EFFICIENT PLACE RECOGNITION WITH CANONICAL VIEWS

Lin Yang, John K. Johnstone, Chengcui Zhang

Computer and Information Sciences
The University of Alabama at Birmingham
{galabing,jj,zhang}@cis.uab.edu

ABSTRACT

We study the problem of place recognition. Given a photo, we estimate its location by scene matching to a large database of internet photos of known locations. Traditional strategies, which involve a linear scan of the database to find matching scenes, fail to scale. On the other hand, internet photos contain a massive amount of noise and redundancy, which is of little help for place recognition. By exploiting the scene distribution of photos, we summarize the database by a set of canonical views. The set of canonical views eliminates the noise and redundancy in internet photos, and provides a compact representation for the database. By restricting scene matching to the set of canonical views, we observe a good tradeoff between efficiency and recall: the average processing time for a query photo is reduced by 97%, while the recall rate for place recognition remains at 75%.

***Index Terms*—** place recognition, canonical view, internet photos

1. INTRODUCTION

Recent years have seen an explosion of internet photos. The explosion was brought about by the vast popularity of photo sharing sites such as Flickr [1]. Millions of internet users upload their personal photos online and share the photos with the public. The volume of internet photos has grown to multiple billions, and it keeps growing at a staggering speed.

The explosion of internet photos inspires research on *place recognition* – estimating the geographic location of a photo based on its photographed scene. Location information serves an important role in indexing and searching internet photos [2]. Despite the recent popularity of GPS devices, the portion of internet photos with GPS tags is relatively small: by querying on Flickr using popular keywords and switching the filter for GPS-tagged photos, we estimate that GPS-tagged photos are fewer than 10% of all photos on Flickr.

Although small in relative number, the absolute number of GPS-tagged photos is gigantic, in the order of hundreds of millions. More over, the collection of GPS-tagged photos provides an excellent coverage of the globe. Given a photo of unknown location, it is very likely that there exist multiple GPS-tagged photos of the same scene. Therefore, recent literature often reduces place recognition to image retrieval: having collected a database of photos of known locations, a query photo is matched to the database to retrieve photos of matching scenes; the locations of the matched database photos provide an estimate of the location of the query photo [3, 4].

However, traditional strategies for image retrieval, which involve a linear scan of the database ([4]), lead to suboptimal performance on internet photo collections. First of all, unlike traditional image databases, which are often collected by a few professionals, internet photos are contributed by millions of internet users, and therefore contain a massive amount of noise for the purpose of place recognition (*e.g.*, photos focusing on people and showing only fragments of the background scene). Secondly, the scene distribution of internet photos is highly non-uniform: a significant portion of photos share redundant scenes of a small number of places (such as landmarks), and there is a long tail of photos with little overlapping scene with others (such as random street corners). During a linear scan of the database, much of the computation is wasted either on matching to noisy photos that reveal little place information, or on matching to photos of redundant scenes over and over again. Little previous work has addressed the noise and redundancy issues.

In this paper, we minimize the impact of noise and redundancy on place recognition by exploiting the scene distribution of database photos and computing an optimal order in which database photos are matched to a query photo.

We leverage recent advances in the study of *canonical views*. Given a photo collection, canonical views are a subset of the photos that aim to summarize the important visual elements in the photo collection. In order to convey a maximal amount of important visual elements by a minimal set of photos, canonical views feature two characteristics: *representativeness* – the scene of each canonical view should represent many photos in the collection, and *diversity* – the scenes among different canonical views should be diverse [5]. Canonical views offer a solution to minimize the noise and redundancy in database photos, because (1) noisy photos are mostly not representative and thus not included in the canonical set, and (2) photos of redundant scenes will have at most one representative in the canonical set. Intuitively, by summa-

978-1-61284-350-6/11/$26.00

rizing the database photos by canonical views and restricting the scene matching of a query photo to the canonical views, we can greatly improve the efficiency of query processing with minimal loss in recall rate.

2. RELATION TO PREVIOUS WORK

With the proliferation of internet photos came a wave of work on scalable image retrieval. The work of Sivic and Zisserman [6] brings text retrieval techniques to the image domain by clustering SIFT features [7] to a finite number of states (termed *visual words*). The performance of visual words is improved by Nister and Stewenius [8] by clustering SIFT features hierarchically in to a *vocabulary tree*. Each image is converted to a bag of words by quantizing its SIFT features. Text retrieval techniques such as inverted index are employed to quickly select a small subset of database images as potential matches to a query image. Therefore the efficiency of query processing is greatly improved. This class of work is widely used for image retrieval in the large scale [3].

While previous work such as [6, 8] focuses on building an efficient search structure across an entire database, our work focuses on compressing a database into a compact representation, where noisy and redundant views are removed. The two approaches are orthogonal to each other and therefore can be combined, which is left for future work.

3. RELATED WORK

The work described in this paper leverages recent advances in canonical view selection for efficient place recognition. We briefly review the state of the art on both topics.

3.1. Canonical views

The study of canonical views has recently gained popularity in the research community with the proliferation of internet photos. In most literature, the selection of canonical views is reduced to a clustering problem: group photos into visually proximal clusters; the photos corresponding to cluster centroids serve for good candidates for canonical views, because each cluster centroid represents a frequently photographed scene (representative), while different cluster centroids have few features in common (diverse).

In [9], Simon *et al*. studies canonical view selection for tourist attractions. Visual proximities among photos are measured by SIFT matches [7]. A photo collection is partitioned into non-overlapping subsets among which no SIFT matches exist. Within each subset, photos are clustered using greedy k-means. The photos corresponding to cluster centroids are chosen as canonical views. In [5], Kennedy *et al*. propose a similar method, but leveraging both visual features and metadata for canonical view selection. The proposed method starts by clustering photos of a landmark using k-means based on global color and texture features of photos. A set of statistics on both photo metadata and visual features is collected to give an importance score to each cluster and all photos within the cluster. Canonical views are selected as top-ranked photos from top-ranked clusters.

A different approach to generating canonical views is to compute a ranking for all photos in the collection, such that the top k photos in the ranking approximate the k canonical views for the photo collection. In [10], photos are clustered hierarchically based on GPS tags. Several heuristics on the metadata (such as the distributions of time and photographers) are used to recursively score each sub-cluster in the hierarchy and therefore rank all photos in the collection. In [11], Yang *et al*. propose to compute a ranking of canonical views in two phases. During the first phase, photos are ranked by representativeness, using an analogue to the PageRank algorithm [12] applied to the image domain [13]. During the second phase, adaptive non-maximal suppression [14] is used to demote redundant views in the ranking, so that top-ranked photos are both representative and diverse.

Unlike clustering-based approaches, which typically require the number of canonical views (clusters) to be fixed, ranking-based approaches offer real-time canonical view selection of various sizes, once the ranking is computed for all photos in the collection.

3.2. Place recognition

Our work is most closely related to [3, 4]. In [3], Schindler *et al*. propose a system for place recognition on the scale of a city. They collect a database of 30K GPS-tagged street-side photos of a city. SIFT features are extracted from all database photos and organized by a vocabulary tree [8] for efficient matching. The location of a query photo is given by its top matched photo in the database using SIFT features. In IM2GPS [4], Hays and Efros leverage a database of over 6 million GPS-tagged internet photos for place recognition on the scale of the globe. A query photo is matched to each photo in the database using a combination of visual features. Mean shift is applied to the locations of the top k matched photos in the database to find the modes (density centers) in their geographic distribution. The locations of the density centers serve as the estimated locations of the query photo. Our work adopts the scene matching approach of [3, 4], but improves the efficiency of query processing by computing an optimal order in which database photos are matched to a query photo.

With the maturity of large-scale image-based modeling, a wave of work adopts structure from motion (SfM) techniques to reconstruct a 3D point cloud from the database of photos, and uses the point cloud as a basis for place recognition [15, 16, 17]. By registering photos in 3D and reconstructing the point cloud, various statistics can be accumulated such as the view count for each 3D point. Therefore, the scene structure can be exploited (*e.g*., which 3D points and associated 2D

views appear more frequently in the photos) to improve the efficiency of query processing by prioritizing 3D points for matching [17], compressing the 3D point cloud to a minimal cover of the location [16], and building an iconic scene graph for matching [15]. Our work is similar to this class of work in that we also exploit the scene distribution of database photos, but our work does not rely on SfM, which often entails more requirements on the photos (such as EXIF tags with the focal length information) and a higher computational cost.

4. CANONICAL VIEW RANKING

We use a modified algorithm of Yang *et al*. [11] to summarize database photos by a set of canonical views. We briefly review the two phases of the algorithm, followed by our modifications to the algorithm.

4.1. Phase 1: ranking representative views

Database photos are encoded by SIFT [7] features and converted to a visual similarity graph, where vertices are photos and edges indicate SIFT matches among photos. The PageRank algorithm [12] is applied to the visual similarity graph. PageRank has gained enormous success in finding authority webpages in a hyperlink network. It treats hyperlinks among webpages as votes and iteratively casts votes among webpages to update their authority scores. Analogously, applying PageRank to a visual similarity graph allows photos of matching scenes to vote for each other (because visual similarities are symmetric, each edge in a visual similarity graph is equivalent to two hyperlinks of opposite directions in a hyperlink network). Upon convergence, a photo of representative view gains more votes and receives a higher PageRank score. Thus the representativeness of photos are measured.

4.2. Phase 2: ranking canonical views

Ranking photos by representativeness leads to redundant views – photos of the same scene have similar representativeness scores and appear in blocks in the ranking. During the second phase, redundant views are demoted using adaptive non-maximal suppression [14]: a suppression radius is determined for each photo by its minimum visual distance (negation of the visual similarity used in Section 4.1) to a more representative photo; photos are re-scored and re-ranked by the count of photos within their suppression radiuses.

The re-scoring/re-ranking effectively demotes redundant views: of all photos of a same scene, only the most representative one remains high-scored, while others are demoted because their suppression radiuses are constrained by a more representative one and therefore have a low count of photos within their suppression radiuses. At the end of phase 2, the top-ranked photos are both representative and diverse.

4.3. Modifications

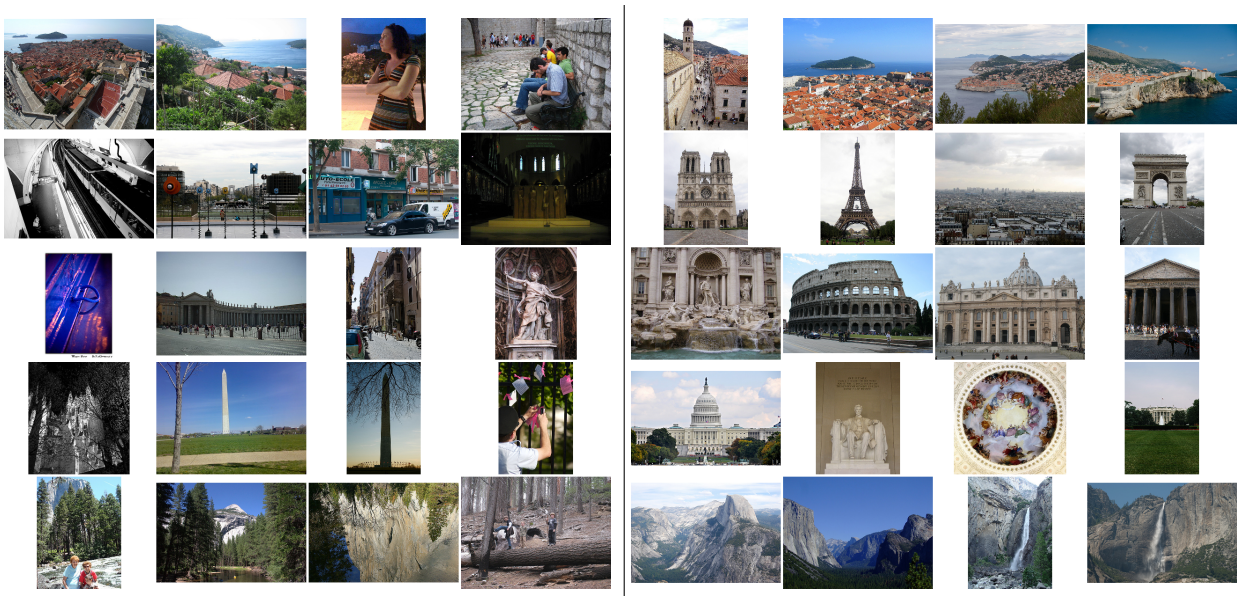
While the original algorithm of Yang *et al*. provides a sound foundation for our work, we find several disadvantages in applying the original algorithm to place recognition. In the sequel we list the disadvantages along with our modifications to the algorithm.

(1) Bias towards productive photographers. If a user takes thousands of photos of a random street corner, then by the original algorithm these photos will become representative after phase 1, and the most representative one will become canonical after phase 2, even though these photos only reflect one person's opinion. We make modifications to both phases to remove the bias. In phase 1, we remove from the visual similarity graph edges that connect photos of the same user, so that votes for representativeness are only propagated among photos of different users. In phase 2, we re-score photos by counting the number of *users*, instead of photos, within their suppression radiuses, so that a single user's opinion is never counted more than once. By applying these modifications, the bias we observed in our early experiments was removed.

(2) Redundancy towards the middle of a canonical view ranking. Top-ranked canonical views are usually diverse. However, redundant views start appearing towards the middle of a canonical view ranking. This is not surprising – the original algorithm does not *remove* any redundant views; it only *demotes* them in the canonical view ranking. For the purpose of place recognition, redundant views bring little new information to a database. In order to remove all redundant views, we add a post-processing step to the algorithm: after the canonical view ranking is computed, we scan the list of canonical views from top to bottom, removing any subsequent photo that have SIFT matches to the current one, until the list is exhausted. No SIFT matches exist among the remaining canonical views.

(3) Long tail of canonical views. Top-ranked canonical views usually have large user counts in their suppression radiuses, but the user count drops sharply towards the middle of a canonical view ranking. Towards the end of a canonical view ranking, there is a long tail of canonical views with tiny user counts in their suppression radiuses, which indicates only a handful of photos sharing a same scene. The popularity of the scenes are so low that these canonical views are rarely matched by query photos in place recognition. Therefore, we treat the long tail of canonical views as noise and remove any canonical view with a user count less than a certain threshold t in its suppression radius. In our experiments, we empirically set $t = 2$, which offers a good tradeoff between the compactness and coverage of canonical views. The threshold indicates that any canonical view must represent a minimum of 3 photos from distinct users (2 in the suppression radius plus 1 for the canonical view itself).

Noise and redundancy abound in internet photos. The last two modifications to the algorithm lead to about 96% com-



(a) Random views (b) Canonical views

Fig. 1. Random views and canonical views for the database. Each row shows four random views and canonical views for a site (from top down: Dubrovnik, Paris, Rome, Washing DC and Yosemite). The random views shed light on the amount of noise in internet image collections and justify our approach of canonical views, in which little noise or redundancy is observed.

pression of a database in our experiments. The remaining 4% of database photos are selected as canonical views and ordered by descending representativeness for place recognition.

5. PLACE RECOGNITION

Having a database of photos of known locations and an ordered list of canonical views, estimating the location for a query photo is straightforward: we scan the list of canonical views from top to bottom, matching each canonical view to the query photo. If a matching scene is found, the scan is terminated, and the location of the matching scene is reported as the estimated location for the query photo. If we exhaust the list of canonical views without a match, the query photo is rejected as not at any location covered by the database.

Since query processing is terminated as soon as a match is found, we aim for a low false positive rate during the scene matching between canonical views and a query photo. We use SIFT features to match photos, followed by a geometric verification using RANSAC [18] on the fundamental matrix [19]. After the geometric verification, if the number of remaining SIFT matches exceeds a certain threshold (16 in our experiments), the pair of photos is deemed to be a match. In a robustness test where we match photos from different sites, we observe zero false positives using the described procedure and threshold (see Section 6).

Table 1. Statistics on database photos and canonical views. The last two columns show the number of canonical views after the second and third modifications are applied to the canonical view ranking. Notice that a consistent $94 - 98\%$ reduction of photos is attained after both modifications.

Keyword(s)	♯ images	M2	M3
Dubrovnik	9350	6059	520
Paris	11997	9854	407
Rome	11959	8951	433
Washington DC	11991	10528	295
Yosemite	5756	3923	257

6. EXPERIMENTS

We evaluate place recognition on five photo collections of different sites: Dubrovnik, Paris, Rome, Washington DC and Yosemite National Park. The database consists of 51053 GPS-tagged photos downloaded from Flickr using keyword searches of corresponding site names (see Table 1 for photo statistics). All photos are downscaled to a maximum dimension of 640 pixels. The algorithm described in Section 4 is applied to each photo collection to select a set of canonical views for the corresponding site. The most time-consuming step – pairwise image matching – is distributed across a cluster of 120 CPUs, while the other steps of canonical view selection are done on a single machine. For all photo collections, canonical view selection takes less than a day to finish.

The top-ranked canonical views are shown in Figure 1.

We keep track of the number of canonical views after the second and third modifications are applied to the original algorithm (See Section 4.3). As shown in Table 1, there is a significant reduction in the number of canonical views after each modification is applied. Together, a 96% reduction is attained. By restricting query processing to this set of canonical views, the maximum processing time for a query photo (proportional to the number of scene matching operations) is also reduced by about 96%.

The significant reduction in processing time leads to some loss in recall rate: some photos that could have been matched to the database may fail to do so because all the corresponding database photos are missing from the canonical set. The loss in recall rate is quantitatively measured by conducting place recognition on a test set of photos against both the original database and the canonical views.

The test set consists of 400 GPS-tagged photos for each of the five sites, yielding 2000 query photos in total. The GPS tags of query photos are only used for localization error analysis. Photos in the test set are downloaded and downscaled in the same manner as database photos. However, a filtering is applied to ensure that the database and the test set share no photo/user in common.

Notice that the ground-truth recall rate will not be 100%. Since the query photos are just as noisy as database photos, a majority of them cannot be matched even by a full scan of the database. Since we are interested in the loss of recall caused by canonical views, our ground-truth recall rate is the one where query photos are matched to the entire database.

For each query photo, we match it against all database photos in random order. This provides ground truth data. Then we apply the method described in Section 5 to match the query photo against the canonical views for all sites – not only the canonical views of the same site, but those of the other sites as well – to test the robustness of place recognition.

Out of 2000 query photos, 280 have at least one match to the original database, and 206 have at least one match to the canonical views. We now analyze efficiency, precision, recall, and localization error in more depth.

Efficiency. Efficiency is measured by the average number of scene matching operations per query. A comparison of efficiency is shown in Table 2, where the method with canonical views saves as much as 97% of scene matching operations. Notice that the canonical views are ordered by descending representativeness. This means that a majority of query photos are expected to match to the top fraction of canonical views, which results in extremely efficient processing. Figure 2 plots the growth of matched query photos as more canonical views are scanned.

Precision. Precision of scene matching is measured by the percentage of correctly matched query photos among all matched query photos. For each matched query photo, we manually inspect its first match to both the entire database and the canonical views (if any) and determine if the matched photos indeed share the same view of the same place with the query photo. Of the 280 matches to the entire database, 5 query photos are matched on indoor objects or street signs of *different places*, causing incorrect estimates of their geographic locations. Therefore the precision of scene matching to the entire database is 98.21%. In comparison, all 206 matches to the canonical view database share the same view of the same place as the query photo: that is, CV scene matching has 100% precision, thanks to the strict procedure of SIFT matching and geometric verification.

Recall. As discussed before, we are interested in the relative recall – among all query photos that can be matched by the entire database, the percentage that are matched by the canonical views. The recalls are shown in Table 2, where the method with canonical views suffers a 25% loss. By inspecting the query photos corresponding to the loss, we find that

Table 2. Comparison of efficiency, precision, and recall. GT refers to the ground-truth method; CV refers to the method with canonical views. Efficiency is measured by the average number of scene matching operations per query.

	GT	CV	CV gain
efficiency (♯ ops)	2540.76	65.216	97.43%
	GT	CV	CV gain
precision (% correct)	98.21%	100%	1.79%
	GT	CV	CV loss
recall (♯ correct matches)	275	206	25.09%

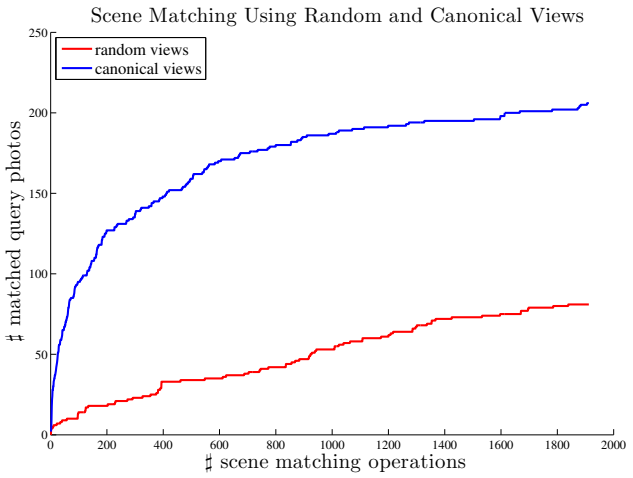


Fig. 2. Scene matching using random views and canonical views. The number of matched query photos is plotted against the number of scene matching operations required by using random views and canonical views. Notice that a majority of query photos are matched by the top fraction of canonical views, which results in extremely efficient query processing.

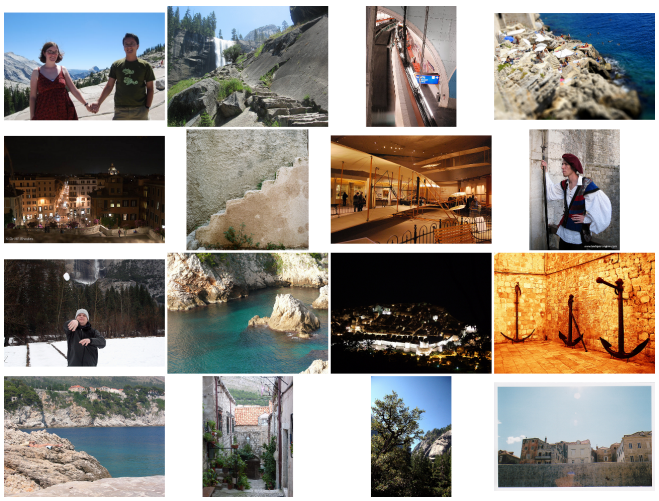


Fig. 3. Samples of difficult query photos. 16 random samples are shown of query photos having ≤ 2 matches in the database, which indicate rarely photographed scenes.

more than 95% of these photos have ≤ 2 matches in the entire database, which indicate rarely photographed scenes (see Figure 3). Place recognition for rarely photographed scenes is inherently a difficult problem. A small change in the database may result in different match/reject decisions. Therefore, we believe it is worth making a sacrifice on such photos in exchange for a significant improvement in efficiency.

Localization error. Each matched query photo has two GPS tags: one of its own (treated as ground truth) and the other predicted by a matched database photo. Localization error is measured by the great-circle distance between the two GPS tags. Among all matched query photos, the median localization error is 30.82m, which is promising given that the typical precision of civilian GPS devices is about 20m [20]. The lower and upper quartiles of localization errors are 14.57m and 103.98m respectively. Only a few predicted GPS tags are far off their ground truth (up to 6km), all of which are caused by incorrect GPS-tagging of either the query photo or the matched database photo.

7. CONCLUSIONS

We present an efficient method for place recognition. The principal novelty of the method is in compressing a database of photos into a compact representation by canonical views. The use of canonical views eliminates noise and redundancy in the database, and enables efficient place recognition with a high recall rate.

8. ACKNOWLEDGEMENTS

The work of the first and second authors was supported in part by a Google Research Award.

References

[1] "Flickr," `http://www.flickr.com/`.

[2] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Int'l Conf. World Wide Web (WWW)*, 2009, pp. 761–770.

[3] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–7.

[4] J. Hays and A. A. Efros, "im2gps: estimating geographic information from a single image," in *CVPR*, 2008, pp. 1–8.

[5] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *WWW*, 2008, pp. 297–306.

[6] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Int'l Conf. Computer Vision (ICCV)*, 2003, vol. 2, pp. 1470–1477.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, vol. 60, pp. 91–110, 2004.

[8] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *CVPR*, 2006, vol. 2, pp. 2161–2168.

[9] I. Simon, N. Snavely, and S.M. Seitz, "Scene summarization for online image collections," in *ICCV*, 2007, pp. 1–8.

[10] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries and visualization for large collections of geo-referenced photographs," in *Int'l Workshop Multimedia Information Retrieval*, 2006, pp. 89–98.

[11] L. Yang, J. K. Johnstone, and C. Zhang, "Ranking canonical views for tourist attractions," *Multimedia Tools and Applications*, vol. 46, no. 2–3, pp. 573–589, 2009.

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Computer Networks and ISDN Systems*, 1998, pp. 107–117.

[13] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, vol. 30, no. 11, pp. 1877–1890, 2008.

[14] M. Brown, R. Szeliski, and S. Winder, "Multi-image matching using multi-scale oriented patches," in *CVPR*, 2005, vol. 1, pp. 510–517.

[15] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *European Conf. Computer Vision (ECCV)*, 2008, pp. 427–440.

[16] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *CVPR*, 2009, pp. 2599–2606.

[17] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *ECCV*, 2010, pp. 791–804.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, 1981.

[19] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[20] "Global positioning system," `http://en.wikipedia.org/wiki/Global_Positioning_System`.