NOT ALL TAGS ARE CREATED EQUAL: LEARNING FLICKR TAG SEMANTICS FOR GLOBAL ANNOTATION

Emily Moxley, Jim Kleban, Jiejun Xu, and B. S. Manjunath

University of California Santa Barbara
ECE/CS Departments
Santa Barbara, CA 93101
{emoxley,kleban,manj}@ece.ucsb.edu, jiejun@cs.ucsb.edu

ABSTRACT

Large collaborative datasets offer the challenging opportunity of creating systems capable of extracting knowledge in the presence of noisy data. In this work we explore the ability to automatically learn tag semantics by mining a global geo-referenced image collection crawled from Flickr with the aim of improving an automatic annotation system. We are able to categorize sets of tags as places, landmarks, and visual descriptors. By organizing our dataset of more than 1.69 million images using a quadtree we can efficiently find geographic areas with sufficient density to provide useful results for place and landmark extraction. Precision-recall curves for our techniques compared with previous existing work used to identify place tags and manual groundtruth landmark annotation show the merit of our methods applied on a world scale.

Index Terms— Image Annotation, Data Mining, Knowledge Discovery

1. INTRODUCTION

Prominent online collaborative communities, like Wikipedia and Flickr, aggregate individual contributions into massive, unorganized datastores with extraordinary research potential. In these communities unstructured annotation of objects, known as tagging, is an important data source to be analyzed along with conventional metadata. Tagging discards the inflexibility inherent in structured labeling but at a cost. The freeform nature of tags brings new challenges as to how to extract a useful signal, or harvest knowledge, in the presence of noise in the labels and semantic uncertainty.

A better understanding of tag semantics would benefit many information-based applications. In this work we explore tag usage in georeferenced photo collections containing images with an associated world coordinate. Such knowledge could benefit applications including automatic extraction of visual examples of events/landmarks [1], [2] and tag-driven image annotation [3]. Analysis by mining a large dataset of photographs for time, location, co-occurrence, and visual information over multiple geographical scales provides valuable knowledge about how tags are applied.

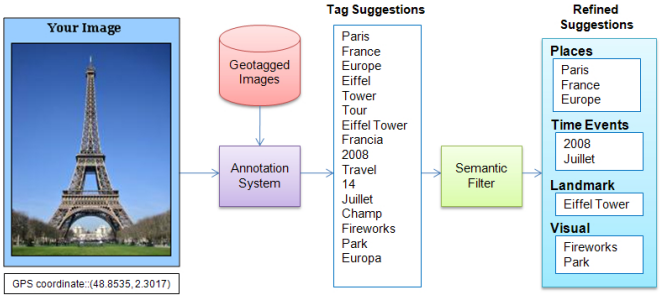

Fig. 1: Knowledge of tag semantics allows for better annotation. Appropriate features can be applied by learning how a tag is employed. For instance, if a tag is deemed a landmark, we can use visual features to determine if it is appropriate. If it is a place, a geotag alone might be the most effective.

Figure 1 shows how learned tag semantics can be employed in an annotation system. Tag suggestions are calculated on a previously unlabeled input image from visual features and world location. However, these tags are noisy as many refer to the same object (the Eiffel Tower) and others can be pruned by examining time or location metadata alone. A semantic post-filtering step which categorizes tags into place, timed event, landmark and visual description provides a cleaner suggestion list to a user or retrieval system.

This work will expand on existing methods for automatically extracting tag semantics and report results on a worldwide database of 1.7 million images downloaded via the Flickr API. The paper is organized as follows. The next section outlines related work. Section 3 provides detail about our dataset and its efficient representation employing a quadtree. Section 4 describes methods for extracting semantics and presents effectiveness results. The last section concludes.

This work was supported by NSF IGERT Grant #DGE-0221713.

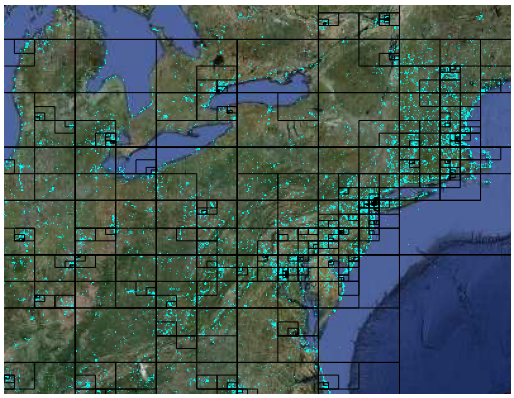

Fig. 2: A quadtree efficiently indexes the geotagged image distribution for efficient search over the 1.7 million images. The tree's node boundaries are the black lines mapped over the light blue dots representing image locations for an area in the eastern United States. Smallest rectangles represent the lowest-level nodes, and are found at dense, primarily metropolitan, areas.

2. RELATED WORK

Exploring the purpose and use of tagging has been an area of active research [4], [5]. Golder and Huberman outline seven functions tags can serve for bookmarking which are a superset of functions served by specifying location and information content in an object. Attempts have been made to extract knowledge on tags through secondary web repositories. Overell et. al. [6] develop a classifier that employs Wikipedia to expand semantic categorization via the WordNet ontology, but they do not consider tag usage statistics and image data.

More similar to our work, Rattenbury and co-authors [7] attempt to automatically extract tag semantics using Flickr data. They establish an entropy-based technique for automatically identifying place and timed event tags and analyze results for a set of roughly 50,000 images with 803 unique tags in the San Francisco Bay Area. As the number of available geotagged photos has increased considerably to 100 million (as of February 2009), a global analysis is now possible. We will extend automatic tag semantic extraction by a) analyzing results over a large worldwide database, b) considering the addition of co-occurrence information, c) identifying visually descriptive tags via mutual information and d) identifying sets of tags which correspond to landmarks.

3. DATA COLLECTION AND REPRESENTATION

3.1. Data Crawl and Feature Extraction

In order to explore the aspects of tag semantic extraction over a world-encompassing dataset we crawled metadata for 1.7 million georeferenced images via the Flickr API on the tag list Hays and Efros selected for their IM2GPS work [8]. For each image we retrieve the owner id, unique Flickr id, time taken, time of upload, title, tags, latitude, longitude, and geotag accuracy as given by zoom level. Our dataset has 65,679 unique users and 436,506 unique tags. The high number of photos per user compared to the Flickr average likely results from the rejection of photos during the crawl with tags indicating personal use. There are 29,652 tags which are employed at least 25 times and by more than one user.

For visual analysis we employ two common descriptors, GIST [9] and SIFT signature [10]:

Gist: The **GIST** descriptor describes the spatial layout of an image using global features derived from the spatial envelope of an image. It is particularly powerful in scene categorization. The final descriptor is 512-dimensional.

Sift Signature: The **SIFT** feature represents descriptors [11] extracted at 5000 random keypoints and pushed through a vocabulary tree with 4 levels and a branching factor of 10.

3.2. Quadtree

A quadtree is a data structure which provides an efficient hierarchical manner to store images adaptively according to their distribution over location. A quadtree is formed by recursively dividing data into four regions until a stopping condition is met. We specified a minimum-support level as the stopping condition: if a node contains fewer than 100 images with unique (owner id, latitude, longitude) triples, subdivision stops. Each of the leaf nodes, then, represents a space that is inversely proportional to the density of photos taken in that area. Figure 2 shows image density and the quadtree overlaid on a map of the eastern United States. Denser regions, like New York City, have deeper nodes than sparse areas. The quadtree allows us to quickly identify dense regions where we can generate tag semantic scores with higher confidence.

4. TAG SEMANTIC IDENTIFICATION

We consider three categories which may be detectable using the signal present in the photo collection data. These are places, visual descriptors, and landmarks.

4.1. Place Extraction

The authors in [7] introduce Scale-structure Identification (SSI) for identifying tags associated with places and events. For each tag they consider the entropy over connected subcomponents on a graph with vertices as photos labeled with that tag. A connection criteria, a maximum distance for which an edge appears, controls the scale. Tags with a tight distribution on a location (or time in the case of events) will generate large clusters over multiple scales. A decision variable, λ, for tag t is computed as a sum of entropy over multiple scales:

$$\lambda_t = \sum_{k=1}^{K} \sum_{Y \in \Psi_{r_k,t}} -\frac{|Y|}{|N_t|} \log \frac{|Y|}{|N_t|} \quad (1)$$

where there are *K* scales, Y is a set of photos connected by distance r_k containing tag t, $\Psi_{r_k,t}$ is the collection of connected component sets Y for t and r_k, and N_t is the set of photos with tag *t*. For event detection, it would be easy to additionally consider periodic time events (e.g. "august") by examining the structure of clusters and appropriately recomputing the graph on modulo time as is done in [7].

We employ a similar idea to SSI, using the quadtree for scale space definition. Each level of the quadtree represents a scale space. We calculate the entropy for each level of the quadtree, as in Equation 1. However, a cluster space is considered *empty* unless the node contains 100 images. Furthermore, we explore the influence of using co-occurrence information. Many place names co-occur with other place names, such as the same place in a different language or a less-specific location (e.g. "Santa Barbara" with "California"). We hope to use this information to influence our decision on a tag. Using the Jaccard coefficient on the image sets containing compared tags (size of the intersection divided by the size of the union), we can measure a normalized co-occurrence of tags and make a prediction using a weighted sum of the scores of tags above a certain threshold of Jaccard similarity: $\lambda_j^{'} = w_0\lambda_j + \frac{1}{w_0+\sum_i J(i,j)} \sum_i J(i,j)\lambda_{t_i}$. Figure 3 shows the results of using a quadtree and incorporating co-occurrence for place identification over 5 scales ranging from 1.1 to 11,100 kilometers. Better initial precision is observed, but when co-occurrence fails (that is, tag t lacks indicative co-occurrent terms), we see lower precision as seen at areas of higher recall.

4.2. Visual Term Extraction

Certain tags refer to qualities which are visually identifiable from the photo's content like "sunset," "sky," and "beach." It is estimated, from a random sampling of the tags contained in our crawled database that were then manually analyzed, that on average however less than 30% of tags belong to this category, identifiable from photo content. These are the tags that could be suggested from analysis of visual features. To find tags which represent visual terms we consider the mutual information between a visual feature random variable *x* and a tag variable *t*. To generate *x*, we discretize the *d*-dimensional computed image feature via K-means clustering. The mutual information, *MI*, is estimated pointwise as:

$$MI(t,x) = \sum_{t,\bar{t}} \sum_{\forall k} p(t,x_k) \log \left(\frac{p(t,x_k)}{p(t)p(x_k)} \right) \quad (2)$$

$p(x)$ is the cluster prior, $p(t)$ is the tag prior, $p(\bar{t})$ = 1 - $p(t)$, and $p(t,x_k)$ is counted as the number of photos with tag *t* in cluster *k* divided by the total number of photos.

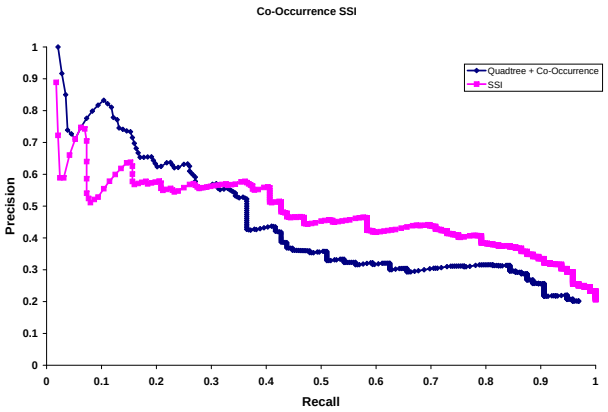

Fig. 3: Graph showing precision/recall for place identification of Flickr tags. SSI approach does not exploit co-occurrence information and does not use a quadtree for scale-space definition. Co-occurrence information provides better initial precision, when co-occurrent terms are indicative. However, when co-occurrence fails, it reduces precision, as evidenced by lower precision at high recall.

We select the GIST feature with K=950 clusters to estimate $p(x)$. GIST's success with scene categorization tasks is pertinent for finding tags that are generally descriptive. Table 1 shows a list of tags with high mutual information discovered in the dataset. Of the top 100 scores for $MI(t,x)$, 57% were judged visually relevant, indicating the success of mutual information as a measure of visual tags.

Tag	MI	Tag	MI
sunset	0.0088	(C)	2.7402e-06
clouds	0.0073	mashup	3.9652e-06
flowers	0.0071	work trip	5.2887e-06
sky	0.0070	sounds	5.935e-06
beach	0.0063	psychiatry	6.0322e-06
nature	0.0053	SFW	5.9390e-06

Table 1: Tags with high (expected to be visually relevant) and low (not expected to be visually relevant) mutual information with visual features.

4.3. Landmark Detection

The strongest evidence for visually descriptive tags is when they occur exclusively on photos taken of the same geographically fixed object. Along these lines Kennedy and Naaman [1] were able to identify representative results for landmarks via clustering, and Quack, Leibe, and Van Gool [2] use matching on Wikipedia images for improved accuracy.

To detect tags used to describe landmarks, we first employ agglomerative clustering on image sets consisting of the members of dense nodes in the worldwide quadtree. Dis-

Fig. 4: Examples of detected landmarks. Tags from images in a cluster generate a name estimate from a list of georeferenced Wikipedia articles. Stricter clustering yields better naming results, as evidenced by the incorrect guess in the last row.

tances between images are computed using L1 distance on the SIFT signature, and images join a cluster when the members are within a distance threshold σ using complete linkage. For each set, clusters exceeding a minimum membership of 5 images are considered possible landmarks. Since the dataset consists of many similar images taken by the same user, we limit membership to one image per user.

To extract the proper name of the landmark, we query a database [12] which lists georeferenced Wikipedia entries. A matching score is calculated for each landmark name in the database within 0.01° latitude and longitude of the median coordinates of cluster members. The score is generated by considering the string similarity between the landmark name and the set of tags and titles for images in the cluster. In an annotation system, if an image is matched to a representative cluster, the landmark title can substitute for similar tags.

By searching breadthwise on the six deepest levels on the quadtree we were able to automatically extract views for 62 identified landmarks. Figure 4 shows an example of the some of the images we were able to correctly match to a proper name. Table 2 shows how the minimum distance threshold and node coarseness effect the results of landmark identification for our data.

5. CONCLUSION

In this paper we have used the signal present in collaborative communities in order to extract knowledge about tag usage. By performing this task automatically we maintain the freeform flexibility inherent to the tagging process. Additionally, the efficient representation on a quadtree provides us with a data-driven confidence metric as to our certainty about the assessment. We have expanded previous work to also consider visually descriptive tags using mutual information, and to identify sets of tags referring to landmarks. In future work,

d	Coherence	Precision	NAcc	Num/Uniques
14	0.86	0.80	0.63	146/51
15	0.90	0.90	0.61	96/36
16	0.90	0.93	0.81	32/11
17	0.86	1.0	1.0	11/5
18	1.0	1.0	1.0	1/1

Table 2: Landmark extraction as a function of quadtree depth d for a fixed clustering threshold σ. Coherence is the fraction of images in a cluster belonging to the correct landmark. Precision is the fraction of clusters which are landmarks. Naming accuracy (NAcc) is the fraction of clusters correctly named. Num/Uniques is the number of landmark clusters found and number of unique landmarks found.

we will consider the interesting problem of learning object tags via detection and study the benefit of tag categorization in a global annotation system.

References

[1] L. Kennedy and M. Naaman, "Generating Diverse and Representative Image Search Results for Landmarks," in *WWW*, 2008.

[2] T. Quack, B. Leibe, and L. Van Gool, "World-scale Mining of Objects and Events from Community Photo Collections," in *CIVR*, New York, NY, USA, 2008.

[3] E. Moxley, J. Kleban, and B. S. Manjunath, "SpiritTagger: A Geo-Aware Tag Suggestion Tool Mined from Flickr," in *Multimedia Information Retrieval*, 2008.

[4] S. Golder and B. Huberman, "The Structure of Collaborative Tagging Systems," 2005.

[5] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read," in *HYPERTEXT*, 2006.

[6] S. Overell, B. Sigurbjornsson, and R. van Zwol, "Classifying Tags using Open Content Resources," in *WSDM*, 2009.

[7] T. Rattenbury, N. Good, and M. Naaman, "Towards Automatic Extraction of Event and Place Semantics from Flickr Tags," in *SIGIR*, 2007.

[8] J. Hays and A. Efros, "IM2GPS: Estimating Geographic Information from a Single Image," in *CVPR*, June 2008.

[9] A. Oliva and A. Torralba, "Building the Gist of a Scene: The Role of Global Image Features in Recognition," in *Progress in Brain Research*, 2006, p. 2006.

[10] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree," in *CVPR*, 2006.

[11] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, 2004.

[12] WikiProjekt Georeferenzierung, "`http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/%Wikipedia-World/en`," Website.