The Great Sphinx is very similar to most of the background (i.e. the desert). To mitigate this problem, we could look into extracting regions mostly related to the scenic themes by using camera calibration technique and incorporating spatial information. In addition, Basil Cathedral scored lowest in both uniqueness and completeness measures. We attribute the low scores to the mismatch between the availability of sufficient data and the expectations of more nice views from the users. In summary, our

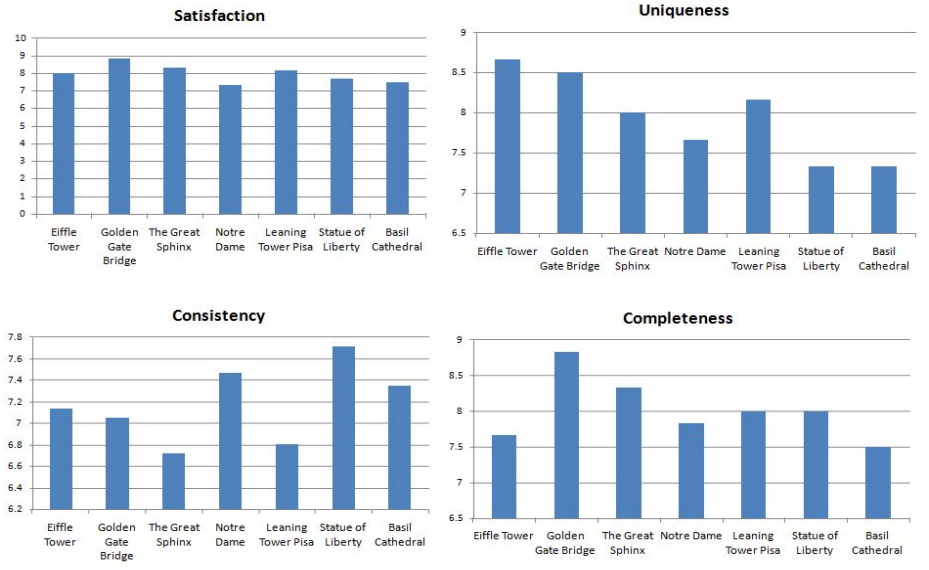

Fig. 7. Average scores for each of the four evaluation questions

proposed model yields satisfactory results and more efforts are required to tackle the problems caused by data sparsity and noises contributed by various factors.

5 Conclusion

We have presented a novel attempt to use Dirichlet Process Gaussian Mixture Model (DPGMM) to generate visual summarization of landmarks via scenic theme detection. Our approach shows promising results in generating satisfactory scenic themes for most of the landmarks. A user study is done with good response in terms of the representativeness and coherence of the scenic theme clusters. In the future, we shall look into building applications which allow users to browse image collections organized by distinct scenery views. In addition, contextual information could be exploited to boost the performance when more accurate meta information and richer web services are available [11].

References

1. Ahern, S., Naaman, M., Nair, R., Yang, J.H.-I.: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: JCDL, pp. 1–10 (2007)
2. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The annals of statistics 2(6), 1152–1174 (1974)
3. Berg, T.L., Forsyth, D.A.: Automatic ranking of iconic images. Technical report (2007)

4. Zisserman, A., Bosch, A., Munoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
5. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: WWW, pp. 761–770 (2009)
6. Escobar, M.D., West, M.: Bayesian Density Estimation and Inference Using Mixtures.. Journal of the american statistical association 90(430) (1995)
7. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. The annals of statistics 1(2), 209–230 (1973)
8. Fritz, M., Schiele, B.: Decomposition, discovery and detection of visual categories using topic models. In: CVPR, vol. 0, pp. 1–8 (2008)
9. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: WWW, pp. 297–306 (2008)
10. Li, F.-F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR, vol. 2, pp. 524–531 (2005)
11. Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., Paepcke, A.: Context data in geo-referenced digital photo collections. In: MM, pp. 196–203 (2004)
12. Neal, R.M.: Bayesian mixture modeling. In: Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, pp. 197–211 (1991)
13. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics 9(2), 249–265 (2000)
14. Rasmussen, C.E.: The infinite Gaussian mixture model. Advances in neural information processing systems 12, 554–560 (2000)
15. Rasmussen, C.E., Ghahramani, Z.: Occam razor. In: Advances in neural information processing systems 13: Proceedings of the 2000 Conference, p. 294. The MIT Press, Cambridge (2001)
16. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: ICCV, pp. 1–8. Citeseer (2007)
17. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Learning hierarchical models of scenes, objects, and parts. In: ICCV, vol. 2, pp. 1331–1338 (2005)
18. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Proceedings of 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 42–51 (1998)
19. Zheng, Y.-T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.-S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: CVPR (June 2009)