Discovering Latent Clusters from Geotagged Beach Images

Yang Wang[1] and Liangliang Cao[2]

[1] Department of Computer Science, University of Manitoba, Canada
ywang@cs.umanitoba.ca
[2] IBM T.J.Watson Research Center, USA
liangliang.cao@us.ibm.com

Abstract. This paper studies the problem of estimating geographical locations of images. To build reliable geographical estimators, an important question is to find distinguishable geographical clusters in the world. Those clusters cover general geographical regions and are not limited to landmarks. The geographical clusters provide more training samples and hence lead to better recognition accuracy. Previous approaches build geographical clusters using heuristics or arbitrary map grids, and cannot guarantee the effectiveness of the geographical clusters. This paper develops a new framework for geographical cluster estimation, and employs latent variables to estimate the geographical clusters. To solve this problem, this paper employs the recent progress in object detection, and builds an efficient solver to find the latent clusters. The results on beach datasets validate the success of our method.

1 Introduction

Geotagged images are receiving more and more research attentions in recent years. A geotagged image is associated with a two dimensional vector, latitude and longitude, representing a unique location on the Earth. The goal of this paper is to use the visual information to estimate the geographical locations even when they are not provided. As evidenced by the success of Google Earth, there is great need for such geographic information among the mass. Many web users have high interests on not only the places they live but also other interesting places around the world. Geographic annotation is also desirable when reviewing the travel and vacation images. For example, when a user becomes interested in a nice photo, he or she may want to know where exactly it is. Moreover, if a user plans to visit a place, he or she may want to find out the points of interest nearby. Recent studies suggest that geo-tags expand the context that can be employed for image content analysis by adding extra information about the subject or environment of the image.

Estimating the geolocation of images is not an easy task. As the earlier work shown in [10] [6], only a quarter of the test images can be located subject to a rough region (approximately 750 km) near their true location. At the metropolitan scale, visual feature based annotations perform no better than chance.

S. Li et al. (Eds.): MMM 2013, Part II, LNCS 7733, pp. 133–142, 2013.

As argued by [4], it is difficult to estimate the exact location at which a photo was taken. Instead, the work in [4] proposes to estimate only the coarse location in terms of geographical clusters. The goal of this paper is to find meaningful geographical clusters corresponding to different geographical regions. The use of geographical clusters can provide group wisdom for trip planing and photo organization applications. It also gathers more training samples to build more reliable classifiers.

In this paper, we focus on estimating rough geo-locations of images in terms of their geographical clusters. In particular, we use beach images in our experiments. Note that our problem is different from that of landmark recognition [23]. A landmark usually corresponds one view or one subject with a unique appearance, while a beach scene may contain a lot of clues including water, boats, people dresses, buildings and plants. Moreover, a landmark is usually limited to a point on the earth, while a beach usually covers a region. It is often inaccurate and also unnecessary to estimate the exact GPS coordinate and we only need to estimate a coarse location for a beach image.

Finding geographical clusters can lead to many applications. If we can correctly assign geolocations to image, we will be able to produce tourist maps using geographical annotation techniques [5]. We can also compare the distribution of different topics, such as cars, food, or landscapes in the world [20]. However, in practice, it is not easy to find meaningful geographical clusters. Country borders that separate the geographical regions are too coarse for large countries but too fine for small ones. [3] proposed to initialize meaningful geographical clusters by spatial clustering refine the cluster by post processing. In this paper, we will discuss a new method to find the geographical clusters using an efficient latent SVM learning.

2 Previous Work

Geographical annotation provides a rich source of information which can link millions of images based on the similarity of their geographical locations. There have been a growing body of work in visual research community investigating geographical information for image understanding [15] [1] [4] [21] [11] [22] [12] [14] [16] [13] [17] [20] [2]. Many applications are motivated by Jim Gray's idea to build a personal Memex which can record everything a person sees and hears, and quickly retrieve any item on request. Moreover, It is more interesting to aggregate information from a large number of users, so that group wisdom can be mined from these media. As suggested by [2], if we know a number of user favored images, we can provide effective tourism recommendation under the premise "If you like this picture, you will also like these places". However, such a personal Memex requires a huge amount of geo-tagged information, which is still not practical given the fact that 99% of Flickr photos do not have related geographical information associated with them.

To address the challenges, one group of research work is devoted to estimating the geographical information from general images. Hays and Efros [10] are

among the first to consider the problem of estimateing the location of a single image using only its visual content. They collect millions of geo-tagged Flickr images. Using a comprehensive set of visual features, they employ nearest neighbor search in the reference set to locate the image. Motivated by [10], Gallagher et al. [9] incorporate textual tags to estimate the geographical locations of images. Their results show that textual tags perform better than visual content and the combination of textual and visual information performs better than either alone. Cao et al. [4] also recognize the effectiveness of tags in estimating the geolocations. They propose a novel model named logistic canonical correlation regression which explores the canonical correlations between geographical locations, visual content and community tags. Unlike [10], they argue that it is difficult to estimate the exact location at which a photo was taken and propose to estimate only the coarse location. Similarly, Crandall et al. [6] only estimate the approximate location of a novel photo. Using SVM classifiers, a novel image is geolocated by assigning it to the best cluster based on its visual content and annotations. In a recent research work [23] supported by Google, Zhen et al. built a web-scale landmark recognition engine named "Tour the world" using 20 million GPS-tagged photos of landmarks together with online tour guide web pages. The experiments demonstrate that the engine can deliver satisfactory recognition performance with high efficiency.

Despite of these research efforts, recognizing the location of a non-landmark image reliably is still an open question. For those non-landmark locations, visual information based classifiers only perform comparable to chance. A recent study [3] propose to discover "geographical clusters" to build classifiers. The use of geographical clusters benefits the problem of localization in two aspects: On the training stage, geographical clusters provide more training samples and hence lead to better recognition accuracy; on the testing stage, estimation of the most possible region for each query photo will be relatively easier than the estimation of exact GPS coordinates, while the information of geographical cluster will be good enough for trip planing and photo organization applications. However, the geographical clusters in [3] are discovered by refined mean-shift clusters, which are not representative enough for visual recognition. In this paper, we aim to develop a more principled approach to find geographical clusters.

This paper is motivated by the some recent progress in object detection [8] and max-margin clustering [18]. In the object detection method of [8], the locations of object parts are unknown, and are treated as latent (hidden) variables in a learning framework called the *latent SVM*. A latent SVM is an extension of regular SVMs to handle latent variables. A latent SVM is semi-convex and the training problem becomes convex once latent information is specified for the positive examples. This leads to an iterative training algorithm that alternates between fixing latent values for positive examples and optimizing the latent SVM objective function. Similar ideas can also be found in [18] which finds maximum margin hyperplanes through data. In this paper, we treat the geographical clusters of training images as hidden labels, and develop a principled learning method that recognizes the geographical clusters of images.

3 Our Approach

The work in [3] finds geographical clusters by clustering the GPS coordinate vectors of training images. Then a SVM classifier based on image features is learned for each cluster. For a new test image, the SVM classifier can be used to assign this image to a corresponding cluster based on its image feature. A limitation of this approach is that clustering and SVM learning are treated as two independent tasks. However, we believe these two tasks should be coupled together. In this paper, we propose a new approach that considers image clustering and model learning in a single unified framework.

3.1 Geo-location Regularized Clustering

Our method is based on the max-margin clustering (MMC) [18]. Naively applying MMC to our dataset is troublesome, since MMC is a generic clustering algorithm and does not take into account of the geo-location information of the data. We propose an extension of MMC that clusters training images so that images in the same cluster are both visually similar and have close GPS locations.

We assume that we are given a training dataset with N instances. Each instance is in the form of (x_i, y_i), where x_i is the i-th image, and y_i is its corresponding geo-location. Our goal is to cluster the training images into C groups in some sensible manner. We would also like to have a discriminative model that can assign an unseen image to one of the clusters. If we ignore the geo-location information y_i in the training data and only consider the image feature x_i, we can use standard clustering algorithms to partition the training images into C clusters. But now the challenge is how to incorporate the GPS location information into the clustering process.

Let us assume that the number of clusters is know to be C. Clustering the training data is equivalent to assigning a binary vector z_i to each image x_i. Here z_i is a vector of length C, where its c-th component z_{ic} is defined as:

$$z_{ic} = \begin{cases} 1 & \text{if } x_i \text{ belongs to cluster } c \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Note that if z_i is observed on training data, we can use this information to learn a multi-class SVM classifier to assign the cluster membership of an unseen image by solving the following optimization problem:

$$\mathcal{P}(w^*) = \min_{w,\xi} \quad \frac{1}{2}||w||^2 + C_1 \sum_i \xi_i \tag{2a}$$

$$\text{s.t.} \quad w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \tag{2b}$$

where w and $\phi(x_i, z_i)$ is a feature vector, ξ_i is the slack variable for handling soft margins in SVM classifiers.

Now since z_i is not observed, we need to simultaneously partition the training data into C groups and learn the multi-class SVM. Using the same reasoning of unsupervised SVM [18,19], we can try to solve the following optimization problem:

$$\mathcal{P}(w^*, \{z_i\} : \forall i) = \min_{w,\xi} \min_{\{z_i\}} \quad \frac{1}{2}||w||^2 + C_1 \sum_i \xi_i \tag{3a}$$

$$\text{s.t.} \quad w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \tag{3b}$$

Note that in Eq. 3, we need to optimize over the variables $\{z_i\}$, since they are unknown on the training data. The optimization problem in Eq. 3 tries to find $\{z_i\}$ so that the resultant SVM has the maximum margin (please refer to [18,19] for details).

Unfortunately, without additional constraints or regularization, Eq. 3 has a degenerate solution. Basically we can assign all training data to the same cluster and learn w to achieve arbitrarily large margin. In [18,19], this problem is addressed by adding a constraint that tries to make sure that the clusters are balanced.

For our application, we have the addition information (i.e. GPS locations) in addition to images. In the following, we will use this additional information to regularize Eq. 3. Intuitively, we would like the clusters to have the following property. If two images are close in terms of their geo-locations, they are more likely to be in the same cluster. One natural way to formalize this intuition is to solve the following optimization problem:

$$\mathcal{P}(w^*, \{z_i\} : \forall i) = \min_{w,\xi} \min_{\{z_i\}} \quad \frac{1}{2}||w||^2 + C_1 \sum_i \xi_i \tag{4a}$$

$$+C_2 \sum_i \sum_j (-|z_i - z_j| d_{ij}) \tag{4b}$$

$$\text{s.t.} \quad w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \tag{4c}$$

where d_{ij} is the distance of two images x_i and x_j in terms of their geo-locations (which can be obtained from y_i and y_j).

Note that $|z_i - z_j| = 0$ if i and j are in the same cluster. So Eq. 4b will try to make the distance (in terms of GPS locations) between images in different clusters to be large.

The optimization problem in Eq. 4 can be solved using an iterative approach:

- Fix $\{z_i\}_{i=1}^N$, optimize over w and ξ.
- Fix w and ξ, optimize over $\{z_i\}_{i=1}^N$.

The first step of this iterative approach is straightforward since it is equivalent to solving a standard multi-class SVM problem. The second step is more

challenging. It involves solving a combinatorial problem which can be shown to be NP-hard.

One possible solution is to use linear program relaxation to get an approximate solution. But the resultant linear program is still too large to be practical. In the following section, we introduce a new formulation that is more amenable to efficient algorithms.

3.2 More Efficient Formulation

The main observation that enables our new formulation is the following. Suppose we know the cluster centers $\{g_c\}_{c=1}^{C}$ (in term of geo-locations), a natural way to solve our problem is to use the following optimization:

$$\mathcal{P}(w^*, \{z_i\} : \forall i) = \min_{w,\xi} \min_{\{z_i\}} \quad \frac{1}{2}||w||^2 + C_1 \sum_i \xi_i \tag{5a}$$

$$+C_2 \sum_i \sum_c (z_{ic}||y_i - g_c||^2) \tag{5b}$$

$$\text{s.t.} \quad w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \tag{5c}$$

Note that Eq. 5b computes the distance (in term of geo-locations) between images and their corresponding cluster centers. When those cluster centers are known, the optimal clustering is obtained by choosing cluster membership that minimizes this distance (i.e. minimizing over $\{z_i\}$).

Now the challenge is that the cluster centers $\{g_c\}$ are also unknown. Using the same reasoning in Sec. 3.1, we propose to treat the cluster centers as yet another set of latent variables in the formulation and use the following iterative method to solve it:

- Fix $\{z_i\}_{i=1}^{N}$ and $\{g_c\}_{c=1}^{C}$, optimize over w and ξ: this step is equivalent to solving a regular multi-class SVM. We use liblinear [7] for it.
- Fix w, ξ and $\{z_i\}_{i=1}^{N}$, optimize over $\{g_c\}_{c=1}^{C}$: it is easy to show that if we use the l_2 distance, the optimal value of the c-th cluster center g_c is the average of the geo-locations of images assigned (based on $\{z_i\}$) to this cluster.
- Fix w, ξ and $\{g_c\}_{c=1}^{C}$, optimize over $\{z_i\}_{i=1}^{N}$: it is easy to show this step is a linear assignment problem.

4 Experiments

We test our approach on a dataset containing images downloaded from Flickr with the tags of "beach" or "coast". Each photo is associated with a two-dimensional GPS coordinate vector. Similar to [3], we use 34558 images for training and 1185 images for testing. We use GIST features to represent images.

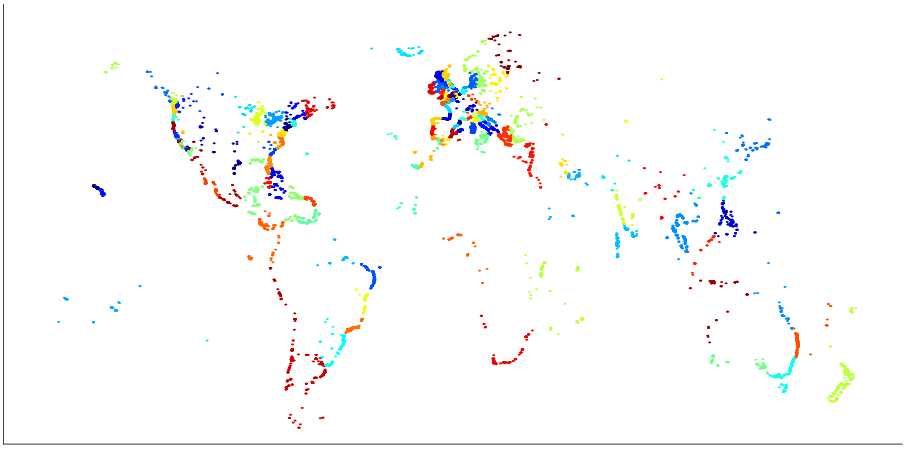

Fig. 1. Visualization of clustering training images using our method. Each color represents a different cluster.

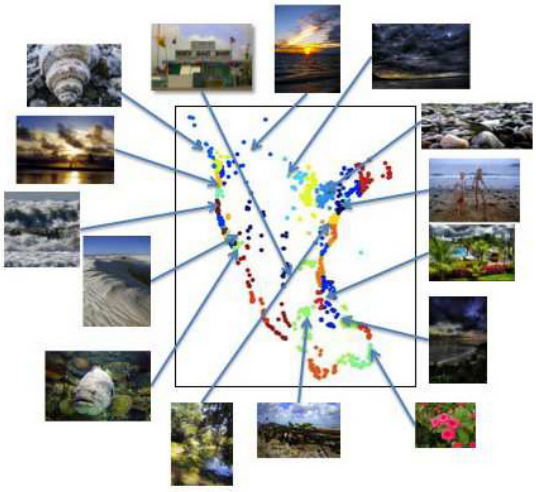

Fig. 2. Visualization of representative images for North America

In Fig. 1, we plot the distribution of training images and their clusters in roughly different colors. In Figs. 2 3 4, we visualize some representative images in some clusters.

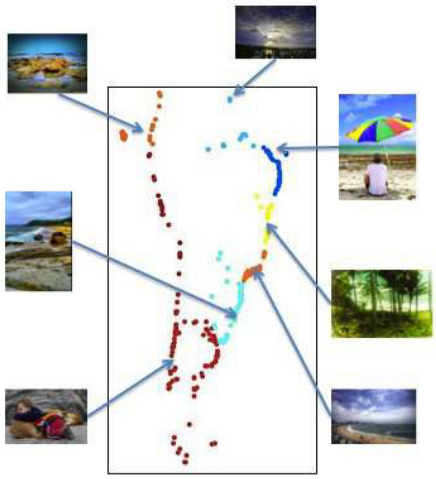

Fig. 3. Visualization of representative images for South America

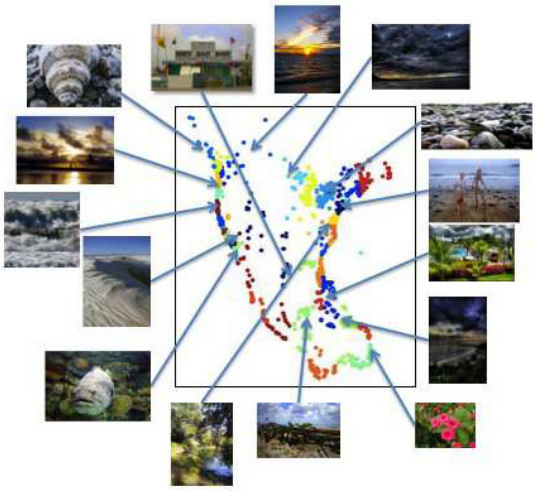

Fig. 4. Visualization of representative images for Asia

5 Conclusion

We have introduced a new framework for geographical cluster estimation. Our approach treats the geographical cluster of an image as a latent variable. Our method jointly clusters training images and learns discriminative classifiers for each cluster in a single framework.

Acknowledgement. Yang Wang is supported by a start-up grant from the University of Manitoba.

References

1. Agarwal, M., Konolige, K.: Real-time localization in outdoor environments using stereo vision and inexpensive GPS. In: International Conference on Pattern Recognition (2006)
2. Cao, L., Luo, J., Gallagher, A., Jin, X., Han, J., Huang, T.: A worldwide tourism recommendation system based on geotagged web photos. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP (2010)
3. Cao, L., Smith, J., Wen, Z., Yin, Z., Jin, X., Han, J.: BlueFinder: Estimate where a beach photo was taken. In: WWW (2012)
4. Cao, L., Yu, J., Luo, J., Huang, T.: Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In: Proceedings of the Seventeen ACM International Conference on Multimedia, pp. 125–134 (2009)
5. Chen, W., Battestini, A., Gelfand, N., Setlur, V.: Visual summaries of popular landmarks from community photo collections. In: ACM International Conference on Multimedia, pp. 789–792 (2009)
6. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: International Conference on World Wide Web, pp. 761–770 (2009)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A Library for Large Linear Classification. JMLR (2008)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
9. Gallagher, A., Joshi, D., Yu, J., Luo, J.: Geo-location Inference from Image Content and User Tags
10. Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
11. Joshi, D., Luo, J.: Inferring generic places based on visual content and bag of geotags. In: ACM Conference on Content-based Image and Video Retrieval (2008)
12. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: Context and content in community-contributed media collections. In: ACM Conference on Multimedia (2007)
13. Luo, J., Yu, J., Joshi, D., Hao, W.: Event recognition: viewing the world with a third eye. In: ACM International Conference on Multimedia, pp. 1071–1080 (2008)
14. Naaman, M.: Leveraging geo-referenced digital photographs. PhD thesis, Stanford University (2005)
15. Naaman, M., Song, Y., Paepcke, A., Garcia-Molina, H.: Automatic organization for digital photographs with geographic coordinates. In: International Conference on Digital Libraries, vol. 7, pp. 53–62 (2004)
16. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: ACM Conference on Image and Video Retrieval, pp. 47–56 (2008)
17. Schindler, G., Krishnamurthy, P., Lublinerman, R., Liu, Y., Dellaert, F.: Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)

18. Xu, L., Neufeldand, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 17, pp. 1537–1544. MIT Press, Cambridge (2005)
19. Xu, L., Wilkinson, D., Southey, F., Schuurmans, D.: Discriminative unsupervised learning of structured predictors. In: Proceedings of the 23th International Conference on Machine Learning (2006)
20. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.: Geographical topic discovery and comparison. In: Proceedings of the 20th International Conference on World Wide Web, pp. 247–256. ACM (2011)
21. Yu, J., Luo, J.: Leveraging probabilistic season and location context models for scene understanding. In: International Conference on Content-based Image and Video Retrieval, pp. 169–178 (2008)
22. Yuan, J., Luo, J., Wu, Y.: Mining compositional features for boosting. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
23. Zheng, Y., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T., Neven, H.: Tour the World: building a web-scale landmark recognition engine. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)