Geo-location Inference from Image Content and User Tags

Andrew Gallagher *Dhiraj Joshi* *Jie Yu* *Jiebo Luo*

Kodak Research Laboratories

Rochester, NY 14615

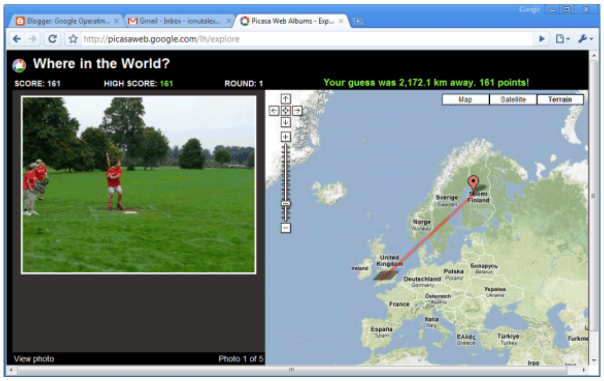
{andrew.gallagher, dhiraj.joshi, jie.yu, jiebo.luo}@Kodak.com

Abstract

Associating image content with their geographic locations has been increasingly pursued in the computer vision community in recent years. In a recent work, large collections of geotagged images were found to be helpful in estimating geo-locations of query images by simple visual nearest-neighbors search. In this paper, we leverage user tags along with image content to infer the geo-location. Our model builds upon the fact that the visual content and user tags of pictures can provide significant hints about their geo-locations. Using a large collection of over a million geotagged photographs, we build location probability maps of user tags over the entire globe. These maps reflect the picture-taking and tagging behaviors of thousands of users from all over the world, and reveal interesting tag map patterns. Visual content matching is performed using multiple feature descriptors including tiny images, color histograms, GIST features, and bags of textons. The combination of visual content matching and local tag probability maps forms a strong geo-inference engine. Large-scale experiments have shown significant improvements over pure visual content-based geo-location inference.

1. Introduction

Human beings, over the years, have constructed rich vocabularies to describe sceneries, objects, people, and places captured in pictures. Most such words instantly strike geographical associations in our minds. These geographical associations may vary from being rather specific (e.g., for Paris) to being fairly general (e.g., for beach). For human beings, building such associations is natural and results from conditioning and education. Additionally, humans possess the unique capability to analyze the visual content of pictures and draw conclusions as to its geographical location. In fact, Google has recently introduced an online game "Where in the World" (Fig. 1) to tap this human potential.



Your guess was 2,172.1 km away. 161 points!

Figure 1: Google "Where in the World" game. Here, a user scores 161 points when the guess is 2171 *km* away.

Making and preserving these geographical associations with pictures is an age-old process. During the "film-camera" days, people would write the place where the picture was taken on the back of the print. Today a user can map his pictures precisely using community image management systems such as Google™ Picasa™, Google Earth, and Yahoo® Flickr.

A fast-emerging trend in digital photography and community photo sharing is *geotagging.* The phenomenon of geotagging has generated a wave of geo-awareness in multimedia[1,2] [5][7][10][12]. Yahoo Flickr has amassed about 2.5 million photos geotagged in the month this paper was written[3]. Geotagging is the process of adding geographical identification metadata to various media such as websites or images and is a form of geospatial metadata. It can help users find a wide variety of location-specific information. For example, one can find images taken near a given location by entering latitude and longitude coordinates into a geotagging-enabled image search engine. Geotagging-enabled information services can also potentially be used to find location-based news, websites, or other resources. A current key limitation to geotagging in photo-sharing websites is the manual labor involved (even though automatic geotagging using GPS receivers is

[1] http://zonetag.research.yahoo.com

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

[3] http://www.flickr.com

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

gaining attraction among early adopters). In other words, geotagging cannot proceed without human intervention and any process that depends on manual labor entirely is not scalable. However, participation of millions of people in the process gives rise to an interesting solution to this problem, as discussed below.

With the massive volume of digital imagery being captured and shared on the Web, and the phenomenon of geotagging having acquired phenomenal proportions, it has become a recent research trend to explore computer vision algorithms to link user-tags, visual content of pictures, and community knowledge with the geographic locations where the pictures were captured. An important research question that motivates our current work is how this massive volume of community geotagged image data can be leveraged to geotag or assign geographic locations to images, especially legacy pictures that were taken before the GPS days.

In a work published in [14], a methodology based on simple K-nearest-neighbor visual search to infer geo-association of images was described. The basic premise explored in the aforementioned work was that visual content of pictures and their geographic locations are correlated. The strength of the system lay in a simple technique and the availability of a very large-scale image database (~6 million images) for search. In our work, we take guidance from [14] and explore how user-tags can be leveraged in addition to K-nearest-neighbor visual search to refine the geo-inference. We work with an image database roughly one-sixth the size of that used in [14] to build location probability maps for user-tags, which will be described in Section 3. Pure tag-based geo-inference forms a baseline against which we compare pure visual search (K-nearest-neighbor) based geo-inference. Finally, we propose a new method where local tag probability maps are exploited to improve the location inference using pure visual search alone. It would be interesting to compare human geo-location prediction performance (as in Google's "Where in the World" game) with that of automatic algorithms.

2. Related Work

Content understanding in images has been studied for decades in the vision research community. Content understanding in images can translate to understanding scene semantics [20][21] or event semantics [21][22]. Invariably, image content understanding algorithms involve building classifiers for a finite number of semantic categories. A potent application of image understanding is image retrieval. However, learning-based retrieval is constrained by the cardinality of semantic categories. Another line of research has, for a long time, explored unsupervised similarity-based search and retrieval using low-level visual features alone [23]. Recent interest in brute force searches using massive image databases has been shown to be useful for image understanding tasks as well [16]. Such methods, which rely on retrieval for semantic understanding, complete a full circle in connecting the fields of image retrieval and image understanding. However, all of the above systems still focus on only the image content. With rapid advances in technologies related to digital imaging, digital cameras also bring with them a powerful source of information little exploited previously for scene classification: camera metadata embedded in the digital image files. Camera metadata (or "data about data") records information related to the image capture conditions and includes values such as date/time stamps, subject distance, and GPS coordinates. They contain rich contextual information that is usually complementary to the image features for the purpose of semantic understanding. The research community increasingly turns to metadata and picture-taking context in the pursuit to solve the semantic understanding problem [1][2].

Important metadata can be collected also as a result of user participation. Online photo-sharing websites such as Flickr have witnessed a surge of collaborative tagging from users, resulting in folksonomies [26][27]. Recently, there have been research efforts to understand user image tagging behavior [3] and to characterize this behavior over time [6]. When users associate geographic content with media on the Web, it becomes an instance of geotagging. With the growing popularity of geotagging, mining, organizing, and making sense of georeferenced data and linking geo-content to visual content has become essential. Initial attempts to identify geo-relevant content on Web pages in order to assign a geographic focus to pages were made in [4]. Retrieval of geographical landmarks from the Flickr dataset was studied using a combination of visual features and geotags in [9]. An algorithm to create summaries of georeferenced collections was proposed in [8] to improve browsing and visualization of images. Season and location context was found to be useful for region labeling in [11]. The problem of finding associations between places and picture semantics was studied in [13][28].

While location context has been used for image understanding, the inverse problem of inferring location from image content is still novel and difficult to address [14][24][25]. An impressive system relying on simple visual search over a massive image database demonstrated good performance for geo-localization task in [14]. Geo-location in a known urban environment was accomplished by matching 3D building facades using SIFT features while the camera pose is recovered [15]. The distinguishing aspect of our work from these two references lies in the novel use of user-tags along with image content for improved geo-localization.

3. Mapping User Tags over the Globe

We build a collection of Flickr images containing geo-location metadata. Our goal is to investigate the relationship between user-tags, geo-location, and image content (appearance). To this end, we want a collection of interesting images that contains both user-tags and geo-locations. Our collection contains 1.2 million images download from Flickr as follows: We observed that geo-located images are nearly always also annotated with user-tags. We query Flickr for the 2500 most interesting geo-located images captured on a specific day, and repeat this process for 504 specific days. Unlike [14], we have no requirements for user-tags except for a set of negative query terms that prevent low-quality (e.g., camera phone) or otherwise objectionable images from being gathered in our query. Note that using the interest level as a filter ensures that the quality and content of the images in the collection are reasonable thanks to the implicit human filtering by Flickr users.

Geo-locations of pictures and the user-tags assigned to the pictures are correlated. This naturally follows from the fact that human vocabulary is countable and only a subset of the vocabulary is likely to be used to tag pictures taken in a given geographic region. Pictures taken in a particular region are also likely to capture similar scenes or objects of interest, hence further limiting the vocabulary used to tag them. For example, pictures taken around the Eiffel Tower are far more likely to be tagged “Eiffel Tower,” “Paris,” and “France” than are pictures taken in New York City. At the same time, it is intuitive to think that pictures bearing different user-tags are distributed differently across the globe. If the entire globe is quantized to represent a finite number of regions, it is possible to probabilistically associate user-tags and geographic regions.

We define the following terms with respect to a corpus of images with user-tags and with known geographic locations. Let $T = \{t_1, t_2, ..., t_N\}$ denote the set of user-tags and $R = \{r_1, r_2, ..., r_M\}$ denote a set of geographic regions. Let us assume that the regions represent roughly equal sized segments on the earth’s surface. If $N_r(t_i)$ is the number of pictures bearing tag t_i and whose geo-locations fall in region r, $N_r = \sum_{t_i} N_r(t_i)$ is the total number of pictures bearing any tag and whose geo-locations fall in region r, we define a location probability map for tag t_i, $p(r \mid t_i)$ over the entire region set $R = \{r_1, r_2, ..., r_M\}$ such that the probability $p(r \mid t_i)$ represents the likelihood of a picture-bearing tag t_i to be found in region r. This can be readily estimated from the corpus as

$$\hat{p}(r \mid t_i) = \frac{N_r(t_i)}{\sum_{r=1}^{M} N_r(t_i)} \qquad (1)$$

A limitation of the above estimation procedure is that the probabilities can be highly unreliable when the denominator is very small (especially for tags with low presence in the corpus). Therefore, we regularize the probability using $p(r)$, the probability of finding a picture in region r, estimated from the corpus as $\hat{p}(r) = \frac{N_r}{\sum_{r=1}^{M} N_r}$.

The regularized estimate for $p(r \mid t_i)$ becomes

$$\hat{p}(r \mid t_i) = \frac{N_r(t_i) + \Lambda p(r)}{\sum_{r=1}^{M} N_r(t_i) + \Lambda} \qquad (2)$$

Here the factor Λ is a combination factor used to ensure that in the limiting case when $\sum_{r=1}^{M} N_r(t_i)$ is very small, the estimate $\hat{p}(r \mid t_i)$ approaches $\hat{p}(r)$.

The usefulness of such location probability maps (with respect to common user-tags) lies in their ability to be used to infer the geographic region of a picture with user-tags. Inferring geo-location involves computing the probability that an image I bearing a set of tags $\{t_1{}^I, t_2{}^I, ...t_k{}^I\}$ was photographed in a region r. Probabilistically, this is written as

$$p(r \mid t_1{}^I, t_2{}^I, ...t_k{}^I) = \frac{P(t_1{}^I, t_2{}^I, ...t_k{}^I \mid r) P(r)}{P(t_1{}^I, t_2{}^I, ...t_k{}^I)}. \qquad (3)$$

In order to find the most likely region, we treat the factors $P(t_1{}^I, t_2{}^I, ...t_k{}^I)$ and $P(r)$ as constants and impose a Naïve Bayes model that assumes conditional independence across tags mainly for computational simplicity, resulting in the reduction of the above to computing the following product

$$p(r \mid t_1{}^I, t_2{}^I, ...t_k{}^I) \approx P(r \mid t_1{}^I) P(r \mid t_2{}^I) ...... P(r \mid t_k{}^I). \qquad (4)$$

In order to visualize geographic patterns for user-tags, we divide the entire globe into 900 × 1800 regions (900 bins along latitudes, 1800 bins along longitudes) such that each region captures 0.2 degrees of granularity in latitude and longitude. For each of the 1500 most frequently occurring user-tags (with respect to our corpus), we estimate the frequencies $N_r(t_i)$ with respect to the above regions. These location frequency maps can be directly visualized as images to see the geographic spread of pictures bearing different tags.

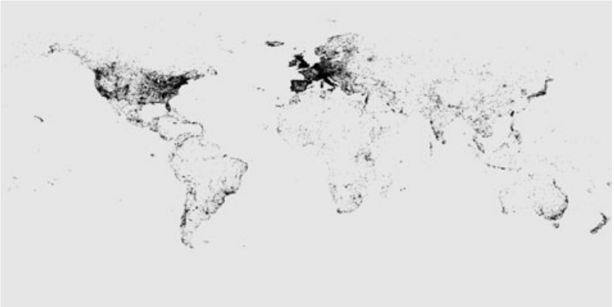
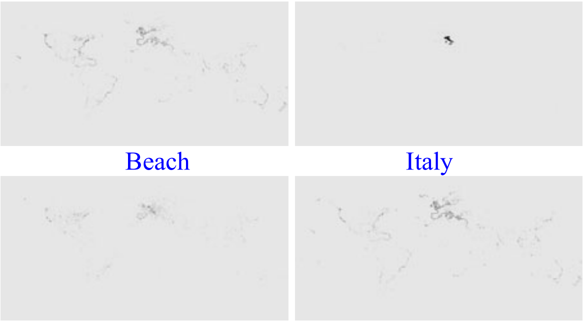


Figure 2: Global location frequency map for all tags.

First, it is interesting to observe the geographical distribution of all pictures regardless of their tags (i.e., $N_r = \sum_{t_i} N_r(t_i)$). This frequency map is shown in Fig. 2. In Fig. 3, we show the frequency maps for eight specific user-tags having varying distributions over the globe. In these figures, each pixel along the Equator represents an area of approximately 22 km^2. Since the frequency maps are expected to be very sparsely distributed over the 900 × 1800 regions, we binarize the images for better visualization. The black regions in Figs. 2 and 3 represent regions in the world where $N_r(t_i) > 0$. While we do not overlay the frequency maps on a world map, the readers can naturally extrapolate the information and visualize the entire globe, e.g., the continents.

It is clear that the map in Fig. 2 (being the cumulative frequency across all user tags) has more non-zero dark regions than any of the tag-specific maps in Fig. 3. In Fig. 2, one can clearly see that the densest regions lie within the continental USA and Europe. Most of the other continents are sparse except for the coastlines. To some extent, Fig. 2 represents the present-day state of picture-taking activities in the world. The predominance of North America and Europe among the denser regions in Figs. 2 and 3 is also testimony to the fact that these continents presently form the hubs of the world's geo-tagging activity.



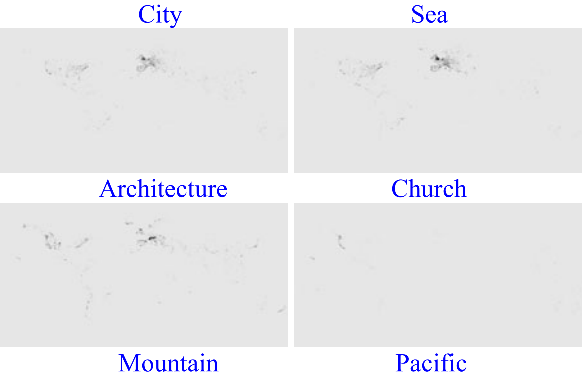


Figure 3: Global location frequency maps for certain selected tags.

Not surprisingly, the maps for "beach" and "sea" (Fig. 3) are dense mainly along the coastlines all over the world. The maps for "Italy" and "Pacific" are localized to Italy and the North American Pacific coast, respectively. The map for "city" is dense at most of the major cities (mainly in North America and Europe). Map for "mountain" concentrates mainly in the Alps in Europe, the Rockies and the Appalachians in North America, and the Andes in South America. Maps for "church" and "architecture" bear a similarity in their geographic spreads. Both are mainly concentrated in Europe where it is not unusual to find historic architectures and old churches.

4. Visual Features and Matching

In our work, we adopt visual feature extraction and matching methodologies from [14]. While [14] used six sets of visual features, here we limit ourselves to using four features only (partly due to computational limitations). Our choice of features was guided by controlled experiments to determine the visual and geo-coherence of individual features on a subset of our training corpus. The four features that we have selected are widely used in computer vision and are effective for matching a large spectrum of visual content.

1. **Tiny Images:** Downsampled (or tiny) images are trivial but useful features. Tiny images were popularized by Torralba et al. [16] for scene classification and object recognition tasks. In our work, similar to [14], we employ 16 × 16 RGB images as one of our features.
2. **Color Histograms:** Use of CIE L*a*b color histograms has been very popular in image retrieval. Similar to [14], we construct histograms with 4, 14, and 14 bins in L, a, and b dimensions, respectively, to form a 784-dimensional feature histogram for each image.

3. **GIST Descriptor:** GIST descriptors encode structural information and have successfully been used for scene categorization and retrieval tasks. As in [14], a 5 × 5 spatial resolution GIST descriptor is created for each image. Each bin is a 24-dimensional (6 orientations and 4 scales) filter response for the respective image region, resulting in a 600- dimensional feature vector for the entire image.
4. **Texton Histograms:** Our texture features consist of histograms over a 120-word texton dictionary extracted from the training corpus. Responses to the 24-dimensional filter response (as in GIST descriptor) are quantized to the nearest texton dictionary word to form texton histograms.

A K-nearest-neighbor search is employed for visual matching and geo-inference. Distances between images are computed differently for different features. The GIST descriptors are compared across images using Euclidean distance. The tiny images are compared using normalized cross correlation. We employ a χ^2 distance measure to match color histograms and texton histograms as these features are inherently probability distributions. A combination of distances using multiple features is performed linearly by using feature-specific weights learned from a small set of controlled data. The feature weights are based on the geometric spread of distance values computed using different features such that distance ranges using different features are comparable.

5. Geographical Location Prediction

We investigate three different approaches to geographic location prediction based on two modalities: visual content and user tags.

5.1. Tag Baseline

Our tag-only baseline assesses the extent of geographical precision we can achieve by employing user-tag location probability maps alone. For an image with user tags $\{t_1^I, t_2^I, ...t_k^I\}$, the product $p(r \mid t_1^I, t_2^I, ...t_k^I)$ is computed for each of the 900 × 1800 geographic regions and the region r^* with the highest probability is selected. The geographic center of region r^* constitutes the geo-location assignment to the image.

5.2. Visual Baseline

Our visual baseline is constructed along the lines of Hays method employing only visual information [14]. Once the *K*-nearest neighbors are retrieved for the query image, their geographical locations are represented as (longitude, latitude) pairs. To predict the geographical location, the mean-shift algorithm [19] is used to segment the data into different clusters. The mode of the cluster with the highest cardinality is predicted as the geographical location of the query image. This is a natural choice in the absence of additional information. The next method we propose draws on additional cues from user tags to achieve improved geo-location inference.

5.3. Integrating Visual Content and Tags

We propose a new method that uses both tag and visual information to find the geographical location of an image. The method draws its power from similarity search in both visual and semantic domains from a database of about 1.2 million images. We do not propose direct fusion of visual and tag baselines for two important reasons:

1. We constructed user-tag location probability maps for only the 1500 most frequent user-tags. The total number of user-tags in the 1.2 million Flickr images is expected to be much higher than 1500. As a result, use of tag location probability maps directly for fusion is a limitation.
2. Tag location probability maps have been built using 1.2 million images. Hence they capture global distributions of images and user-tags. For image-specific geo-assignment, a visually local location distribution map is expected to be more useful. In our case, local location distribution maps are constructed by using a composite tag and visual similarity in the K-nearest-neighbor search.

For any two images, a composite distance measure is defined as $d(i,j) = d_V(i,j) + d_T(i,j)$, where $d_V(i,j)$ is the distance measure between the visual content of images *i* and *j*, while $d_T(i,j)$ describes the similarity of the tag sets associated with the two images. Measuring the semantic distance between tags or tag sets is an active research topic in its own right. The *tf-idf* method has been widely used to build ranking functions in information retrieval and text mining tasks. However, it cannot be simply adopted for our problem because tags of the same image are not repeated. Normalized Google Distance [17] measures the semantic correlation of two words by measuring their co-occurrence in web page search results. WordNet can be used to derive the semantic distances from two words based on ontological knowledge [18]. Methods using external knowledge, such as NGD and WordNet, do not suit our task mainly because of extra computational cost.

In this paper, we use a text retrieval-based method to measure the tag similarity efficiently

$$d_T(i,j) = \begin{cases} 0 & \text{if i and j have at least one common tag} \\ MAX & \text{if i and j have different tags} \end{cases} \qquad (5)$$

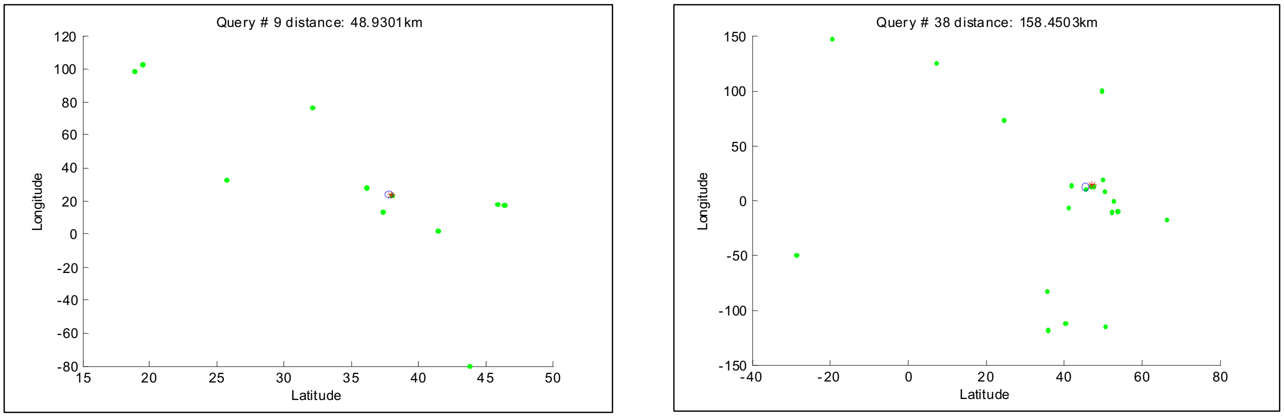


Figure 4: Visualization of the geo-location inference results. Green dots are the top ranked images based on combined visual and tag similarity matching. The blue circles represent the predicted locations of the query images while the red asterisks are the ground truth.

Here *MAX* is a constant that is larger than the upper limit of the visual distance $d_V(i, j)$, e.g., 10^{100}.

Clearly, sorting the combined distances of the query image to all of the images in the database is equivalent to i) first ranking the visual distances of all of the images that share at least one tag as the query image, and ii) then ranking the visual distances of all of the images that do not share any tag with the query image.

Integrating visual and tag information using the above-mentioned method has several advantages:

1. Complimentary information is extracted from the tags, which significantly improves the inference compared with using the visual content alone;
2. The text-retrieval-based distance is more efficient than other state-of-the-art text distance measures;
3. For a *K-NN* based method, the number of images needed to evaluate the visual similarity to the query image is greatly reduced if more than *K* images have the same tags as the query image; and
4. Our system can be readily incorporated with text search engines, which also use this method as the front end.

6. Experiments and Analysis

To evaluate the performance of our methods, we test them on two image data sets and use our 1.2 million geo-tagged images for training. The results are quantitatively analyzed in the following way: If the distance of the predicted location of a query image to the ground truth location is within a specific range, it is considered a *hit*, otherwise a *miss*. The accuracy is defined as the number of *hits* over the total number of query images. The ranges are set to 200, 300, 500, and 1000 kilometers, respectively. The neighborhood size K is set to 20 in this study.

6.1. Experiments on Image2GPS Test Set

The Image2GPS test data set in [14] contains 237 images. In Fig. 4, we show the local location distribution maps using combined tag and visual similarity for two selected query images.

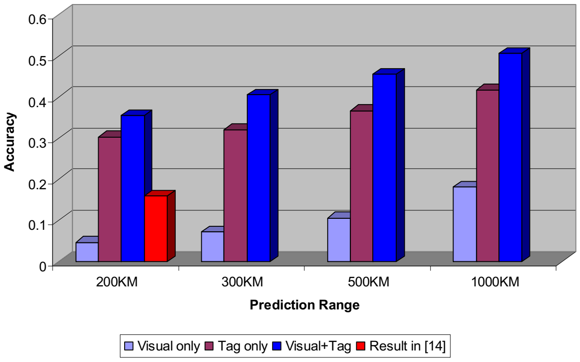


Figure 5: Prediction accuracy for Image2GPS data set.

The geo-prediction accuracies of the two baselines and the combined method are shown in Fig. 5. We also include the performance from [14] as the red bar, which is reported for the 200 km range. One can notice that our visual baseline is not as strong as that reported in [14] for the 200 km range. The reasons are as follows:

1. The training database in our experiment is smaller than that used in [14] (1.2 million versus 6 million images).
2. In [14], all of the training and test images are pre-filtered such that only images from 500 metropolitan areas in the world and bearing tags corresponding to these cities are included. In our work, the images are collected from all around the globe without any such restriction, thus further reducing the concentration of images in urban areas. As a result, the image-matching problem that we address is much less constrained and

thus more challenging.

. Each of these factors has the effect of reducing the accuracy of the visual matches. This result corroborates [14] Figure 4, in that a smaller search space reduces the visual match quality, from 14% on 6 million reference images to around 7% on 1 million ones. Despite the effect, our key result is that the combination of text and visual matches outperforms the previous result in [14]. In essence, this shows that by using tags, we gain the advantage that the training database does not need to be so large to achieve equivalent or better performance.

In all of the prediction ranges, the integration of visual content and tags achieved significant performance improvement over the visual baseline. In the 200 km range, it outperforms the accuracy reported in [14] by 16%. In all prediction ranges, it performs better than the tag baseline.

6.2. Experiments on Flickr Image Test Set

In this experiment, we test our methods on 2000 interesting images downloaded from Flickr. They are independent from the training set.

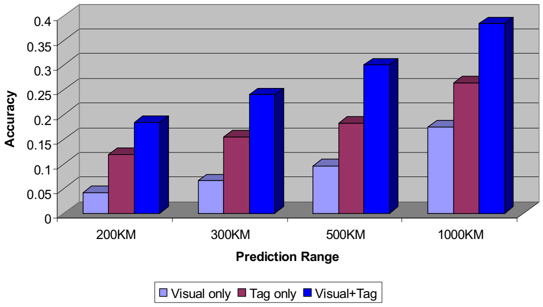


Figure 6: Prediction accuracy on Flickr data set.

The result in Fig. 6 indicates that it is very difficult to infer the geographical information from the visual baseline because of the extremely rich content of the query images (as they are not limited to 500 metropolitan cities). Compared with the visual baseline, the tag baseline predicts the geographical location of the images more precisely. Most importantly, the integration of visual content and tags significantly improves the prediction accuracy.

Figure 7 (a) shows a few cases when tags may provide more precise geographical information than visual content, and vice versa. For example, the visual baseline performs poorly for the inconspicuous street scene in Fig. 7 (a). As demonstrated in Figs. 7 (b) and (c), integrating visual and tag information significantly improves geo-prediction performance over the visual and tag baselines. Overall, tags and visual content provide complementary information that is leveraged by the fusion of the two modalities.

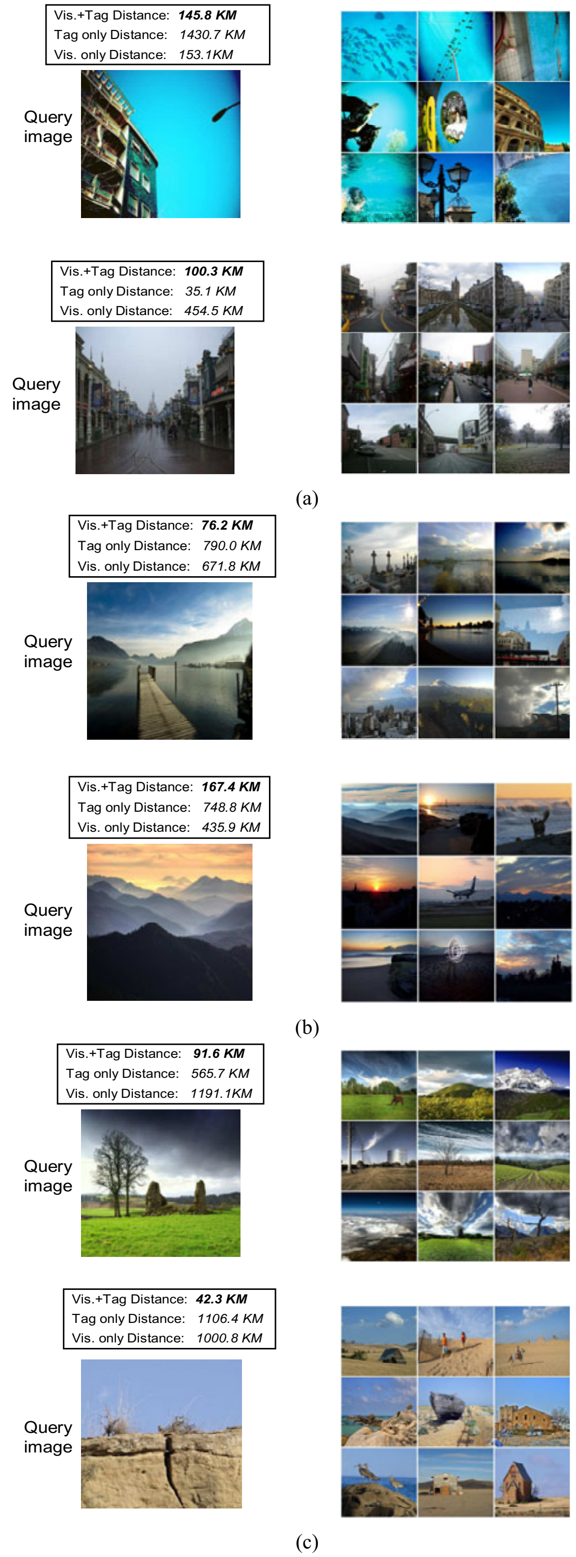

Vis.+Tag Distance: 145.8 KM
Tag only Distance: 1430.7 KM
Vis. only Distance: 153.1KM
Query image
Vis.+Tag Distance: 100.3 KM
Tag only Distance: 35.1 KM
Vis. only Distance: 454.5 KM
Query image
(a)
Vis.+Tag Distance: 76.2 KM
Tag only Distance: 790.0 KM
Vis. only Distance: 671.8 KM
Query image
Vis.+Tag Distance: 167.4 KM
Tag only Distance: 748.8 KM
Vis. only Distance: 435.9 KM
Query image
(b)
Vis.+Tag Distance: 91.6 KM
Tag only Distance: 565.7 KM
Vis. only Distance: 1191.1KM
Query image
Vis.+Tag Distance: 42.3 KM
Tag only Distance: 1106.4 KM
Vis. only Distance: 1000.8 KM
Query image
(c)

Figure 7: Examples of query images, predicted distances, and the nearest neighbors by the integration of visual content and tags.

7. Discussions and Conclusions

In this paper, we explored the benefit of employing user-tags along with image content to infer geo-location of Web images. We constructed global location probability maps for common user-tags using a large collection of 1.2 million geotagged photographs. Visual content matching was performed using multiple image features and K-nearest-neighbor similarity search. The combination of brute force visual content matching and local tag probability maps is shown to outperform baselines based on single modalities. An important future direction will be to integrate more advanced feature descriptors for visual matching and scalable image-matching methods. We also plan to incorporate tag co-occurrences into tag-based geo-inferencing.

References

[1] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 68(1):43-52, 2006.

[2] J. Luo, M. Boutell, and C. Brown, Pictures are not taken in a vacuum: An overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine*, 23(2):101–114, 2006.

[3] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007.

[4] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: Geotagging web content. *In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[5] Y. Chen, X. Y. Chen, F. Y. Rao, X. L. Yu, Y. Li, and D. Liu, LORE: An infrastructure to support location-aware services. *IBM J. Res. Devel*. 48(5/6):601-616, 2004.

[6] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. *In Proceedings of the World Wide Web*, 2006.

[7] A. Hinze and A. Voisard. Location and time-based information delivery in tourism. *Advances in Spatial and Temporal Databases. Lecture Notes in Computer Science*, 2750:489-507. 2003

[8] A. Jaffe, T. Tassa, and M. Davis, Generating summaries and visualization for large collections of geo-referenced photographs. *In Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, 2006.

[9] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How Flickr helps us make sense of the world: context and content in community-contributed media collections. *In Proceedings of the ACM International Conference on Multimedia*, 2007.

[10] L. Liu, O. Wolfson, and H. Yin. Extracting semantic location from outdoor positioning systems. *In Proceedings of the IEEE International Conference on Mobile Data Management*, 2006.

[11] J. Yu and J. Luo. Leveraging probabilistic season and location context models for scene understanding. *In Proceedings of the ACM International Conference on Image and Video Retrieval*, 2008.

[12] J. H. Schiller and A. Voisard. Location-based services. *Morgan Kaufmann*, 2004.

[13] D. Joshi and J. Luo. Inferring generic activities and events from image content and bags of geo-tags. *In Proceedings of the ACM International Conference on Image and Video Retrieval*, 2008.

[14] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[15] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert. Detecting and Matching Repeated Patterns for Automatic Geo-tagging in Urban Environments. *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[16] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images, *Technical Report MIT-CSAIL-TR-2007-024*, 2007.

[17] R. L. Cilibrasi and P. M. B. Vitanyi. The Google Similarity Distance, *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370-383, 2007.

[18] T. Pedersen. S. Patwardhan. J. Michelizzi, WordNet Similarity - measuring the relatedness of concepts, *In Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 2004.

[19] D. Comaniciu, P. Meer. Mean Shift: A robust approach toward feature space analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603-619, 2002.

[20] L. -J. Li and L. Fei-Fei. What, where and who? Classifying event by scene and object recognition. *In Proceedings of International Conference on Computer Vision*, 2007.

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *In Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2006.

[22] V. Jain and A. Singhal. Selective hidden random fields: Exploiting domain specific saliency for event classification. *In Proceedings of International Conference on Computer Vision and Pattern Recognition*, 2008.

[23] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Survey*, 40(65) (2008).

[24] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless, Geolocating static cameras. *In Proceedings of International Conference on Computer Vision*, 2007.

[25] J. Kosecka and W. Zhang. Video compass. *In Proceedings of European Conference on Computer Vision*, 2002.

[26] I. Simon, N. Snavely, and S. Seitz, Scene Summarization for Online Image Collections, *Proc. of IEEE Intl. Conf. on Computer Vision*, 2007.

[27] T. Quack, B. Leibe, and L. Van Gool, World-scale mining of objects and events from community photo collections, *Proc. of ACM Conf. on Image and Video Retrieval* 2008.

[28] G. Schindler, M. Brown, and R. Szeliski, City-Scale Location Recognition, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.