How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections

Lyndon Kennedy[*] , Mor Naaman, Shane Ahern, Rahul Nair, Tye Rattenbury[†]
Yahoo! Research Berkeley
Berkeley, CA, USA
{lyndonk,mor,sahern,rnair,tye}@yahoo-inc.com

ABSTRACT

The advent of media-sharing sites like Flickr and YouTube has drastically increased the volume of community-contributed multimedia resources available on the web. These collections have a previously unimagined depth and breadth, and have generated new opportunities – and new challenges – to multimedia research. How do we analyze, understand and extract patterns from these new collections? How can we use these unstructured, unrestricted community contributions of media (and annotation) to generate "knowledge"?

As a test case, we study Flickr – a popular photo sharing website. Flickr supports photo, time and location metadata, as well as a light-weight annotation model. We extract information from this dataset using two different approaches. First, we employ a location-driven approach to generate aggregate knowledge in the form of "representative tags" for arbitrary areas in the world. Second, we use a tag-driven approach to automatically extract place and event semantics for Flickr tags, based on each tag's metadata patterns.

With the patterns we extract from tags and metadata, vision algorithms can be employed with greater precision. In particular, we demonstrate a location-tag-vision-based approach to retrieving images of geography-related landmarks and features from the Flickr dataset. The results suggest that community-contributed media and annotation can enhance and improve our access to multimedia resources – and our understanding of the world.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

geo-referenced photographs, photo collections, social media

[*]Also affiliated with Columbia University Dept. of Electrical Engineering

[†]Also affiliated with UC Berkeley Computer Science Dept.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

1. INTRODUCTION

The proliferation of digital photo-capture devices, and the growing practice of online public photo sharing, have resulted in large public pools of consumer photographs available online. Billions of images shared on websites such as Flickr[1] serve as a growing record of our culture and environment. Searching, viewing, archiving and interacting with such collections has broad social and practical importance. However, due to their magnitude, these collections are increasingly difficult to understand, search and navigate. In particular, automated systems are largely incapable of understanding the semantic content of the photographs. Thus, the prospects of 'making sense' of these photo collections are largely dependent on metadata and information that is manually assigned to the photos by the users.

Clues as to the content of the images can be found in text (such as labels or "tags") that is associated with the images by users. Researchers have previously analyzed user-supplied tags in multimedia collections to extract trends and visualization data [6], as well as suggest annotations for unannotated images [14, 20]. However, the associated tags presents an additional set of challenges for multimedia systems. As used on Flickr and other photo-sharing website, tags and other forms of text are freely entered and are not associated with an ontology or any type of categorization. Tags are therefore often innacurate, wrong or ambiguous. In particular, due to the complex motivations involved that drive usage of tags and text descriptions [2], tags do not necessesarily describe the content of the image [11].

Location information associated with the photos can prove valuable in understanding photos' content. Exceedingly, images are geo-referenced (or, "geotagged"): associated with metadata describing the geographic location in which the images were taken [24]. For instance, more than twenty million photos with location metadata are now available on Flickr – the first major collection of its kind. Location metadata will be exceedingly available, primarily from location-aware camera-phones and digital cameras, and initially from user input [24]. The location metadata alone was shown to be beneficial in browsing and organizing these collections [15, 18, 24]. In addition, location can sometimes suggest the semantic content of the images [14, 20].

Ultimately, systems would benefit from applying computer vision techniques to improve our understanding of images in community-contributed media collections. Applying

[1]http://flickr.com

computer vision in unconstrained domains is a difficult problem that is sure to be a research topic for years to come [21]. However, visual pattern recognition approaches can be used for some well-defined tasks in such unstructured collections.

The key contribution of this paper is combining tag-based, location-based and content-based analysis to improve the automated understanding of such large user-contributed media collections. First, an analysis of the tags associated with images using a location-driven approach helps us generate "representative tags" for arbitrary areas in the world. The selected tags often correspond to landmarks or geographic features inside the areas in question. Second, we employ a tag-driven approach to automatically extract place and event semantics for Flickr tags, based on each tag's metadata (location and time) patterns.

Using the patterns we extract from tags and location, vision algorithms can be employed with greater precision. In particular, we demonstrate a location-tag-vision-based approach to retrieve images of geography-related landmarks and features from the Flickr dataset.

In particular, this context-annotation-content analysis has the potential to assist in various critical tasks involving media collections, including:

- Improving precision and breadth of retrieval for landmark and place-based queries.
- Soft annotation of photos, or suggesting tags to un-annotated geo-referenced photos uploaded by users.
- Generating summaries of large collections by selecting representative photos for places and identified landmarks.

Our work is motivated and designed by the characteristics of an actual, existing dataset of more than 20,000,000 geo-referenced photos currently available on Flickr. We do not rely on gazetteers, or existing lists of landmarks, ontologies of tag semantics, or any other manual classification. This realistic dataset offers a huge opportunity, accompanied, of course, by new challenges and requirements for multimedia research.

The metadata model used throughout this paper is defined in Section 3. We describe the location-driven analysis in Section 4. In Section 5, we describe how we extract semantics from Flickr tags by their metadata distributions. Section 6 provides the details on incorporating vision algorithms in our analysis; a short evaluation is presented in Section 7. We begin, of course, with related work.

2. RELATED WORK

We report below on related work in metadata and multimedia fusion, metadata-based models of multimedia, and computer-vision approaches to landmark recognition.

The topic of "landmark recognition" has been studied in the last few years, but applied to limited or synthetic datasets only. In particular, analysis of context and content in photo collection has been studied in [5, 17, 25, 27] and more. The work of Tsai et al. [25], 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 that assumed an existing database of landmark. This assumption is at best exremely limiting, and perhaps unrealistic. O'hare at al. [17] 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. This method is of course different than our work to automatically identify landmarks and their locations. Davis et al. [5] had a similar method that exposed the similarity between places based on content and context data, but did not detect or identify landmarks.

Other work has addressed building models of location from the context and annotation of photographs. In [14], Naaman et al. 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 the computer vision field, in [11], 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 are trained on images from web search results and then applied to consumer photos.

In [3], Berg and Forsyth present an approach to ranking "iconic" images from a set of images with the same tag on Flickr. Our work also examines ranking the most representative (or iconic) 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.

Snavely et al. [22] have presented a system which can register point-wise correspondences between various images of the same location and iteratively approximate the camera angles from which the various images were collected. This system, however, is intended for exploration and has no mechanism for selecting a few "representative" images to summarize the location. The system is also computationally expensive and currently impractical for running over the wide range of landmarks. Our system can serve as input and automatic filter for the Snavely et al. algorithm.

3. MODEL AND REQUIREMENTS

This section formalizes the properties of the dataset used throughout this paper. We expand the research problem definitions and proposed solutions in the respective sections.

Our dataset consists of three basic 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. The set of all tags over all photos is defined as: $\mathbb{X} \triangleq \cup_{p \in \mathbb{P}} \mathbb{X}_p$. We can use the equivalent notation to denote the set of tags that appear in any subset $\mathbb{P}_S \subseteq \mathbb{P}$ of the photo set as $\mathbb{X}_S$. For convenience, we define the subset of photos associated with a specific tag as: $\mathbb{P}_x \triangleq \{p \in \mathbb{P} \mid x \in \mathbb{X}_p\}$. Accordingly, photos with the tag x in a subset $\mathbb{P}_S$ of $\mathbb{P}$ are denoted $\mathbb{P}_{S,x} \triangleq \{\mathbb{P}_S \cap \mathbb{P}_x\}$.

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.

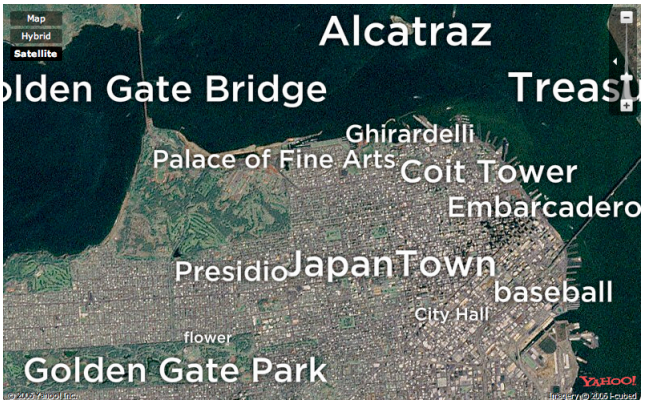

Figure 1: Representative tags for San Francisco

4. EXTRACTING KNOWLEDGE ABOUT LOCATIONS

How do we extract knowledge about geographic regions from community contributions of images and metadata? Using the data described in Section 3 we wish to automatically identify tags that are "representative" for each given geographical area. It is important to note that these representative tags are often not the most commonly used tags within the area under consideration. Instead, we wish to surface tags that uniquely define sub-areas within the area in question. For example, if the user is examining a portion of the city of San Francisco, then there is very little to be gained by showing the user the `San Francisco`[2] or `Bay Area` tags, even if these tags are the most frequent, since the tags apply to the entire area under consideration. Instead, we would ideally show tags such as `Golden Gate Bridge`, `Alcatraz` and `Fisherman's Wharf` which uniquely represent specific locations, landmarks and attractions within the city.

Before we can determine the "representativeness" of a tag, we need to have an intuition of what the term implies. We follow some simple heuristics that guide us in devising the algorithms. The heuristics attempt to capture the human attention and behavior as represented in the photos and tag dataset. Our heuristics are aimed toward both finding important locations and identifying represntative tags. For example, the number of photographs taken in a location is an indication of the relative importance of that location; a similar indication is found in the number of individual photographers that have taken photos in a location. Looking at tags, users are likely to use a common set of tags to identify the objects/events/locations that occur in photographs of a location; and tags that occur in a concentrated area (and do not occur often outside that area) are more representative than tags that occur diffusely over a large region.

We start by assuming that the system considers a single given geographic area G, and the photos that were taken in this area, $\mathbb{P}_G$. The system attempts to extract the representative tags for area G. This computation is done in two main steps: in the first step, we cluster the set of photos $\mathbb{P}_G$ using the photos' geographic locations. In the second step, we score the tags in each cluster for their "representativeness".

In the first step, the system geographically clusters the set of photographs $\mathbb{P}_G$. We use the k-Means clustering algorithm, based on the photos' latitude and longitude. Geographical distance is used as the distance metric, and the stopping condition for the k-Means algorithm is when each cluster's centroid movement drops below 50 meters. The number of seed points used for the k-Means algorithm is based on $|\mathbb{P}_G|$, the number of photographs in the area under question. Based on empirical observation of the results, we set the seed value to range from three for sparse areas (under 100 photographs) to 15 for denser areas (greater than 4000 photographs).

Once the clusters have been determined, the system scores each cluster's tags to extract representative tags. In other words, we consider each cluster C, and the set of tags $\mathbb{X}_C$ that appear with photos from the cluster. We score each tag $x \in \mathbb{X}_C$ according to the factors defined below.

One of the factors we use is TF-IDF (term frequency, inverse document frequency). This metric assigns a higher score to tags that have a larger frequency within a cluster compared to the rest of the area under consideration. Again, the assumption is that the more unique a tag is for a specific cluster, the more representative the tag is for that cluster. Of course, we do not wish to use tags that only appear a few times in the cluster; the term frequency element prefers popular tags.

The TF-IDF is computed with slight deviation from its regular use in Information Retrieval. The term frequency for a given tag x within a cluster C is the count of the number of times x was used within the cluster: $\text{tf}(C,x) \triangleq |\mathbb{P}_{C,x}|$. The inverse document frequency for a tag x, in our case, computes the overall ratio of the tag x amongst all photos in the region G under consideration: $\text{idf}(x) \triangleq |\mathbb{P}_G|/|\mathbb{P}_G, x|$. Note that we only consider a limited set of photos ($\mathbb{P}_G$) for the IDF computation, instead of using the statistics of the entire dataset. This restriction to the current area, G, allows us to identify local trends for individual tags, regardless of their global patterns.

While the tag weight is a valuable measure of the popularity of the tag, it can often be affected by a single photographer who takes a large number of photographs using the same tag. To guard against this scenario, we include a user element in our scoring, that also reflects the heuristic that a tag is more valuable if a number of different photographers use it. In particular, we factor in the percentage of photographers in the cluster C that use the tag x: $\text{uf} \triangleq \mathbb{U}_{C,x}/\mathbb{U}_C$.

The final score for tag x in cluster C is computed by $\text{Score}(C,x) = \text{tf} \cdot \text{idf} \cdot \text{uf}$. The higher the tf-idf score, and the user score, the more distinctive the tag is within a cluster.

[2]We use this format to represent tags in the text.

For each cluster, we retain only the tags that score above a certain threshold. The threshold is needed to ensure that the selected tags are meaningful and valuable for the aggregate representation. We use an absolute threshold for all computed clusters to ensure that tags that are picked are representative of the cluster.

A sample set of representative tags for San Francisco is shown in Figure 1. In [1, 9] we supply more details on the algorithm, and on how we extend the computation to support multiple regions and zoom levels; we also evaluate the algorithmic results. Using this algorithm, we had created a live visualization[3] of the world. The details and evaluation of this system can also be found in [1].

We return to this algorithm in Section 6. We note that the representative tags often correspond to landmarks and geographic features. In Section 6, we use these computed landmark tags to seed a vision-based system that attempts to idenfity representative images for each tag.

5. IDENTIFYING TAG SEMANTICS

How do we extract knowledge about specific tags or textual terms, using community contributions of images and metadata? Using the same data, as described in Section 3, we wish to identify tags that have event or place semantics. Based on the temporal and spatial distributions of each tag's usage on Flickr, we attempt to automatically determine whether a tag corresponds to a "place" and/or "event". As mentioned above, extraction of event and place semantics can potentially assist many different applications in the photo retrieval domain and beyond. These applications include improved image search through inferred query semantics; automated creation of place and event gazetteer data; generation of photo collection visualizations by location and/or event/time; support for tag suggestions for photos (or other resources) based on location and time of capture; and automated association of missing location/time metadata to photos, or other resources, based on tags or caption text.

We loosely define "place tags" as tags that are are expected to exhibit significant spatial patterns. Similarly, "event tags" are ones that are expected to exhibit significant temporal patterns. Example place tags are `Delhi`, `Logan Airport` and `Notre Dame`. Sample event tags are `Thanksgiving`, `World Cup`, `AIDS Walk 2006`, and `Hardly Strictly Bluegrass`. Interestingly, `Hardly Strictly Bluegrass` is a festival that takes places in San Francsico, and thus represents both an event and a place. Spatial and temporal distributions for `Hardly Strictly Bluegrass` are shown in Figure 2. Examples of tags not expected to represent events or locations are `dog`, `party`, `food` and `blue`.

Formally, we can define the location and time usage distributions for each tag x: $\mathcal{L}_x \triangleq \{\ell_p \mid p \in \mathbb{P}_x\}$ (locations of all photos with tag x) and $\mathcal{T}_x \triangleq \{t_p \mid p \in \mathbb{P}_x\}$ (time of all photos with tag x). We now show how place semantics for a tag x can be derived from the tag's location distribution, $\mathcal{L}_x$. Similarly, time semantics can be derived from $\mathcal{T}_x$.

The method we use to identify place and event tags is *Scale-structure Identification* (or SSI). This method measures how similar the underlying distribution of metadata is to a single cluster at multiple scales. For example, examining the location distribution $\mathcal{L}_x$ for the tag `Hardly Strictly Bluegrass`, it appears as a single strong cluster at the city scale; but appears as multiple clusters at a neighborhood scale (see Figure 2). SSI identifies place tags by: (1) clustering the usage distribution $\mathcal{L}_x$ at multiple scales; (2) measuring how similar the clustering at each scale is to a single cluster by calculating the information entropy; and (3) summing the entropy calculations at each scale to produce a single score that indicates how similar the usage data is to a single cluster over multiple scales. The process for identifying event tags is equivalent, using the time distribution $\mathcal{T}_x$. In the case of `Hardly Strictly Bluegrass`, we expect SSI to identify the tag both as an event, and as a place.

In [19] we provide additional details and evaluation on SSI and other methods for extracting these semantics from Flickr tags. While not quite perfect, we consider all geotagged photos in San Francisco and show that tags that represent events and places can be identified with reasonable precision and recall.

In other words, the metadata patterns of tags in community-contributed media collections can be used to extract semantics of these tags. These semantics can be used when combining content-based analysis with context-based tools. For example, attempting to detect landmarks in our dataset, we can rule out tags that have do not have place semantics, or have place and event semantics: the tag `dog` is not likely to be a geographic-based landmark (because it does not have place semantics). Similarly, `Hardly Strictly Bluegrass` is not likely to represent a landmark (because of its identified event semantics). This semantic extraction can assist us to select tags for which we can generate computer vision models, as discussed next.

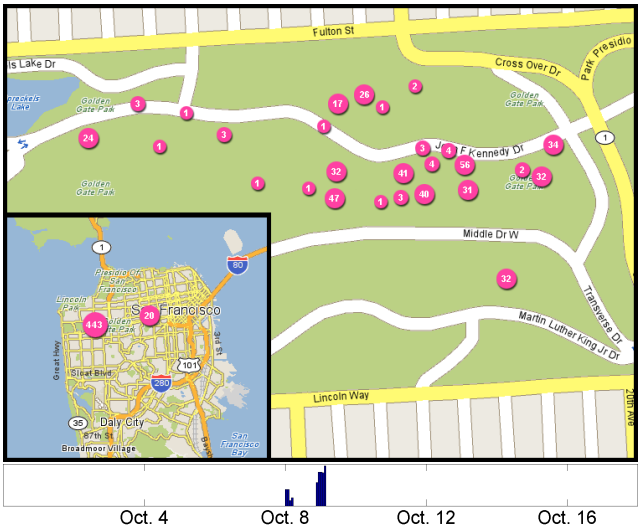

Figure 2: Location (top) and time (bottom) metadata distributions for the tag `Hardly Strictly Bluegrass` in the San Francisco Bay Area.

6. COMBINING VISION WITH TAG-LOCATION ANALYSIS

If we can identify tags that represent places and landmarks, can we apply computer vision techniques to get relevant and diverse set of images for these tags? In previous

[3]http://tagmaps.research.yahoo.com

sections we described how we use tag and location metadata to extract and identify tags that represent landmarks and places. In this section, we assume a given tag x represents a landmark or place, and show how to find a relevant, representative and diverse set of images for that tag. The opportunity here is to improve both image search and visualization in such community media collections. For example, when a user searches for photos of the `Golden Gate Bridge`, our system will be able to detect that the search refers to a landmark; the precision of the result set can be improved, and the set of returned images more diverse and complete.

Current image search in community-annotated datasets is far from satisfactory: issues of precision, recall and diversity in the result set are acute. These issues exist in the Flickr photo dataset, and may persist even when the images have been tagged as being *of* a place or landmark, and geotagged *at* their location. In all these cases, the quality and representativeness of images can be highly variant.

The precision problem in community-contributed datasets is a reflection of the fact that tags associated with any individual image are not guaranteed to be "correct" [2, 3, 11]. For example, previous work [11] reports that, over a set of tagged Flickr photographs for a few hand-selected New York City landmarks, the precision of images with a given landmark tag is alarmingly low (around 50%) – only half of the images tagged with a landmark name are actually images of that landmark. This phenomenon is likely due to the variety of contexts in which tags could be used: it is important to recognize that tags are not always used to annotate the content of the image in the traditional sense [2].

A more subtle precision issue can arise with these personal photographs that are shared online. For example, people often take photographs of themselves or their family members standing in front of a visited landmark. Other users browsing the collection may be more interested in the landmark than in photos of strangers in front of it. Similarly, many users on Flickr, for example, are photo enthusiasts who take a decidedly artistic approach to personal photography. In these cases, photos tagged with a given location or landmark may actually be of the landmark, but are framed in such a manner (such as extreme close-ups) that they are more abstract in appearance and hardly recognizable as the object in question.

The diversity of the retrieved images also poses a problem. Often, the set of retrieved images for a specific landmark or location, even if precise, can be homogenous (e.g., showing many photos from the same view point). Ideally, a search or a selection of photos for a landmark would return multiple views and angles to better cater to the user's specific need.

Recall, of course, is also an issue in community-annotated datasets: there is no guarantee that all images will be retrieved for any search query. Individual images may not be annotated with the "correct" or appropriate tags. In this paper we do not tackle the recall issue directly. We note that our methods can be used for the purpose of "soft annotation", which could ultimately improve recall.

6.1 Problem Definition

We pose the task of finding representative images from a noisy tag-based collections 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. Our focus here is on unsupervised methods, where the statistics 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. The resulting models could also be applied to enhance indexing by suggesting additional tags for images or to refine queries for search.

Formally, our problem involves identifying a tag x as representative of some landmark or geographic feature, and computing a set of photos $\mathbb{R}_x \subseteq \mathbb{P}_x$ that are representative of that landmark. 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). However, we do not handle this case in our current work.

The architecture of the system for finding representative images is shown Figure 3. Given a set of photographs, we first determine a set of tags that are likely to represent landmarks and geographic features, and the geographic areas where these tags are prominent. Then, we extract visual features from the images that correspond to each tag in its respective areas. For each tag, we cluster its associated images based on their visual content to discover varying views of the landmark. The clusters are then ranked according to how well they represent the landmark; images within each cluster are also ranked according to how well they represent the cluster. Finally, the most representative images are sampled from the highest-ranked visual clusters to return a summary of representative views to the user.

6.2 Identifying Tags and Locations

The key idea behind our system is that location metadata comes into play in both finding tags that represent landmarks, and in identifying representative photos. To identify tags that represent landmarks and places we use a variant of the algorithm described in Section 4, looking for tags that mostly occur in a concentrated area and used by a number of users in the same area. These "representative tags" often correspond to place names and landmarks.

As detailed in Section 4, the process involves location-based clustering of the photo set, followed by scoring of the tags that appear in the dataset. We first cluster all photos $\mathbb{P}$ in a certain region, using the photos' location coordinates ℓ_p. For each resulting photo cluster C, we score each tag x in that cluster (see Section 4) to find a set of representative tags for each cluster. This process results in a list of (x, C) pairs – (tag, cluster) pairs indicating tag x to be a geographic feature in the cluster C. An important point is that the same cluster could have multiple associated tags. Similarly, the same tag x can appear in multiple clusters. For example, the tag `Golden Gate Bridge` may appear multiple times in our list, in a different clusters, potentially representing different viewpoints of the bridge.[4] To summarize using another example, if we examine all photos from San Francisco, our system would ideally find tags such as `Golden Gate Bridge`, `Alcatraz` and `Coit Tower` that represent landmarks and geographic features within the city. Furthermore, the system

[4]For further refinement of the resulting list of tags, we can use the algorithms described in Section 5 to retain only tags that have been identifyed as places and not as events.

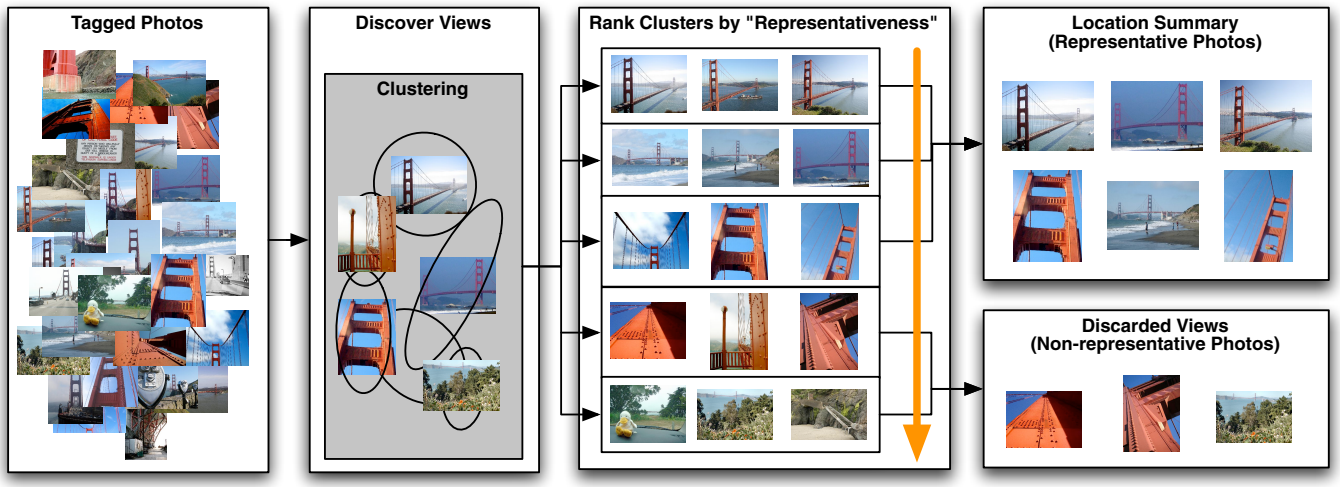

Figure 3: System architecture for choosing representative images from tag/location clusters using visual content.

would identify the different areas where people take photos of each of these landmarks.

More formally, let us assume we identified the tag x as a prominent geographic feature in a set of photo clusters $\mathbb{C}_x \triangleq C_{x,1}, C_{x,2}, ...$ The corresponding set of photos we are interested in is $\mathbb{P}_{x,\mathbb{C}_x} \triangleq \mathbb{P}_x \cap \mathbb{P}_{\mathbb{C}_x}$.

Next, we show how we perform visual analysis on the photos in $\mathbb{P}_{x,\mathbb{C}_x}$.

6.3 Visual Features

This step involves the task of extracting visual features for a tag x. This task is now easier because we hopefully restricted the corresponding set of images for the tag: not only is the tag now likely to represent the same object (e.g., same windmill), but the viewpoints from which the photos were taken are limited – all the photos were taken in a restricted set of geographic clusters. Given the set $\mathbb{P}_{x,\mathbb{C}_x}$, we attempt to select a set of images $\mathbb{R}_{x,\mathbb{C}_x}$ that is most represtative of the visual features for the landmark represented by tag x. We will then select the images from $\mathbb{R}_{x,\mathbb{C}_x}$ to generate our final set of representative images, $\mathbb{R}_x$.

We first describe the visual features extracted from the images in the set $\mathbb{P}_{x,\mathbb{C}_x}$.

We use a range of complementary features to capture the color, texture, and local point characteristics of images. These features are chosen because of their proven effectiveness in a range of recognition tasks for both generic objects and specific scenes. The aspects of the visual content that they capture have also been shown to be complementary.

- **Color**. We use grid color moment features [23] to represent the spatial color distributions in the images. These features are calculated by first segmenting the image into a 5-by-5 grid of non-overlapping blocks. The first three moments of the distributions of the three color channels (in LUV color space) within each block are then calculated and retained as a feature vector for the image. Using three moments for each of three channels in 25 blocks result in a 225-dimensional feature vector.
- **Texture**. The texture of the image is represented with Gabor textures [13]. These features are calculated as the mean and standard deviation of Gabor transformations in 4 scales and 6 orientations over the grayscale image, yielding a 48-dimensional feature vector.
- **Interest Points**. 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. This yields a 128-dimensional SIFT feature vector for each interest point in an image. Typical images in our data set have a few hundred interest points, while some have thousands. SIFT features have received much interest due to their invariance to scale and rotation transforms and their robustness against changes in viewpoint and illumination. The SIFT features have been found to be extremely powerful in a number of vision applications, from generic object recognition [7] to alignment and registration between various photographs of a single scene [22].

We now discuss how we use these visual features to find visual clusters $\mathbb{V}_x$ that represent different views for the landmark x.

6.4 Visual Clustering

Many landmarks and locations can be frequently captured from a number of distinct viewpoints. Famous bridges, like the Golden Gate Bridge or the Brooklyn Bridge, are frequently photographed from a few distinct points, such as the banks on either side of the bridge or while walking across the bridge, looking up at the towers. Similarly, famous architectural sites, like the San Francisco MOMA or the Guggenheim, are frequently photographed from the outside, showing the facade, and from the inside, showing an inner atrium or skylight. In such cases, the selection of a single viewpoint to encapsulate the location may be insufficient. If we are given the chance to present the user with multiple images of the location, repeating images from a single most-representative view may be less productive than showing images from a variety of angles.

Discovering various classes of images from within a set is

a prime application for clustering. We perform clustering using k-means, a standard and straight-forward approach, using the concatenated color and texture feature vectors, described above, to represent the images. SIFT features are not used for clustering due to their high dimensionality, but are later incorporated for ranking clusters and images.

6.4.1 Number of clusters

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 with using only 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 apply an approach where the number of clusters is selected such that the average number of photos in each resulting cluster will be around 20. Future work may investigate the effects of other strategies for selecting the number of clusters and incorporating geographic cues to seed and guide clustering towards finding viewpoints.

6.4.2 Ranking clusters

Given the results of a clustering algorithm, $\mathbb{V}_x$, we would like to rank the visual clusters according to how well the clusters represent the various views associated with a given tag or location. This will allow 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 views, $\mathbb{R}_x$. Lower-ranked clusters can be completely ignored and hidden from the user, since they are presumed to contain less-representative photographs. We determine the ranking of the clusters through a combination of four different scoring mechanisms, designed to ensure the selection of strong, useful clusters.

We use a fusion of the following four cluster scoring methods to generate a final ranking of the clusters:

- **Number of users**. If a large number of photographs from many different users are found to be visually similar, then it is likely that the cluster V is an important view angle for the location. We use the number of users, $|\mathbb{U}_V|$ instead of the number of photos $|\mathbb{P}_V|$ since many photos from a single user may bias the results.
- **Visual coherence**. We measure the intra-cluster distance, or the average distance between photos within the cluster V and all other photos within the cluster, and the inter-cluster distance, or the average distance between photos within the cluster and photos outside of the cluster. We take 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 fairly noisy and may not be visually coherent.
- **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 6.5.3.) If a cluster's photos are linked to many other photos in the same cluster, then the cluster is likely to be representative. Clusters with fewer linked images are less likely to be representative. We count the average number of links per photo in each cluster and use the result to score the clusters.
- **Variability in dates**. We take the standard deviation of the dates that 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 in dates indicates that the photos in the cluster were taken around the same time and that the cluster is probably related to an event, rather than a geographic feature or landmark. We can also use the techniques described in Section 5 to filter those images from $\mathbb{P}_x$ that include tags (other than x) that correspond to events.

To combine these various cluster scores, 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.

To select images from clustered results and present them to the user for browsing, we rank the representative images within each cluster, $\mathbb{P}_V$ using the methods described in the following section, so we have a ranking of the most representative images for each cluster. Then, we sample the highest-ranking images from the clusters. The clusters are not sampled equally, however. The lowest-ranking clusters have no images sampled from them, and the higher-ranking clusters have images sampled proportionally to the score of the cluster.

6.5 Ranking Representative Images

Given the visual clusters, $\mathbb{V}_x$ and their associated rankings, in order to generate a set of representative images, $\mathbb{R}_x$, we further need to rank the images within each cluster, according to how well they represent the cluster. To achieve this, we apply several different types of visual processing over the set of images $\mathbb{P}_V$ to mine the recurrent patterns associated with each visual cluster V.

6.5.1 Low-Level Self-Similarity

Perhaps the most straight-forward approach to discovering and ranking the representative images out of a set is to find the centroid for the set and rank the images according to their distance from the centroid. We start by joining the color and texture features for each image into one long feature vector. We statistically normalize along each dimension of the vector such that each feature has a mean of zero, and unit standard deviation over all images within the set. The centroid is the point determined by the mean of each feature dimension. The images in the set are then ranked according to their Euclidean distance from the centroid.

6.5.2 Low-Level Discriminative Modeling

One shortcoming of the low-level self-similarity method mentioned above is that each example image and each feature dimension is considered to be equally important for centroid discovery and ranking. While this approach can still be quite powerful, recent efforts have suggested that sampling pseudo-negative examples from outside of the initial candidate set and learning light-weight discriminative models can actually greatly improve the performance of image ranking for a number of applications [16, 8, 10]. 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 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 concatenated and normalized feature vector 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 [26, 4] 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.

6.5.3 Point-wise Linking

The above-mentioned low-level self-similarity and discriminative modeling methods do not make use of the SIFT interest point descriptors that we have extracted. The most powerful approach for our application, where we are modeling specific locations to find representative images, is most likely a matching of images of the same real-world structure or scene through the identification of correspondences between interest points in any two given images. Given two images, each with a set of interest points and associated descriptors, we can use a straight-forward approach to discover correspondences between interest points. For each interest point in an image, we can take the Euclidean distance between it and every interest point in the second image. The closest point in the second image is a candidate match for the point if the distance between it and the original interest point is significantly less than the distance between the second-closest point and the original interest point, by some threshold. This matching from points in one image to another is asymmetric, however, so the process can then be repeated, finding candidate matches for each point in the second image through comparison with each point in 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 can take the candidate match as a set of corresponding points between the two images.

Once these correspondences are determined between points in various images in the set, we can establish links between images as coming from the same real-world scene when three or more point correspondences exist between the two images. The result is a graph of connections between images in the candidate set based on the existence of corresponding points between the images. We can then rank the images according to their rank, or 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 can be linked across various images. Non-representative views (such as extreme close-ups or shots primarily of people), on the other hand, will have fewer links across images.

6.5.4 Fusion of Ranking Methods

The ranking methods described above each 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 results. Each method effectively 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.

The end result of this process is a ranked list of clusters, representing different views for each location, from which the most representative images can be sampled to arrive at a set of images summarizing the location. This resulting set of images is our representative set, $\mathbb{R}_x$.

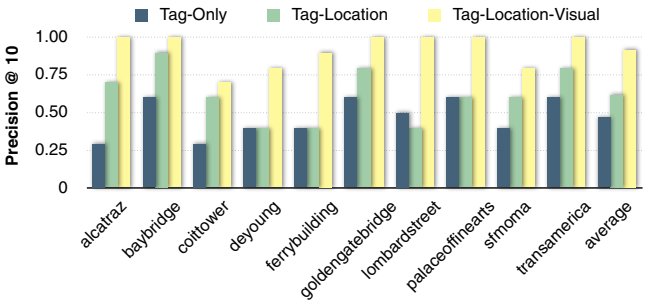

Figure 4: Precision at 10 for representative images selected for locations using various methods.

7. EVALUATION

The goal of the system is to generate a set of representative images for automatically discovered tagged locations. To evaluate the performance, we use a set of over 110,000 geo-referenced photos from the San Francisco area. We discover landmark locations via location-based clustering of the photos, generating 700 clusters (the number was chosen as a tradeoff 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. Tags with scores exceeding a threshold, α, are retained as a tag/location pair, (x, C). For the purposes of this paper, we evaluate the system using only a subset of 10 manually selected landmarks (listed in Figure 4), though, in principle the system could be applied to all of the discovered tag/location pairs. Representative images for each location are extracted using three different techniques:

- **Tag-Only**. In the baseline version of representative image selection, we choose representative images randomly from the set of all images with the corresponding tag.
- **Tag-Location**. In this second baseline, we choose representative images for a tag randomly from all images that are labeled with the tag *and* fall within a location cluster where the tag is found to be representative.
- **Tag-Location-Visual**. This is the approach detailed in the previous section.

We use each of the three above-described methods to extract ten representative images for each of the ten landmarks and evaluate the results in terms of precision at 10 (P@10). This metric measures the percentage of the top-ten selected images that are indeed representative of the land-

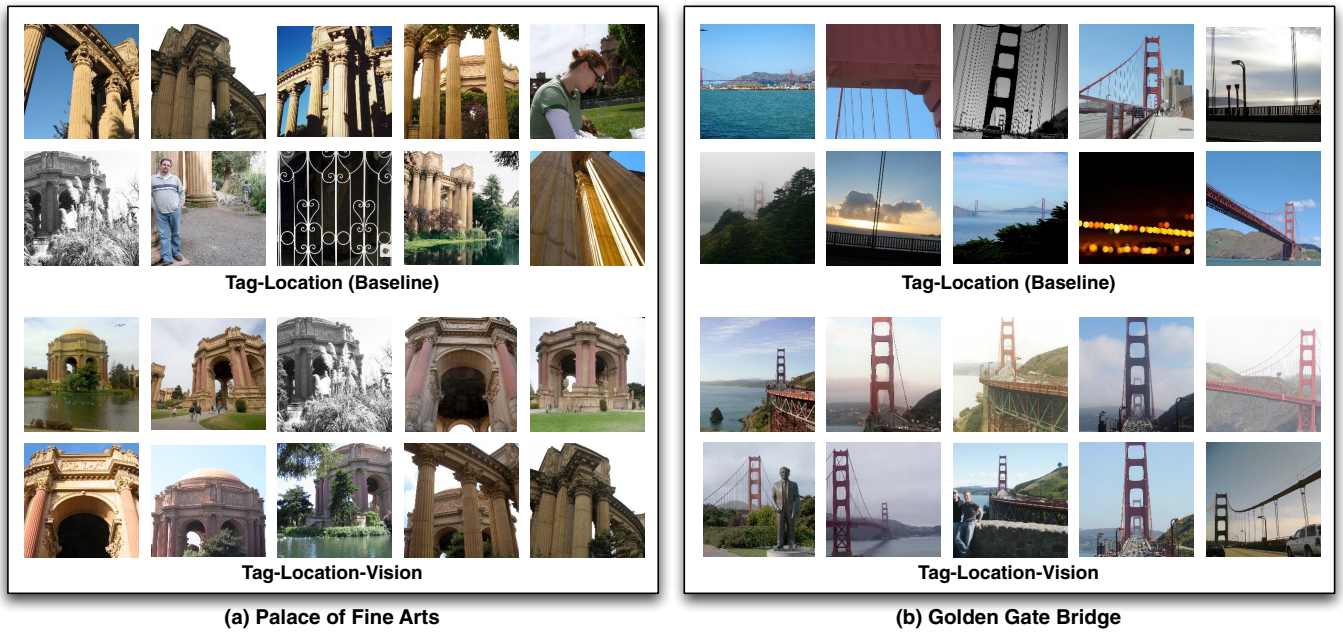

Figure 5: Comparison of recommended representative images resulting from the tag-location filtering and Fixed-size clustering approaches for the Palace of Fine Arts and the Golden Gate Bridge.

mark. The ground-truth judgments of image representativeness are defined manually by human evaluators. If 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 4.

The results show a clear added benefit of location-based constraints for the selection of representative images. In the baseline case, the tag-only approach, the P@10 is slightly less than 0.5, on average. This finding confirms many recent observations about the accuracy of tags. Constraining the selection of representative images to come only from images associated with a tag-location pair (as in the tag-location approach) yields a 30% increase in the precision of the selected images, which indicates that location information can help refine the representativeness of a set of related images.

More striking, though, is the improvement we get with the technique that utilizes the visual analysis (tag-location-visual). On average, across all of the landmarks, there is a clear relative increase in precision of more than 45% gained over the tag-location baseline by adding visual features. Across most of the locations, the visual-based selection of representative images gives perfect P@10 score (all top-ten ranked images are representative). A comparison of the summaries from the best non-visual technique (tag-location) and the visual technique (tag-location-visual) for two sample tags is shown in Figure 5.

For some geographic features, the visual-based methods still do not provide perfect precision in the summaries. 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 city skyline taken from the observation deck at the top of the tower. Similarly, for museums, such as `De Young` and `SF MOMA`, the representative views are defined to be outside views of the building and recognizable internal architectural aspects; however, users also like to photograph particular artworks and other non-representative views while at museums. The trend across these failure 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 desireable for representation of the location. 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?

A related issue is the fact that precision does not capture all of the aspects that could impact the perceived quality of a set of representative images. For example, the notion of representativeness is not binary. If we compare the top-left images for each set of results in Figure 5a, it is fair to argue that the visual-based result, which shows the entire structure, is more representative than the tag-location result, which shows only a close-up. The precision metric does not capture this aspect, since both images are technically considered to be representative. Also, repetition in a set of images can impact the perceived quality of the summary. Repeated, nearly-identical images will not convey additional information to the viewer, so it may be preferable to display images from diverse views. These issues of relative quality of representative images and the diversity of results can be evaluated with human subjects. We leave such an evaluation for future work.

8. CONCLUSIONS

We have shown how community-contributed collections of

photographs can be mined to successfully extract practical knowledge about the world. We have seen how geographical labels and tagging patterns can lead us to summaries of important locations and events. We further introduce the use of visual analysis in a controlled manner, using the knowledge extracted from tags and locations to constrain the visual recognition problem into a more feasible task. We have shown that the use of visual analysis can increase the precision of automatically-generated summaries of representative views of locations by more than 45% over approaches in the absence of visual content. All of these benefits are observed despite the fact that the user-provided labels in such collections are highly noisy.

In future work, we plan to further explore the impact of visual content on retrieval and summarization of georeferenced photographs. In particular, we will investigate the perceived quality of a wider variety of approaches to discovering and presenting related views for a single landmark. Beyond that, we will examine whether visual analysis can help in the discovery of meaningful locations and tags, perhaps by eliminating geographical clusters that are too visually diverse to be a single landmark, or by using visual diversity as part of the criteria used in distinguishing between landmark- and event-oriented tags. We will also explore automatically tagging photographs or suggesting tags to the user based on the visual content of the image, a difficult task to perform based on visual content alone, but one that can be simplified with contextual and geographical cues.

9. REFERENCES

[1] 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*. ACM Press, June 2007.

[2] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human Factors in computing systems*, New York, NY, USA, 2007. ACM Press.

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

[4] 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.

[5] 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.

[6] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 193–202, New York, NY, USA, 2006. ACM Press.

[7] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. *Proc. ECCV*, pages 242–256, 2004.

[8] W. Hsu, L. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *ACM Multimedia*, Santa Babara, CA, USA, 2006.

[9] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 89–98, New York, NY, USA, 2006. ACM Press.

[10] L. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Conference on Image and Video Retrieval*, Amsterdam, 2007.

[11] 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.

[12] D. 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] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries*, 2004.

[16] 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 annual ACM international conference on Multimedia*, pages 598–607, 2005.

[17] 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.

[18] A. Pigeau and M. Gelgon. Organizing a personal image collection with statistical model-based ICL clustering on spatio-temporal camera phone meta-data. *Journal of Visual Communication and Image Representation*, 15(3):425–445, September 2004.

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

[20] R. Sarvas, E. Herrarte, A. Wilhelm, and M. Davis. Metadata creation system for mobile images. In *Proceedings of the 2nd international conference on Mobile systems, applications, and services*, pages 36–48. ACM Press, 2004.

[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] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics (TOG)*, 25(3):835–846, 2006.

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

[24] 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.

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

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

[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.