Describing Geographical Characteristics with Social Images

Huangjie Zheng[1,2](✉), Jiangchao Yao[2], and Ya Zhang[2]

[1] SJTU-ParisTech Elite Institute of Technology, Shanghai Jiao Tong University, Shanghai, China
[2] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China
{zhj865265,sunarker,ya_zhang}@sjtu.edu.cn

Abstract. Images play important roles in providing comprehensive understanding of our physical world. When thinking of a tourist city, one can immediately imagine pictures of its famous attractions. With the boom of social images, we attempt to explore the possibility of describing geographical characteristics of different regions. We here propose a Geographical Latent Attribute Model (GLAM) to mine regional characteristics from social images, which is expected to provide a comprehensive view of the regions. The model assumes that a geographical region consists of different "attributes" (e.g., infrastructures, attractions, events and activities) and "attributes" are interpreted by different image "clusters". Both "attributes" and image "clusters" are modeled as latent variables. The experimental analysis on a collection of 2.5M Flickr photos regarding Chinese provinces and cities has shown that the proposed model is promising in describing regional characteristics. Moreover, we demonstrate the usefulness of the proposed model for place recommendation.

Keywords: Geographic characteristics · Recommender systems · Latent variable models · Region description

1 Introduction

Geotagged images are pervasive, and they also provide an intuitive and objective view of our life. Thanks to these properties, images can easily reflect personal, regional, even social characteristics, and plenty of research works have been conducted with social images to facilitate people's life. Geographical analysis from social media has been widely investigated in the recent years. While most of existing studies focus their analysis on landmarks with the assumption that they are representative to regions [1–4], other perspectives such as local festivals

Y. Zhang—The work is partially supported by the High Technology Research and Development Program of China 2015AA015801, NSFC 61521062, STCSM 12DZ2272600.

L. Amsaleg et al. (Eds.): MMM 2017, Part I, LNCS 10132, pp. 115–126, 2017.
DOI: 10.1007/978-3-319-51811-4_10

and events could also be essential for profiling a region. We thus study the problem of forming comprehensive description of geographical characteristics from social media. With the description of geographical characteristics in one specific region, we could better recognize this region and boost a number of utilities such as tourist advertising, etc.

While some existing applications such as tourist recommendation and location retrieval could also extend to this problem [5–8], they mainly rely on the textual information, e.g., social tags. To our best knowledge, geotagged photos help understand intuitively a specific region and it can boost plenty of applications in several domains. For example, it is interesting that systems could generate a recommendation based on its understanding of images, which leads us free from taking effort to find a proper word for the description of the region. Therefore, since the goal is to understand a region from images, the challenge lies in how to map low level visual features to semantic characteristics.

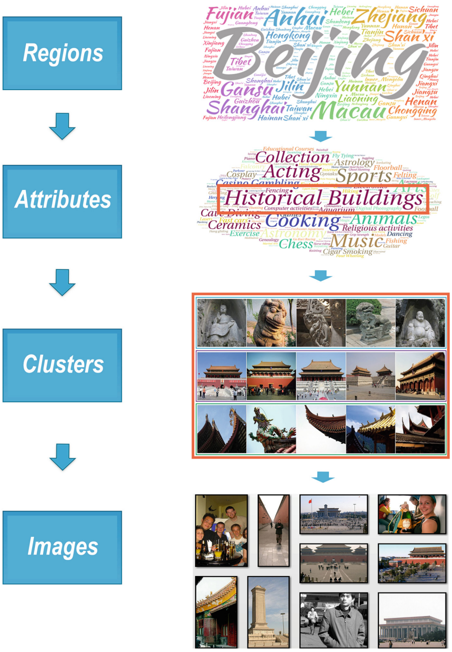


Fig. 1. Motivation of the model. We assume that in every geographic area, people's life consists of several aspects, e.g. sports, music, etc. These aspects could be presented by several clusters, while clusters are formed by vast images.

In this paper, we propose a Geographical Latent Attribute Model (GLAM) to learn geographical characteristics from photo collections. We assume that each region consists of some latent "attributes" (considered as characteristics) and each "attribute" consists of image "clusters". The motivation of our model is illustrated in Fig. 1 using Beijing as an example. A city may be described by several aspects (e.g., historical buildings), and each aspect includes different image clusters (e.g., antiques, temples, sculptures). These clusters are summarized from images taken in Beijing. Following the idea of the generative model, we introduce corresponding latent variables to formalize this procedure. By learning the latent parameters, a comprehensive view about geographical regions is formed.

The major contributions of this paper could be summarized as follows:

- We propose a Geographical Latent Attribute Model (GLAM) to learn geographical characteristics from photo collections without utilizing any textual information.
- We validate the proposed model with 2.5M Flickr photos taken in China to demonstrate its effectiveness in both qualitative and quantitative ways.
- As one of the potential applications, a region recommendation strategy is proposed based on the similarity between region's characteristics and user's interest according to his/her photo album.

The rest of paper is organized as follows: In Sect. 2, we review the related work. Section 3 explains our model and its inference technique. The experiment results will be displayed in Sect. 4, and we conclude our paper in Sect. 5.

2 Related Work

Plenty of works have been conducted in geographical analysis. Ji, et al. [2] propose a hierarchical structure to mine city landmarks from view, scene and city layers. [9] analyzes the attribute at region level for region exploration and [10] handles the urban understanding with CNN. Livia Hollenstein and Ross S. Purves [11,12] focus on social media to find out how people generate their understanding for a city. Similarly, [1] extract the tags representing landmarks to better present and extract view of one region. In [3,4], the authors find the popular landmarks using mean shift.

This work is also related to several applications such as location retrieval, tourist recommendation, etc. [5] shows the same viewpoint that users are more interested in a geographic area than the precise GPS coordinate. Our work thus pay more effort into recommending users with a proper geographic area rather than location estimation with exact geographic coordinates. [6,7] give personalized tourist recommendation based on users' interest and their similarity, while our work focus more on the similarity between user's interest and geographic characteristics.

3 Model

3.1 Geographical Latent Attribute Model

The plate notation of GLAM is illustrated in Fig. 2. Assuming that we have M regions and each region has N_m images, we target to learn the regional attribute distributions $\{\theta_m\}_{m=1,...,M}$ from these images. We first use GoogLeNet to extract one D dimensional feature vector v_{mn} for each image. Then our problem could be formalized to learn $\{\theta_m\}_{m=1,...,M}$ from the feature collection $\{v_{11},...,v_{MN_M}\}$.

We transform this problem into a generative procedure and consider that each region has a distribution over characteristics and each characteristic has a distribution over clusters which are modeled by a series of Gaussian mixtures. Both "characteristic" and "cluster" are introduced as latent variables in this hierarchical structure and could be inferred by the observed variables $\{v_{11},...,v_{MN_M}\}$. The generative procedure is summarized as follows:

- Choose regional characteristic proportion $\theta_m \sim Dir(\alpha)$.
- Choose the characteristic of one image $i_{mn} \sim Multinomial(\theta_m)$.
- Choose the cluster $z_{mn} \sim Multinomial(\phi_{i_{mn}})$, where $i_{mn} \in \{1,2,...,K\}$.
- Choose each visual vector $v_{mn} \sim \mathcal{N}(\mu_{z_{mn}}, \sigma_{z_{mn}}\mathbf{I})$, where $z_{mn} \in \{1,2,...,K'\}$.

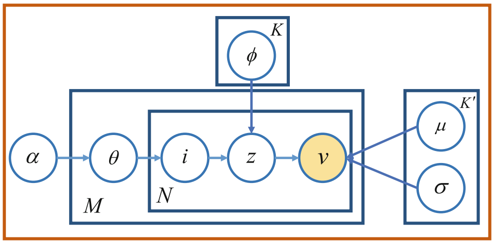


Fig. 2. The plate notation of GLAM

In our model, $\{(\mu_{k'}, \sigma_{k'})\}_{k'=1,...,K'}$ constitute the visual space and $\{\Phi_k\}_{k=1,...,K}$ are used to capture the characteristic-cluster distributions. Latent variables z_{mn} and i_{mn} are decided by v_{mn} and reversely affect the regional characteristic distribution θ_m. In short, we use a topic model structure to learn the high level concepts at the top layer and facilitate Gaussian mixture model to cluster low level visual features at the bottom layer.

3.2 Inference and Learning

In this part, we present our inference algorithm. The key inferential problem of our model is to compute the posterior distribution of latent variables given data as Eq. 1.

$$p(\theta, i, z|\alpha, \phi, \mu, \sigma, v) = \frac{p(\theta, i, z, v|\alpha, \phi, \mu, \sigma)}{p(v|\alpha, \phi, \mu, \sigma)} \tag{1}$$

Above equation is intractable due to the non-integrable denominator and an alternative method, e.g., Gibbs sampling or variational approximation [13], could be employed. In this paper, we adopt a mean field variational bayes method [14] (variational EM) to deal with our model. Following its methodology, we assume that the variational distribution is defined as

$$q(\theta, i, z) = q(\theta|\gamma)q(i|\psi)q(z|\varPhi), \tag{2}$$

where γ is the Dirichlet parameter and ψ, $\varPhi$ are the multinomial parameters. With this specification, the latent variables could be approximated by minimizing the Kullback-Leibler (KL) divergence between Eqs. 1 and 2.

$$\arg\min_{(\gamma,\psi,\varPhi)} D(q(\theta, \psi, \varPhi)|p(\theta, \psi, \varPhi)) \tag{3}$$

By setting the derivative of free parameters γ, ψ, $\varPhi$ in Eq. 3 to zero, we obtain the following equations.

$$\varPhi_{mnk'} \propto \exp(\sum_k \psi_{ijk} \log \varPhi_{kk'})\mathcal{N}(v_{ij}|\mu_{k'}, \sigma_{k'}) \tag{4}$$

$$\psi_{ijk} \propto \exp(\varPsi(\gamma_{ik})) \exp(\sum_{k'} \varPhi_{ijk'} \log \phi_{kk_\prime}) \tag{5}$$

$$\gamma_{ik} = \alpha_k + \sum_j \psi_{ijk} \tag{6}$$

The most frequent approach to estimate the model parameters is maximizing the likelihood of observed variables, i.e., $p(v|\alpha, \phi, \mu, \sigma)$. Although there is no analytical integral for this likelihood, Jensen's inequality could be used to get an adjustable lower bound.

$$\begin{aligned}
&\ln p(v|\alpha, \phi, \mu, \sigma)) \\
&= \ln \int_\theta \sum_{i,z} p(v, \theta, i, z|\alpha, \phi, \mu, \sigma)d\theta \\
&= \ln \int_\theta \sum_{i,z} \frac{p(v, \theta, i, z|\alpha, \phi, \mu, \sigma)q(\theta, i, z)}{q(\theta, i, z)} d\theta \\
&\geqslant E_q(\ln p(v, \theta, i, z|\alpha, \phi, \mu, \sigma)) - E_q(\ln q(\theta, i, z)) \\
&\triangleq L(\alpha, \phi, \mu, \sigma)
\end{aligned} \tag{7}$$

With previous optimal free parameters γ, ψ, $\varPhi$, we could maximize the lower bound L by setting the derivatives to zero with respect to the parameters ϕ, μ, σ respectively. Then, we have following solutions:

$$\phi_{kk'} \propto \sum_i \sum_j \psi_{ijk} \varPhi_{ijk'} \tag{8}$$

$$\mu_{k'} = \frac{\sum_i \sum_j \varPhi_{ijk'} v_{ij}}{\sum_i \sum_j \varPhi_{ijk'}} \tag{9}$$

$$\sigma_{k'} = \frac{\sum_i \sum_j \Phi_{ijk'}(\mu'_k - v_{ij})^{\mathrm{T}}(\mu'_k - v_{ij})}{D \sum_i \sum_j \Phi_{ijk'}} \tag{10}$$

And for Dirichlet prior α, we use Newton-Raphson method to update it like LDA [15]. Iterating the inference and parameter estimation procedure, we would gradually acquire the solution of our model.

4 Experimental Results

To validate GLAM for geographical analysis, we evaluate it on a Flickr dataset of 2.5M photos in both qualitative and quantitative ways. In addition, we show its potential to retrieve the regions of interest.

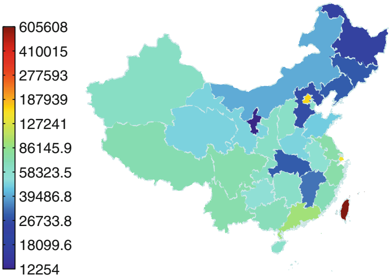


Fig. 3. The color map of data distribution in China. The warmer the color is, the more images are taken there. Taiwan possesses the most amount of data, while Ningxia possesses the least. The average amount in each province is about 85K.

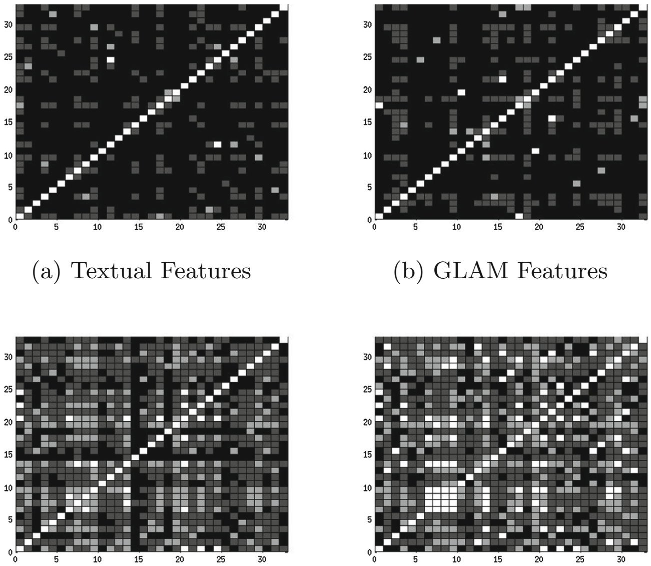
4.1 Experimental Settings

We crawled 6.5M photos that had the GPS information in the YFCC100M dataset [16]. Then with the database of GADM[1], which is a database containing the boundary geo-coordinates of each administration region, we filter out the photos not taken in China and the 2.5M remaining photos are divided into 34 groups according to the administration regions as shown in Fig. 3. One feature vector is extracted for each image from the dropout layer (the second last layer) of GoogLeNet [17].

4.2 Quantitative Evaluation

In this section, we provide a quantitative evaluation for our GLAM model. The GLAM aims to find a better description for regions based on social images. As we know, textual content is good at delivering semantic information. Thus,

[1] https://www.gadm.org/.



(a) Textual Features

(b) GLAM Features

(c) GMM Features

(d) Average CNNs Visual Features

Fig. 4. Region's similarity computed with different features. We can observe the results of text feature and our model are quite coherent, while the results of the others are difficult to determine the similar regions. Presented with $n = 20$, $K = 15$, $K' = 500$.

we employ the documents from the online tour guide "TravelChinaGuide"[2], the largest and most authoritative online tour operator in China, for comparison. Each document covers general introduction, facts, even life details for each region. We build topic models with LDA [15] from the textual document. The Euclidean distance between regions is computed based on the learned topic model. Similarly, we compute the distance between regions based on visual features learned by GLAM, Gaussian Mixture Model (GMM), and average visual features extracted directly from GoogLeNet. The corresponding distance matrix are shown in Fig. 4, where brighter colors mean higher similarity. It can be seen that our model presents more similar results as textual features, suggesting that our model generates a better semantic description for regions.

To test the effectiveness of our model, we employ the Kernel Canonical Correlation Analysis (KCCA) to compute the correlation between the distance matrix obtained from the textual feature and the other three types of visual features. As shown in Table 1, from textual feature we learn respectively 5, 10, 15, 20, 25 and 30 topics. Meanwhile, GLAM is severally trained with 200 and 500 clusters, and the number of characteristics K is set to 10, 15 and 20 respectively in the experiments. Distance matrix built from GMM and average visual features lead

[2] https://www.travelchinaguide.com/.

Table 1. Comparing the correlation between ground truth and the three types of features.

	$GLAM_{K'=200}$			GMM	$GLAM_{K'=500}$			GMM	θ_{avg}
	θ_{10}	θ_{15}	θ_{20}	$K' = 200$	θ_{10}	θ_{15}	θ_{20}	$K' = 500$	
$Text_{5topics}$	0.5548	0.5945	0.5835	0.3904	0.5910	0.6010	0.6192	0.3912	0.3484
$Text_{10topics}$	0.6191	0.6515	0.6568	0.3920	0.6310	0.6571	0.6780	0.4040	0.3726
$Text_{15topics}$	0.6764	0.7414	0.7251	0.4304	0.7021	0.7827	0.7574	0.4467	0.4038
$Text_{20topics}$	0.7550	0.8014	0.7842	0.5064	0.7704	0.8212	0.8195	0.5163	0.4595
$Text_{25topics}$	0.7253	0.7843	0.7725	0.4739	0.7502	0.8130	0.7982	0.4973	0.4510
$Text_{30topics}$	0.7181	0.7838	0.7670	0.4865	0.7446	0.8056	0.7941	0.4836	0.4477

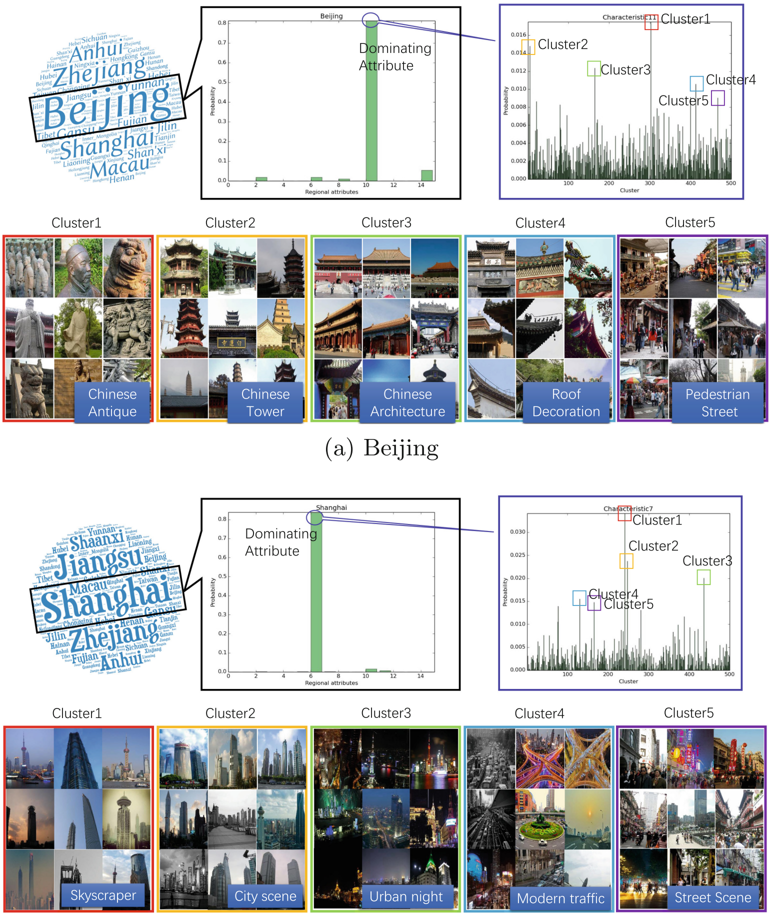
to a weak correlation to that of textual feature, with the highest correlation at 0.52 and 0.46, respectively, while the highest correlation for GLAM is 0.82, confirming it has a higher similarity to textual features in terms of semantic region description. This superiority is due to that geographical characteristics is abstract and semantic, while GMM and CNN features lack the mechanism to model the semantic features, which makes them difficult to discover complex patterns.

4.3 Qualitative Evaluation

We illustrate here an example (Fig. 5). A region is described by its dominant characteristics and each characteristic is described by the corresponding top 5 clusters. Here we only present one set of experiment results for qualitative evaluation, where the number of characteristics and number of clusters are respectively set to 15 and 500 with the strongest correlation in Table 1. The rest of results can be accessed at: https://sites.google.com/site/geolatentim/.

Take Beijing and Shanghai, two famous cities in China as an example. As shown in Fig. 5, according to Beijing's characteristic distribution, the characteristic 11 dominates, which can be regarded as the main descriptor for Beijing. To interpret this characteristic, the top 5 representative clusters are picked out to describe it. We manually summarize these five clusters, which correspond to Chinese antique, Chinese tower, Chinese architecture, Chinese roof decoration and pedestrian street, indicating people in Beijing prefer a Chinese traditional atmosphere. This conclusion is well-aligned with Beijing because Beijing is the national center of Chinese history and culture and the historical sites are quite common. Similarly, we can see that Shanghai, the economic center of China, is a modern city with large population, as its characters are mainly described by skyscraper, city scene, urban night, modern traffic, and street scene with people crowd. Among all these regions[3], it is remarkable that some cities are

[3] To see other examples with different parameter sets, please go to our website: https://sites.google.com/site/geolatentim/.



(a) Beijing

(b) Shanghai

Fig. 5. Analysis of the region "Beijing" and "Shanghai".

dominated by one single characteristic (e.g. Beijing, Shanghai) while others possess diverse characteristics (e.g. Sichuan, Shandong) because of geographical and cultural reasons.

4.4 City Recommendation

In this section, we introduce a strategy for region recommendation based on user's photo album. We evaluate the effectiveness of GLAM for recommendation with the Mean Reciprocal Rank metric (MRR).

A photo collection could reflect a user's interest since it contains snapshots of things that the user adores. Here we design a strategy based on the similarity between a user's interest and a region's geographical characteristics for recommendation. First, we compute an interest distribution θ_{new} for a photo collection by Eq. 6. Then, we measure the similarity between this distribution and a region's characteristics with the following distance metric:

$$d_i = ||\theta_i - \theta_{new}||^2_{i=1,\ldots,M}$$

where θ_i is the characteristic distribution in the i^{th} region. The smaller the distance is, the more similar the collection and the region are. The top 3 similar provinces are picked as a recommendation. In our experiments, we crawled additional photos with GPS information from Flickr community[4] (not included in our training data) for both quantitative and qualitative evaluation.

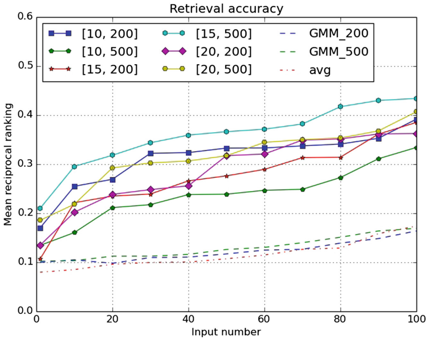


Fig. 6. The recommendation accuracy. In this figure, we can observe that the recommendation accuracy increases as the input number of images increases and GLAM features outperform than GMM and visual features.

For quantitative evaluation, according to the GPS information, we choose 100 images from a province to form a virtual album and the province is regarded as the label of this album. Then we input different amount, accumulating gradually until 100, of images for each album and compute the average MRR to show the recommendation accuracy. Figure 6 presents the average recommendation accuracy with different parameters. The best average MRR performance of GLAM region feature ($K = 15, K' = 500$) is over 40% when input number is more than 70, and according to the property of MRR, we can infer that the label region appears in the top 3 recommended regions, which provide us a reliable recommendation result. Compared with GMM features and visual features, they possess close performance when the input number is small. Nevertheless, it is clear that our model could better perform with more input images and outperform GMM feature and average CNNs visual features because more images could better cover the personal characteristics. For qualitative evaluation, we randomly pick several users, and in each user's photo collection, we randomly select 100 images to form test photo albums. Since the parameter set as 15 "attributes"

[4] https://www.flickr.com/.

and 500 "clusters" provide the best performance (Fig. 6), we here employ this parameter setting. Figure 7 present one example: the photo collection containing mostly nature scenes which present mountain and waterside. This indicates the owner of the photo collection may be a fan of traveling in nature. Our recommendation result shows Yunnan, Chongqing and Jiangxi, which are famous for their landscape. Browsing the photos in these regions, we observe the scenery is similar to the photo collection.

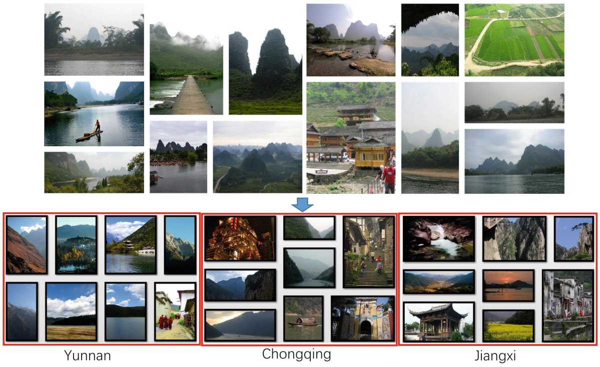


Fig. 7. The album recommendation. It is clear that the recommended regions possess the similar natural scene like the input ones.

5 Conclusion

In this paper, assuming "attributes" as the descriptors of regional characteristics, we have attempted to find the characteristic relevance of a region and use the high-relevant ones to describe this region. Meanwhile, representative clusters, formed by social images, are picked out to present the attributes of regions. The experiments on photos in China qualitatively and quantitatively demonstrate our model has the capacity to semantically describe a region with image content. Based on our model, the regional features could be extracted, from which the recommendation strategy profits to provide reliable results and outerperform GMM features, as well as average CNNs features in the experiments. Therefore, our model is promising for plenty of applications and could be further developed in future work related to geographical characteristics.

Acknowledgments. The work is partially supported by the High Technology Research and Development Program of China 2015AA015801, NSFC 61521062, STCSM 12DZ2272600.

References

1. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. ACM, New York, April 2008
2. Ji, R., Xie, X., Yao, H., Ma, W.-Y.: Mining city landmarks from blogs by graph modeling. ACM, October 2009
3. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. ACM, New York, April 2009
4. Crandall, D., Snavely, N.: Modeling people and places with internet photo collections. Commun. ACM **55**, 52–60 (2012)
5. Cao, L., Jie, Y., Luo, J., Huang, T.S.: Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 125–134. ACM (2009)
6. Clements, M., Serdyukov, P., de Vries, A.P., Marcel, J.T.: Reinders: using flickr geotags to predict user travel behaviour. ACM, New York (2010)
7. Popescu, A., Grefenstette, G.: Mining social media to create personalized recommendations for tourist visits. In: Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications, COM.Geo 2011, pp. 37:1–37:6. ACM, New York (2011)
8. Li, J., Qian, X., Lan, K., Qi, P., Sharma, A.: Improved image GPS location estimation by mining salient features. Image Commun. **38**(C), 141–150 (2015)
9. Fang, Q., Sang, J., Changsheng, X.: Giant: geo-informative attributes for location recognition and exploration. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 13–22. ACM (2013)
10. Porzi, L., Bulò, S.R., Lepri, B., Ricci, E.: Predicting and understanding urban perception with convolutional neural networks. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 139–148. ACM (2015)
11. Hollenstein, L., Purves, R.: Exploring place through user-generated content: using Flickr tags to describe city cores. J. Spat. Inf. Sci. **2010**, 21–48 (2010)
12. Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N.: The livehoods project: utilizing social media to understand the dynamics of a city. In: International AAAI Conference on Weblogs and Social Media, p. 58 (2012)
13. Blei, M.D.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
14. Xing, E.P., Jordan, M.I., Russell, S.: A generalized mean field algorithm for variational inference in exponential families. In: Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, pp. 583–591. Morgan Kaufmann Publishers Inc. (2002)
15. Blei, M.D., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
16. Thomee, B., Elizalde, B., Shamma, D.A., Ni, K., Friedland, G., Poland, D., Borth, D., Li, L.-J.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
17. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9. IEEE (2015)