Multimed Tools Appl
DOI 10.1007/s11042-009-0345-1

Ranking canonical views for tourist attractions

Lin Yang · John Johnstone · Chengcui Zhang

© Springer Science + Business Media, LLC 2009

Abstract Online photo collections have become truly gigantic. Photo sharing sites such as Flickr (http://www.flickr.com/) host billions of photographs, a large portion of which are contributed by tourists. In this paper, we leverage online photo collections to automatically rank canonical views for tourist attractions. Ideal canonical views for a tourist attraction should both be representative of the site and exhibit a diverse set of views (Kennedy and Naaman, International Conference on World Wide Web 297–306, 2008). In order to meet both goals, we rank canonical views in two stages. During the first stage, we use visual features to encode the content of photographs and infer the popularity of each photograph. During the second stage, we rank photographs using a suppression scheme to keep popular views top-ranked while demoting duplicate views. After a ranking is generated, canonical views at various granularities can be retrieved in real-time, which advances over previous work and is a promising feature for real applications. In order to scale canonical view ranking to gigantic online photo collections, we propose to leverage geo-tags (latitudes/longitudes of the location of the scene in the photographs) to speed up the basic algorithm. We preprocess the photo collection to extract subsets of photographs that are geographically clustered (or geo-clusters), and constrain the expensive visual processing within each geo-cluster. We test the algorithm on two large Flickr data sets of Rome and the Yosemite national park, and show promising results on canonical view ranking. For quantitative analysis, we adopt two medium data sets and conduct a subjective comparison with previous work. It shows that while both algorithms are able to produce canonical views of high quality, our algorithm has the advantage of responding in real-time to canonical view retrieval at various granularities.

Keywords Canonical view ranking · Photo collections · Page Rank · Adaptive non-maximal suppression · SIFT · The wisdom of crowds

L. Yang (✉) · J. Johnstone · C. Zhang
Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: galabing@cis.uab.edu

J. Johnstone
e-mail: jj@cis.uab.edu

C. Zhang
e-mail: zhang@cis.uab.edu

Published online: 15 September 2009

1 Introduction

The proliferation of online photo collections has posed a great challenge to browsing gigantic amount of photographs. Current photo sharing sites simply retrieve all the photographs relevant to a keyword search and organize them into pages and pages of thumbnails, which leads to a tedious browsing experience (see Fig. 1a for example).

In this paper, we focus on a significant component of online photo collections – photographs contributed by tourists, and propose an algorithm that can automatically rank canonical views for tourist attractions. Ideal canonical views for a tourist attraction should both be representative of the site and exhibit a diverse set of views [18]. A small set of canonical views can compose a visual summary, which provides to the user a big picture

Fig. 1 **a** The first page of almost two million search results on keyword "Rome" on Flickr. Nearly two million photographs are retrieved. Irrelevant views abound. A user is likely to comb through pages and pages of thumbnails without getting the big picture of Rome. **b** The top five canonical views ranked by our algorithm on a fraction of the same photo collection. The tiny canonical view set captures a diverse set of popular views in Rome. More canonical views can be retrieved in real-time if requested

of the site instead of a huge number of noisy and redundant views (see Fig. 1b for example).

We take a completely data-driven approach and rely on the wisdom of crowds to rank canonical views. If a site is photographed many times by different people, the distribution of photographed views indicates the opinion of the crowds on which spots are more important than the others. Inspired by [15], we apply PageRank [3] over the collection of photographs to rank popular views for a tourist attraction. Analogous to Google search of websites [9], applying PageRank over a photo collection allows photographs of similar visual content to vote for each other, and upon convergence, a photograph with a popular view gains more votes and receives a higher PageRank score.

Although suitable for search applications, the PageRank results on photographs cannot directly serve as canonical views, because the top-ranked photographs are often dominated by a few most popular views and therefore lack diversity. The goal of canonical view ranking is to achieve both representativeness and diversity. Therefore we adopt a suppression scheme to demote duplicate views from the PageRank results. Based on the PageRank scores, photographs are re-ranked by their *visual dominance* so that duplicate views are spread out and the final ranking naturally approximates canonical views in all granularities (see Section 6 for details).

A novel contribution of our work is a ranking of the entire photo collection such that the top-k photographs in the ranking approximate the top-k canonical views for the tourist attraction. The philosophy of ranking distinguishes our work from most of the previous work that is clustering-based [23, 24]. While clustering can generate a set of canonical views (*e.g.*, by grouping photographs of similar content into clusters and selecting photographs corresponding to cluster centers), it cannot generate various numbers of canonical views in real-time, which requires the clustering algorithm to be rerun with a different parameterization. Ranking, on the other hand, frees us from recomputation when a different number of canonical views are requested, in which case we just return a different number of top-ranked photographs.

Another contribution that distinguishes our algorithm from clustering is that our algorithm requires no parameter tuning during ranking, while most clustering algorithms requires laborious parameter tuning in order to generate an appropriate number of clusters and cannot generalize well to various geographic regions. In our experiments, we apply the identical procedure to rank canonical views for Rome and the Yosemite national park, two tourist attractions with drastically different geographic features (architecture *vs.* nature) and distributions of landmarks (dense *vs.* sparse). We are able to achieve promising results on both sites. An unsupervised algorithm is especially beneficial for processing tourist attractions, because there are an enormous number of tourist attractions with different geographic features in the world.

The third contribution of our work is an optional preprocessing step to speed up canonical view ranking. The bottleneck of canonical view ranking is caused by pairwise matching of visual content over the photo collection (see Section 5 for details). The preprocessing leverages geo-tags (latitudes/longitudes of the location of the scene in the photographs), which have become widely available thanks to GPS-embedded camera devices. We observe that, for almost all tourist attractions, photographs are not uniformly distributed, but clustered at several spots (see Fig. 2 for an example). Therefore we preprocess the photo collection to extract subsets of photographs that are geographically clustered (or *geo-clusters*), and constrain the expensive pairwise matching within each geo-cluster. Given the wide availability and increasing popularity of geo-tags, geo-clustering can be applied to most online photo collections. However,

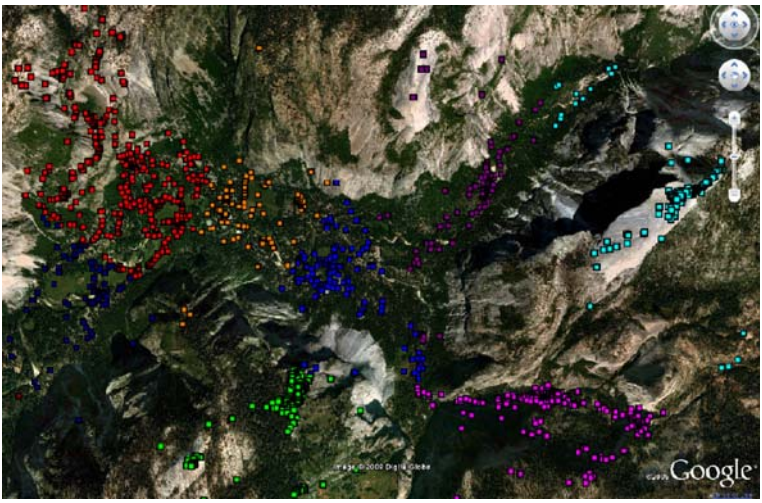


Fig. 2 Geo-clusters on a subset of the Yosemite photo collection. We can clearly observe several densely photographed geo-locations. Mean Shift tends to delineate geo-clusters through sparse regions, and many of the geo-clusters actually capture the iconic spots at Yosemite. For example, the Yosemite Valley (*red*), Glacier Point (*green*), Vernal Fall (*purple*) and Half Dome (*cyan*). We adopt the Google Earth API [10] for visualization

notice that the preprocessing is not required if geo-tags are not available, in which case it is equivalent to having a single geo-cluster containing all photographs. The basic algorithm for canonical view ranking operates on images only, with no requirement on any form of metadata.

The rest of the paper is organized as follows. In Section 2 we review previous work on canonical view selection. Section 3 gives a formal definition for ranking canonical views for a tourist attraction. In Section 4 we discuss geo-clustering. Sections 5 and 6 provide details for ranking view popularity for photographs and suppressing duplicate views to form a canonical set. The proposed algorithm is evaluated in Section 7 and we conclude in Section 8.

2 Related work

The explosion of online photo collections has triggered a number of attempts to summarize photo collections by a subset of canonical views. Below we give a review of the state of the art.

In [16], Jing et al. use SIFT features [20] to measure similarity between images and propose an algorithm for selecting a single iconic view for an image category of a commercial product. A visual similarity graph is formed with nodes being images and edges indicating similarity between images. The node (image) whose accumulated similarity with other nodes is the highest is selected as the iconic view for the category. Their work can be regarded as simplified canonical view selection, where a single canonical view is selected for each collection. In [15], Jing and Baluja generalize the algorithm by adopting the same similarity graph but using PageRank to give an importance score to all images in the category. Inspired by their work, we also adopt PageRank to rank popularity of

photographs, but our algorithm is followed by a suppression scheme to demote duplicate views so that the final results exhibit diversity. In [23], Raguram and Lazebnik propose to compute iconic views for images of an abstract concept (*e.g.*, images tagged with "love"). They use gist features [22] to encode the scene structure and pLSA [13] on image tags to encode textual themes for each image and perform joint clustering based on visual and textual features. The clustering results are subsets of images that are both perceptually and semantically similar. Iconic views are selected by choosing the image with highest visual quality [17] from each subset.

Apart from collections of commercial products or abstract concepts, tourist attractions provide a significant source of online photographs, and several recent attempts focus on canonical view selection for tourist attractions.

The work of Jaffe et al. [14] is the only previous work to our knowledge that computes a ranking of canonical views for the entire photo collection. However, their approach is purely based on the metadata of photographs (tags, geo-tags, author and time information, *etc.*). They first cluster photographs hierarchically based on geo-tags. Then they use several heuristics on metadata (such as distributions of time and photographers) to recursively score each sub-cluster in the hierarchy and therefore rank the entire photo collection. However, since their method only considers metadata but no visual information from photographs, their top-ranked views tend to include many duplicate and unrepresentative photographs.

Yang et al. [26] propose a canonical view selection algorithm based on greedy selection. They use quantized SIFT features (or visual words) [25] and encode each photograph with a distribution of visual words. Then they analyze the distribution of visual words and compute a coverage score for each photograph. Given a photo collection, they use a greedy scheme to select canonical views iteratively. During each round, the photograph with the highest coverage score is selected into the canonical set, corresponding visual words covered by the selected photograph are removed from consideration, and coverage scores for the rest of photographs are updated using the smaller set of visual words. The selected views are diverse because they contain a disjoint set of visual words. However, the ranking soon starts to include irrelevant views because visual words for representative views can only be seen once as they are discarded by greedy selection. See Section 7 for a detailed comparison with their algorithm.

In [24], a pure vision-based approach is proposed by Simon et al. Photographs are matched pairwise using SIFT features, and the original photo collection is divided into several non-overlapping subsets among which no matches exist. Within each subset of photographs, they cluster visually similar photographs using greedy k-means and select centroids from each cluster as canonical views for the subset. However, since k-means-based clustering is sensitive to outliers, which abound in online photo collections, selected centroids (canonical views) often do not correspond to any densely photographed viewpoints. Moreover, as a clustering-based algorithm, parameter tuning is required for a desired number of canonical views. See Section 7 for a detailed comparison with their algorithm.

Kennedy et al. [18] leverage both metadata and visual features to form a hybrid approach to canonical view selection. Their method starts by learning spatio/temporal patterns of user-provided tags and discovering tags related to landmarks. For each discovered landmark, photographs with corresponding tags are retrieved and clustered using k-means based on global color and texture features. Then they use a set of statistics on both metadata and visual features to give an importance score to each cluster and all photographs within the cluster. Canonical views are selected as top-ranked photographs from top-ranked

clusters. As they employ clustering during canonical view selection, parameter tuning is inevitable for different tourist attractions.

3 Problem definition

The goal of canonical view ranking can be formally defined as follows. Given a photo collection P of a tourist attraction, let C_k be a subset of photographs that can best summarize P out of all subsets of photographs of size k. Our goal is to re-rank photographs in P to generate P^*, such that $C_k=P^*(1\cdots k)$ for all positive integers k. Therefore, once P^* is generated, any number of canonical views can be generated simply by returning the top-ranked photographs in P^*.

4 Geo-clustering

Thanks to the wide availability of GPS devices, a considerable number of photographs in online photo collections include geo-tags. Flickr hosts more than 3 million geo-tagged photographs with 16 levels of accuracy from "world level" to "street level". In our data collection, we download photographs from Flickr that have geo-tags with "street-level" accuracy. Even though we require geo-tags of the highest accuracy, we are able to download tens of thousands of photographs for both of our experimental data sets. Since most of the "street-level" geo-tags are automatically saved in the image EXIF header when captured by a GPS-embedded camera, few photographs are tagged with a wrong geo-tag. Therefore, we believe geo-tags are both practical and reliable for the preprocessing step.

The goal of the preprocessing step is to extract from the original photo collection P subsets of photographs $\{G_1,G_2,\cdots G_M\}$ that are geographically proximal (geo-clusters). Because of the immense variety of the world's tourist attractions, it is important to note that geo-clustering is not dependent on any prior knowledge of tourist attractions (*e.g.*, the number of geo-clusters).

We adopt Mean Shift [8], a non-parametric clustering algorithm, to generate geo-clusters. Mean Shift treats data points (in our case, geo-tags of photographs) as samples from a density function and locates density peaks (modes) of the density function. Data points that converge to the same mode are associated with the same cluster. Since the extracted modes correspond to density peaks of data points, Mean Shift is well suited to the clustering of photographs of tourist attractions, which are densely clustered at iconic spots.

In order to use a simple distance metric, we convert coordinates of latitude/longitude to 3D coordinates centered at the earth center for Mean Shift clustering as suggested in [12]. After Mean Shift clustering, extracted clusters in 3D coordinates are projected back to the earth's surface to form geo-clusters. For robustness, we remove weak geo-clusters that consist of fewer than 100 photographs or 20 photographers. In all of our experiments, weak geo-clusters contain photographs taken by one or a few users at an unusual site.

Depending on the area and distribution of iconic spots, a typical tourist attraction can generate from several to tens of geo-clusters, with the number of photographs in each geo-cluster ranging from a few hundred to thousands. Because the bottleneck complexity is pairwise matching of visual content, which grows quadratically in the number of photographs, constraining the pairwise matching to each geo-cluster can lead to a large leap in efficiency. For example, if geo-clustering divides the original photo collection evenly into 25 geo-clusters, then visual processing of the geo-clusters will cost $(1/25)^2\times 25=4\%$

of the time of processing the original photo collection. In practice, even though photographs are not evenly divided into geo-clusters, we still observe a typical tenfold speedup.

On the other hand, the information loss caused by geo-clustering is insignificant. Most photographs in different geo-clusters are too far away to share any visual content, so it is safe to skip matching. Even in the exceptional case when photographs on the boundary of adjacent geo-clusters share visual content, the number of missing matches is negligible compared to the total matches among the photo collection, and has little effect on the following PageRank procedure. Besides, Mean Shift tends to delineate clusters along sparse regions, which further reduces the number of missing matches.

In Fig. 2, geo-clustering is illustrated for a subset of the Yosemite photo collection. Notice both the non-uniform distribution of photographs and the correspondence between iconic spots at Yosemite and the extracted geo-clusters.

5 Ranking the popularity of photographs

In this section we compute a popularity score for each photograph in $P=\{G_1,G_2\cdots G_M\}$. We depend on the wisdom of crowds to measure the popularity of photographs: if a view is preferred by a lot of photographers (*i.e.*, there are a lot of photographs of the view), it is a strong cue that the view is iconic for the tourist attraction, and therefore all photographs of the view should receive a high popularity score.

The ranking model of this section is inspired by the work of Jing and Baluja [15]. There are two components to the ranking model: a metric to measure similarity of content between photographs, and an algorithm to compute popularity scores based on the similarity metric.

For the similarity metric, we intentionally avoid textual tags in our current work. Although textual tags are used for indexing photographs by most of the current search engines such as Flickr, text annotations of user-contributed photo collections are noisy. Most photo sharing sites allow users to annotate a photograph with tags, which has no predefined dictionary or ontology. A recent survey [19] shows that only 50% of tags provided by Flickr users actually reflect the concepts in their photographs.

We use SIFT [20] to encode the visual content of photographs and to evaluate similarity between photographs. SIFT locates feature points on an image by detecting local extrema in an octave of difference of Gaussian (DoG) functions over scale space. Sub-pixel accuracy for feature points is achieved by fitting a quadratic function and interpolating the location of the maximum. For each feature point, a dominant orientation is computed so that the SIFT descriptor is invariant to image rotation. A SIFT feature descriptor is formed by aggregating gradient values in the rotated neighborhood of the feature point weighted by a Gaussian window. For a typical photograph, SIFT is able to generate hundreds to thousands of feature points. We refer readers to [20]; [21] for a detailed description of SIFT and a complete survey on local interest point descriptors.

The match for a SIFT feature point is the nearest neighbor of its descriptor in Euclidean space. In practice, because of the large volume of feature points and the high dimensionality (128 dimensions) of SIFT feature descriptors, brute-force matching based on exhaustive search is very inefficient. Beis and Lowe [2] propose a modified kd-tree structure, Best Bin First (BBF). BBF stores feature points in a kd-tree and checks only a small portion of bins in increasing order of Euclidean distance to the query feature point. It can only locate an approximate nearest neighbor for a query feature point, but costs a fraction of the time compared to brute-force matching. We use the approximate nearest neighbor library ANN of Mount and Arya [1] to match SIFT features, which implements an algorithm similar to

BBF. After a set of point matches are obtained between two photographs, we detect and remove false matches using a robust estimator RANSAC [6] based on the constraints of epipolar geometry [11].

We extract SIFT features for all photographs in $P=\{G_1,G_2\cdots G_M\}$, but only match pairs of photographs in the same geo-cluster. The similarity metric between two photographs is defined to be the number of their SIFT matches divided by their average number of SIFT features:

$$s(P_i,P_j)=\frac{\left\|matches(P_i,P_j)\right\|}{\left(\left\|features(P_i)\right\|+\left\|features(P_j)\right\|\right)/2}. \tag{1}$$

The output of this step is a set of square symmetric similarity matrices $\{A_1,A_2\cdots A_M\}$ where $A_k(i,j)$ measures the similarity between the i^{th} and j^{th} photographs of geo-cluster G_k. Since photographs from different geo-clusters are skipped for matching, the similarity between any pair of photographs from different geo-clusters is set to be 0. We can accumulate the similarity matrices $\{A_1,A_2,\cdots A_M\}$ to form a single square symmetric similarity matrix A that measures the similarity between photographs in P:

$$A=\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \cdots & \\ & & & A_M \end{pmatrix}. \tag{2}$$

From this similarity matrix, we can form a similarity graph over P, where nodes are photographs and edges connect photographs with a positive similarity score between them. Matrix A is therefore the matrix representation of the similarity graph.

We adopt PageRank [3] to rank the popularity of photographs based on the similarity graph. PageRank is a successful algorithm for off-line ranking of websites. It analyzes the link structure of the web and treats links as votes between websites. A score is computed for each website iteratively, determined by both the number of votes the website receives and the importance of its voter websites. When applying PageRank to the image domain, websites are replaced by photographs, and links between websites are replaced by visual similarity between photographs. The philosophy of ranking is: if two photographs are visually similar, they cast a vote for each other. A photograph receiving a lot of votes indicates that its view is preferred by a lot of photographers and is therefore likely to be iconic for the site.

Algorithmically, all nodes (photographs) in the similarity graph are initialized to the same popularity score. During each round of PageRank iteration, the popularity score of a node is distributed across all of its edges to its neighboring nodes in the similarity graph. Mathematically, the PageRank algorithm can be written in the matrix form:

$$R^{(t+1)}=A^{*}R^{(t)}, \tag{3}$$

where A^{*} is a column-normalized version of A, and $R^{(t)}$ is a column vector containing the popularity scores for all nodes after the t^{th} round of iteration. $R^{(0)}$ is initialized to $\left(\frac{1}{n}\right)_{n\times 1}$ and $R^{(\infty)}$ gives the true PageRank scores. In practice, PageRank iteration is terminated when the change of popularity score for all nodes is below a certain threshold.

PageRank scores correspond to the dominant eigenvector of A^{*}, which is unique if and only if the similarity graph is connected [5]. Since our similarity graph is at least

disconnected between different geo-clusters, strong connectivity is guaranteed in practice by adding a damping factor to Eq. 3 as in [3]:

$$R^{(t+1)} = dA^{*}R^{(t)} + (1-d)\left(\frac{1}{n}\right)_{n\times 1}, \tag{4}$$

where d is the damping factor and n is the number of nodes in the similarity graph. The new equation is equivalent to adding a lightly weighted edge between all pairs of nodes in the similarity graph to make it connected. Damping factor d=0.85 is often chosen in practice, which is the setting for all experiments in this paper. With Eq. 4, popularity scores R can be computed iteratively using the following procedure:

Algorithm *Popularity Ranking*

1: $R = \left(\frac{1}{n}\right)_{n\times 1}$

2: **do**

$R' = R$

$R = dA^{*}R' + (1-d)\left(\frac{1}{n}\right)_{n\times 1}$

while $\|R - R'\| \geq \varepsilon$

Upon convergence, each photograph in P has a popularity score, which measures the representativeness of the photograph of the tourist attraction.

Photographs with poor quality (*e.g.*, blurry or over/under-exposed photographs) are implicitly demoted in the ranking of PageRank, because few stable SIFT features can be extracted and matched on such photographs and therefore they receive few votes during PageRank iteration.

Local interest point descriptors such as SIFT are known to have difficulties in handling drastic lighting or viewpoint changes, in which case few stable matches can be found between the two views [21]. However, these difficulties have little effect on the PageRank results under the context of the wisdom of crowds: SIFT is robust enough to detect and match stable features on a majority of photographs, and the heavy redundancy of views in online photo collections is more than sufficient to infer the true popularity of each spot of the site.

In Fig. 3, our popularity ranking is demonstrated for the Rome data set. Flickr has a "most relevant" search option, which ranks retrieved photographs in decreasing relevance to the search keywords. Figure 4 shows the advantage of our algorithm by comparing our results to Flickr's ranking of the most relevant photographs.

6 Ranking canonical views

The popularity scores given by PageRank provide a ranking of photographs, in which top-ranked photographs tend to correspond to iconic spots for a tourist attraction. However, top-ranked views cannot be used directly as canonical views for the site, because they often

Fig. 3 Popularity ranking for the Rome data set. The first rows shows the top five photographs and the second row shows the bottom five photographs ranked by decreasing popularity score. It is obvious that the top-ranked photographs capture iconic views of Rome, while the bottom-ranked photographs hardly convey any information about Rome

contain duplicate content (see Fig. 3 for example) and therefore may not exhibit a diverse set of views within the top several photographs.

In this section we provide a re-ranking scheme that suppresses duplicate views. The re-ranking scheme is inspired by adaptive non-maximal suppression (ANMS) [4], where a photograph with a high popularity score suppresses all photographs that are visually similar but have a lower popularity score, and photographs are re-ranked according to the number of photographs they suppress, or their *visual dominance*. Because duplicate views are suppressed by the photograph with a higher popularity score, only that suppressing photograph will remain top-ranked and the rest of the duplicate views are demoted. Meanwhile, photographs of other popular views are promoted to the top of the ranking.

Since we encode each photograph by visual features and derive similarity metrics between photographs, we can consider photographs to be points in a high dimensional feature space, each with an associated popularity score. ANMS computes for each photograph P_i a suppression radius r_i within which its popularity score is the maximum:

$$r_i = \max\{r : R_i > R_j \text{ for all } j \text{ such } d(P_i, P_j) < r\} \tag{5}$$

where R_i and R_j are the popularity scores of photographs P_i and P_j, respectively and $d(P_i, P_j)$ measures the distance between two photographs. We define $d(P_i, P_j)=1-s(P_i, P_j)$, where

Fig. 4 A comparison to Flickr's relevance measure. The left view shows Flickr search results with the "most relevant" option applied to the Rome data set. The right view shows our popularity ranking. Due to the noise of metadata (such as user provided title, tags, *etc.*), only a few of the top views returned by Flickr are recognizable as views of Rome. On the other hand, the top-ranked photographs from our popularity ranking successfully capture several iconic spots of Rome

$s(P_i, P_j)$ is the similarity between two photographs P_i and P_j defined in Eq. 1. After the suppression radius for a photograph P_i is computed, the visual dominance of P_i is defined as the number of photographs that fall into the radius (those are the photographs that are suppressed by P_i). In practice, by sorting all photographs in decreasing order of popularity score, a photograph needs only to be compared to higher-ranked photographs to compute the suppression radius, and lower-ranked ones to count the number of photographs it suppresses:

Algorithm *ANMS*

1: sort photographs in P in decreasing order of popularity score
2: $C=(0)_{1\times n}$ // C_i is the visual dominance of photograph P_i
3: **for** i from 1 to n

```
r=Infinity
for j from i – 1 to 1
    if d(P_i, P_j)<r then
        r=d(P_i, P_j)
for j from i+1 to n
    if d(P_i, P_j)<r then
        C_i=C_i+1
```

4: sort P in decreasing order of C

After ANMS, the top-ranked views still have high popularity scores but duplicate views are suppressed, so the top-ranked views contain a diverse set of popular views and can be used as canonical views for the site. The visual dominance metric naturally approximates canonical views at all granularities: in the top few canonical views, photographs exhibit a diverse set of the most popular spots; as more canonical views are retrieved, photographs with slight duplicate content (maybe a different aspect of the same popular spot) start to appear in the ranking. Duplicates are inevitable as more and more canonical views are requested. According to a study in [18], people prefer duplicate but representative views over irrelevant views when browsing landmarks. Figure 1 shows the top five canonical views for the Rome data set.

7 Experimentation

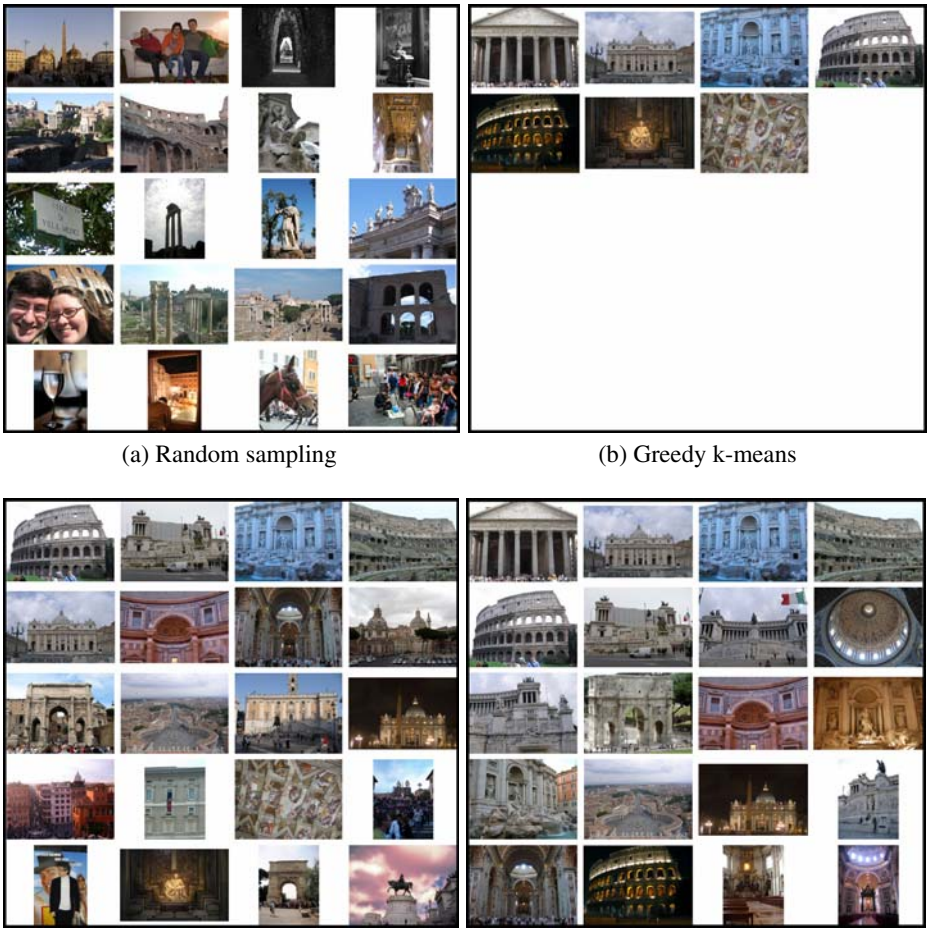
We downloaded approximately 40,000 photographs from Flickr using the search keywords "Rome Italy" and "Yosemite". Photographs without street-level geo-tags were skipped for downloading, and all downloaded photographs were downscaled to a maximum dimension of 512 pixels. The Rome data set contains approximately 30,000 photographs and the Yosemite data set contains approximately 10,000 photographs. Around 20 geo-clusters are extracted for each collection. We prune all the geo-clusters at a 1km radius to further reduce the number of pairwise matching. Since iconic spots typically occupy a small region, this pruning preserves most views of interest. The bottleneck computation – pairwise matching of SIFT features within each geo-cluster – was carried out on a cluster of 200 CPUs, and all other steps were done on a single machine. The total runtime from geo-clustering to canonical view ranking was around 3 days for the Rome data set and 1.5 days for the Yosemite data set. For both data sets, the photographs were ranked by popularity using PageRank and duplicate views were demoted using ANMS. The output is a ranking of

Fig. 5 Left: a random sampling of nine photographs from the Yosemite data set. Right: the top nine canonical views computed by our algorithm. These canonical views capture various aspects of the Yosemite national park, and do not contain photographs that focus on people or non-iconic views

photographs such that the top-k photographs correspond to the top-k canonical views for the tourist attraction. The right of Fig. 5 shows our top nine canonical views for the Yosemite data set, compared to nine random photographs on the left of Fig. 5.

We further evaluate our ranking of canonical views by comparing it with the canonical views produced by two other methods on the Rome data set. The first method is greedy k-means as used in [24] (Fig. 6b), which builds the canonical view set by iteratively adding a photograph that minimizes an objective function. The objective function is designed such that each photograph should be represented by one of the canonical views, and the SIFT features of selected canonical views should not overlap. The second method is a greedy winner-take-all strategy similar to [26] (Fig. 6c), except that we use original SIFT features instead of quantized SIFT features: given a list of photographs ranked by decreasing popularity score (PageRank result), the photograph with the highest popularity score is iteratively added to the canonical view set, and all photographs that share the same SIFT features are removed from the list, until the list of photographs is empty.

In Fig. 6b–d, the top canonical views are shown for each method, along with a random sampling of photographs in (a) as base-line. All three methods achieve significant improvement over random sampling. Since the number of canonical views selected by greedy k-means is determined by parameters α and β of the objective function, the algorithm must be run again with a different setting whenever a different number of canonical views is required. We adopted a similar setting to the one suggested in the paper [24], and only seven canonical views were generated. The greedy winner-take-all strategy is able to generate reasonable results for the first several canonical views. However, every time an iconic photograph is added to the canonical view set, all SIFT features relevant to the view are removed: therefore, after several rounds, all photographs of iconic views are out of consideration, and noise starts to appear in the canonical view set (see the last few of the top 20 canonical views). Our algorithm, on the other hand, achieves reasonable quality at all granularities. The top several views have considerable overlap with the two other methods yet exhibit a diverse set of views, and as the number of canonical views grows, less iconic but still representative views of Rome start to appear in the set along with a few duplicates. (Notice that the duplicate of the Roman Coliseum shows this site at night, adding valuable context to the earlier daytime view.)



(a) Random sampling (b) Greedy k-means

(c) Greedy winner-take-all (d) Our algorithm

Fig. 6 Comparison of different canonical view selection methods on the Rome data set. **a** shows 20 randomly selected photographs. **b** selects canonical views using the greedy k-means algorithm of Simon et al. [24] with α=2 and β=100. Only seven canonical views are generated. **c** selects canonical views using a greedy winner-take-all algorithm similar to [26]. Irrelevant photographs start to appear toward the last several canonical views. **d** shows the top 20 canonical views found using our algorithm. Only iconic views for Rome appear in the canonical view set

It is difficult to quantitatively evaluate the quality of canonical views, because there is no ground-truth answer to the question "which subset of photographs *best* summarizes a tourist attraction?" Since the goal of canonical view selection is to create a better browsing experience for the user, we conduct a subjective study to evaluate our ranking of canonical views against human judgment.

We obtained the Pantheon data set (1,112 photographs) and Bay Bridge data set (566 photographs) from the project website of Simon et al. [24] and let five human judges manually cluster the photo collection. The judges were asked to group photographs into a hierarchy of viewpoints with no further restrictions. Since the data sets are relatively small, and access to other resources was available, the judges had a good knowledge of the tourist

attraction before clustering a photo collection. The manual clustering produced a hierarchy of photographs with each level of the hierarchy being a finer division of viewpoints of its parent's viewpoint. We collect the leaf level clusters as the ground-truth clustering for the photo collection.

Treating the manually generated clusters as ground-truth, we can quantitatively measure the quality of a canonical view set by its coverage of view clusters: the more view clusters are covered in a canonical set, the more information the canonical set conveys to the user. In Fig. 7, we compute the view cluster coverage against all ground-truth clustering and plot the average view coverage of our top-k canonical views as a function of k. We also acquire the canonical view results of Simon et al. [24] from their project website for comparison. Our method exhibit a slightly higher coverage of view clusters at the beginning. More importantly, the results of Simon et al. soon terminate at a small value of k with only about half of the view clusters covered, while our canonical view set continues to grow at a steady rate until all view clusters are covered.

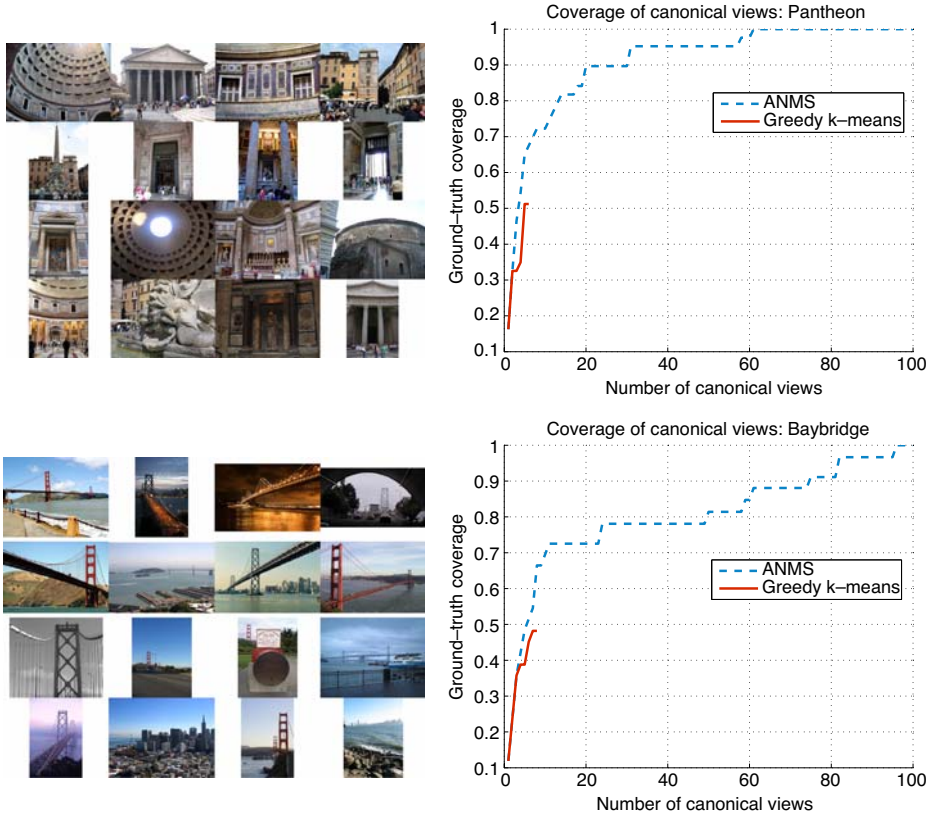


Fig. 7 A quantitative study of canonical view selection. The left column shows the manual clustering of a photo collection into a number of view clusters by one of our judges. The right column shows the average coverage of ground-truth clusters by top-k canonical views as a function of k. Our canonical view set achieves slightly better results on small k values. More importantly, greedy k-means terminates with only about half of view clusters covered while our canonical view set can continue growing to cover all view clusters

8 Conclusions and future work

In this paper we have presented an algorithm for ranking canonical views of a photo collection from a tourist attraction. We encode photographs with SIFT features and use PageRank to compute a popularity score for each photograph, leveraging the wisdom of crowds. Photographs are then ranked in decreasing popularity, and we use adaptive non-maximal suppression to demote the ranking of duplicate views, yielding top-ranked views that present a diverse but representative set of views.

In order to conquer the computational bottleneck of pairwise matching of photographs, we leverage geo-tags of photographs to preprocess the photo collection. We use a non-parametric clustering algorithm, Mean Shift, to extract subsets of photographs that are geographically proximal, and constrain the pairwise matching to each geo-cluster without sacrificing the quality of canonical views.

The proposed algorithm is unsupervised and requires no parameter tuning on different tourist attractions. Compared to clustering-based methods that only extract a pre-set number of canonical views, our algorithm computes a ranking of photographs such that the top-k photographs in the ranking correspond well to the top-k canonical views for the tourist attraction. This allows various numbers of canonical views to be retrieved without re-initiating the algorithm.

A limitation of our work is the lack of explicit quality assessment of the canonical views. User-contributed photographs vary drastically in visual quality. When multiple photographs exist for an iconic view, an ideal choice would be the one with the highest visual quality so that the selected canonical views have a guidebook look. Currently, our algorithm implicitly demotes photographs with extremely poor quality (*e.g.*, blurred, under/over-exposed), because such photographs have few SIFT matches with other photographs and therefore will receive a low popularity score during popularity ranking. A future direction of our work is to consider vision or metadata based quality measures for photographs (such as the visual quality assessment proposed by Ke et al. [17]) and incorporate visual quality into canonical view selection.

Acknowledgments The authors would like to thank Wei-bang Chen, Srinivasa Datla, Sagar Thapaliya, Richa Tiwari and Liping Zhou for generating ground-truth view clusters on two of the test data sets.

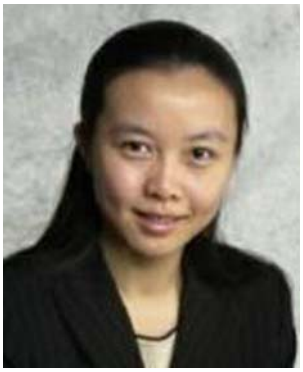
References

1. ANN. from http://www.cs.umd.edu/~mount/ANN/.
2. Beis J, Lowe DG (1997) Shape indexing using approximate nearest-neighbor search in high dimensional spaces. IEEE Comp Vision Patt Recog 1000–1006
3. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Systs 30(1–7):107–117
4. Brown M, Szeliski R, Winder S (2005) Multi-image matching using multi-scale oriented patches. IEEE Comp Vision Patt Recog 510–517
5. Bryan K, Leise T (2006) The $25, 000, 000, 000 eigenvector: the linear algebra behind Google. SIAM 48(3):569–581
6. Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24:381–395
7. Flickr. from http://www.flickr.com/
8. Georgescu B, Shimshoni I, Meer P (2003) Mean shift based clustering in high dimensions: a texture classification example. IEEE Int Conf Comp Vis 456–463
9. Google. from http://www.google.com/

10. GoogleEarthAPI. from http://code.google.com/apis/earth/
11. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision. Cambridge University Press
12. Hays J, Efros AA (2008) IM2GPS: estimating geographic information from a single image. IEEE Comp Vision Patt Recog 1–8
13. Hofmann T (1999) Probabilistic latent semantic analysis. Uncertainty Artif Intell
14. Jaffe A, Naaman M, Tassa T, Davis M (2006) Generating summaries and visualization for large collections of geo-referenced photographs. MIR 89–98
15. Jing Y, Baluja S (2008) Pagerank for product image search. International Conference on World Wide Web 307–316
16. Jing Y, Baluja S, Rowley H (2007) Canonical image selection from the web. CIVR 280–287
17. Ke Y, Tang X, Jing F (2006) The design of high-level features for photo quality assessment. CVPR 419–426
18. Kennedy LS, Naaman M (2008) Generating diverse and representative image search results for landmarks. WWW 297–306
19. Kennedy L, Chang S, Kozintsev I (2006) To search or to lable? Predicting the performance of search-based automatic image classifiers. MIR 249–258
20. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
21. Mikolajczyk K, Schmid C (2004) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
22. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
23. Raguram R, Lazebnik S (2008) Computing iconic summaries for general visual concepts. IV
24. Simon I, Snavely N, Seitz S (2007) Scene summarization for online image collections. ICCV 1–8
25. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. IEEE Int Conf Comput Vis 2:1470–1477
26. Yang Y, Wu P, Lee C, Lin K, Hsu W, Chen H (2008) ContextSeer: context search and recommendation at query time for shared consumer photos. ACM Multimedia

Lin Yang is a Ph.D. student of Computer and Information Sciences at University of Alabama of Birmingham. His research is on mining useful information from community-contributed multimedia data using data mining, machine learning and computer vision techniques. He received the B.S. degree in Computer Science from Fudan University in 2006.

John K. Johnstone is an Associate Professor of Computer and Information Sciences at UAB. He received his B.Sc. in Mathematics from the University of Saskatchewan, and his M.S. and Ph.D in Computer Science from Cornell University. He has also been on the faculty at Johns Hopkins University. His primary research interest is shape modeling, with recent interest in the interface of graphics and vision. His research has been supported by several NSF grants.



Chengcui Zhang is an Assistant Professor of Computer and Information Sciences at University of Alabama at Birmingham (UAB) since August, 2004. She received her Ph.D. from the School of Computer Science at Florida International University, Miami, FL, USA in 2004. She also received her bachelor and master degrees in Computer Science from Zhejiang University in China. Her research interests include multimedia databases, multimedia data mining, image and video database retrieval, bioinformatics, and GIS data filtering. She is the recipient of several awards, including the IBM Unstructured Information Management Architecture (UIMA) Innovation Award, UAB ADVANCE Junior Faculty Research Award from the National Science Foundation, and UAB Faculty Development Award.