Scene Summarization for Online Image Collections

Ian Simon Noah Snavely Steven M. Seitz
University of Washington

Abstract

We formulate the problem of scene summarization as selecting a set of images that efficiently represents the visual content of a given scene. The ideal summary presents the most interesting and important aspects of the scene with minimal redundancy. We propose a solution to this problem using multi-user image collections from the Internet. Our solution examines the distribution of images in the collection to select a set of canonical views to form the scene summary, using clustering techniques on visual features. The summaries we compute also lend themselves naturally to the browsing of image collections, and can be augmented by analyzing user-specified image tag data. We demonstrate the approach using a collection of images of the city of Rome, showing the ability to automatically decompose the images into separate scenes, and identify canonical views for each scene.

1. Introduction

How can the visual splendor of Rome be conveyed in a few images? While a good guidebook can provide a lot of information and context to plan your trip, guidebooks tend to be far less efficient at conveying what you should expect to see. This paper addresses the problem of automatically selecting images that best summarize a scene by analyzing vacation photos for a large population of people.

If a site is visually interesting, it's almost certain that there are several photos of it on the Internet, uploaded by people who have visited that site in the past. Hence, the collection of photos on the Internet comprises an extremely rich and increasingly comprehensive visual record of the world's interesting and important sites. However, the unorganized nature of this collection makes finding relevant photos very difficult. For example, a search for "rome" on the photo-sharing site Flickr [2] returns several hundred thousand thumbnails, listed in seemingly random order.

Our objective is to automatically derive, from thousands of photos downloaded from Internet sharing sites, a one page visual summary of a scene or city that captures the key sites of interest. In an interactive setting, a user can see "canonical views" of each site of interest, and browse photos on the Internet that correspond to each canonical view. When textual user "tag" data is available, we show how it can be used to augment scene summaries by analyzing the tag statistics.

Our approach to scene summarization involves three problems. The first is to partition the image set into groups of images, each corresponding to a different representative view of that scene. The second is to identify a canonical view to represent each group. The third is to compute textual tag information that best represents each view. Computing a city summary further requires identifying all of the distinct sites of interest in that city. (See Figure 1 for an example summary of the Vatican.)

At a technical level, our approach works by applying clustering techniques to partition the image set into groups of related images, based on SIFT feature co-occurrences. The clustering is performed using a greedy method that we have found outperforms k-means for this application. Canonical views are found by using a likelihood measure, also defined based on feature co-occurrences. Descriptive textual tags are computed using probabilistic reasoning on histograms of image-tag co-occurrences. Due to the large amount of noise in user tags, obtaining high quality tags turns out to be a surprisingly difficult problem, on which we show promising initial results.

2. Problem Statement

We begin by defining some terminology. Throughout the paper, we use the term *photo* interchangeably with *image* and *view*, all of which refer to an ordinary 2D image. We define a *collection* as a set of photos, and we consider two photos to be *connected* if at least one object is visible in both photos. We define a *scene* in a collection as a set of connected photos of rigid 3D geometry.

In its most basic form, a *summary* is a set of photos that represents the most interesting visual content of a scene. The purpose of a summary is to quickly give a viewer an accurate impression of what a particular scene looks like. In Section 5, we augment summaries to handle photo collections containing multiple scenes.

Our goal, then, given a set of photos $\mathcal{V}$ of a single scene

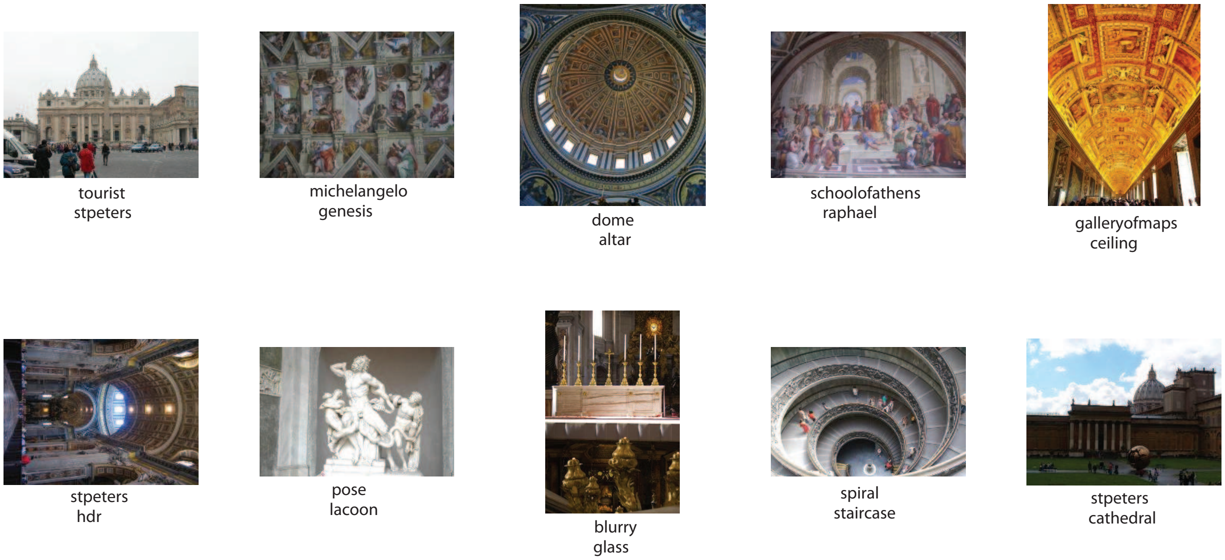


Figure 1. A 10-image summary of 2000 images of the Vatican computed by our algorithm.

S, is to compute a summary $\mathcal{C} \subseteq \mathcal{V}$ such that most of the interesting visual content in $\mathcal{V}$ is represented in $\mathcal{C}$.

3. Related Work

Our scene summarization approach draws on related work from three main areas: canonical views, summarization, and modeling linked visual and textual data.

3.1. Canonical Views

The problem of selecting views that capture the essence of a geometric object has been studied for over twenty years in the human vision and computer vision communities. Defining precisely what makes a view "canonical" is still a topic of debate. In their seminal work, Palmer et al. [21] proposed four different criteria, which we paraphrase as follows:

1. Given a set of photos, which view do you like the best?

2. When taking a photo, which view do you choose?

3. From which view is the object easiest to recognize?

4. When imagining the object, which view do you see?

In a series of experiments designed to compare these criteria, Palmer *et al*. found significant correlation between all four tasks, concluding that human observers choose the same types of views regardless of the task. Subsequent studies provide further support for many of these conclusions, although recent experiments by Blanz *et al*. [6] provide conflicting conclusions for the fourth criterion, finding that people tend to imagine objects in plan views, but prefer looking at off-axis views. From the standpoint of scene summarization, however, the first three tasks are most relevant, and these perceptual experiments suggest that recurring views in multi-user photo collections (criterion 2) can enable summaries that are both visually appealing (criterion 1) and facilitate recognition (criterion 3).

There is a significant literature on computing canonical views in the computer vision literature. Note that the criteria of Palmer *et al*. are not applicable for computer vision tasks as they are defined in terms of a human observer. Hence, canonical views work in the computer vision community has sought to identify principles and algorithms baesd on the geometry of an object or a set of photos. For example, Freeman [16] and Weinshall *et al*. [28] quantify the likelihood of a view by analyzing the range of viewing conditions that produce similar views, using knowledge of object geometry. More closely related to our work are methods that take as input a set of photographs. In particular, Denton *et al*. [12] use a semidefinite programming relaxation to select canonical views from a larger set of views, choosing views which are as similar as possible to the non-canonical views while being dissimilar to each other. Hall and Owen [17] define canonical views as images with *low* likelihood, while being orthogonal to each other. They compute a 10- to 20-dimensional eigenmodel of a set of images (represented as vectors of grayscale values), then iteratively extract the least likely view that is nearly orthogonal to all previously selected views.

In our work, we take a fundamentally different approach to computing canonical views that is more directly related to the original principles in Palmer *et al*. Instead of attempting to infer such views from the geometry or from a set of uniformly sampled views, we use photo sharing websites to sample the distribution of views from which people choose to take photographs. Hence, we are relying on a population of photographers to provide a likelihood distribution over camera viewpoints (as in criterion 2), and our task reduces to computing clusters and peaks of this distribution.

A second fundamental difference between our work and prior work on canonical views is our focus on *large scale scenes* rather than individual objects. While many objects can be represented effectively with a single canonical view, the same is not true of scenes. (Consider a church, for example, which has both an interior and an exterior, and may require several images to capture the interesting aspects.) And while most prior work on canonical views considered only a limited range of viewpoints (e.g., views on a hemisphere), images from photo sharing websites tend to have a broad sampling of positions, orientations, and focal lengths, sampling a 7D viewing space.

3.2. Summarization

With the recent proliferation of images and other shared data accessible online, techniques for summarizing large data sets for human consumption have garnered some interest. Rother *et al*. [23] summarize a set of images with a "digital tapestry". They synthesize a large output image from a set of input images, stitching together salient and spatially compatible blocks from the input image set. Wang *et al*. [27] create a "picture collage", a 2D spatial arrangement of the images in the input set chosen to maximize the visibility of salient regions. In both of these works, the set of images to appear has already been chosen, and the visual layout is to be determined. We ignore issues of layout and focus on selecting the set of images to appear in the summary. Once selected, these images could be arranged in a digital tapestry or picture collage.

Clough *et al*. [9] construct a hierarchy of images using only textual caption data, and the concept of subsumption. A tag t_i subsumes another tag t_j if the set of images tagged with t_i is a superset of the set of images tagged with t_j. Schmitz [24] uses a similar approach but relies on Flickr tags, which are typically noisier and less informative than the captions. Jaffe *et al*. [19] summarize a set of images using only tags and geotags. By detecting correlations between tags and geotags, they are able to produce "tag maps", where tags and related images are overlaid on a geographic map at a scale corresponding to the range over which the tag commonly appears. All of these approaches could be used to further organize our summaries. However, none of them take advantage of the visual information in the images to fill in for bad or missing metadata.

3.3. Modeling Linked Visual and Textual Data

Barnard and Forsyth [5] explored the relationship between visual and textual data for several problems in computer vision, including object recognition, image search, and image and region auto-annotation. Further work by Duygulu *et al*. [14] and Barnard *et al*. [4] explored various generative models for images with associated text. Also, Blei and Jordan [7] extend the generative model from latent Dirichlet allocation [8] to handle annotated data.

Our approach for combining visual and textual data is perhaps simpler than the previous approaches, as this is not the main focus of our paper. In addition, we are not attempting to learn a model to apply to unseen images. We only use textual tag data to enhance the scene summaries, which involves selecting tags that are likely to apply to large clusters of the images we already have.

4. Scene Summarization Algorithm

Given a set of views $\mathcal{V}$ of scene S (see Figure 2), we wish to compute a summary $\mathcal{C} \subseteq \mathcal{V}$ that represents the most interesting visual content in $\mathcal{V}$. Before discussing the algorithm, we describe our representation of views and scenes:

Scene S is represented as a set of visual features $f_1, f_2, \ldots, f_{|S|}$. Each visual feature corresponds to exactly one point in the 3D environment. (However, it is possible that due to large differences in lighting or viewing direction, the same 3D point corresponds to multiple features.) A view $V \in \mathcal{V}$ is represented as the subset of S corresponding to the features which are visible in the view. Therefore, the set of photos $\mathcal{V}$ can be represented by an $|S|$-by-$|\mathcal{V}|$ Boolean matrix. This type of term-document matrix is often used as input for systems dealing with text documents [11], and more recently images [22]. Note that in many previous cases, each entry (i, j) in the term document matrix is a tally (how many times term/feature i appears in document/image j). In our case, since a feature corresponds to an actual 3D point, it can only be present or absent.

4.1. Computing the Feature-Image Matrix

We first transform the set of views into a feature-image incidence matrix. To do so, we use the SIFT keypoint detector [20] to find feature points in all of the images in $\mathcal{V}$. The feature points are represented using the SIFT descriptor. Then, for each pair of images, we perform feature matching on the descriptors to extract a set of candidate matches. We further prune the set of candidates by estimating a fundamental matrix using RANSAC and removing all inconsistent matches, as in [26]. After the previous step is complete for all images, we organize the matches into tracks, where a track is a connected component of features. We remove tracks containing fewer than two features total, or at least two features in the same image. At this point, we consider each track as corresponding to a single 3D point in S. From the set of tracks, it is easy to construct the $|S|$-by-$|\mathcal{V}|$ feature-image incidence matrix.

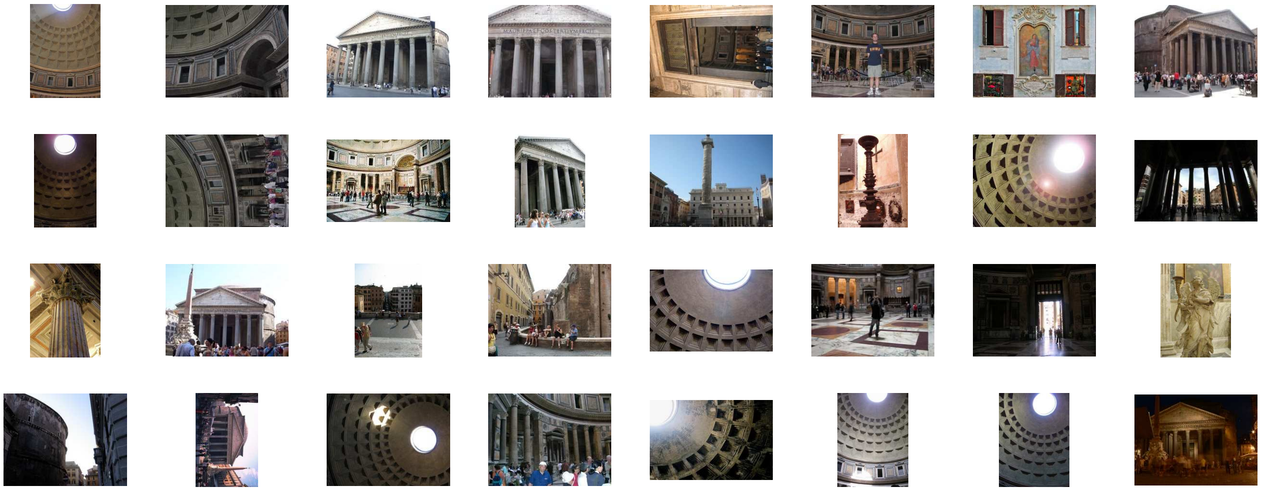


Figure 2. A random set of 32 images of the Pantheon. Our algorithm takes an unsorted image set like this one, but containing thousands of images, and selects a set of canonical views to serve as a summary.

4.2. Selecting the Summary Views

There are a number of possible criteria for choosing views to include in the summary, some of which are:

likelihood - An image should be included if it is similar to many other images in the input set.

coverage - An image should be included if it covers a large number of visual features in the scene.

orthogonality - Two images should not both be included if they are similar to each other.

We focus mainly on likelihood, as we are interested in harnessing the consensus of users of photo sharing sites for selecting canonical views.

4.2.1 Image Likelihood

The most popular criteria in previous work on canonical views are likelihood ([16], [28]) and orthogonality ([12], [17]). However, in previous work, the likelihood of an image referred to the range of viewing parameters that produces similar views. We, on the other hand, have a set of images distributed according to the viewpoint preferences of human photographers. Our likelihoods are measured on this distribution and not inferred solely from geometry (or using a uniform distribution over viewing directions). We define the *similarity* between two views as:

$$\mathrm{sim}(V_i, V_j) = \frac{|V_i \cap V_j|}{\sqrt{|V_i||V_j|}} \tag{1}$$

Equation (1) measures the cosine of the angle between the normalized feature incidence vectors for the two images. If both views have the same number of features, this is simply the fraction of features that are shared. If the two views do not share any features, the similarity is zero. In a slight abuse of notation, we will use V to refer to the set of features in a view as well as the normalized Boolean feature incidence vector. So:

$$\mathrm{sim}(V_i, V_j) \quad = \quad V_i \cdot V_j$$

A simple definition of *likelihood* is then:

$$\mathrm{lik}(V) = \sum_{V_i \in \mathcal{V}} (V_i \cdot V) \tag{2}$$

This definiton of likelihood is closely related to the log likelihood of the set of images $\mathcal{V}$ being drawn from a von Mises-Fisher distribution (a spherical analogue of a Gaussian) with the normalized feature incidence vector for V as mean parameter μ:

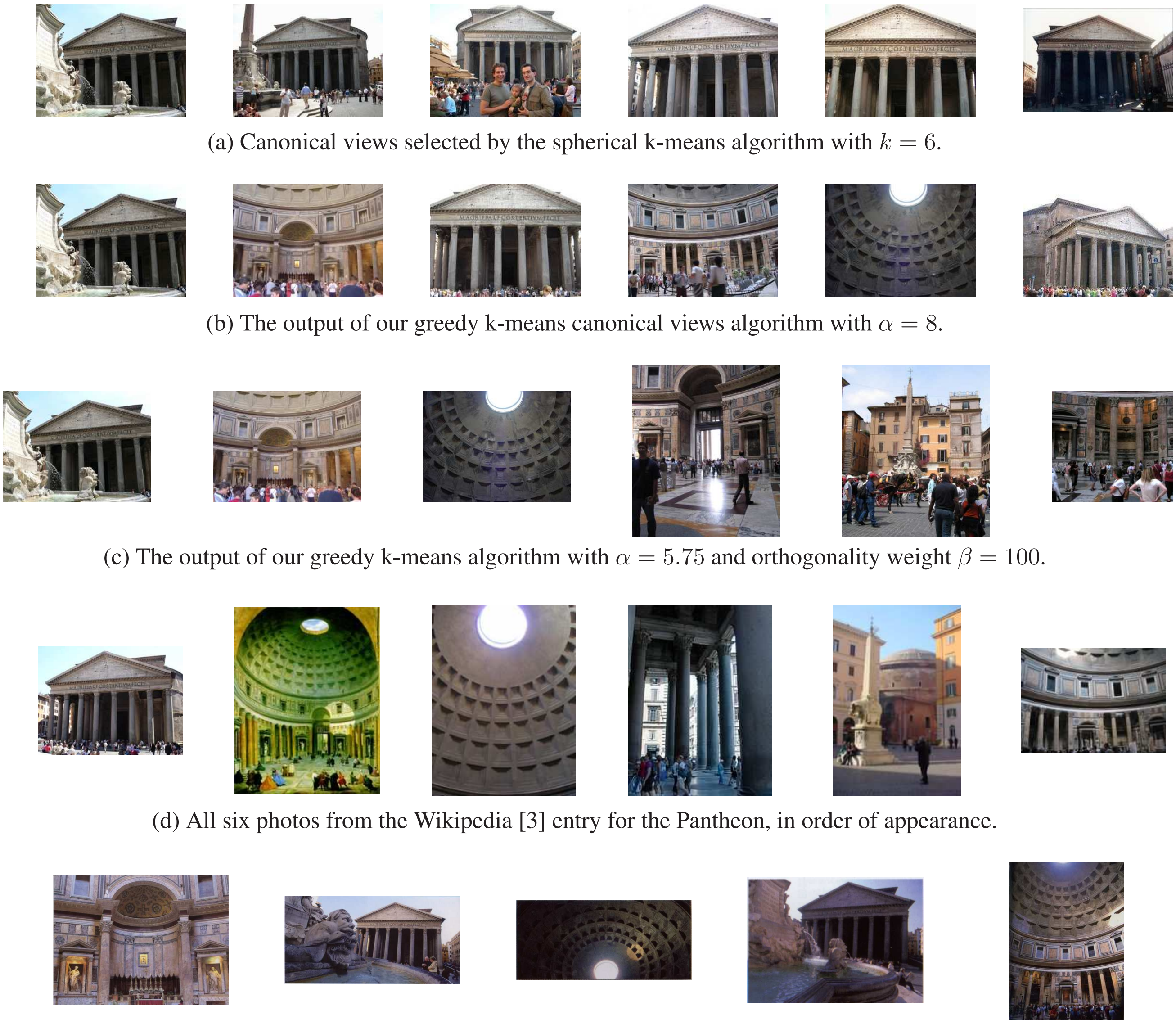
$$p(X|\mu, h) \quad = \quad \prod_{x \in X} f(h) e^{h(x \cdot \mu)} \tag{3}$$

$$\log p(X|\mu, h) \quad = \quad h \sum_{x \in X} (x \cdot \mu) + \log f(h) \tag{4}$$

where h is the nonegative concentration parameter and $f(h)$ is the normalizing constant chosen so that $p(x|\mu, h)$ integrates to one. Note that h only specifies a linear transformation on the sum of similarities, and can often be ignored.

4.2.2 Clustering Objective for Canonical Views

Because our goal is to represent the target image set $\mathcal{V}$, we include a quality term for each view $V_i \in \mathcal{V}$ expressing the similarity between V_i and its closest canonical view $C_{c(i)}$ in $\mathcal{C}$, where c contains the mapping of views to canonical



(a) Canonical views selected by the spherical k-means algorithm with $k = 6$.

(b) The output of our greedy k-means canonical views algorithm with $\alpha = 8$.

(c) The output of our greedy k-means algorithm with $\alpha = 5.75$ and orthogonality weight $\beta = 100$.

(d) All six photos from the Wikipedia [3] entry for the Pantheon, in order of appearance.

(e) Left to right: one Pantheon photo from the Berlitz [25] and Lonely Planet [18] guidebooks, and three from Fodor's [15]. These are the only images of the Panthon in the three guidebooks.

Figure 3. Comparison of several summaries of the Pantheon. Summary (a) illustrates the failure of the spherical k-means algorithm to find meaningful clusters. Summaries (b) and (c) are typical of our results, and demonstrate the effect of the explicit orthogonality constraint. Hand-created summaries (d) and (e) are included for comparison. Note that our summary views are quite similar to those in Wikipedia and the guidebooks. When we produce larger summaries, we often select interesting views which are left out of Wikipedia and typical guidebooks (see our project web page [1]).

views. Also, we want to penalize solutions with too many canonical views, as our summaries are meant to be readable quickly, so we include a cost term α for each canonical view. Our algorithm attempts to maximize the following quality function:

$$Q(\mathcal{C}) \quad = \quad \sum_{V_i \in \mathcal{V}} \left(V_i \cdot C_{c(i)}\right) - \alpha|\mathcal{C}|$$

The summation term is closely related to the log likelihood of the set of views $\mathcal{V}$ being drawn from a mixture of von Mises-Fisher distributions with equal mixture weights and common concentration parameter h. The $-\alpha|\mathcal{C}|$ term can be thought of as enforcing a geometric prior on the number of canonical views.

This objective function implicitly encourages the canonical views to be orthogonal, as each view V need only be explained by one canonical view. In cases where orthogonality is more important, we add an extra term to the objective function:

$$Q(\mathcal{C}) = \sum_{V_i \in \mathcal{V}} \left(V_i \cdot C_{c(i)}\right) - \alpha|\mathcal{C}| - \beta \sum_{C_i \in \mathcal{C}} \sum_{C_{j>i} \in \mathcal{C}} \left(C_i \cdot C_j\right)$$

This explicitly penalizes pairs of canonical views for being too similar (see Figure 3(c)).

Without the $-\alpha|\mathcal{C}|$ term, the function could be optimized by a simple modification of the spherical k-means algorithm ([13]) in which the means are restricted to views in the data set. However, even in the simplified case where $|\mathcal{C}|$ is known, the spherical k-means algorithm performs poorly when the dimension is large and is extremely sensitive to the initial configuration (see Figure 3(a)). We avoid this problem by using the following greedy algorithm, beginning with $\mathcal{C} = \emptyset$:

1. For each view $V \in \mathcal{V} \setminus \mathcal{C}$, compute $Q_V = Q(\mathcal{C} \cup \{V\}) - Q(\mathcal{C})$.
2. Find the view V^* for which Q_{V^*} is maximal.
3. If $Q_{V^*} > 0$, add V^* to $\mathcal{C}$ and repeat from step 1. Otherwise, stop.

At each iteration, we choose the view that will cause the largest increase in the quality function and add it to the set of canonical views, as long as this increase is at least α. If not, we stop. Cornuejols *et al.* [10] proved that this greedy algorithm always yields a solution that has quality at least $\frac{e-1}{e}$ times the optimal solution, where e is the base of the natural logarithm. We find that the greedy algorithm (Figure 3(b,c)) also performs much better in practice than the standard spherical k-means algorithm (Figure 3(a)), which has an arbitrarily bad approximation ratio. This algorithm also frees us from having to choose the number of canonical views in advance, though we do need to specify α. When using explicit orthogonality penalties, the proof of approximation bound no longer applies, though we find the algorithm still works well in practice. We have also experimented with running the standard spherical k-means algorithm initialized with the means chosen by the greedy algorithm. For most of our data sets, this changes very few of the means, and changes them to nearly identical views, and we therefore omit this step for the results included in the paper.

In the next section, we will construct an image browsing application on top of this basic scene summarization method, extending it to photo collections containing many scenes, and incorporating user-specified tag data into the summaries.

5. Image Browsing Application

The photo sharing website Flickr [2] contains over 500,000 photos of the city of Rome, spanning such sites as the Colosseum (over 11,000 photos), St. Peter's Basilica (over 8000 photos), and the Trevi Fountain (over 7000 photos). The photos are organized by user-specified tags and, in some cases, timestamps and geotags. A system for summarizing and browsing photos using only this data is handicapped in several ways, as illustrated in Figure 5:

- Some photos are missing relevant tags. A photo of the Colosseum may be tagged with "rome" but not "colosseum".
- Some photos have tags that are misleading. A user might tag a set of photos with "vatican" even if some of the photos were not taken in the Vatican.
- Some photos have tags that are uninformative. A user might tag a set of photos with "vacation2005", which may be useful for the user's own photo organization, but useless for creating a summary of Rome that is of value to multiple users.
- Tags are essentially useless for summarizing or browsing a single scene. In the case of the Trevi Fountain, a flat index of over 7000 photos is too large to browse, and the variation among the photos is not reflected in the tags.

We present a prototypical browsing application using our scene summaries that resolves all of these issues. The application can function in the complete absence of tags, but when tags are provided, we can extract tags which are likely to be correct and use them to enhance the browser.

5.1. Organizing the Photos for Browsing

For a single scene, our set of canonical views $\mathcal{C}$, along with the mapping from each image in $\mathcal{V}$ to its most similar canonical view, can serve as a simple two-level hierarchy for image browsing. The top level of the hierarchy contains the canonical views, and beneath each canonical view C is the set of images $V \in \mathcal{V}$ such that C is the most similar canonical view to V.

For larger image collections that span multiple scenes, we add another level to the top of the hierarchy. We construct the three-level hierarchy in two steps:

1. Find the connected components of the image collection; each connected component comprises a *scene*, as defined in Section 2.
2. For each scene, use the greedy algorithm from Section 4.2.2 to compute the canonical views.

We provide a browseable summary of Rome on our project web page [1]. A smaller version of the top level of this summary is shown in Figure 4. Note that image connectivity does not correspond with the semantic concept of a scene, and connected components is prone to oversegmentation and undersegmentation. In the next section, we introduce a technique for avoiding this problem by exploiting tag data.

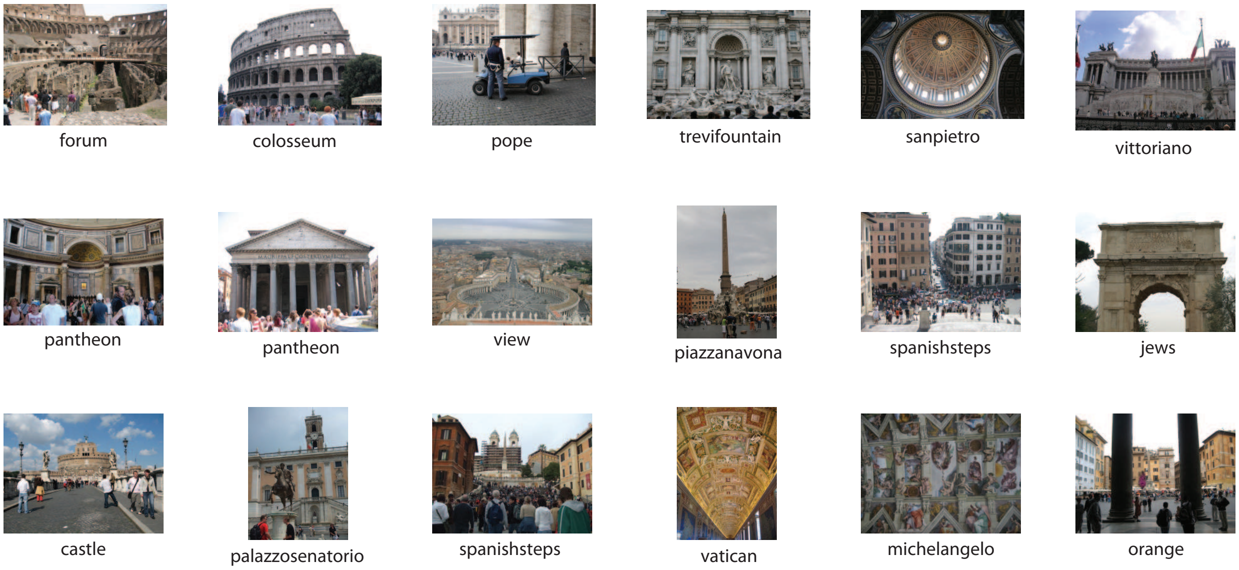


Figure 4. A segmentation of a 20,000 image Rome data set into the 18 largest scenes, with the best tag associated with each scene. The tags are computed according to Equation 5.

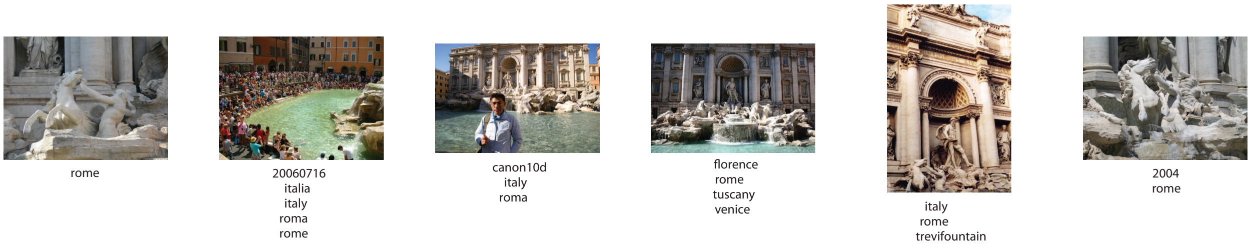


Figure 5. Six randomly selected images of the Trevi Fountain, and tags given to each image by Flickr [2] users.

5.2. Incorporating Tag Data

At either the scene or cluster level, we enhance our summaries by displaying one or more tags for each canonical view. As the user-specified tags associated with each image may be unreliable (see Figure 5), we look at all images in the scene or cluster to choose the tags to display. Two main difficulties arise in selecting appropriate tags:

1. The most popular tags in the cluster may be associated with a broader concept than the cluster itself. For example, the most popular tags for a cluster containing images of the Pantheon may be "italy" and "rome".

2. The occurrence of a tag may be highly correlated with the cluster because of the behavior of a few users. Tags like "anniversary2005" or "jason" could be strongly associated with a cluster, but do not help describe the scene.

We define a function $\mathrm{score}(c, t)$ that measures how well tag t describes cluster c. A first approach might be to choose tags with large values of $P(t|c)$. However, this falls into the first trap above, and assigns tags that are correct, but not very discriminative, like "rome" or "italy", to most clusters. One might also consider choosing tags with large values of $P(c|t)$. This avoids the first problem, but ends up choosing useless tags that happen to be discriminative, like "anniversary2005", for most clusters. To resolve both issues, we compute the score as the conditional probability of the cluster given the tag, independent of the user u (the Flickr [2] member who took the photograph):

$$\mathrm{score}(c, t) = \sum_{u \in U} P(c|t, u) P(u) \qquad (5)$$

For all probabilities, we treat each image as a sample and count the number of co-occurrences. We therefore define:

$$\begin{aligned} P(c|t, u) &= \frac{\left|\{V \in \mathcal{V} \mid c(V) = c, t \in T(V), u(V) = u\}\right|}{\left|\{V \in \mathcal{V} \mid t \in T(V), u(V) = u\}\right|} \\ P(u) &= \frac{\left|\{V \in \mathcal{V} \mid u(V) = u\}\right|}{|\mathcal{V}|} \end{aligned}$$

where $c(V)$ and $u(V)$ are the cluster and user associated with view V, respectively, and $T(V)$ is the set of tags associated with V. Note that strictly speaking, P represents frequencies instead of probabilities, as we are only measuring the former. A small issue arises, in that $P(c|t, u)$ will be undefined when user u never uses tag t. In this case (which happens frequently), we replace $P(c|t, u)$ by $P(c|u)$. We use this score function at both the scene and cluster level. However, for most clusters, accurate tags do not exist or are rare enough to be indistinguishable from user-specific tags.

Using tags for browsing can avoid problems associated with the connected components segmentation. For example, in Figure 4, the Pantheon is split among multiple segments, since connecting images are missing. However, in our browseable index [1], we also allow a user to view the set of clusters associated with a given tag. Under the "pantheon" tag, clusters from both segments appear in the index. Note that this is not the same as ordinary browsing by tags, for example on Flickr [2], as in our index many of the images browseable under the tag "pantheon" were not given the tag by any Flickr user.

6. Conclusions

We defined the problem of scene summarization, and provided an algorithm that solves this problem on large image sets. When textual tags are associated with each image, we can use them to enhance our summaries, in spite of the large amount of noise in the tags. We also demonstrate an image browsing application that uses our summarization approach, and show how scene summaries can serve as portals into an interactive 3D browser (see our project web page [1]). Our summaries and image browser allow a user to quickly navigate a huge collection of images in a way that was previously impossible, and has the capability to greatly enhance the experience of browsing photo collections on Flickr [2] and other photo sharing sites.

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-0413198, the University of Washington Animation Research Labs, Adobe, Microsoft, and an endowment by Rob Short and Emer Dooley.

References

[1] http://grail.cs.washington.edu/projects/canonview/.

[2] http://www.flickr.com.

[3] http://www.wikipedia.org.

[4] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan. Matching words and pictures. *J. of Machine Learning Research*, 3(6):1107–1135, 2003.

[5] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. *Proc. ICCV*, pages 408–415, 2001.

[6] V. Blanz, M. Tarr, and H. Bülthoff. What object attributes determine canonical views? *Perception*, 28(5):575–600, 1999.

[7] D. Blei and M. Jordan. Modeling annotated data. *Proc. SIGIR Conf. on Information Retrieval*, pages 127–134, 2003.

[8] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.

[9] P. Clough, H. Joho, and M. Sanderson. Automatically Organising Images using Concept Hierarchies. *Proc. SIGIR Workshop on Multimedia Information Retrieval*, 2005.

[10] G. Cornuejols, M. Fisher, and G. Nemhauser. Location of Bank Accounts to Optimize Float: An Analytic Study of Exact and Approximate Algorithms. *Management Science*, 23(8):789–810, 1977.

[11] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. for Information Science*, 41(6):391–407, 1990.

[12] T. Denton, M. Demirci, J. Abrahamson, A. Shokoufandeh, and S. Dickinson. Selecting canonical views for view-based 3-D object recognition. *Proc. ICPR*, pages 273–276, 2004.

[13] I. Dhillon and D. Modha. Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, 42(1):143–175, 2001.

[14] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Proc. ECCV*, pages 97–112, 2002.

[15] Fodor's. *See It Rome*. Fodor's, 2006.

[16] W. Freeman. The generic viewpoint assumption in a framework for visual perception. *Nature*, 368(6471):542–545, 1994.

[17] P. Hall and M. Owen. Simple canonical views. *Proc. BMVC*, pages 839–848, 2005.

[18] A. Hole. *Best of Rome*. Lonely Planet. Lonely Planet Publications, 2006.

[19] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries for large collections of geo-referenced photographs. *Proc. Int. Conf. on World Wide Web*, pages 853–854, 2006.

[20] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. of Computer Vision*, 60(2):91–110, 2004.

[21] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. *Attention and Performance IX*, pages 135–151, 1981.

[22] P. Praks, J. Dvorsky, and V. Snasel. Latent semantic indexing for image retrieval systems. *Proc. SIAM Conf. on Applied Linear Algebra*, 2003.

[23] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry. *Proc. CVPR*, pages 589–596, 2005.

[24] P. Schmitz. Inducing ontology from Flickr tags. *Collaborative Web Tagging Workshop at WWW2006*, 2006.

[25] P. Schultz. *Rome*. Berlitz Pocket Guide. Berlitz Publishing, 2003.

[26] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *SIGGRAPH Conf. Proc.*, pages 835–846, 2006.

[27] J. Wang, J. Sun, L. Quan, X. Tang, and H. Shum. Picture Collage. *Proc. CVPR*, pages 347–354, 2006.

[28] D. Weinshall, M. Werman, and Y. Gdalyahu. Canonical Views, or the Stability and Likelihood of Images of 3D Objects. *Image Understanding Workshop*, 1994.