Global Annotation on Georeferenced Photographs

Jim Kleban
Vision Research Lab
Harold Frank Hall
Santa Barbara, CA 93106
kleban@ece.ucsb.edu

Emily Moxley
Vision Research Lab
Harold Frank Hall
Santa Barbara, CA 93106
emoxley@ece.ucsb.edu

Jiejun Xu
Vision Research Lab
Harold Frank Hall
Santa Barbara, CA 93106
jiejun@cs.ucsb.edu

B.S. Manjunath
Vision Research Lab
Harold Frank Hall
Santa Barbara, CA 93106
manj@ece.ucsb.edu

ABSTRACT

We present an efficient world-scale system for providing automatic annotation on collections of geo-referenced photos. As a user uploads a photograph a place of origin is estimated from visual features which the user can refine. Once the correct location is provided, tags are suggested based on geographic and image similarity retrieved from a large database of 1.2 million images crawled from Flickr. The system effectively mines geographically relevant terms and ranks potential suggestion terms by their posterior probability given observed visual and geocoordinate features. A series of experiments analyzes the geocoordinate prediction accuracy and precision-recall metric of tags suggestions based on information retrieval techniques. The system is novel in that it fuses geographic and visual information to provide annotations for uploaded photographs taken anywhere in the world in a matter of seconds.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Design, Experimentation, Theory

1. INTRODUCTION

The problem of image annotation has received significant attention in recent years. With the coming pervasiveness of GPS-enabled camera devices, further metadata to inform annotation decisions is becoming available. How to effectively utilize this information on a world-wide scale given computation time constraints has yet to be demonstrated. Freely-offered community image repositories, such as Flickr and Picasa, offer a ripe base of knowledge. As of February 2009 there were more than 100 million geotagged images in Flickr. Given such a large amount of data it is tempting to determine the extent to which signal can be mined given the often noisy annotations present in the form of tags. This paper presents a system which effectively suggests tags using geocoordinates to inform the annotation decision process.

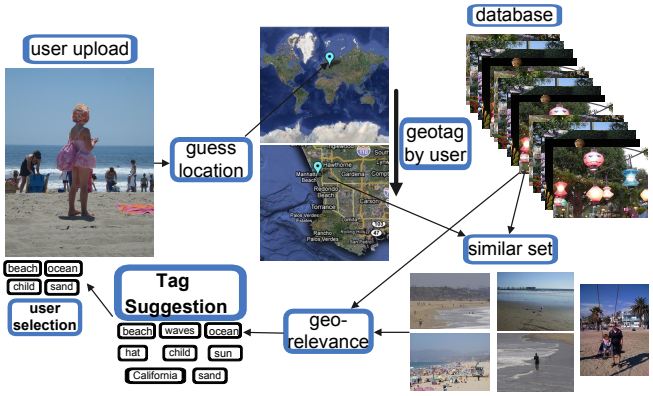

Figure 1: System diagram. First, user uploads a photo. A geotag is predicted using only the visual content of the photo. Then the user selects the actual location for the geotag. A similar set of images is identified that are close to the visual content and geotag of the upload. The system mines the similar set of photos for geographically relevant annotations and suggests them to the user. The user can then select those which it wants to apply to the photo.

A system flow diagram is shown in Figure 1. The user can either upload a photo or supply a URL for a web image. Using quickly extracted visual features, a geolocation estimate is made for the photo to aid the user in placing it on a map. The user adjusts a map provided by the Google API centered at the system's location estimate in order to correctly place the origin. Once the user has done this, the system offers relevant annotations predicted by ranking estimated posterior probabilities derived from the geocoordinate and visual features of the 1.2 million global images in the database. Finally, the user can choose tags so the annotated image can be added to a collection.

The paper will proceed by reviewing relevant work in an-

CIVR '09, July 8-10, 2009 Santorini, GR

notation and online media communities in the next section. Section 3 provides details about how the data was crawled and features were extracted. The algorithm for predicting geocoordinates is in Section 4, and the the rubric for annotating a geotagged photo follows in Section 5. Experiments on these algorithms and conclusions wrap up the paper in Sections 6 and 7.

2. RELATED WORK

This paper addresses the problem of multimedia annotation of images driven by leveraging online community-based data. Therefore, principles from research on collaborative tagging and social media sites, such as YouTube, Flickr, and other multimedia repositories, motivate our work.

While extensive research has been done on the annotation problem [1], [2], most computer vision systems employ descriptors derived from image content which treat annotations as a detection problem. The usefulness of such methods can be limited by difficulties presented by the cluttered natural scenes often found in tourist photos. Other methods have approached the image analysis problem by working with large sets of photos and considering geographical metadata. Notable works have included event and place identification [3], extraction of canonical landmark views [4], and geocoordinate prediction from a singular image [5].

Directly relevant to our aim to use geo-references to infer labels, Joshi and Luo have presented work that quantifies the probability of a particular activity or event that have a relevant geographic footprint (i.e., events such as "hiking" but not "party") by learning the likelihood of the event conditional on geotags, text tags, and visual features [6]. Their experiments on a Flickr dataset, from which noisy data has been manually removed, show that fusion of geographic and visual information can improve results for classification of geotagged photos *in some cases*.

In contrast to typical annotation systems based on computer vision models, our system does not rely on learning particular vocabulary terms. It is therefore scalable to any dictionary size. It is also the only work to our knowledge which specifically addresses the problem of efficiently tagging images on a worldwide scale. The technique of applying tag propagation by mining large datasets to annotate is related to work in image annotation by mining [7] as well as with videos [8].

Previous work by the authors serves as a proof of concept for such a system [9]. In that work, annotations are derived from image similarities constrained to a geographic radius, and a comparison of the local frequency of terms to their global frequency is used to weigh terms that occur frequently in a local area. The local area in [9] is defined as a box bounded by a set of handpicked geographic coordinates, and the system is only tested for two general areas, "Los Angeles" and "Southern California." This paper reformulates the same problem on a worldwide scale and explores the effect of dataset density on the results. The contributions of this paper are fourfold. This work:

- extends the concept of georelevant term propagation to a worldwide database.
- offers a method for choosing smartly which annotations of geotagged photos have visual relevances and for effectively combining them with geography-based annotations.
- formulates the annotation decisions in a Bayesian framework of maximizing posterior annotation probabilities given a geographic and visual feature space.
- analyzes the ability of basic features to predict geolocation using the method in [5] with real-time system constraints.

3. DATA CRAWL AND FEATURE EXTRACTION

In order to learn and test an annotation system for geotagged photos, we first crawled 1.75 million georeferenced images using the Flickr API and the methodology from [5] covering the globe. Of the 1.75 million images, we were able to retain and extract features for 1.2 million images found to be of suitable resolution and aspect ratio. Additionally, for each image we retrieve the following metadata: owner id, Flickr id, time taken, time of upload, title, tags, latitude, longitude, geotag accuracy as given by zoom level of map when geographically annotated, and public license information.

We employ the following five types of visual features extracted from each photo:

Edge Histogram Descriptor: The **EHD** is an 80-dimensional feature consisting of histograms of gradient orientations computed from the image tiled in a 4x4 grid, as described in the MPEG-7 standard [10]. Each histogram contains 5 bins and consists of the magnitude response of a filter.

Homogeneous Texture Descriptor: The **HTD** feature captures the statistics (mean, variance) computed across the image from the response of a bank of 24 oriented Gabor filters [10]. The resulting descriptor has 48 dimensions.

Color Layout Descriptor: The **CLD** is characterized by an 18-dimensional descriptor, consisting of three 6-dimensional coefficients from the DCT of each color channel in YCbCr space [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 [11].

Sift Signature: The **SIFT** feature represents the SIFT descriptors [12] extracted at 5000 random keypoints [13] and pushed through a vocabulary tree with 4 levels and a branching factor of 10, as advanced by Nister and Stewenius [14].

The MPEG-7 descriptors (EHD, HTD, CLD) are extracted using slightly altered code available from [10]. We employ a C implementation of the code provided by Torralba [15] for GIST extraction, and a modification of the code provided by Vedaldi [16] for SIFT signature extraction.

4. ESTIMATING GEOLOCATION

In order to provide geographically relevant tag suggestions, we first need to know the location at which a photograph was taken. Although some photographs contain this information embedded as metadata in EXIF format garnered via GPS device or cell tower triangulation, this information is still not available in the majority of cases. Since a user can rarely be expected to know the latitude and longitude coordinates directly, we provide a map interface from Google to allow placing of the photo. Many of the georeferenced images in Flickr were placed using a similar map

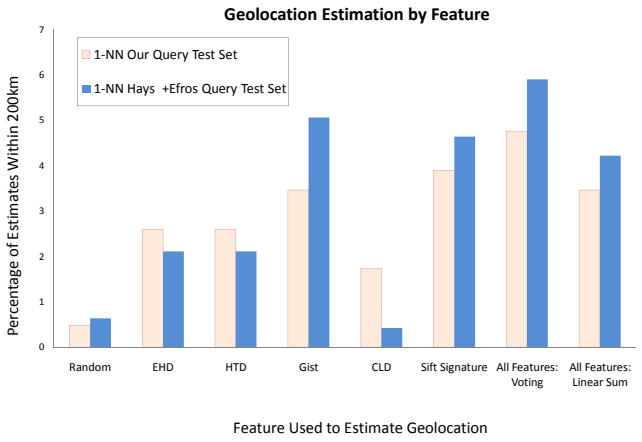

Figure 2: Performance of geolocation estimation from a single image using the different features described in Section 3. Note that using only GIST performs nearly as well as using a voting method to combine all features.

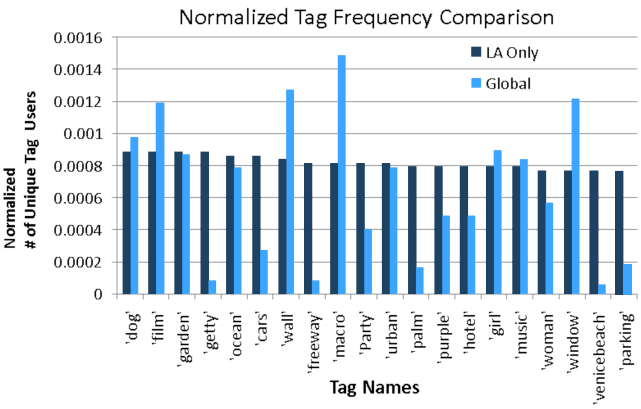

Figure 3: Twenty ordered tags shown to demonstrate tag frequency differences between Los Angeles region and globally. Tags such as "getty," "cars," "freeway," and "palm" with a higher normalized frequency in Los Angeles are more applicable to the region.

interface. As an initial estimate of where the photograph was taken we employ a simple nearest neighbor scheme to infer a location as detailed in the IM2GPS work by Hays and Efros [5]. Essentially, a geographic estimate is made by using the coordinates associated with the nearest neighbor in feature space over the 1.2 million crawled images.

Figure 2 shows the results for first nearest neighbor run over our 1.2 million images on the query set provided in [5] and from a set of 230 random images selected in a similar manner. Photos by the same user were removed from the dataset. The figure shows estimation results using the five employed features separately, and two methods for fusing features: linear sum on normalized feature similarities and voting. Feature similarity was computed using L1 distance for all cases. For each separate feature, the mean and standard deviation of the distribution of distances between 25,000 random images was calculated in order to normalize the feature distances to a standard Gaussian. The linear sum score was then computed on the normalized feature distances. A Borda voting scheme ranked candidate images in the top 120 nearest neighbors for each feature as $\sum_f 120 - r(f) + 1$ where r is the rank of f.

Out of the features used, it is observed that the SIFT signature and the GIST are the most effective. It is not surprising to see SIFT perform well for this retrieval task, although it is relatively expensive to compute. The color layout descriptor (CLD) proved surprisingly ineffective when compared with previous reported success with color histograms. This is perhaps due to its reliance on positioning of colors within the image. Our best method, a voting scheme to combine features, produced an estimate within 200 kilometers 5.9% of the time. The authors [5] report approximately 9% performance for a dataset with 1.2 million images. We believe the discrepancy comes from using a set of features not as effective overall for this task.

Since GIST performs well and has a modest computation cost (on the order of 1 second) we select this feature for world-scale geolocation estimation. Exact nearest neighbor estimation would require a linear scan over 1.2 million images, and since this is too time consuming we explore the effects of partitioning the dataset using K-means clustering in Section 6.1.

5. IMAGE ANNOTATION

We note that the vocabulary used to describe photos is biased by the geographic region associated with the photo. To motivate this assumption, Figure 3 compares unique tag user frequencies for a set of tags crawled from Flickr in an area restricted to Los Angeles and a set of tags crawled globally. Examples of tags with high frequency in LA but lower global frequency are words such as "cars," "freeway," and "palm." We use the premise of a geographic bias in the tag distribution as a method for reranking tag suggestions in a way that reflects the local spirit of a place and improves the relevancy of top annotations. Suggestions may be especially useful for tourists as a quick way to annotate their vacation photos with distinctive labels. While previous work utilized tag distributions in a geographic area in order to find representative tags for visualization and knowledge extraction [17], we extend the idea by framing the problem in terms of optimal Bayesian Maximum A Posteriori (MAP) probability estimation for a set a tag candidates.

The image is described by a set of visual feature primitives, described in Section 3, which we will call $\mathbf{x}$, and geographic information $\mathbf{g}$ indicating the location that the photo was taken. Thus, for each tag b we can derive a probability that the tag b is applicable to the image as

$$p(b|\mathbf{x},\mathbf{g}) = \frac{p(b,\mathbf{x},\mathbf{g})}{p(\mathbf{x},\mathbf{g})} = \frac{p(\mathbf{x},\mathbf{g}|b)p(b)}{p(\mathbf{x},\mathbf{g})} \quad (1)$$

Several methods exist for calculating the posterior $p(b|\mathbf{x},\mathbf{g})$. We prefer non-parametric techniques which extend flexibility by allowing us to avoid expensive model calculations. Density estimation using k-nearest neighbors allows a direct calculation of the posterior using a fixed number of closest data points rather than by searching over a fixed volume [18]. We adopt kNN density estimation to calculate the posterior imposing a kernel on each candidate within the search space. This method is described in the next section.

5.1 Non-Parametric kNN Density Estimation

Non-parametric density estimation using k-Nearest Neighbors (kNN) provides a way to estimate the posterior $p(b|\mathbf{x},\mathbf{g})$ and has been used in segmentation [18], video motion classification [19], and object classification in images [20]. We can reformulate the posterior as:

$$p(b|\mathbf{x},\mathbf{g}) = \frac{p(b,\mathbf{x},\mathbf{g})}{p(\mathbf{x},\mathbf{g})} = \frac{p(b,\mathbf{x},\mathbf{g})}{p(\mathbf{x},\mathbf{g}|b) + p(\mathbf{x},\mathbf{g}|\bar{b})} \quad (2)$$

Simple kNN algorithms treat the k-nearest neighbors identically and derive a probability from the number of kNN that are of class b, for instance, $p(b|\mathbf{x},\mathbf{g}) = \frac{\sum_{i=1}^{k}\mathbf{I}_b(X_i)}{k}$ where X_i is the i^{th} closest image to $(\mathbf{x},\mathbf{g})$ and $\mathbf{I}_b(X_i)$ is an indicator function denoting whether image X_i is of class b. More sophisticated algorithms apply a penalty, parameterized by cost function K, for the k^{th}-nearest neighbor's distance from the given parameters $\mathbf{x}$ and $\mathbf{g}$, leading to a formulation

$$p(b|\mathbf{x},\mathbf{g}) = \frac{\sum_{i=1}^{k}\mathbf{I}_b(X_i)K(x,X_i)}{\sum_{i=1}^{k}K(x,X_i)} \quad (3)$$

$K(x,X_i)$ is often formulated as $K(\frac{x-X_i}{h})$ where h is a bandwidth or smoothing parameter.

Our algorithm first employs a rectangular window, $K_{\mathbf{g}} = \mathbf{I}_{\hat{\mathbf{g}}}(X_i)$, whose size is a function of dataset density, around $\mathbf{g}$ indicating whether X_i is within a certain distance of $\mathbf{g}$. The geographic region of influence, $\hat{\mathbf{g}}$, is determined by the quadtree described in Section 5.2. A Gaussian with mean zero and unit variance is another commonly used kernel, and we use a Gaussian kernel, $K_{\mathbf{x}}(x,X_i) = \frac{1}{\sqrt{2\pi}h}e^{-\frac{(x-X_i)^2}{2h}}$ around the visual feature space $\mathbf{x}$ to apply a contribution to the density estimate which drops with the distance between x and X_i. We formulate the multivariate Gaussian for each feature isotropically, denoted as $K(\mathbf{x}) = \prod_{j=i}^{d}K(x_j)$ if $\mathbf{x} = [x_1,...,x_d]^T$. This leads to $K(\mathbf{x}) = \prod_{j=1}^{d}\frac{1}{\sqrt{2\pi}h_j}e^{-\frac{x_j^2}{2h_j}}$, and finally, the formulation becomes:

$$p(b|\mathbf{x},\mathbf{g}) = \frac{\sum_{i=1}^{k}\mathbf{I}_b(X_i)\mathbf{I}_{\hat{\mathbf{g}}}(X_i)\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi}h_j}e^{-\frac{(x_j-X_{i,j})^2}{2h_j}}}{\sum_{i=1}^{k}\mathbf{I}_{\hat{\mathbf{g}}}(X_i)\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi}h_j}e^{-\frac{(x_j-X_{i,j})^2}{2h_j}}} \quad (4)$$

We choose the bandwidth parameters, h_j, as uniform over each dimension per feature after normalizing the feature dimensions according to their standard deviation. We attempted to learn optimal h per feature by employing gradient descent designed to move in a direction that minimized the tag selection error on a held-out validation set. However, we did not see much difference in the various local maxima that resulted from different starting conditions. The actual values of h used were 4.5 for the EHD, 6.0 for the HTD, 12.0 for the GIST, 2.35 for the CLD, and 18.2 for SIFT signature. Smoothing was proportional to feature dimensionality.

5.2 Regional Representation Using a Quadtree

In order to efficiently retrieve the image objects considering their global distribution in world coordinates we employ a quadtree. A quadtree is a data structure formed by recursively dividing data into four regions until a stopping condition is met. A quadtree adapts to the source data, growing in areas where the data is rich and terminating where data is sparse. Quadtrees have been used previously for watermarking [21] and object recognition through inexact image matching [22]. Wu *et al.* present a system that performs content-based image retrieval by searching an updating quadtree that effectively represents segmented region features [23]. Grady and Schwartz use a quadtree to segment medical images [24]. In most of these works, quadtree decomposition is used for sub-image definition.

We build a quadtree on the worldwide image database of 1.2 million geotagged Flickr images, using the geocoordinate tags for branching. The quadtree is grown by branching a central node into four equal-sized geographic quadrants until a stopping condition is met. We specified a minimum-support level of 100 images as the stopping condition: if a node contains fewer than 100 images with unique (user, 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. For popular geographic locations (e.g., New York), the leaf node has a small geographic footprint, while for less popular locations (e.g., parts of Africa), the geographic footprint of the leaf node is large. Each of these terminal nodes is considered to have enough images to characterize the space robustly in the presence of noisy geo- and text-tagging that results from use of voluntary user content. The geotag, $\hat{\mathbf{g}}$, used in this paper refers to the region covered by the terminal node that contains $\mathbf{g}$, and represents a discretization of a previously continuous quantity.

The algorithm for posterior calculation using kNN density estimation is provided below:

Input: a query image Q

1. Extract image features $\mathbf{x}$ for Q
2. Identify the appropriate quadtree node and geotag $\hat{\mathbf{g}}$ for Q
3. Collect the set of images I that share $\hat{\mathbf{g}}$
4. Collect the set of tags B associated with I, and compute over each feature f and each tag tag b,
$$p_f(b|\mathbf{x},\mathbf{g}) = \frac{\sum_{i=1}^{k}\mathbf{I}_b(X_i)\mathbf{I}_{\hat{\mathbf{g}}}(X_i)\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi}h_j}e^{-\frac{(x_j-X_{i,j})^2}{2h_j}}}{\sum_{i=1}^{k}\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi}h_j}e^{-\frac{(x_j-X_{i,j})^2}{2h_j}}}$$
5. For each $b \in B$, compute $p(b|\mathbf{x},\mathbf{g}) = \prod_{f=1}^{5}p_f(b|\mathbf{x},\mathbf{g})$

Output: a list of tag scores

5.3 Baseline Methods

We will compute two baseline methods for assessing the quality of annotation suggestions. A visual baseline will employ content based analysis alone and a geographic baseline will employ the prior distribution of tags present in the node specified by $\hat{\mathbf{g}}$.

5.3.1 *Visual Baseline*

The visual baseline assumes that the tags can be predicted

by visuals alone. This formulation reduces to:

$$p(b|\mathbf{x}) = \frac{p(b,\mathbf{x})}{p(\mathbf{x})} = \frac{\sum_{i=1}^{k} \mathbf{I}_b(X_i) \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}h_j} e^{-\frac{(x_j - X_{i,j})^2}{2h_j}}}{\sum_{i=1}^{k} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}h_j} e^{-\frac{(x_j - X_{i,j})^2}{2h_j}}} \tag{5}$$

The visual baseline amounts to a tag suggestion agent that is ignorant of the image's geotag. It thus reduces the problem to the standard image annotation approach.

5.3.2 *Geographic Baseline*

This baseline assumes that the tags can be predicted by geography alone. This formulation reduces to:

$$p(b|\mathbf{g}) = \frac{p(b,\mathbf{g})}{p(\mathbf{g})} \tag{6}$$

This formulation amounts to offering tag suggestions ranked by their prior for an area. While such an algorithm provides mainly place names, common objects of photographic interest can be well represented.

5.4 Smart Fusion

As the results from Joshi and Luo indicate, a fundamental problem in annotation of geotagged images is that some tags are relevant to visual features (e.g., sunset, beach), while others are not (e.g., vacation, California). A smart way to know when to fuse them has yet to be established. We propose finding the mutual information between the distribution of tags in visual feature clusters as a way to determine if visual features help assign the tag.

To estimate the pointwise mutual information we first cluster all of the images using k-means (K=950) on the GIST feature space. For each tag b, we calculate the mutual information between b and the visual feature space using the formulation:

$$MI(b,c) = \frac{1}{K} \sum_{c=1}^{K} \quad p(b,c_i) \log \frac{p(b,c_i)}{p(b)p(c_i)} + \\ p(\bar{b},c_i) \log \frac{p(\bar{b},c_i)}{p(\bar{b})p(c_i)} \tag{7}$$

Table 1 shows tags with b with high $MI(b,c)$. The tag "sunset" had the maximum estimated value. The mutual information of sorted terms decreases exponentially, so only a few terms have potential visual use. We apply late fusion of the dual method described in Section 5.1 with the Geographic Baseline in Section 5.3.2 to make an effective system. If tag b has one of the 1,250 highest values of $MI(b,\mathbf{x})$ as estimated by $MI(b,c)$, we choose to use the score from the dual method for tag b. Otherwise, we choose the geographic baseline score for tag b.

Table 1: Pointwise Mutual Information

Tag b	MI $MI(b,c)$
sunset	0.00884583
clouds	0.0073968
flowers	0.00713516
beach	0.00633382
underwater	0.00429288
car	0.00398256

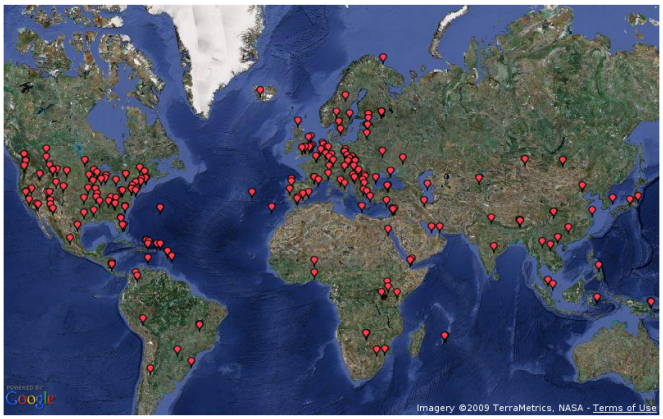

Figure 4: Distribution of the 230 test images randomly selected from the 1.2 million worldwide. Most examples are in the United States and Europe.

6. EXPERIMENTS

A series of experiments were performed to examine explicitly the performance of the system as a geocoordinate predictor and a tag suggestion agent. They were performed on the data and features described in Section 3 using the geocoordinate prediction algorithm described in Section 4 and the annotation algorithm described in Section 5. An analysis of smart fusion as well as the choice of $\hat{\mathbf{g}}$, the geographic search region, are also considered.

6.1 Geo-Coordinate Prediction Scalability

In an effort to make geocoordinate prediction scalable to millions of images, clustering was performed on the 512-dimensional GIST feature that was found to have the best performance for this task, as seen in Figure 2. Using fewer clusters results in feature comparison with more of the 1.2 million images, which we expect to lead to better accuracy but slower prediction time. Figure 5 shows the tradeoff between the number of clusters and the accuracy of the prediction, that is, the percentage of test images that were placed within 200km of the owner-supplied geotag. Results indicate performance of estimation holds up to 200 clusters, while improvement in computation time begins to saturate at this point.

6.2 Precision-Recall of Tag Suggestions

For the remaining experiments, 230 images were reserved as a testset. These images and *all images from the same owner* were removed from the collection and the learning used in this paper is done without the benefit of the 104,000 images from owners of test images. A web interface was provided for a team of judges to select correct tags for the test images from a randomized subset of the tags suggested by any of the methods. The web interface provided the analyst with the test image along with the owner-supplied tags and a map centered at the geotag of the image. The analyst could then click on the relevant tags and submit. Tags not clicked but that had been offered were considered incorrect. In the following subsections various experiments on the database are presented that judge the performance of an annotation method by precision/recall. Precision is taken to be the number of tags provided by the algorithm that were judged correct, while recall is the number of total

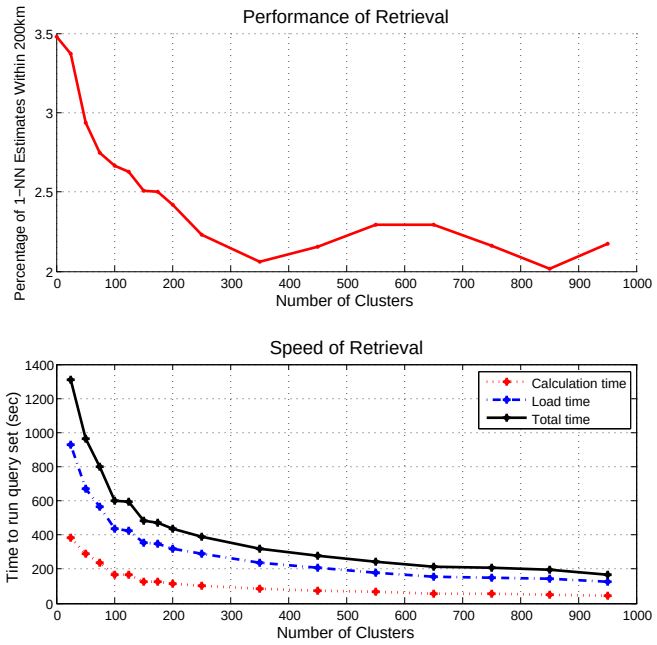

Figure 5: Results for geolocation system in terms of performance and speed with increasing number of clusters ranging from 25 to 950.

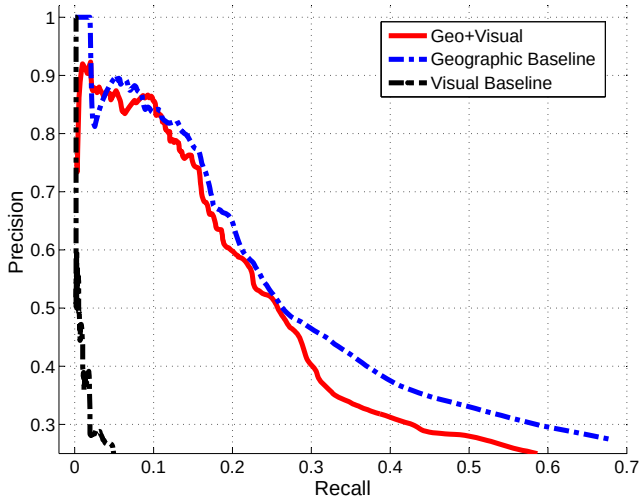

Figure 6: Precision vs. recall curve comparing dual method to two baselines. Geographic prior is found to outperform dual method suggesting visual information is best employed only for tags where it helps.

relevant tags discovered using *any* method that are covered by the *particular* method.

6.2.1 Dual Method vs. Baselines

In this experiment, we compare the annotation probabilities offered by the dual algorithm to the visual and geographic baseline methods. The visual baseline compares values of $p(b|\mathbf{x})$, and the geographic baseline values of $p(b|\mathbf{g})$, as compared to the suggestions offered by $p(b|\mathbf{x}, \mathbf{g})$. Figure 6 presents the discouraging results from this experiment, which show a combination method that calculates the probability of a tag as $p(b|\mathbf{x}, \hat{\mathbf{g}})$ performs near or worse than the geographic baseline, which formulates the problem as only maximizing $p(b|\mathbf{g})$. As is expected in an extremely noisy, non-object oriented dataset such as tourist photos, the visual-only baseline performs extremely poorly. Indeed, most "correct" tags were place names and a formulation that maximizes only $p(b|\mathbf{x})$ covers few place names.

Research from Joshi shows similar difficulties in seeing gains by fusing visual information with a geographic baseline for words such as "vacation," "university," and "stadium" [6].

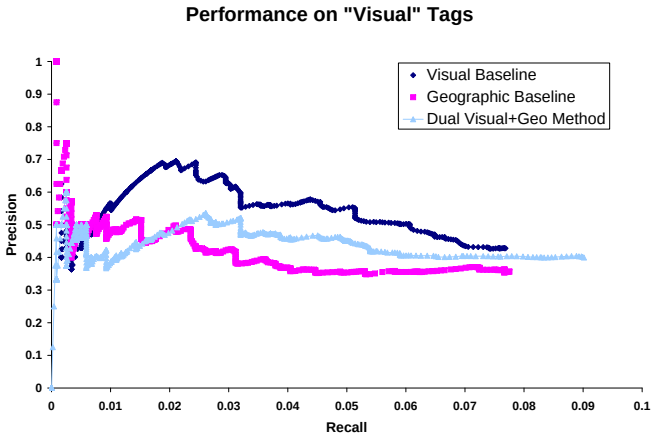

Figure 7: Performance of various algorithms on a subset of hand selected visual tags such as "beach," "sunset," "nature," "wildlife." This subset of 9% of the tags performs best when using only a visual based coder, and accordingly, the proposed visual/geographic algorithm shows improvements for this subset of tags as compared to the geographic prior-only baseline.

6.2.2 Visual Tags

A subset of tags was selected manually to verify the situations where a dual visual and geographic approach would outperform a geographic baseline. The tags are listed in Table 2. Results of algorithms on *only* these keywords are given in Figure 7.

Table 2: Manually selected keywords expected to be visually relevant. Results of algorithms on *only* these keywords are given in Figure 7.

nature, outdoors, mountain, sky, tree, water, sea, bridge, beach, jungle, park, animals, tower, flower, river, trees, boats, ship, architecture, wildlife, clouds, palm, overcast, door, desert, highway, house, street, city, building, skyline, ocean, snow, steam, forest, sunset, tropical, church

Figure 8 shows the results of using this mutual information to fuse smartly the dual method with the geographic baseline, and Figure 9 provides examples of the tags suggested using this method compared to the visual and geographic baseline. A performance gain, while not pronounced, is seen in the higher ranked tags. Tags with visual information occur in low frequency compared with place names and hence there is a limit to improvement.

6.2.3 Geographic Baseline vs. Reverse Geo-Coding

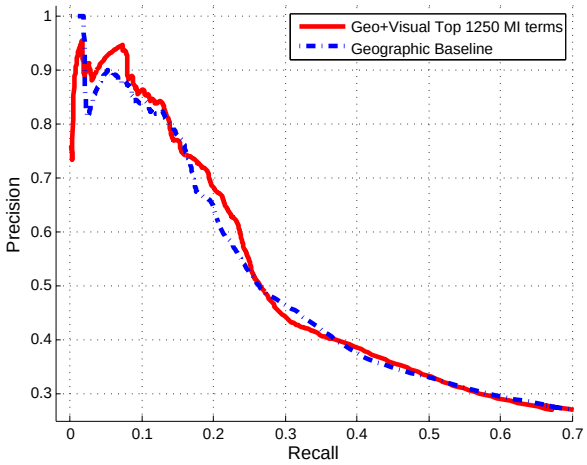

Figure 8: Performance of smart fusion against geographic baseline.

FlickrID	994514	536489	147754
image			
visual baseline	**Orlando**, **florida**, **United States**, **themepark**, **night**, cathedral, view, Seoul, nad conference, national association for the deaf, Landscape, Great Salt Lake, stars	**bridge**, **California**, **nocal**	**beach**, OIF, africa, vacation, **water**, **boat**, **sea**, trip, **ocean**, **clouds**
geographic baseline	**Orlando**, **florida**, **Universal Studios**, **Universal city Walk**, Islands of Adventure, vacation, **usa**, **FL**, **sign**, citywalk, microsoft, flickr, CSC National Sales Meeting, TechEd2007, bar, **Night Photography**, Royal Pacific Resort, laser show, **nightlife**, teched, Cinco de Mayo, **Universal Orlando**, **Walt Disney**, **United States**, 2004, Landscape, Club, fireworks, Red Coconut, lounge	**san francisco**, **California**, **United States**, **Golden Gate Bridge**, **bridge**, **nocal**, **fortpoint**, San Francisco Coastal Trail, gigi4791, **golden gate**, **Presidio**, **August**, Seacoast Fortifications, North Vista	**Vancouver**, **Canada**, **bc**, **bay**, fireworks, celebration, light, finale, English, 150, **British Columbia**, iPhone, iPhone, AirMe, WeatherBug, **beach**, City, sunset, stanley park, building, **ocean**, Mostly Sunny, **sea**, thelastminute, thelastminuteblog, **clouds**, tree, Granville Island
Dual visual and geographic method	**florida**, **United States**, **Orlando**, **Universal Studios**, **usa**, vacation, flickr, fireworks, bar, **Night Photography**, disney, Club	**California**, **United States**, **bridge**, **Golden Gate Bridge**, **golden gate**, **August**	**British Columbia**, **Canada**, **Vancouver**, **bc**, City, building, tree, **sea**, park, **boat**, **bay**, House, fireworks, celebration, light, celebration, **beach**, sunset, **ocean**, trip, flower, **sand**, aerial, stanley park, **BritishColumbia**, helijet

Figure 9: Example tags suggested by each method. Suggestions are in order of decreasing confidence. Bold red for correct, black plaintext for incorrect.

An experiment was done that performed a reverse geocoding on the query geocoordinates to examine the performance of the geographic baseline. The reverse geo-coder used the Flickr API [25] to lookup a geotag and then suggest the associate city, region (e.g., state or province), and country as annotations with maximal score. The performance increase that can be seen from the geographic baseline to one that has undergone this reverse geocoding is evident in Figure 10.

6.2.4 *Definition of Geographic Search Region, $\hat{\mathbf{g}}$*

A comparison was made between the performance of the algorithms based on the area of the geographic footprint of the node. Performance was measured for two groups based on density, split at the median of geographic footprint area. We examine the performance of the baselines with the composite algorithm to determine their performance as a function of image *density*. The results are shown in Figure 11. The dual geo+visual method performs better initially for

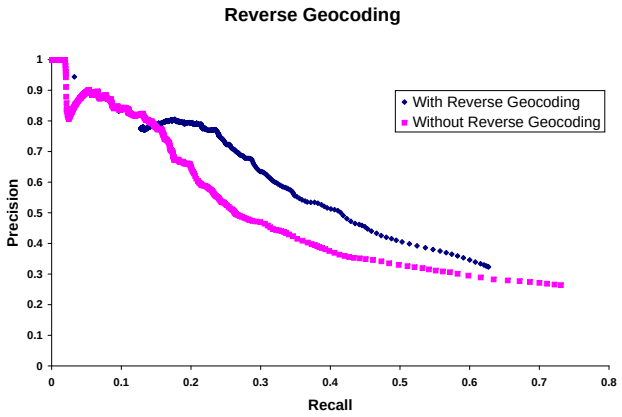

Figure 10: Improvement on geographical baseline by providing location terms from reverse geocoder.

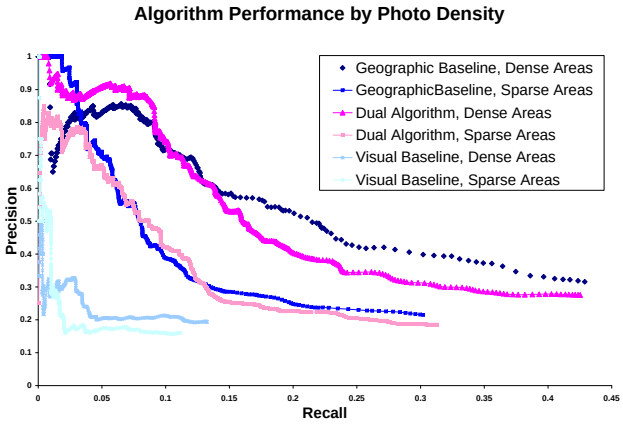

Figure 11: Comparison of performance based on density of images at geocoordinate of image. While the dual geo+visual method and the geographic baseline *are* significantly affected by the density of images in the area (they perform better in dense areas), the visual baseline exhibits a mixed tradeoff when we separate by image density.

denser regions, suggesting the system can be tuned to offer high ranking visual-based annotations in such areas.

Additionally, a comparison was made between using the quadtree to find an appropriate region and a fixed geographic area. This study analyzes whether the use of a quadtree, which terminates at a level of minimum support in a geographic area, is a robust way of determining a geographic area of relevance. Figure 12 shows:

- 1) geographic baseline using the quadtree against
- 2) a formulation of $p(b|\hat{\mathbf{g}})$ that fixes a radius of 7km around the geocoordinates of the target.

If the image was taken in a dense area (e.g., New York City), then method 2) will include more images in calculation $p(b|\hat{\mathbf{g}})$; if it was taken in a less popular area (e.g., Antarctica) then method 2) will include fewer images in measuring the probability of a particular tag. The fixed radius of 7km was chosen because it was the average of the corresponding radii taken at the level of minimum support for method 1).

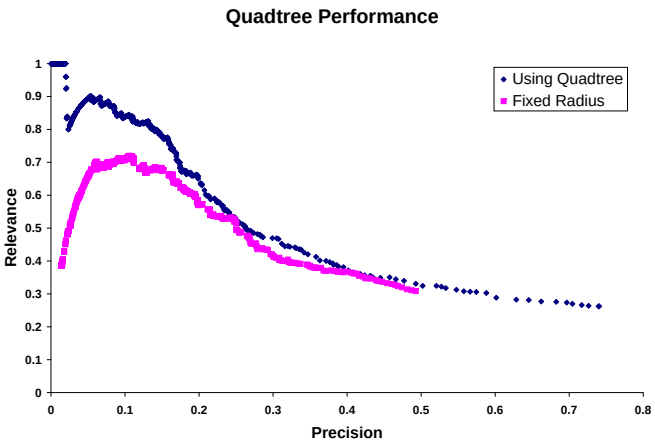

Figure 12: Comparison of performance using a quadtree to determine geographic area versus a fixed radius. The quadtree defines an area of influence based on sample distribution, while a fixed radius allows for many cases with too few images.

7. CONCLUSIONS

In this work we have presented a world-scale tag suggestion system which employs a database of 1.2 million geotagged images in order to provide annotations for input photographs taken anywhere in the world. Geolocation estimation can be quickly provided to aid the user interface. Relevant annotations were found to be highly geographically dependent, as seen by the performance of a baseline derived from representing the geographic tag distribution on a quadtree. Tag suggestions which are aided by visual analysis can be determined via estimating mutual information, and we found that visual methods hold the most promise for densely sampled regions. In the future we will explore efficient methods for scaling the system up to a larger dataset.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[2] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," in *CVPR*, 2004.

[3] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr Helps Us Make Sense of the World: Context and Content in Community-Contributed Media Collections," in *ACM Multimedia*, 2007.

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

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

[6] D. Joshi and J. Luo, "Inferring Generic Activities and Events from Image Content and Bags of Geo-tags," in *CIVR*, 2008.

[7] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: Image Auto-Annotation by Search," in *CVPR*, 2006.

[8] E. Moxley, T. Mei, X.-S. Hua, W.-Y. Ma, and B. Manjunath, "Automatic Video Annotation Through Search and Mining," in *ICME*, 2008.

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

[10] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, B. Manjunath, Ed. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[11] 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.

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

[13] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification," in *ECCV*, 2006.

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

[15] "http://people.csail.mit.edu/torralba/code/spatialenvelope," Torralba GIST implementation.

[16] "http://vision.ucla.edu/~vedaldi/code/bag/bag.html), note = UCLA Bag of Features ."

[17] S. Ahern, M. Naaman, R. Nair, and J. H. Yang, "World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections," in *Conference on Digital Libraries*, 2007.

[18] T. Tran, R. Wehrens, and L. M. Buydens, "KNN Density-Based Clustering for High Dimensional Multispectral Images," *Computational Statistics and Data Analysis*, 2006.

[19] M. Asefi, "Classification-Based Adaptive Search Algorithm for Video Motion Estimation," Ph.D. dissertation, University of Waterloo, Waterloo, Ontario, Canada, 2006.

[20] O. Boiman, E. Shechtman, and M. Irani, "In Defense of Nearest-Neighbor Based Image Classification," in *CVPR*, 2008.

[21] S. Tsu and W. Hsieh, "Quadtree-Based Perceptual Watermarking Scheme," in *ASIACCS*, 2006.

[22] L. A. Consularo and R. M. Cesar, "Quadtree-Based Inexact Graph Matching for Image Analysis," in *SIBGRAPI*, 2005.

[23] S. Wu, M. Rahman, and T. Chow, "Content-Based Image Retrieval Using Growing Hierarchical Self-Organizing Quadtree Map," *Pattern Recognition*, 2005.

[24] L. Grady and E. L. Schwartz, "Faster Graph-Theoretic Image Processing via Small-World and Quadtree Topologies," *CVPR*, 2004.

[25] "http://www.flickr.com/services/api/flickr.places.findByLatLon.html," Flickr Places API Services.