Multi-view Landmark Recognition in Large-scale Image Collections

Dongyeop Kang
Language Technology Institute
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
dongyeok@cs.cmu.edu

Abstract

From the large amount of web photos and other meta knowledge, recognizing landmark images has been actively studied in recent years. Previous landmark recognition research based on visual similarities have several limitations in terms of sparseness, efficiency and functionality. In this paper, I proposed a landmark hierarchy which consists of images, geographical clusters, visual clusters and topics and cluster images into the hierarchy to reduce sparseness in image retrieval for landmark recognition. Especially, I discovered topics from landmark visual clusters from multiple features such as texts and images. To discover hidden relationship between them, I applied multi-view topic analysis on landmark recognition, and inference topics using max-marginal approach. Moreover, to deal with million of images and their meta information, I have deployed whole clustering and multi-view learning procedure on distributed machines. Finally, I validate landmark clusters and topics in the experiment section.

1 Introduction

Popularization of personal digital cameras, together with Internet, can provide billions of photos manipulated by Web users. For example, Flickr has been reporting millions of images per month, and they contain not only images but also additional meta information such as geographical locations, time, textual tags, photographers, comments and so on. From the large amount of web photos and other meta knowledge, recognizing landmark images such as Eiffel Tower and Statue of Liberty has been actively studied in recent years. Previous landmark recognition research has been focusing on searching similar images called nearly duplicated [4] or classifying similar images [3] based on their visual similarities. However, the near-dup or classification based approaches have several limitations in terms of effectiveness, efficiency and functionality as follows:

- First of all, existing visual features such as SIFT [7] are not enough to distinguish two landmark images. While SIFT features are typically effective to retrieve similar images of detail patterns such as grid, landmark images are quite simple and similar.
- For millions of images, retrieving similar images given a query image is very inefficient. Moreover, the dimension of visual images are much larger than text so indexes for landmark retrieval will be very sparse.
- Showing similar landmarks are not enough for real needs of users. For example, when a user visits Eiffel Tower and take a photo of the tower as a input query, she or he may want to find various perspectives of Eiffel Tower such as Eiffel Tower at night, Eiffel Tower with sunset, famous cafes or restaurants near Eiffel Tower, or another landmarks similar to Eiffel Tower rather than just to show similar images of the query image. By doing so, users will interact with others with different views, topics or feelings.
In this paper, in order to those limitations, my contributions are as follows:

First of all, I believe that there are strong relationships not only between images but also between texts which are coupled with images, and between images and tags, and between images correlated by other metadata such as time and location. To discover such internal relationships between different sources of information (e.g., time, location, images, texts), I applied a state-of-the-art technique called multi-view subspace learning [2] to the landmark problem. Specifically, the paper proposed a max-margin learning to discover and predict subspaces from multiple data sources simultaneously. Compared to previous approaches such as MedLDA [8], this method uses undirected graph of multi-view data (see fig 1 (a)) so hidden relationship between heterogeneous types of data are simultaneously discovered.

Second, in order to reduce sparseness of image collections for landmark retrieval, I have generated a landmark hierarchy (see fig 1 (b)). First I cluster original image collections into geographical clusters based on their GPS information, and each Geo cluster correspond each landmark. Then, for each Geo cluster, similar views of images in a landmark cluster are clustered into a view cluster based on their visual contents. Finally, latent topics are discovered through their undirected network between images and texts using multi-view topic analysis. In addition, in this project, I have used distributed computing power through hundreds of machines. Especially, such high computational image processing tasks such as SIFT feature extraction, vector quantization, and geo and view clustering are partitioned over multiple machines in clusters. Also, the main algorithm of multi-view topic analysis is deployed into multiple machines during EM inference.

The purpose of our system is either displaying multiple perspective of landmarks or retrieving similar landmarks given a input query image. For displaying function, I show top-k landmarks with multiple perspectives such as different views and topics. For example, if a user explore a Brooklyn Bridge, then our system shows not only different views of London Bridge (e.g., daytime, night, fog, sunset) but also other bridge such as Golden Gate Bridge associated with same topic “Bridge” through texts and images. The perspectives also can be extended not only views and topics but mood, feelings, and so on. For retrieving function, when a image query by a user is given, the system finds the most similar view clusters.

2 Large Scale Landmark Recognition

During the experiment, I found previous near-dup based approaches shows very poor performance of finding duplicated images of landmark in terms of accuracy and search time efficiency. This preliminary experiment shows that our hierarchical clustering approach is indispensable for large scale landmark recognition.

In the fig 1 (b), I reduce large number of image instances into Geo cluster space, and each Geo cluster is divided by view clusters according to their image contents. By doing so, I could reduce dimensionality of image collections to be efficient for landmark retrieval task. In this section, I will describe details of my algorithm for geo and view clustering and other heuristic strategies to filter some noisy images.
2.1 Geographical Clustering

In my dataset, every images contain their GPS information. Thus, I can easily merge those images which have similar GPS locations. However, existing clustering techniques that I learned from the class is not scalable for those large number of instances. For example, classical K-means and spectral clustering does not work on this problem due to its scalability. Thus, I proposed a simple but effective clustering method based on K-means with location constraint in the algorithm 1. First, the algorithm takes new instance continuously, and if the instance is close to the existing clusters with Threshold_{dist} distance, then the instance is added to the cluster, and the centroid is re-calculated again with the new instance. Otherwise, new cluster which contains only the new instance is created. After sequential clustering, I removed the landmark clusters which number of instances are below than certain threshold Threshold_{landmark}.

Algorithm 1 The sequential clustering: K-means clustering with location constraint

**Input:** longitude and latitude of each instance (image)

- \(C_g\): central point (latitude, longitude) of each cluster
- \(L_i\): cluster label of each instance

**for each instance** \(i\) **do**

**for each** cluster \(c \in C_g\) **do**

- calculate distance \(c\) and instance point \(i\) on sphere
- find minimum distance, \(\text{Dist}_{\text{min}}\), and its label \(\text{Label}_{\text{min}}\).

If \(\text{Dist}_{\text{min}} \leq \text{Threshold}_{\text{dist}}\) then

- \(L_i = \text{Label}_{\text{min}}\)
- update \(C_{\text{Label}_{\text{min}}}\) with new instance \(i\)

else

- create new centroid \(C_{g+1}\) with new instance \(i\)

**for each cluster** \(c\) **do**

- If size of cluster \(\leq \text{Threshold}_{\text{landmark}}\) then

  remove the cluster \(c\)

With \(N\) number of instances, the algorithm takes \(O(N^2)\) time complexity which is faster than K-Means and spectral clustering. One limitation in this algorithm is two threshold values, \(\text{Threshold}_{\text{dist}}\) and \(\text{Threshold}_{\text{landmark}}\). In this problem, I set \(\text{Threshold}_{\text{dist}}\) as 100m which is reasonable distance of two different real landmarks and \(\text{Threshold}_{\text{landmark}}\) as its average number of instance for each cluster which ignores lower number of clusters below than their average. This sequential clustering algorithm also can be extended to other problems which have some constrains over clusters.

2.2 View Clustering

When a geo cluster is given, I regard it as a landmark, and cluster its different views of the landmark such as different angles, different landscapes, different time views and so on. First of all, every images in geo clusters should be processed in the form of view clustering. In this project, I follow the bag-of-words [5] approach which the majority of image retrieval systems have adopted. First, I extracted SIFT descriptors [7] from photos in the training set. Then, to build a visual vocabulary, those training SIFT descriptors are clustered using randomized k-d trees [6]. The descriptors are then vector quantized into the vocabulary of visual words.

Each Geo cluster contains below than 10 thousand images. In this reduced dimension in a geo cluster, K-means works efficiently with reasonable parameter of \(k\). Since each image in the geo cluster is represented as a vector with about 10 thousand key-points dimensions, view clustering with K-means is reasonable approach. In order to find the appropriate \(k\), I empirically run the algorithm several times, and find out 1/5 of total images in a geo cluster is the most reasonable \(k\).

One practical challenge in view clustering is to clean up noisy images which are irrelevant to landmarks such as human faces. In case of public dataset collected by common users, the large amount of noises is critical issue in landmark recognition task. Referred from previous studies [5], here is the heuristics used in this project:
• For each geo cluster, if a cluster is collected by less number of users (e.g., 10) or contains less number of images (e.g., 50), ignore those Geo clusters because a famous landmark should be taken by many users.

• For each view cluster, if the average distance between centroid of view cluster and any other data points is less than certain threshold, ignore the view cluster because those visual cluster is not visually coherent.

• For each view cluster, if the distance of certain image to the centroid is larger than certain threshold, filter out the image from the view cluster because it is a noisy image such as face.

3 Multi-View Subspace Learning

Besides SIFT features and geo-tags, I collect other meta features related to image such as textual information such as tags, title and descriptions and other image features such as , and color histogram. The multi-view approach [2] is based on an undirected latent space Markov network that fulfills a weak conditional independence assumption rather than previous directed topic models. It assumes that multi-view observations and response variables are independent given a set of latent variables. The fig 1(b) shows that the two set of input views, X and Z and the set of latent variables H are linearly connected. Based on the paper [2], those latent variables are inferred by large-margin idea into the learning of supervised multi-view latent spaces for multi-view data analysis.

For landmark multi-view analysis, for those images in view clusters, three types of features such as SIFT (S), color histogram (Z), and textual information (X) are given (see fig 2). The model is constructed based on an underlying conditional independence assumption that given the latent variables H, the three views X, Z and S are independent. For each view, prior probability is considered as first-order Markov network. Based on the random filed theory, each component has following marginal probabilities:

\[
p(x) = \exp\left\{ \sum_i \theta_i \phi(x_i, x_{i+1}) - A(\theta) \right\}
\]

\[
p(z) = \exp\left\{ \sum_j \eta_j \psi(z_j, z_{j+1}) - B(\eta) \right\}
\]

\[
p(s) = \exp\left\{ \sum_i \delta_i \tau(s_i, s_{i+1}) - C(\delta) \right\}
\]

\[
p(h) = \prod_k p(h_k) = \prod_k \exp\left\{ \gamma_k \phi(h_k) - D_k(\gamma_k) \right\}
\]

where \( \phi, \psi, \tau \) and \( \varphi \) are feature functions, A, B, C and D are log partition functions. Then, the joint model probability is defined by combining the above components in the log-domain and introducing additional terms that couple the random variables X, Z, S, and H:

\[
p(x, z, s, h) \propto \exp\left\{ \sum_i \theta_i \phi(x_i, x_{i+1}) + \sum_j \eta_j \psi(z_j, z_{j+1}) + \sum_i \delta_i \tau(s_i, s_{i+1}) + \sum_k \gamma_k \varphi(h_k) + \sum_{ik} \phi(x_i, x_{i+1}) W^k \varphi(h_k) + \sum_{jk} \psi(z_j, z_{j+1}) U^j \varphi(h_k) + \sum_{ik} \tau(s_i, s_{i+1}) V^i \varphi(h_k) \right\}
\]

Then, the conditional distribution on each view are directly written as follows:

\[
p(x|h) = \exp\left\{ \sum_i \hat{\theta}_i \phi(x_i, x_{i+1}) - A(\hat{\theta}) \right\}, \text{ where } \hat{\theta}_i = \theta_i + \sum_k W^k \varphi(h_k)
\]

\[
p(z|h) = \exp\left\{ \sum_j \hat{\eta}_j \psi(z_j, z_{j+1}) - B(\hat{\eta}) \right\}, \text{ where } \hat{\eta}_j = \eta_j + \sum_k U^j \varphi(h_k)
\]

\[
p(s|h) = \exp\left\{ \sum_i \hat{\delta}_i \tau(s_i, s_{i+1}) - C(\hat{\delta}) \right\}, \text{ where } \hat{\delta}_i = \delta_i + \sum_k V^i \varphi(h_k)
\]

\[
p(h|x, z, s) = \Pi_k \exp\left\{ \tilde{\gamma}_k \phi(h_k) - D_k(\tilde{\gamma}_k) \right\}, \text{ where } \tilde{\gamma}_k = \gamma_k + \sum_i W^k \phi(x_i, x_{i+1}) + \sum_j U^j \psi(z_j, z_{j+1}) + V^i \tau(s_i, s_{i+1})
\]

Since the data likelihood is intractable to compute, parameter estimation is based on the efficient contrastive divergence technique [2, 8]. Then, the data likelihood can be variationally approximated with two variables: a variational distribution of observed components clamped to their observed values and a distribution with all variables free. The approximated data likelihood, then, can be
iteratively minimized using alternating minimization method. After those two variables are inferred, parameters are estimated through E and M steps with a multi-class SVM and sub-gradient descent method, respectively. After the sub-gradient, $\theta$, $\eta$, $\gamma$, $W$, $U$, and $V$ are computed. Please refer the original papers [2, 8, 9] to understand details of estimation and underlying theories.

In this paper, I don’t have experimented on predictive performance of multi-view learning because the Flickr datasets are collected from users without any label information. However, when some labels are given, I can also predict new landmark photos using the objective function, $Y$, learnt from the model. One of challenge in existing multi-view algorithm was high computation in initialization and learning part. However, by reducing the instances of landmark photos from our landmark hierarchy, I could obtain a few thousands of clean landmark photos which can be applied to multi-view analysis.

However, another practical challenge is high dimensions in the combined features. For example, SIFT feature dimensions are usually 100K, texts are 15K, and color histograms are 1K. When they are combined, the total size of feature dimensions are 120K. The large number of feature dimension is still bottleneck of multi-view algorithm especially in SVD initialization and parameter estimation. At this time, I just split data into multiple machines and process EM iterations through 200 hundreds of machines by calculating gradient descent to optimize over $W$, $U$, and $V$. In the next version of paper, I will thoroughly discuss the scalability of multi-view analysis by dividing high feature dimensions.

4 Experiments

In this project, I will use the Flicker image which consist of visual information such as images and color histogram, text information such as tags, title, and description, and GPS and time information. For convenience, I used the pre-crawled dataset described in the paper [1], which images are crawled by geographical keyword based queries such as name of countries, states, and cities. However, since the dataset only contains geographical metadata, I additionally textual tags for those images in the representative view clusters for multi-view analysis, and the total number of images are 6, 471, 706.

4.1 World Landmark Recognition

Due to the large number of computation of image pre-processing, GEO clustering and view clustering, I used the distributed computing power for speed-up whole procedure. For image pre-processing, I extracted SIFT key-points from 6.4 millions of images by splitting them into hundreds of sub parts and parallelly compute them with 200 machines. Due to the high resolution of images crawled from Flickr, I down-sampled images and then extract key-points so each image approximately has 700 key-points. From the sampled 300 thousands images, I generated 130, 405 size of visual vocabulary using hierarchical k-means approach. Then, again I calculated vector quantization of 6.4 million images into the vocabulary with those 200 machines.

![Figure 3: Landmark locations in the world.](image-url)
Then, I have finished GEO and view clustering part in the same of distributed computation. Finally, I could collect 97 thousand Geo clusters. The table 1 shows of how I resolve sparsity problem throughout hierarchy of Geo and View clusters and topics. The landmark means clean geo clusters after heuristics on view clustering. The Figure 3 shows landmark points in the world map. The total number of view clusters in the landmark cluster are 8, 145, and each landmark contains 7.1 average number of view clusters.

<table>
<thead>
<tr>
<th>Images</th>
<th>Geo clusters</th>
<th>Landmarks</th>
<th>View clusters</th>
<th>topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,471,706</td>
<td>97,421</td>
<td>1,138</td>
<td>8,145</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 1: The effect of dimensionality reduction.

For more statistics, the table 2 shows top-10 locations in the world which contains largest number of landmarks. Interestingly, they matches with general mind of famous places to travel or visit in the world. Unfortunately, a few countries (e.g., Japan, France, Germany, California, Canada, Seattle) are still on processing due to their large size, and they would be updated in a week.

<table>
<thead>
<tr>
<th>Ranks</th>
<th>Location</th>
<th># Landmarks</th>
<th>Ranks</th>
<th>Location</th>
<th># Landmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Berlin</td>
<td>91</td>
<td>6</td>
<td>Florida</td>
<td>33</td>
</tr>
<tr>
<td>2</td>
<td>Barcelona</td>
<td>63</td>
<td>7</td>
<td>Beijing</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Washington</td>
<td>44</td>
<td>8</td>
<td>Hawaii</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>HongKong</td>
<td>35</td>
<td>9</td>
<td>Melbourne</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>LosAngeles</td>
<td>34</td>
<td>10</td>
<td>Minnesota</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 2: The top-10 locations which contain largest number of landmarks in the world. For each row, two consecutive photos mean a view cluster for the landmark.

Finally, the figure 4 shows view clusters for 7 landmarks in the world. As you can see, for each landmark, view clusters show different views of the landmark. For example, in the Beijing 379th landmark, each view cluster shows different views such as bridge, river, front-view and detail view of the main building. Also, in the Australia 209th landmark, different angles and views of bridge are shown. Our initial clustering results reflect the power of crowds and diversity views collected by many public users. Furthermore, in the next section, I will show how latent topic analysis with other types of information such as texts can improve existing landmark result.

**4.2 Multi-view Topic Analysis**

From the landmark clustering, I got 8,908 clean images as a training set for multi-view analysis. Then, I crawled textual information such as title, description and tags and color histogram for the training images. In summary, the dimension of SIFT feature is 130, 405, the dimension of text feature is 12, 790, and the dimension of color histogram is 784. Also, I used the number of topics as 50 obtained empirically. In this section, I will evaluate my results in terms of clustering and classification.

The figure 5 shows top-4 topics discovered. As you can see, different semantically similar view clusters are gathered. Even though previous geo and view clusters are geographical and visually constrained, respectively, latent topics allow allow photos to be flexible to be grouped with other landmark photos without any constraint. These result is very useful when you want to travel with similar perspective. For example, if I want to some place like HongKong, then our system can recommend other landmark places in Chicago and Barcelona. Moreover, this is useful for architectures to find similar appearance of landmarks. For example, if an architect wants to find tower which is similar with Beijing 73, then our system can recommend other tower landmarks in Egypt 99 and London 20. In landmark perspective, a certain landmark such as Beijing 379 contains one more topics such as bridge, river and building.

The figure 6 shows the convergence of objective values during EM iterations. It converges in reasonable number of iterations. However, I also could find many noisy topics because of biased features. For example, current equal weight for color histogram, SIFT, and text features is biased into SIFT feature which has relatively large numbers. Thus, in order to discover the effectiveness of each
Figure 4: A illustration of 10 photos for 7 landmarks.

Figure 5: (a) A illustration of 4 latent topics. Each topic is shown with the top 3 images extracted from the most related representative photos in a visual cluster and its location and ID.

feature, I have experiment topic analysis with different weights of feature values. Before that, to choose best weighting scheme, I needed gold standard set of photos. Then, I manually label the first representative photo of top 100 view clusters according to their type such as bridge, river, building, tower, park, sculpture, landscape, sea, stadium, sun, and antique. Throughout the experiments, the task of classifying types of landmark depends on SIFT, text, and color histogram in order of importance. Specifically, SIFT features are effective to detect similar view patterns of different landmarks.
Though SIFT features work well only for complicated texture of images. Tags are more likely to
be the names of cities, states, and countries, and many of them are not useful. Finally, since similar
type of view clusters can be seen different in terms of colors, color histogram features usually make
the performance worse. The classification result with best weight setting is 34%, and I think there
are a lot of spaces to improve landmark type classification.

![Figure 6: Convergence in EM iteration.](image)

## 5 Conclusion

To improve landmark recognition from large scale image collections which consist of multiple types
of information such as texts and images, I propose applying multi-view max marginal approach to
landmark recognition problem. Before that, to alleviate the sparsity of large scale data and increase
semantics, I clustered geo clusters, view clusters and topics hierarchically. I developed actual system
of world level landmark recognition system. The view clusters for top landmarks are well clustered,
and they are combined with latent topic spaces discovered by multi-view learning. Especially, I have
studied characteristics and effects of each feature for landmark type classification and clustering.

Even though my landmark results and topics are recognizable with multiple perspectives, there are
still many practical limitations to improve. First, multi-view analysis is not scalable to more than
millions of images. If I did not scale original image dataset down to the geo and view cluster spaces,
the algorithm may not work due to its computation. Also, data partition for multi-view learning
is not enough so feature division should be considered in the future work. The parallel coordinate
descent [10] is one possible way to consider.

The perspective also can be extended not only views and topics but mood, feelings, and so on. By
adding more features to learn such as social network information, time, interaction between users,
I will explore more perspectives for photo sensing activities. Also, I will generate clean, accurate
labeled photos for landmark recognition or its types.

## References

keypoints, In Workshop on Statistical Learning in Computer Vision, ECCV 2004
ICML2009.
[9] Eric Xing, Rong Yan, Alexander Hauptmann, Mining Associated Text and Images with Dual-Wing Harmoniums, UAI2005
ICML2011