Generating Diverse and Representative Image Search Results for Landmarks

Lyndon Kennedy *
Dept. of Electrical Engineering
Columbia University, New York, NY
lyndon@ee.columbia.edu

Mor Naaman
Yahoo! Inc.
Berkeley, CA
mor@yahoo-inc.com

ABSTRACT

Can we leverage the community-contributed collections of rich media on the web to automatically generate representative and diverse views of the world's landmarks? We use a combination of context- and content-based tools to generate representative sets of images for location-driven features and landmarks, a common search task. To do that, we using location and other metadata, as well as tags associated with images, and the images' visual features. We present an approach to extracting tags that represent landmarks. We show how to use unsupervised methods to extract representative views and images for each landmark. This approach can potentially scale to provide better search and representation for landmarks, worldwide. We evaluate the system in the context of image search using a real-life dataset of 110,000 images from the San Francisco area.

Categories and Subject Descriptors: H.4 [Information Systems Applications]:Miscellaneous

General Terms: Algorithms, Human Factors

Keywords: geo-referenced photographs, photo collections, social media

1. INTRODUCTION

Community-contributed knowledge and resources are becoming commonplace, and represent a significant portion of the available and viewed content on the web. In particular, popular services like Flickr [8] for images and YouTube [28] for video have revolutionized the availability of web-based media resources. In a world where, to paraphrase Susan Sontag [22], "everything exists to end up in an (online) photograph", many challenges still exist in searching, visualizing and exploring these media.

Our focus in this work is on landmarks and geographic elements in these community datasets. Such landmarks enjoy a significant contribution volume (e.g., over 50,000 images on Flickr are tagged with the text string `Golden Gate Bridge`), and are important for search and exploration tasks [2]. However, these rich community-contributed datasets pose a significant challenge to information retrieval and representation. In particular, the annotation and metadata provided by users is often inaccurate [10] and noisy; photos are of varying quality; and the sheer volume alone makes content hard to browse and represent in a manner that improves rather than degrades as more photos are added. In addition, hoping to capture the "long tail" of the world's landmarks, we can not possibly train classifiers for every one of these landmarks. We attempt to overcome these challenges, using community-contributed media to improve the quality of representation for landmark and location-based searches. In particular, we outline a method that aims to provide precise, diverse and representative results for landmark searches. Our approach may lead not only to improved image search results, but also to better systems for managing digital images beyond the early years [21].

Our approach in this paper utilizes the set of geo-referenced ("geotagged") images on Flickr: images whose exact location was automatically captured by the camera or a location-aware device (e.g., [1]) or, alternatively, specified by the user (the Flickr website supports this functionality, as do other tools – see [23] for a survey of methods for geo-referencing images). There are currently over 40,000,000 public geotagged images on Flickr, the largest collection of its kind. With the advent of location-aware cameraphones and GPS-integrated cameras, we expect the number of geotagged images (and other content) on Flickr and other sites to grow rapidly.

To tackle the landmark problem, we combine images analysis, tag data and image metadata to extract meaningful patterns from these loosely-labeled, community-contributed datasets. We conduct this process in two stages. First, we use tags (short text labels associated with images by users) and location metadata to detect tags and locations that represent landmarks or geographic features. Then, we apply visual analysis of the images associated with discovered landmarks to extract representative sets of images for each landmark. This two-stage process is advantageous, since visual processing is computationally expensive and often imprecise and noisy. Using tags and metadata to reduce the number of images to be visually processed into a smaller, more coherent subset can make the visual processing problem less expensive and more likely to yield precise results.

Given the reduced set of images, our approach for generating a diverse and representative set of images for a landmark is based on identifying "canonical views" [20, 18]. Using various image processing methods, we cluster the landmark images into visually similar groups, as well as generate links between those images that contain the same visual objects. Based on the clustering and on the generated link structure, we identify canonical views, as well as select the top

*This work was done while the first author was at Yahoo!.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

representative images for each such view.

Our contributions therefore include:

- An algorithm that generates representative sets of images for landmarks from community-contributed datasets;
- A proposed evaluation method for landmark-driven and other image search queries;
- A detailed evaluation of the results in the context of image search.

We define the problem and the data model more specifically in Section 3. In Section 4 we shortly describe possible methods for identifying tags and locations that correspond to landmarks or geographic features. Section 5 describes the analysis of the subset of photos that corresponds to each landmark to generate a ranking that would support representative and diverse search results. We evaluate our algorithm on ten San Francisco landmarks in Section 6. Before we do all that, we report on important related work.

2. RELATED WORK

The main research efforts related to our work here are computer-vision approaches to landmark recognition, as well as metadata and multimedia fusion, and metadata-based models of multimedia. We also report on some of the latest research that addresses web image search.

Most closely related to our work here is the research from Simon et al. [20] on finding a set of canonical views to summarize a visual "scene". The authors' approach, similarly to ours, is based on unsupervised learning. Given a set of images for a given scene (e.g., "Rome" or "San Francisco Bay Bridge"), canonical views are generated by clustering images based on their visual properties (most prominently, SIFT features [12], which we are using here). Once clusters are computed, Simon et al. propose an "image browser" where scenes can be explored hierarchically. The researchers extract representative tags for each cluster given the photographs' tags on Flickr. Our approach is somewhat different, as we start from the tags that represent landmarks, and generate views for these landmarks (and not just "a scene"). Starting with tag data does not entail a great difference in how the two systems work; however, in practice, using the tag data and other metadata before applying image analysis techniques may prove more scalable and robust. For instance, Simon et al. do not specify how such initial "scene" sets will be generated; we propose to automatically identify the tags to be analyzed, and provide the details on how to construct the set of photos for each such tag. In addition, we show how to select representative photographs once the "canonical views" were identified. Finally, we evaluate our system in the context of a traditional web task (image search) and suggest a user-driven evaluation that is meant to capture these difficult themes.

In [3], the authors rank "iconic" images from a set of images with the same tag on Flickr. Our work similarly examines ranking the most representative (or iconic, or canonical as [20] suggests) images from a set of noisily labeled images which are likely of the same location. A key difference is that in [3], the locations are manually selected, and it is assumed that there is one iconic view of the scene, rather than a diverse set of representative views as we show in this work.

Beyond visual summaries and canonical views, the topic of "landmark recognition" has been studied extensively, but mostly applied to limited or synthetic datasets. Various efforts ([7, 16, 24, 27] and more) performed analysis of context metadata together with content in photo collections. The work of Tsai et al. [24], for example, attempted to match landmark photos based on visual features, after filtering a set of images based on their location context. This effort serves as an important precursor for our work here. However, the landmarks in the dataset for Tsai et al. were pre-defined by the researchers, assuming the existence of a landmark gazetteer. This assumption is certainly limiting, and perhaps unrealistic when gearing towards performance in a web-based, long-tailed environment. O'hare et al. [16] used a query-by-example system where the sample query included the photo's context (location) in addition to the content, and filtered the results accordingly, instead of automatically identifying the landmarks and their views as we do here. Davis et al. [7] had a similar method that exposed the similarity between places based on content and context data, but did not detect or identify landmarks. Naaman et al. [14] extract location-based patterns of terms that appear in labels of geotagged photographs of the Stanford campus. The authors suggest to build location models for each term, but the system did not automatically detect landmarks, nor did it include computer vision techniques.

In [10], the authors investigated the use of "search-based models" for detecting landmarks in photographs. In that application, the focus was the use of text-based keyword searches over web image collections to gather training data to learn models to be applied to consumer collections. That work, albeit related to our work here, relies upon pre-defined lists of landmarks; we investigate the use of metadata to automatically discover landmarks. Furthermore, the focus of that work is on predicting problems that would emerge from cross-domain learning, where models trained on images from web search results are applied to consumer photos.

Jing et al. proposed an algorithm to extract representative sights for a city [9] and propose a search and exploration interface. The system uses a text-based approach, ranking phrases that appear in photos associated with a city and selecting the top-ranked phrases as "representative sights". Both the exploration and analysis techniques described in this work could be used in concert with the system described in this paper.

Naturally, the topic of web image search has been explored from both algorithmic and HCI perspectives. Clustering of the results was suggested in a number of papers [4, 26]. Most recently, Wang et al. [26] used a clustering-based approach for image search results; searching for "San Francisco" images in their system returns clusters of related concepts. Such exploration avenues are now built into most popular search engines, often showing derived concepts for narrowing or expanding the search results.

Finally, we also had initially reported on work towards a landmark search system in [11]. The current work exceeds and extends [11], which gave a general overview of the system and did not supply the details of the visual analysis, or the deeper evaluation we perform here.

3. MODEL AND PROBLEM DEFINITION

We first describe the data model we use in this work. We then point out several of the salient features and issues that arise from the data and the model. Finally, we define the research problem that is the focus of this paper.

Formally, our dataset consists of three major elements: photos, tags and users. We define the set of photos as $\mathbb{P} \triangleq \{p\}$, where p is a tuple $(\theta_p, \ell_p, t_p, u_p)$ containing a unique photo ID, θ_p; the photo's capture location, represented by latitude and longitude, ℓ_p; the photo's capture time, t_p; and the ID of the user that contributed the photo, u_p. The location ℓ_p generally refers to the location where the photo p was taken, but sometimes marks the location of the photographed object. The time t_p generally marks the photo capture time, but occasionally refers to the time the photo was uploaded to Flickr.

The second element in our dataset is the set of tags associated with each photo. We use the variable x to denote a tag. Each photo p can have multiple tags associated with it; we use $\mathbb{X}_p$ to denote this set of tags. For convenience, we define the subset of photos associated with a specific tag as: $\mathbb{P}_x \triangleq \{p \in \mathbb{P} \mid x \in \mathbb{X}_p\}$. We use similar notation to denote any subset $\mathbb{P}_S \subseteq \mathbb{P}$ of the photo set.

The third element in the dataset is users, the set of which we denote by the letter $\mathbb{U} \triangleq \{u_p\}$. Equivalently, we use $\mathbb{U}_S \triangleq \{u_p \mid p \in \mathbb{P}_S\}$ and $\mathbb{U}_x \triangleq \{u_p \mid p \in \mathbb{P}_x\}$ to denote users that exist in the set of photos $\mathbb{P}_S$ and users that have used the tag x, respectively.

Note that there is no guarantee for the correctness of any image's metadata. In particular, the tags x are *not* ground-truth labels: false positive (photos tagged with landmark tag x but do not actually contain the landmark) and false negatives (photos of the landmark that are not tagged with the landmark name) are commonplace. Prior work had observed that landmark tags are about 50% precise [10]. Another issue with tags, as [20] points out, is that the sheer volume of content associated with each tag x makes it hard to browse and visualize all the relevant content; other metadata that can suggest relevance, such as link structure, is not available.

Our research problem over this dataset can therefore be described in simple terms: given a 'landmark tag' x, return a ranking $\mathbb{R}_x \subseteq \mathbb{P}_x$ of the photos such that a subset of the images in the top of this ranking is a precise, representative, and diverse representation of the tag x. Or, to paraphrase [20]: given a set of photos $\mathbb{P}_x$ of a single landmark represented by the tag x, compute a summary $\mathbb{R}_x \subseteq \mathbb{P}_x$ such that most of the interesting visual content in $\mathbb{P}_x$ is represented in $\mathbb{R}_x$ for any number of photos in $\mathbb{R}_x$.[1]

4. DETECTING TAGS AS GEOGRAPHIC FEATURES

This section briefly describes potential approaches for extracting tags that represent geographic features or landmarks (referred to in this paper as "landmark tags") from the dataset. What are geographic features or landmarks tags? Put differently, these are tags that represent highly local elements (i.e., have smaller scope than a city) and are not time-dependent. Examples may be `Taj Mahal`, `Logan Airport` and `Notre Dame`; counter examples would be `Chicago` (geographically specific but not highly localized), `New York Marathon` (representing an event that occurs in a specific time) and `party` (does not represent any specific event or location). While this is quite a loose definition of a landmark tag, in practice we show that our approach can reasonably detect tags that are expected to answer these criteria.

The approach for extracting landmark tags is based on two parts. In the first part, we identify representative tags for different locations inside a geographic area of interest G. In the second part, we can perform a check to see if these tags are indeed location-specific within area G, and that they do not represent time-based features.

The first part of the process is described in detail in [2], and consists of a geographic clustering step followed by a scoring step for each tag in each cluster. The scoring algorithm is inspired by TF-IDF, identifying tags that are frequent in some clusters and infrequent elsewhere. The output of this step is a set of high-scoring tags x and the set of location clusters $\mathbb{C}_x$ in which the tag x has scored higher than some threshold. Thus, given a geographic region as input, these techniques can detect geographic feature tags as well as the specific locations where these tags are relevant. For example, in the San Francisco region, this system identifies the tags `Golden Gate Bridge`, `Alcatraz`, `Japan Town`, `City Hall` and so forth.

The second part of our proposed landmark identification is identifying individual tags as location-driven, event-driven or neither. We can then use the already-filtered list of tags and their score (from the first part of the computation), and verify that these tags are indeed location-driven, and that the tags do not represent events. The approach for identifying these tag semantics is based on the tag's metadata patterns; the system examines the location coordinates of all photos associated with x, and the timestamps of these photos. The methods are described in more detail in [19]. For example, examining the location and time distribution for the tag `Hardly Strictly Bluegrass` (an annual festival in San Francisco), the system may decide that the tag is location-specific, but that the tag also represents an event.

To summarize, our combined methods allow us to map from a given geographic area G to a set of landmark tags; for each landmark tag x, we extract a set of location clusters $\mathbb{C}_x$ in which x is relevant. These tags x indeed often represent landmarks and other geographic-driven features like neighborhood names. This set of tags and their location clusters is the input for our image analysis effort of creating representative views, as discussed next.

5. SYSTEM DESCRIPTION: GENERATING REPRESENTATIVE VIEWS

Once we have discovered a set of landmark-associated tags and locations, we turn to the task of mining the visual content of the images associated with these landmark tags x to extract sets of representative photos $\mathbb{R}_x$ for each. Our approach is based on the fact that despite the problematic nature of tags, the aggregate photographing behavior of users on photo sharing sites can provide significant insight into the canonical views of locations and landmarks. Intuitively, tourists visit many specific destinations and the photographs that they take there are largely dictated by the few photo-worthy viewpoints that are available. If these repeated views

[1]Theoretically speaking, the set $\mathbb{R}_x$ could include photos that were not annotated with the tag x (i.e., $\mathbb{R}_x \not\subseteq \mathbb{P}_x$). In other words, there could be photos in the dataset that are representative of a certain landmark/feature defined by x but were not necessarily tagged with that tag by the user (thus improving recall). We do not handle this case in our current work.

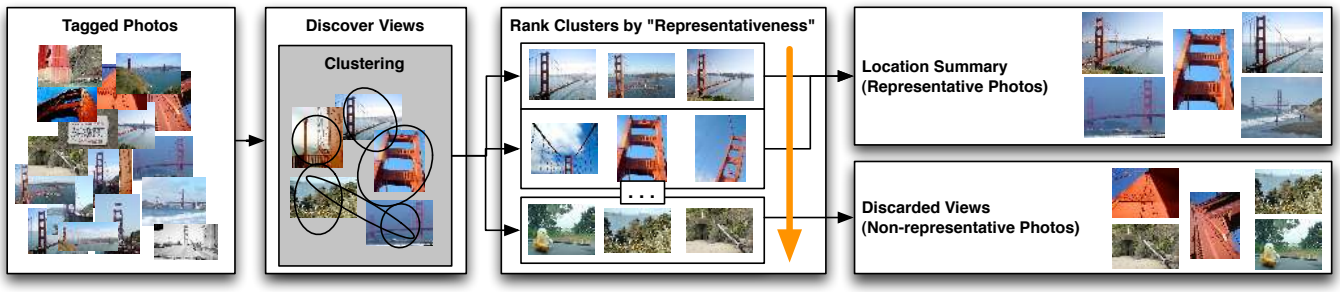


Figure 1: System architecture for generating representative summaries of landmark image sets.

of the location can be learned automatically from the data that users provide, then we can easily build visual models for landmarks and apply them to generate reliable visual summaries of locations.

We treat the task of finding representative images from a noisy tag-based collection of images as a problem of selecting a set of actual positive (representative) images from a set of pseudo-positive (same-tag or same-location) images, where the likelihood of positives within the set is considered to be much higher than is generally true across the collection. In particular, we expect that the various positive views of a landmark will emerge as highly clustered regions within the set of photos, while the actual negative (irrelevant) photos will be somewhat evenly distributed across the space as noise. We focus on unsupervised methods, where visual models of representative images can be learned directly from the noisy labels provided by users, without the need for explicitly defining a location or manually relabeling the images as representative or not (such manual effort cannot be expected in long-tailed community-contributed datasets). The resulting models could also be applied to enhance indexing by suggesting additional tags for images or to refine queries for search.

The general approach for our visual location summarization framework is illustrated in Figure 1. First, given a set of images (and their extracted visual features) associated with a landmark, we perform visual clustering across the set of images to find various common views of that landmark. Then, we apply a set of heuristics over these visual clusters to order them according to their representativeness of the landmark. Also, *within* each visual cluster, we rank the individual images according to their representativeness. In the end, we extract a set of summary images by selecting the highest-ranked images from the highest-ranked clusters and discarding low-ranked clusters and low-ranked images.

5.1 Extracting Visual Features

Before we report on the process, we briefly introduce the features that we extract to model the visual content of images. In this work, we use a mix of global color and texture descriptors and local geometric descriptors to provide a robust multi-level representation of the image content. Such mixed global and local representations have been shown to provide a great deal of complementary information in a variety of recognition tasks [6]. In particular, global color and texture features can capture the recurrent spatial layouts of typical photographs. For example, in photographs of Coit Tower, we would expect a shot of a white structure centered against a blue sky. However, many other locations have similar patterns, such as the TransAmerica Building, for example. Local feature descriptors can help to identify the actual structural elements of the real-world object and ensure that the intended object is actually contained in the photograph; however, these local descriptors do little to help us identify the common photographic compositions used to portray these landmarks. Each type of descriptor can help to fill in the shortcomings of the other. By combining these two types of descriptors, we can ensure that the photos we select (1) have both the expected photographic composition and (2) actually contain the target landmark. The specific features used are as follows:

- **Global Features**. We extract two types of features to capture the global color and texture content of the image. We use grid color moment features [17] to represent the spatial color distributions in the images and Gabor textures [13] to represent the texture. We concatenate these two feature sets together to produce a single feature vector for the global color and texture content of each image in the data set.
- **Local Features**. We further represent the images via local interest point descriptors given by the scale-invariant feature transform (SIFT) [12]. Interest points and local descriptors associated with the points are determined through a difference of Gaussian process. Typical images in our data set have a few hundred interest points, while some have thousands.

We now describe the different steps in our process of generating representative views for the landmark x given these visual features.

5.2 Step 1: Clustering on Visual Features

We use visual features to discover the clusters of images within a given set of photos for landmark x. The hope is that the clustering will expose different views of the landmark: a variety of angles, different portions of the structure, and even exterior vs. interior photos. We perform clustering using k-means, a standard and straight-forward approach, using the global (color and texture) features, described above. Local (SIFT) features are not used for clustering due to their high dimensionality, but are later incorporated for ranking clusters and images.

In any clustering application, the selection of the right number of clusters is important to ensure reasonable clustering results. While some principled methods do exist for selecting the number of clusters, such as Bayesian Information Criterion (BIC), we proceed by using a simple baseline

method. Since the number of photos to be clustered for each location varies from a few dozen to a few hundred, it stands to reason that an adaptive approach to the selection of the number of clusters is appropriate, so we select the number of clusters such that the average number of photos in each resulting cluster is around 20.

The result of Step 1 is a set of visual clusters $\mathbb{V}_x$ for each landmark x.

5.3 Step 2: Ranking clusters

Given the results of the clustering algorithm, a set of clusters $V \in \mathbb{V}_x$, we rank the clusters according to how well they represent the various views associated with a landmark. This ranking allows us to sample the top-ranked images from the most representative clusters and return those views to the user when we are generating the set of representative images, $\mathbb{R}_x$. Lower-ranked clusters can be discarded and hidden from the user, since they are presumed to contain less-representative photographs.

We use several heuristics to identify representative clusters, hypothesizing that such clusters should (1) contain photos from many different users (i.e., there is a broad interest in the photos from this cluster), (2) be visually cohesive (the same objects are being photographed or the same type of photos taken) and (3) contain photos that are distributed relatively uniformly in time (there is an on-going interest in the cluster's visual subjects – the cluster does not represent photos from one specific event at the landmark's location).

We design the following four cluster scoring mechanisms to capture the above-described criteria:

- **Number of users**. We use the number of users that are represented in photos from cluster V, or $|\mathbb{U}_V|$. We chose this metric instead of the number of photos $|\mathbb{P}_V|$ to avoid having a single user bias the results.
- **Visual coherence**. We use the visual features described above to measure the intra-cluster distance (the average distance between photos within the cluster V), and the inter-cluster distance (the average distance between photos within the cluster and photos outside of the cluster). We compute the ratio of inter-cluster distance to intra-cluster distance. A high ratio indicates that the cluster is tightly formed and shows a visually coherent view, while a low ratio indicates that the cluster is noisy and may not be visually coherent, or is similar to other clusters.
- **Cluster connectivity**. We can use SIFT features to reliably establish links between different images which contain views of a single location (this process is discussed in greater detail in Section 5.4.3.) If a cluster's photos are linked to many other photos in the same cluster, then the cluster is likely to be representative, as these links may imply a similar view or object that appears in many photos. The metric is based on the average number of links per photo in the cluster.
- **Variability in dates**. We take the standard deviation of the dates on which the photos in the cluster were taken. Preference is given to clusters with higher variability in dates, since this indicates that the view is of persistent interest. Low variability indicates that the photos were taken around the same time and the cluster is related to an event, rather than a geographic feature. We can also use the techniques described in [19] to filter images from $\mathbb{P}_x$ that include tags related to events.

To combine these various cluster scores for a cluster V, we first normalize each of the four scores, such that the L1-norm of each of the scores over the clusters is equal to one. Then, we average the four scores to reach a final, combined score for V. A higher score suggests that photos in V are more representative of the landmark.

5.4 Step 3: Ranking Representative Images

Given the visual clusters, $\mathbb{V}_x$ and their associated rankings, we rank the images within each cluster according to how well they represent the cluster. Given this ranking, we generate a set of representative images, $\mathbb{R}_x$, by sampling photos using the ranked order of clusters and photos.

To rank photos in each cluster V, we apply several different types of visual processing over the set of images $\mathbb{P}_V$ to mine the recurrent patterns associated with the cluster. In particular, we propose that representative images will exhibit a mixture of qualities: (1) representative images will be highly similar to other images in the cluster, (2) representative images will be highly dissimilar to random images outside the cluster, and (3) representative images will feature commonly-photographed local structures from within the set. Notice that these criteria are somewhat parallel to ones we used to rank clusters.

We therefore extract scores for each image, based on low-level self-similarity, low-level discriminative modeling, and point-wise linking. We explain each of these factors below; we then report on how we combine all these scores to generate an image score.

5.4.1 Low-Level Self-Similarity

To measure whether images are similar to other images in the cluster, we take the centroid of all of the images in low-level global (color and texture) feature space and rank images bu to their distance from the centroid. Each feature dimension is statistically normalized to have a mean of zero and unit standard deviation and the centroid is the mean of each feature dimension. The images within each cluster are then ranked by their Euclidean distance from the centroid.

5.4.2 Low-Level Discriminative Modeling

To measure the dissimilarity between a given image within a cluster and images outside of a cluster, we apply a discriminative learning approach by taking the images within the cluster to be pseudo-positives and the images outside the set to be pseudo-negatives. Recent efforts have suggested that such light-weight discriminative models (fused with low-level self-similarity) can actually greatly improve the performance of image ranking for a number of applications [15]. Intuitively, centroids can be adversely affected by the existence of outliers or bi-modal distributions. Similarly, the distances between examples in one dimension may be less meaningful (or discriminative) than the distances in another dimension. Learning a discriminative model against pseudo-negatives can help to alleviate these effects and better localize the prevailing distribution of positive examples in feature space and eliminating non-discriminative dimensions. In our implementation, we take the photos $\mathbb{P}_V$ from within the candidate set and treat them as pseudo-positives for learning. We then sample images randomly from the global pool, $\mathbb{P}$, and treat these images as pseudo-negatives. We take the same normalized low-level global feature vector (consisting of color and texture) from the previous distance-

ranking model as the input feature space. We randomly partition this data into two folds, training a support vector machine (SVM) classifier [5, 25] with the contents of one fold and then applying the model to the contents of the other fold. We repeat the process, switching the training and testing folds. The images can then be ranked according to their distance from the SVM decision boundary.

5.4.3 Point-wise Linking

The above-mentioned low-level self-similarity and discriminative modeling methods use global low-level features and mostly capture recurrent global appearances and patterns. These metrics do not necessarily capture whether or not any two images are actually of the same real-world scene, or contain the same objects. We use SIFT descriptors to discover the presence of these overlaps in real-world structures or scenes between two photographs.

The overlap between any two given images can be discovered through the identification of correspondences between interest points in these images. Given two images, each with a set of SIFT interest points and associated descriptors, we use a straight-forward approach, sometimes known as ambiguity rejection, to discover correspondences between interest points. Intuitively, in order to decide if two SIFT descriptors indeed capture the same real-world object, we need to measure the distance between the two descriptors and apply some threshold to that similarity in order to make a binary match/non-match decision. In ambiguity rejection, this threshold is set on a case by case basis, essentially requiring that, for a given SIFT descriptor in an image, the nearest matching point in a second image is considered a match only if the Euclidean distance between the two descriptors is less than the distance between the first descriptor and all other points in the second image by a given threshold. To ensure symmetry, we also find matching points using a reverse process, matching from the second image against the first image. When a pair of points is found to be a candidate both through matching the first image against the second *and* through matching the second image against the first, then we take the candidate match as a set of corresponding points between the two images. The intuition behind this approach is that matching points will be highly similar to each other and highly dissimilar to all other points.

Once these correspondences are determined between points in various images in the set, we establish links between images as coming from the same real-world scene when the number of point-wise correspondences between the two images exceeds a threshold. In our experiments, we have set this threshold equal to three, since some of our initial observations have shown that this yields precise detection. The result is a graph of connections between images in the candidate set based on the existence of corresponding points between the images. We then score the images according to their rank in the graph – the total number of images to which they are connected. The intuition behind such an approach is that representative views of a particular location or landmark will contain many important points of the structure which will be linked across various images. Non-representative views (such as close-ups or shots of people), on the other hand, will have fewer links across images.

5.4.4 Fusion of Ranking Methods

The ranking methods described above capture various complementary aspects of the repeated views of the real-world scenes. To leverage the power of each of the methods, we apply each of them independently and then fuse the resulting scores. Each method returns a score for each of the images in the set. We normalize the results returned from each method via a logistic normalization and then take the average of the scores resulting from each method to give a fused score for each image. For each cluster V the images within each cluster, $\mathbb{P}_V$, we now have a list of photos R_V, ranked by their representativeness within cluster V.

Once the ranking is done, the system generates the final ranked list of representative photos $\mathbb{R}_x$. We do that by sampling the highest-ranking images in R_V from the set of clusters $V \in \mathbb{V}_x$. The clusters are not sampled equally: as noted above, the lowest-ranking clusters are simply discarded, and the higher-ranking clusters have images sampled proportionally to the score of the cluster. The end result is a ranked list of images, which hopefully captures varying representative views for each landmark. How well does the system work?

6. EVALUATION

We used a number of different methods to evaluate the system results in generating representative views. All the different methods were based on input from human judges, and were driven by an "image search" use case. The goals of the evaluations included:

- Verifying that the generated views for landmarks are representative, but still diverse and precise.
- Confirming that our methods improve on performance of naïve methods.
- Tuning the parameters used throughout the system.
- Assessing the contribution of the different factors (tags, metadata, image analysis) to the results.

To this end, we ran two different experiments, described below: a simple test to measure the precision of search results using the system, and a more elaborate experiment designed to evaluate more difficult metrics such as "representativeness" and diversity. First, though, we provide some details on the dataset and the analysis.

6.1 Dataset and Processing

To evaluate the system's performance, we use a set of over 110,000 geo-referenced photos from the San Francisco area. The photos were retrieved from the dataset of geotagged photos available on Flickr [8]. We discovered landmark tags in the dataset and their locations, using the methods described above. In particular, we generated 700 location clusters (the number was chosen as a trade-off between span of geographic coverage and the expected number of photos per cluster). For each location cluster, representative tags are determined by scoring frequent tags within the cluster. For the tags chosen by the system, we retain the information about the tag and the clusters where the tag scored well – a set of (tag, cluster set) tuples $(x, \mathbb{C}_x)$.

To make the evaluation feasible, we consider a subset of ten manually selected landmarks tags (listed in Figure 2) and their clusters. Representative images for each tag are extracted using four different techniques:

- **Tag-Only**. This method serves as a baseline for the system performance, randomly selecting ten images with the corresponding tag from the dataset (i.e., from $\mathbb{P}_x$).

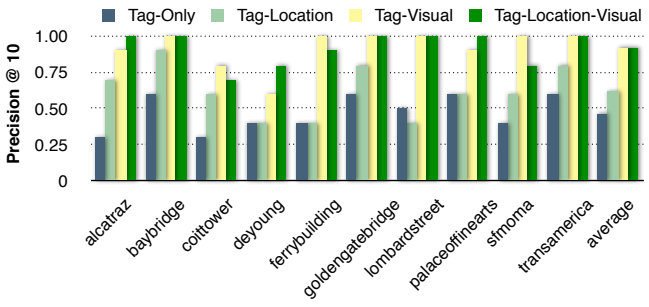


Figure 2: Precision at 10 for representative images selected for locations using various methods.

- **Tag-Location**. In this second baseline, the system randomly chooses ten images with the corresponding tag that fall within one of the tag's extracted location clusters (i.e., from $\mathbb{P}_x$ photos that fall in one of the extracted clusters $\mathbb{C}_x$.
- **Tag-Visual**. Images are selected by our system, running the visual analysis described above on all photos $\mathbb{P}_x$.
- **Tag-Visual-Location**. Images are selected by our system, running the visual analysis as described above on photos in $\mathbb{P}_x$ that fall in one of the extracted clusters $\mathbb{C}_x$.

Consequentially, for each of our selected landmarks tags x, we generated four different rankings $\mathbb{R}_x$ for the photos in $\mathbb{P}_x$. We further look at the top ten images in $\mathbb{R}_x$ for our evaluation, simulating image search results for the landmark. Next, we perform a simple precision evaluation on these ten sets of ten images for each of the four methods. We then describe a more elaborate evaluation of these results.

6.2 Initial Evaluation: Precision

As a first step in our evaluation, we examine the potential benefit of using the location metadata and and image analysis to improve the precision of tag-based retrieval for landmark queries.

We used the four different methods to select ten representative images for each of the ten evaluated landmarks and evaluate the precision (P@10) of each set of results. This metric measures the percentage of the images that are indeed representative of the landmark. The ground-truth judgments of image representativeness are defined manually by human evaluators. The precision evaluation criteria was rather simple for a human to evaluate: images contain views of the location that are recognizable to viewers familiar with the location, then they are marked as representative, otherwise, they are marked as non-representative.

The results of the evaluation are shown in Figure 2. In this figure, the X-axis shows the ten selected landmarks, the right-most column shows averaged results over all landmarks. For each landmark, we show the P@10 score for each of the four methods. For example, for `Bay Bridge`, six of the ten images retrieved using "Tag-Only" were recognizeable images of the bridge, compared to all 10 of the images retrieved using "Tag-Location-Visual".

Overall, Figure 2 shows a clear added benefit of location and vision constraints for the selection of representative landmark images. In the baseline case, the tag-only approach, the average P@10 is 0.47, a finding that confirms other recent observations about the accuracy of tags [10]. The Tag-Location condition yields, on average, a 32% relative increase in the precision of the selected images, which indicates that location is a strong predictor of image content. The Tag-Visual and Tag-Location-Visual condition both have similar performances, improving upon the Tag-Location condition by 48% on average (or, a 96% improvement over the tag-only baseline). This indicates that visual processing is equally robust on subsets found by tags-only or by constraining on tags and locations and that the vision-driven analysis significantly improves precision.

The precision metric does not capture other critical elements of search results evaluation. In particular, we want to verify that providing representative results does not influence other desired criteria such as diversity and overall quality of representation. Next, we describe a wider evaluation that was designed to measure these other metrics.

6.3 Experimental Setup

Ideally, a set of image search results for a landmark or a geo feature will have a diverse set of photos, demonstrating different aspects of the landmark using highly-representative images. We designed a user-driven evaluation to help us assess these qualities in our system.

We compared each of the conditions mentioned in Section 6.1: Tags-only, Tag-Location, Tag-Visual and Tag-Location-Visual. For each of the conditions we applied the method to produce a page of "image search results" for each of the ten San Francisco landmarks mentioned above. In total, then, we had 40 different pages to be evaluated by human judges.

The core of the evaluation addressed these pages of results produced by one of the different conditions for one of the landmarks. Each page contained the name of the landmark, and ten images that were selected by the applied method. For each such page, the user had to answer four evaluation questions (appearing in a slightly abbreviated form here):

- **Representative**. How many photos in this set are representative of the landmark (0-10 scale)?
- **Unique**. The question was posed as "How many of the photos are redundant (0-10)?", but for our analysis below, we used the measure of "unique" photos, which is simply the number of representative photos minus the number of redundant photos for each judgment.
- **Comprehensive**. Does this set of results offer a comprehensive view of the landmark (1-5)?
- **Satisfying**. How satisfied are you with this set of search results (1-5)?

For the purpose of the evaluation, we provided futher explanations for the differnet categories. For example, we explained the "representative" question as "pictures you might have chosen if you were asked to create a representative set of the landmark's photos". The evaluation users viewed pages in random order, without repetition, switching between landmarks and methods as they hit "next".

We solicited via email a total of 75 judges to participate in this web-based evaluation. We did not force any minimum number of evaluations from each judge, but rather let them go through as many of the 40 pages as they cared to do. This way, we could get many participants while still ensuring that the judges are relatively engaged in the evaluation. We did, however, discard all data from judges whom evaluated fewer than 5 pages. For the results, then, we used judgments from

30 judges on a total of 649 pages, answering four questions per page and yielding a total of 2596 data points.

6.4 Results

Table 1 shows a summary of the results of our user evaluation. The conditions are marked T (Tag-Only), T+L (Tag-Location), T+V (Tag-Visual) and T+L+V (Tag-Location-Visual) and results are shown for each of the questions described above. The results for each condition are averaged over all users and landmarks. For example, for the *representative* photos test, our judges ruled that an average of 8.8 out of 10 photos chosen using the tag-location-visual method were indeed representative of the landmark; compared to 8.6 using the tag-visual condition, 7.1 using the tag-location method, and 6.2 representative photos using the tag-only condition. We tested for statistical significance in the changes in the tag-location, tag-visual, and tag-location-visual systems over the baseline Tag-Only system using a paired T-test. Statistically significant improvements ($p < .1$) are shown in boldface in the table; the *representative* improvement in both T+L+V and T+V over Tag-only was significant with $p < .05$.

The *representative* photos test, then, shows that the visual analysis clearly improves the quality of the visual summary presented for each landmark over the baseline methods. Note that this test should roughly correspond to the precision evaluation discussed above (where evaluation was executed by a single judge). Interestingly, the average scores of the judges agree with the evaluation of precision for the visual-based approaches; but the judges' scores are higher for the tag-only and tag-location methods. Seemingly, our judges were more tolerant of semi-representative images, such as those where the landmark is obscured by people posing in front or where it is harder to recognize the landmark due to it being photographed in extreme close-up.

In general, we see in Table 1 that the application of visual processing provides significant gains in *representative* score and *satisfaction* but yields little (if any) difference in the *unique* and *comprehensiveness* measures. This is still, though, a promising result, indicating that the visual processing increases the total number of relevant photos in a summary by replacing irrelevant photos with relevant (but sometimes redundant) photos. The results for the *satisfaction* test show that the users do prefer this trade-off, or in other words, the presence of relevant redundant photos is preferable to the presence of irrelevant photos. Indeed, the 22% improvement in the *satisfaction* metric, from a score of 2.7 in the tags-only condition, to 3.3 in tag-location-visual and tag-visual, is the most encouraging.

Question	T	T+L	T+V	T+L+V
Representative	6.2	**7.1** ***14.5%***	**8.6** ***38.7%***	**8.8** ***41.9%***
Unique	5.5	6.0 *9.0%*	5.9 *7.2%*	5.5 *0%*
Comprehensive	3.2	3.3 *3.1%*	3.5 *9.4%*	3.5 *9.4%*
Satisfying	2.7	3.0 *11.1%*	**3.3** ***22.2%***	**3.3** ***22.2%***

Table 1: Average scores for each of the four evaluation questions on each of the test conditions: tags-only (T), tags and locations (T+L), tags and visual (T+V), and tags and locations and visual processing (T+L+V). Relative improvements over the tags-only condition are shown in italics. Statistically significant changes ($p < 0.1$) are shown in boldface.

6.5 Discussion

This section lays out several additional observations that follow from our results. What can our results tell us about the views of different landmarks? What are users looking for in these results? We also briefly note how quality metrics could be added to the processing, and provide ideas about how to incorporate this processing in a live search system.

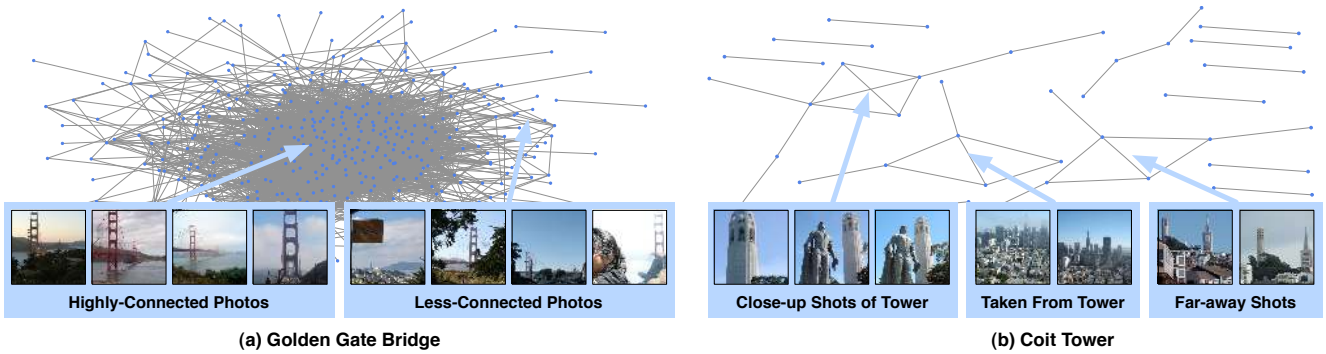
6.5.1 Scene Views and the Image Link Graph

For some of the landmarks (or geographic features) we extract, the visual-based methods still do not provide perfect precision. A few complicating issues arise from the nature of landmarks, and the way users apply tags to photos. For instance, some geographic landmarks can act as a point from which to photograph, rather than the target of the photo; such photographs are often tagged with the geographic landmark which is the source of the photo. For example, Coit Tower is a frequently-photographed landmark, but many of the photographs associated with the tag `Coit Tower` are actually photographs of the San Francisco skyline, taken from the observation deck at the top of the tower. Similarly, for museums and other buildings such as `De Young` and `SF MOMA`, the expected representative views are split between outside views of the building, as well as recognizable internal architectural aspects. However, users might also photograph particular artworks and other non-representative interior views of such landmarks.

The trend across these cases is that some of the frequently-taken photograph views associated with the landmark are not necessarily representative of the landmark. It is arguable, and could be left for human evaluation, whether these images are desirable for representation of the landmark. Do users wish to see images taken from Coit Tower when they search for that phrase? Do they want to see images from inside the De Young?

Our analysis of SIFT-based links between photos (Section 5.4.3) can potentially detect such cases of truly disparate views. We briefly discuss the structure of these graphs to give insight into the efficacy of this approach and suggest ways in which we can better leverage the approach in future work to improve the overall performance of our system. Figure 3 shows graphical representations of the link structures discovered among photos using the point-wise linking method discussed in Section 5.4.3. In the figure, each node is an image and the edges are point-wise links discovered between two photos according to the criteria specified above.

In Figure 3a, we see the visual-link graph of the photos tagged with `Golden Gate Bridge`; a nearly fully-connected graph emerges. Examining the degree of each node (or the number of connections discovered from a given photo), we verify that our proposed point-wise linking scheme for image ranking is performing as we expect: highly-connected images (closer to the center of the graph) tend to be qualitatively more iconic and encompass more of the landmark, while less-connected images (closer to the edge of graph) tend to be qualitatively less iconic, with many photos having portions of the landmark occluded or otherwise obscured. Not depicted in the link-graph structure are a large portion of images for which no connections to other images were discovered. These images mostly have no portion of the landmark visible at all.



(a) Golden Gate Bridge (b) Coit Tower

Figure 3: Visualizations of graphs of point-wise links resulting for Golden Gate Bridge and Coit Tower.

	R	U	C	S
(R) Representative	1	0.5548	0.4672	0.5506
(U) Unique	0.5548	1	0.4482	0.5381
(C) Comprehensive	0.4672	0.4482	1	0.7639
(S) Satisfying	0.5506	0.5381	0.7639	1

Table 2: Pearson correlation values between responses for each of the four evaluation questions. All scores are significantly correlated ($p < 0.0001$, $N \sim 1000$).

On the other hand, for photos tagged with `Coit Tower` (shown in Figure 3b), we find that a substantially different graph structure emerges. There are a number of large disjoint sets that appear, each of which encapsulates a different view of the structure. One set shows close-up views of the exterior of the tower, while another set shows far-away views of the tower exterior from an opposing direction. Still another set contains photos taken *from* the observation deck inside the tower, so the photos actually do not physically show the tower, despite being tagged `Coit Tower` and exhibiting aspects of the same real-world scene. Each of these disjoint subsets captures unique, but common, views associated with the landmark. Interestingly, these views are difficult to capture using low-level global (color and texture) features, since they all appear fundamentally the same in that space, with blue skies on top and building-like structures in the center and foreground.

The fact that point-wise (SIFT) descriptors can successfully discriminate between these views might suggest that this point-wise linking strategy may discover more meaningful views of locations than the k-means clustering approach that we have employed in this work.

6.5.2 *What Users Value in Landmark Image Search*

Our results indicate some interesting aspects of the human evaluation of image search results. We have processed the results of the user evaluation to check for correlations between the scores that users provided for each of the four questions that we asked. Table 2 shows the Pearson correlation values for each of the question pairs. Not surprisingly, the resulting scores for the various questions are significantly correlated ($p < 0.0001$, $N \sim 1000$). However, it is noteworthy that the answer to question 4 ("how satisfied") is slightly more correlated with question 1 ("how many representative") than with question 2 (which we transform into a positive-valued "how many unique representative images" score). This correlation suggests that users may be more tolerant of redundant (but relevant) results than they are of irrelevant results. Interestingly, the answer to question 3 ("how comprehensive") is again slightly more correlated with question 1 than with question 2, even though the latter is a more direct measure of a "comprehensive" quality. This finding might suggest that the presence of irrelevant images has a more negative impact than the presence of relevant (but redundant) images on the users' perception of the comprehensiveness of a set. We do stress that these findings are not conclusive and are just reported here as a path for future exploration.

6.5.3 *Introducing Photo Quality Metrics*

Ideally, the representative photos returned by our system are not only accurate, and diverse, but will also visually compelling and of high quality. While measures of photo quality are hard to extract from photo content, they could be readily mined from activity patterns in community-driven environment like Flickr. In Flickr, for example, photos are assigned an "interestingness" score that is based in part on the number of views for each image, the number of people who marked the photo as a "favorite", and the count of comments left on the image by other users. Such a measure, or any other measure of photo quality, could be easily incorporated into the result set to bias the system towards displaying more visually compelling images, that are still ranked high according to our other metrics and processing.

7. CONCLUSIONS AND FUTURE WORK

We have demonstrated that rich information about locations and landmarks can be learned automatically from user-contributed media shared on the web. In particular, a collection's locations of interest can arise from geo-spatial photographing patterns. Meaningful tags that represent these these locations and landmarks can be learned from tags that users frequently associate with these images. Finally, visual models of landmarks and geographic features can be learned through mining the photos shared by many individuals, potentially generating a summary of the frequently-photographed views by selecting canonical views and rejecting outliers. Evaluating visually-filtered summaries in the context of image search shows a significant increase in the representativeness of the selected photos when compared against sets derived from tags and metadata alone, suggesting potential for search and exploration tasks.

Future work might explore the best approaches for incorporating such a system into a standard web-based image search engine. Can our learned sets of location/landmark tags be applied as a pre-filter for web image queries to decide when to apply further visual re-ranking? How will the results be merged with traditional web-based results? What kind of new result presentation technique can be used to leverage the knowledge of visual clusters and map locations? Some answers are easier than others, but it is all certainly quite promising.

In general, our results suggest that tag-based and community-driven media sites are not a 'lost cause'. Despite the many issues that arise from the loosely-annotated media in these web sites (false positive and false negatives in tag data are just one example), rich and useful information about some domains can be derived. In addition, despite the noisy data, vision algorithms can be employed effectively, and without training. Applying such techniques in other domains, beyond landmarks and geographically-driven features, would even further improve our knowledge of the world.

8. ACKNOWLEDGMENTS

We would like to thank Rahul Nair for his contribution to the data collection and analysis and our dedicated "judges" who spent valuable time tirelessly evaluating our results.

9. REFERENCES

[1] S. Ahern, S. King, M. Naaman, R. Nair, and J. H.-I. Yang. ZoneTag: Rich, community-supported context-aware media capture and annotation. In *Workshop on Mobile Spatial Interaction (MSI) at the SIGCHI conference on Human Factors in computing systems (CHI '07)*, 2007.

[2] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries*, May 2007.

[3] T. L. Berg and D. A. Forsyth. Automatic ranking of iconic images. Technical report, U.C. Berkeley, January 2007.

[4] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the 12th International Conference on Multimedia (MM2004)*, pages 952–959, New York, NY, USA, 2004. ACM Press.

[5] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] S. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D. Zhang. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. *NIST TRECVID Workshop, Gaithersburg, MD, November*, 2005.

[7] M. Davis, M. Smith, F. Stentiford, A. Bambidele, J. Canny, N. Good, S. King, and R. Janakiraman. Using context and similarity for face and location identification. In *Proceedings of the IS&T/SPIE 18th Annual Symposium on Electronic Imaging Science and Technology*, 2006.

[8] Flickr.com, yahoo! inc. http://www.flickr.com.

[9] F. Jing, L. Zhang, and W.-Y. Ma. Virtualtour: an online travel assistant based on high quality images. In *Proceedings of the 14th International Conference on Multimedia (MM2006)*, pages 599–602, New York, NY, USA, 2006. ACM Press.

[10] L. Kennedy, S.-F. Chang, and I. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258, 2006.

[11] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th International Conference on Multimedia (MM2007)*, pages 631–640, New York, NY, USA, 2007. ACM.

[12] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] B. Manjunath and W. Ma. Texture features for browsing and retrieval of image data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(8):837–842, 1996.

[14] M. Naaman, A. Paepcke, and H. Garcia-Molina. From where to what: Metadata sharing for digital photographs with geographic coordinates. In *10th International Conference on Cooperative Information Systems (CoopIS)*, 2003.

[15] A. Natsev, M. Naphade, and J. Tešić. Learning the semantics of multimedia queries and concepts from a small number of examples. *Proceedings of the 13th International Conference on Multimedia (MM2005)*, 2005.

[16] N. O'Hare, C. Gurrin, G. J. Jones, and A. F. Smeaton. Combination of content analysis and context features for digital photograph retrieval. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2005.

[17] M. S. M. Orengo. Similarity of color images. *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 2420:381–392, 1995.

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

[19] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, July 2007.

[20] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *ICCV '07: Proceedings of the 11th IEEE international Conference on Computer Vision*. IEEE, 2007.

[21] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[22] S. Sontag. *On Photography*. Picador USA, 2001.

[23] K. Toyama, R. Logan, and A. Roseway. Geographic location tags on digital images. In *Proceedings of the 11th International Conference on Multimedia (MM2003)*, pages 156–166. ACM Press, 2003.

[24] C.-M. Tsai, A. Qamra, and E. Chang. Extent: Inferring image metadata from context and content. In *IEEE International Conference on Multimedia and Expo*, 2005.

[25] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[26] S. Wang, F. Jing, J. He, Q. Du, and L. Zhang. Igroup: presenting web image search results in semantic clusters. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 587–596, New York, NY, USA, 2007. ACM Press.

[27] Y. Wu, E. Y. Chang, and B. L. Tseng. Multimodal metadata fusion using causal strength. In *Proceedings of the 13th International Conference on Multimedia (MM2005)*, pages 872–881, New York, NY, USA, 2005. ACM Press.

[28] Youtube.com, google inc. http://www.youtube.com.