Chapter 14
Mining Regional Representative Photos from Consumer-Generated Geotagged Photos

Keiji Yanai and Qiu Bingyu

14.1 Introduction

The development of World Wide Web, and the popularization of digital photography, as well as the advent of public media-sharing websites such as Flickr and Picasa, have led to tremendous growth in large online multimedia resource. As a result, the problem of managing, browsing, querying and presenting such collections effectively and efficiently has become critical. However, these rich community-contributed collections are usually organized in an irregular way which makes it difficult to obtain relevant, accurate and complete results. For example, a search for the "noodle" images on the photo sharing site Flickr as well as many other similar sites returns results which contain many visually unrelated photos to the target category.

Community-contributed photo sharing sites collect metadata in addition to photos. While keywords and comments are common as metadata, recently some users attach "geotags" to their uploaded photos. A "geotag" means metadata which represents a location where the corresponding photo was taken, which is usually expressed by a set of a latitude and a longitude.

An accurate geotag can be obtained with a GPS device or a location-aware camera-phone. However, since it forces us to use relatively special devices, GPS-based geotags have not been common so far. Instead, map-based geotags have become common, after Flickr, which is the largest photo sharing site in the world, launched an online geotagging interface in 2006. Then, Flickr also became the largest "geotagged" photo database in the world. According to the Flickr official blog, the number of geotagged photo stored in Flickr exceeded 100,000,000 in February 2009, which corresponds to 3.3% of the total number of Flickr photos. These geotagged photos would be valuable not only for browsing and finding individual concepts, but also for helping us understand how specific objects or scenes are distributed and different over the world.

K. Yanai (✉)
Department of Computer Science, The University of Electro-Communications,
Chofugaoka 1–5–1, Chofu-shi, Tokyo 182–8585, Japan
e-mail: yanai@cs.uec.ac.jp

B. Furht (ed.), *Handbook of Social Network Technologies and Applications*,
DOI 10.1007/978-1-4419-7142-5_14,

Our objective is thus to facilitate a system which can automatically select relevant and representative photographs for the general object or scene categories corresponding to given keywords in the worldwide dimensions. In particular, we consider the geotagged photos on Flickr, identify the representative image groups, and generate an aggregate representation based on locations that allows navigation, exploration and understanding of the general concepts.

From a technical perspective, our approach for selecting the representative images is constituted of three main stages. First, we apply clustering techniques to partition the image set into similar groups, based on bag-of-visual-words feature vectors [1]. By evaluating the intra-cluster densities as well as the cluster member numbers, we discard most of the irrelevant images and obtain a reduced set of images which are visually similar each other. This stage could be regarded as the "Filtering Stage". Then, we geographically cluster the reduced set of images and select large geographic clusters as representative regions. Here we use the *k*-means clustering algorithm based on the photos' geographic latitude and longitude. Finally, for each representative region, we perform the Probabilistic Latent Semantic Analysis (PLSA) [2] to identify the distinct "topics," and do additional clustering on the entire topic vectors and select the "significant" cluster as the representative results for this geographic region. In addition, with the help of map service, a map-based UI is designed to support the browsing photos in context and understanding of the general object concepts.

The remaining part of this paper is organized as follows. Section 14.2 gives an overview of the related work. Section 14.3 describes our approach for region-based selection of the representative photos. Experimental results are reported and analyzed in Sect. 14.4. Finally, we present our conclusions and discuss the future work in Sect. 14.5.

14.2 Related Work

Until several years ago, researches on geotagged images focused on only location-based photo browsing for a personal geotagged photo collection [3, 4], since it is almost impossible to obtain a large number of geotagged images. The situation has been changed after Flickr launched an online geotagging interface in 2006. At the present, Flickr has become the largest geotagged photo database in the world. Geotagging with GPS devices is too expensive to spread, but Flickr online geotagging system allows users to indicate the place where photos are taken by clicking the online map. In addition, Flickr database is open to everyone via FlickrAPI which allows users' program to search Flickr photo databases for geotagged images.

Therefore, some works on geotagged image recognition with huge Flickr geotagged image database has been proposed recently. Cristani et al. [5] and Cao et al. [6] proposed methods on event recognition of geotagged images by integrating visual features and geographical information. In general, a geotag represents a pair of values on latitude and longitude. It is a just 2-dimensional vector. To convert a

2-d vector into more rich representation, [7] and [8] converted geotags into visual information from the sky using aerial images, and [9] transformed geotags to words using reverse geo-coding technique. On the other hand, [10] used GPS trace data which is a series of geotags instead of using just a pinpoint geotag in order to classify images into several pre-defined events. [11] used time and seasons for geotagged image recognition in addition to visual information and geo-location data. While event or scene recognition on geotagged images is common, "IM2GPS" project [12] proposed a unique idea of estimating a place from just one non-geotagged image with 6 million geotagged images gathered from Flickr.

As extension of location-based photo browsing, several recent researches have considered the problem of selecting representative or canonical photographs for online image collections. Jaffe et al. [13] select a summary set of photos from a large collection of geotagged photographs based on only tags and geotags. By analyzing the correlations between tags and geotags, a map-based visualization "Tag Map" is developed to help indicate the most important regions and the concepts represented in those regions. Our work similarly identifies the most important regions and select representative photos for these regions. A key difference is that in [13], the concepts are learn which could be mostly affected by users' photographic behavior. While in our work, we aim to select representative photographs for particular concepts by applying computer vision techniques. Simon et al. [14] have proposed a method to select canonical views for the landmarks by clustering images based on the visual similarity between two views. Like [14], Kennedy et al. [15] attempt to generate representative views for the world's landmarks based on the clustering and on the generated link structure. Unlike the works [14] and [15], we choose the general category objects or scenes as our target, but not the identical objects like landmarks which rely on 3D structure or viewpoint. Crandall et al. [16] extended the work by Kennedy et al. [15] to the worldwide dimension by using 35 million geotagged photos. In addition, Crandall et al. [16] introduced the meanshift clustering [17] to detect representative regions instead of *k*-means clustering, which allows users to give a scale of representative regions in a natural way instead of the number of regions.

The work by Raguram et al. [18] another similar work. They aim to select iconic images to summarize general visual categories, like "love," "beauty," "closeup" and "apple." Since general visual or abstract concepts usually have many semantic "themes," their canonical view selection is hence defined as select a small number of salient images for each semantic "theme." Our goal is different as we aim to select representative photos for geographic regions in the worldwide dimensions, and we concentrate on the general concrete concepts such as "noodle" and "waterfall" as well [19]. [20] also treated with generic concepts. However, we select canonical images on generic concepts regarding several regions in the worldwide dimension, while they treated with general concepts within only given regions.

14.3 Proposed Approach

In this section, we propose a novel method to select representative photographs for regions in the worldwide dimensions, which helps detect cultural differences over the world.

14.3.1 Overview

Our approach for selecting the representative images for representative local regions from geotagged images consists of three main stages as shown in Fig. 14.1: (1) removing irrelevant images to the given concept, (2) estimating representative geographic regions, and (3) selecting representative images for each region.

First, we apply clustering techniques to partition the image set into similar groups, based on bag-of-features (BoF) vectors [1]. By evaluating the intra-cluster densities as well as the cluster member numbers, we discard most of the irrelevant images and obtain a reduced set of images which are visually similar each other. This stage could be regarded as the "Filtering Stage." The method employed in this stage is based on the method proposed by Raguram et al. [18].

Then, we geographically cluster the reduced set of images and select large geographic clusters as representative regions. Here we use the k-means clustering algorithm based on the geographic latitude and longitude of photos to obtain representative regions in the world for the given concept.

Finally, for each representative region, we perform the Probabilistic Latent Semantic Analysis (PLSA) [2] to identify the distinct "topics," do additional clustering on the entire topic vectors, and select the "significant" cluster as the representative results for this geographic region. In addition, with the help of map service, a UI is designed to support the browsing photos in context and understanding of the general object concepts.

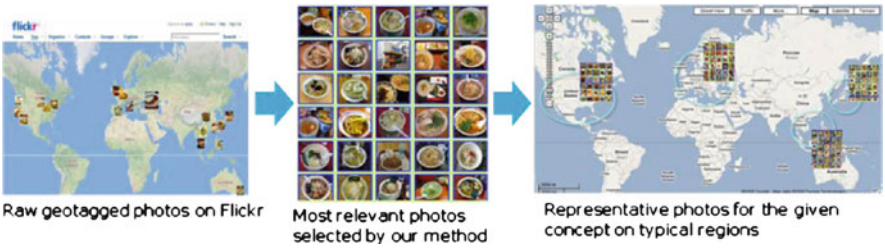

Fig. 14.1 After collecting geotagged photo related to the given concept by the tag-based search, we remove noise images, cluster regions and select regional representative images

14.3.2 Filtering Irrelevant Images

At most of the public photo sharing sites, the photos are organized by textual tags. Therefore, collecting photos using an input of query keywords normally returns many irrelevant or visually unrelated photos to the query concepts. Hence, in this stage we attempt to select most relevant photos to the concept from a large raw image set. As noted above, this stage can be regarded as "Filtering Stage" which based on clustering on bag-of-visual-words feature vectors and the intra-cluster similarity evaluation.

As follows, we describe bag-of-visual-words representation, visual clustering, and selecting the most relevant clusters by the evaluation of intra-cluster similarity in a detail way.

14.3.2.1 Image Representation

We adopt bag-of-visual-words model from [1] as the image representation. This model was first proposed for the text document analysis and recently applied in visual object recognition which has been found to be extremely powerful in tasks of representing the image features. The construction of bag-of-visual-words feature vectors for images involves several steps: (1) a set of points of interest are automatically detected in the image and local descriptors are computed over each point; (2) all the descriptors are quantized to form visual words; (3) for each image, we count the occurrences of each visual word to form a histogram of visual words which can be regarded as a bag-of-visual-words feature vector.

In our experiment, we first apply grid-based policy to detect the points of interest, and then compute the local descriptors by the Scale Invariant Feature Transform (SIFT) descriptor [21]. The SIFT descriptors are computed at 8 orientation directions over a 4×4 parts of spatial location, forming 128-dimensional vectors. Then we apply the k-means clustering algorithm over all extracted descriptors and computer the means to form visual words. Here we tried $k = 500$ and form a vocabulary of size 500. Finally, for each image, we assign all SIFT vectors to the nearest visual word and convert these vectors into one k-bin histogram which represents the bag-of-visual-words feature vector.

In our experiment, we still have used the color-based feature representation for the images as the comparison to the bag-of-visual-words representation. The performance of both methods is shown in the results part.

14.3.2.2 Visual Clustering

After building bag-of-visual-words representation for all raw images, we perform clustering using k-means algorithm over the bag-of-visual-words feature vectors to partition images into similar groups. In order to ensure a clear partition, we choose a high number of clusters k ($\approx$200 clusters for a dataset of about two thousand

images). Since most irrelevant and visually unrelated photos tend to fall into the small clusters, we can discard such small clusters based on a minimum threshold (usually less than 10 cluster members in our experiment).

14.3.2.3 Selecting the Most Relevant Clusters

Since there may still exist some clusters with large noises (irrelevant images), in order to detect such irrelevant clusters and select the most relevant clusters, we employ the method of evaluating the intra-cluster similarity for the remaining clusters. The intra-cluster similarity is the average similarity between the images that belong to the cluster and the similarity between two images P_i and P_j can be calculated using the cosine metric between two image vectors V_i and V_j:

$$sim(P_i, P_j) = \frac{V_i \cdot V_j}{\sqrt{\mid V_i \mid\mid V_j \mid}} \tag{14.1}$$

The (1) indicates that, if the cosine angle between two image vectors equals to 0°, the two images are very similar, whereas if the cosine angle between two image vectors equals to 90°, the two images are very dissimilar.

Then given a cluster of n photos, $\mathbb{C} = \{P_1, \ldots, P_n\}$, we can define the intra-cluster similarity as:

$$SIM(\mathbb{C}) = \frac{\sum_{P_i, P_j \in \mathbb{C}, i \neq j} sim(P_i, P_j)}{{}_nC_2} \tag{14.2}$$

which denotes the average similarity between two photos within one cluster.

By computing the intra-cluster similarity value for each cluster and sorting all clusters in the descending order of the SIM values, we select several top ones as the most relevant clusters (We selected 40 clusters in our experiments).

14.3.3 Detecting Representative Regions

In this stage, given the remaining most relevant photos, we attempt to detect representative regions based on the photos' geographic locations. For simplicity, we perform *k*-means clustering algorithm, based on the photos' geographic latitude and longitude (with the help of geotags), using geographical distance as the distance metric. Then we select several largest geo-clusters to form the representative regions since they have more relevant photos and the number of photos taken in a region is an indication of the relative importance of that region for the particular concept. (In our experiment, for simplicity, we generally select about four or five representative regions for each concept.)

14.3.4 Generating Representative Photographs

At this point, we have obtained the most relevant or visually similar photos, and the corresponding representative regions. To generate a set of representative photos for these representative regions, we explore the Probabilistic Latent Semantic Analysis (PLSA) [2] model, which is recently applied to recognize object categories in an unsupervised manner.

As a generative model, PLSA was originally used to discover latent topics in the text documents represented by bag-of-words. In a similar consideration, since images can be regarded as "documents" and represented by bag-of-visual-words, hence PLSA can be applied to images for discovering the object categories in each image. In terms of images, suppose we have a set of images $D = (d_1, \ldots, d_n)$, each containing the visual words from the visual vocabulary $W = (w_1, \ldots, w_m)$. By introducing a mediator known as latent topics $Z = (z_1, \ldots, z_k)$, we can build a joint probability model over images and visual words, defined as:

$$P(w, d) = P(d) \sum_{z \in Z} P(w|z) P(z|d) \tag{14.3}$$

where every image is modeled as a mixture of topics, $P(z|d)$, and $P(w|z)$ represents probability occurrence of visual words within a topic. We can learn the unobservable mixture parameters $P(z|d)$ and topic distributions $P(w|z)$ by the EM algorithm. Refer to [2] for a full explanation of the PLSA model.

As in our experiment, for each representative region, we apply the PLSA method to all the photos belonging to the region with a given number of topics, and get the probability distributions of all topics over each image, $P(Z|d)$, which can be regarded as topic vectors to represent an image. In the experiment, the number of topics was set to 20. After that, we aggregate photos according to the distributions of mixture topics by doing an additional step of clustering the topic vectors, $P(Z|d)$. In our experiments, we obtained the best results by applying k-means clustering with $k = 5$. Then the set of photos in the largest cluster are selected as representative photos of the given region, which is the final output of the proposed system.

14.4 Experimental Results

To test and verify if our approach works in practice, we conducted experiments with photos collected directed from Flickr. In order to ensure that the results would make a significant impact in practice, we concentrated on the most popular concepts including "noodle," "flower," "castle," "car," "waterfall," and "beach." The first four concepts are "object" concepts, while the rest are "scene" concepts. For each concept, we collected about 2000 most relevant geotagged photos distributed evenly in the world wide areas. As follows, we first provide numeric evaluation

on our proposed method for extracting the most relevant photos. In the second part, we primarily demonstrate the representative set of photos selected for several representative regions.

14.4.1 Quantitative Evaluation

To evaluate our method for extracting the most relevant photos described in Sect. 14.3.1, we use the precision, which is defined as $N_R/(N_R + N_{IR})$, and the recall, which is defined as $N_{R_{sel}}/N_{R_{col}}$, where N_R, N_{IR}, $N_{R_{sel}}$, and $N_{R_{col}}$ are the number of relevant photos, the number of irrelevant photos, the number of relevant photos in selected photos, and the number of relevant photos in raw collected photos, respectively.

For comparison, we still have applied color-based method to this selection task. This method is based on the color-feature representation [22] for the images. First, we quantized the RGB color space into 64 (4 for each axis) bins, and made a color distribution histogram for each image. Based on the distance (histogram intersection) between images, we clustered all images into groups using k-means algorithm, and finally selected the largest clusters to form the most relevant image set. In addition, for more intuitively comparing, we kept the number of images selected by this method is almost equivalent to selecting by the proposed method.

In Table 14.1, we present the evaluation results which show the precision of raw photos directly collected from Flickr, the precision and recall of photos selected by color-based method and our proposed method. We obtain an average precision of 80% and an average recall of 74% by using our proposed method, which outperform the 59% and 54% by using color-based method. It also shows that our method can identify and select most of relevant photos effectively, though the raw dataset has many irrelevant ones. For a clear comparison on the precision, we also show the evaluation in Fig. 14.1.

Table 14.1 Evaluation results. This table describes the number of raw photos directly collected from Flickr (the numerical value in the parenthesis represents the precision), the number of photos selected by color-based method and our proposed method (two numerical values in the parenthesis represents the precision and the recall)

Concepts	Raw photos from Flickr	Selection by color-based method	Selection by proposed method
noodle	2,080 (42)	769 (60, 54)	752 (90, 80)
flower	2,225 (60)	703 (71, 37)	705 (85, 45)
castle	1,848 (35)	780 (52, 61)	761 (70, 81)
waterfall	1,901 (39)	689 (63, 59)	672 (78, 70)
beach	1,917 (38)	824 (51, 58)	813 (80, 90)
car	1,908 (43)	817 (56, 55)	800 (77, 75)
TOTAL/AVG.	11,879 (43)	4,582 (59, 54)	4,503 (80, 74)

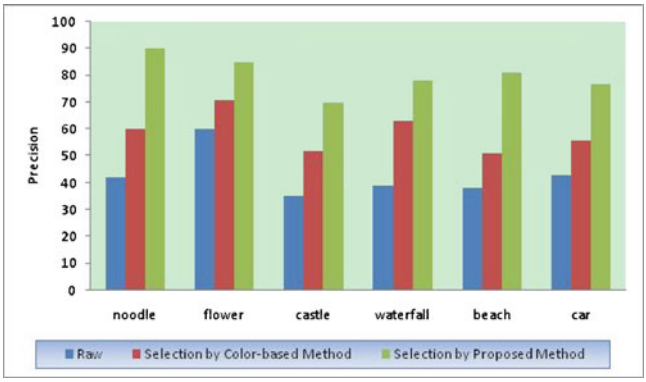

Fig. 14.2 Precision of raw photos directly collected from Flickr, photos selected by color-based method and our proposed method

14.4.2 Examples of Regional Representative Photos

In this subsection, we show the representative photos selected for several representative regions, while these regions were generated automatically based on geographic locations of the most relevant photos selected in the Sect. 14.4.1.

Figs. 14.3–14.7 show the results for the concept "noodle," each presents the most representative photos generated for the approximate regions: Japan, South East Asia, Europe as well as Mideast US, respectively. Without doubt, these results can help us understand about the "noodle" in these local areas. For example, Fig. 14.3 demonstrates many "ramen" photos in Japan, Fig. 14.4 demonstrates "spaghetti" photos in the European area, and Fig. 14.5 shows many noodles in the South East Asia area containing some Taiwanese style noodles and spicy Thai noodles, while others presents all kinds of peculiar "noodle" photos in other local areas.

We show the results for the concept "flower" in Figs. 14.8–14.12. The results demonstrate many species of flowers from five main approximate areas in the world: South East Asia, Europe, Central and South America, and South America. By combining the concept "flower" with geographic location, our system can not only help visualize varied and colorful appearance of the flowers, but also explore representative flowers for a concrete location, like Netherlands national flower "Tulip" (Fig. 14.9), North Carolina State flower "American Dogwood," Kansas State flower "Sunflower" (Fig. 14.11), and so on. With a larger set of "flower" photos, it is convinced that more representative regional flowers can be discovered.

While for the scene concept "waterfall," we extracted the representative photos for four large regions: Asia, Europe, North America, and South America. Figure 14.13 for the region of Asia and Fig. 14.14 for the region of South America. From the results, it is clear to find that waterfalls in South America seem to be more powerful, while waterfalls in the Asian area are somehow more beautiful. Such kinds of implications would be of importance in guiding travels around the world.

Fig. 14.3 Representative "noodle" photos for the region of Japan. Chinese-style noodle "ramen" is popular

Fig. 14.4 Representative "noodle" photos for the region of Europe. Most of the photos are "Spaghetti"

Fig. 14.5 Representative "noodle" photos for the region of South East Asia

Fig. 14.6 Representative "noodle" photos for the region of Mideast US

Fig. 14.7 Representative "noodle" photos generated for the region of Western US

Fig. 14.8 Representative "flower" photos for the region of South East Asia

Fig. 14.9 Representative "flower" photos for the region of Europe

Fig. 14.10 Representative "flower" photos generated for the region of Oceania

Fig. 14.11 Representative "flower" photos for the region of Central and South America

Fig. 14.12 Representative "flower" photos for the region of South America

Fig. 14.13 Representative "waterfall" photos for the region of Asia. They are somewhat beautiful

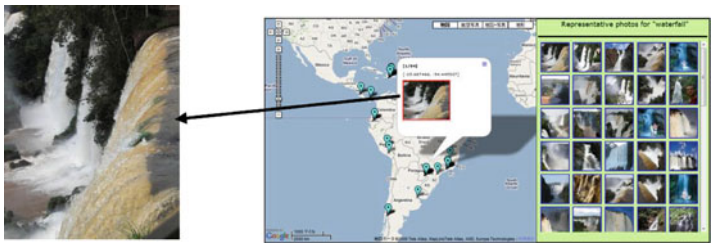

Fig. 14.14 Representative "waterfall" photos for the region of South America. They are more powerful

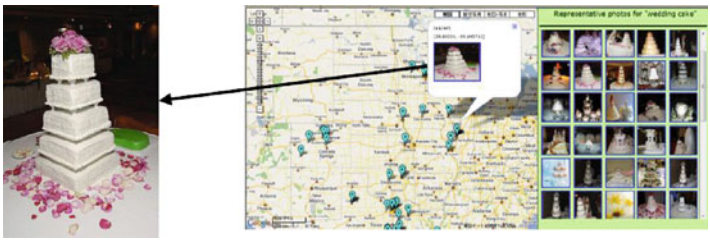

Fig. 14.15 "Wedding cake" in Mid US. Tall cakes are common. This is five-layered

Fig. 14.16 "Wedding cake" in Europe. They are much shorter and simpler than US

Figs. 14.15 and 14.16 correspond to "wedding cake" in Europe and in Mid US, respectively. We can find many of the wedding cakes in Mid US are much taller than ones in Europe. To see more results for other concepts, please visit the following Web page: `http://img.cs.uec.ac.jp/yanai/ASRP/`

14.5 Conclusion and Future Work

In this paper, we proposed a novel method to select representative photographs for typical regions in the worldwide dimensions, which helps detect cultural differences over the world regarding given concepts.

For future work, we plan to propose a method to discover cultural differences for many concepts from a geotagged image database automatically, and think out some other strategies in detecting more representative regions with a more precise scope. In addition, we will conduct some evaluations on the representativeness of the photos selected for the corresponding region.

References

1. G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pages 59–74, 2004
2. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001
3. K. Toyama, R. Logan, A. Roseway, and P. Anandan. Multiple instance learning for sparse positive bags. In *Proc. of ACM International Conference Multimedia*, pages 156–166, 2003
4. M. Naaman, Y.J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proc. of ACM International Conference Multimedia*, pages 53–62, 2004
5. M. Cristani, A. Perina, U. Castellani, and V. Murino. Geo-located image analysis using latent representations. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008
6. L. Cao, J. Luo, H. Kautz, and T. Huang. Annotating collections of geotagged photos using hierarchical event and scene models. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008
7. J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: Viewing the world with a third eye. In *Proc. of ACM International Conference Multimedia*, 2008
8. K. Yaegashi and K. Yanai. Can geotags help image recognition? In *Proc. of Pacific-Rim Symposium on Image and Video Technology*, 2009
9. D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. In *Proc. of ACM International Conference on Image and Video Retrieval*, 2008
10. J. Yuan, J. Luo, H. Kautz, and Y. Wu. Mining GPS traces and visual words for event classification. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2008
11. J. Yu and J. Luo. Leveraging probabilistic season and location context models for scene understanding. In *Proc. of ACM International Conference on Image and Video Retrieval*, pages 169–178, 2008
12. J. Hays and A. A. Efros. IM2GPS: Estimating geographic information from a single image. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2008
13. A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2006
14. I. Simon, N. Snavely, and S.M. Seitz. Scene summarization for online image collections. In *Proc. of IEEE International Conference on Computer Vision*, 2007
15. L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. of the International World Wide Web Conference*, pages 297–306, 2008
16. D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proc. of the International World Wide Web Conference*, 2009

17. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):603–619, 2002
18. R. Raguram and S. Lazebnik. Computing iconic summaries of general visual concepts. In *Proc. of IEEE CVPR Workshop on Internet Vision*, 2008
19. B. Qiu and K. Yanai. Objects over the world. In *Proc. of Pacific-Rim Conference on Multimedia*, 2008
20. T. Quack, B. Leibe, and L.V. Gool. World-scale mining of objects and events from community photo collections. In *Proc. of ACM International Conference on Image and Video Retrieval*, pages 47–56, 2008
21. D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004
22. M.J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991