

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 use 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 representative images for each such view.

^{*}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.

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 ef-

forts ([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 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 Air-**

¹Theoretically 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.

port 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

