World-Wide Scale Geotagged Image Dataset for Automatic Image Annotation and Reverse Geotagging

Conference Paper · March 2014

DOI: 10.1145/2557642.2563673

Hatem Mousselly Sergieh
Universität Passau
17 PUBLICATIONS **122** CITATIONS
SEE PROFILE

Mario Doeller
University of Applied Sciences Kufstein
80 PUBLICATIONS **485** CITATIONS
SEE PROFILE

Harald Kosch
Universität Passau
266 PUBLICATIONS **1,392** CITATIONS
SEE PROFILE

World-Wide Scale Geotagged Image Dataset for Automatic Image Annotation and Reverse Geotagging

Hatem Mousselly-Sergieh
Université Lyon
20 Av. Albert Einstein
69621 Villeurbanne, France
hatem.mousselly-sergieh@insa-lyon.fr

Daniel Watzinger
Universität Passau
Innstr. 43
94032 Passau, Germany
watzinger.daniel@gmail.com

Bastian Huber
Universität Passau
Innstr. 43
94032 Passau, Germany
huber.baste@gmail.com

Mario Döller
FH Kufstein
Andreas Hoferstr. 7
6330 Kufstein, Austria
mario.doeller@fh-kufstein.ac.at

Elöd Egyed-Zsigmond
Université Lyon
20 Av. Albert Einstein
69621 Villeurbanne, France
elod.egyed-zsigmond@insa-lyon.fr

Harald Kosch
Universität Passau
Innstr. 43
94032 Passau, Germany
harald.kosch@uni-passau.de

ABSTRACT

In this paper, a dataset of geotagged photos on a world-wide scale is presented. The dataset contains a sample of more than 14 million geotagged photos crawled from Flickr with the corresponding metadata. To guarantee the spatial representativeness of the dataset, a crawling approach based on the small-world phenomena and the Flickr friendship's graph is applied. Furthermore, the noisiness of user-provided tags is reduced through an automatic tag cleaning approach. To enable efficient retrieval, photos in the dataset are indexed based on their location information using quad-tree data structure. The dataset can assists different applications, especially, search-based automatic image annotation and reverse geotagging[1].

1. INTRODUCTION

In the era of web 2, collaborative system for photo sharing become ubiquitous tools. Nowadays, an increasing number of users upload their photos, annotate them using keywords called tags and share them with each other. This led to an explosion in the amount of photos contributed to the web everyday. For instance, the photo sharing website Flickr[2] announced on their blog that more than 3,000 photos are upload every minutes. In 2011 Flickr reached 6 billion photos[3].

The availability of such amounts of user-tagged image led to a new research direction in the field of automatic image annotation, namely, search-based image annotation. In contrast to the traditional approach which employs machine learning (e.g. [5]) , search-based image annotation exploits the collective knowledge represented by user-tags to predict tags for new unlabeled images [20, 21, 19]. The idea is to determine a neighborhood[4] for the input image in a collection of already tagged images. Consequently, tags of the neighbors can be analyzed and used to annotate the input image.

Most recently, a considerable amount of user-contributed photos are assigned location information, i.e., geotagged. A geotag consists of the longitude and latitude of the location of image capture. Geotags can be automatically added to the EXIF descriptor of the image through built-in GPS receivers of modern cameras or smart phones. It is also possible to assign location information manually using an interactive map as provided by Flickr. The number of geotagged photos on the web is also increasing constantly. A study curried out 2010 by Doherty and Smeaton [4] shows that there are over 95 million geotagged photos on Flickr with a daily growth rate of around 500,000 new geotagged photos.

Geotagged images provide an additional context for search-based image annotation. The location information can be used to narrow the search space, thus, identifying the neighborhood of a to-be-annotated image can be done more efficiently (e.g. [17, 15]). Furthermore, datasets of geotagged images can also assist the task identifying the location of non-geotagged ones. This process, called reverse geotagging, exploits the different features of community photos, such as textual metadata (tags), location information (geotags), and visual features to mine the location of an input image (e.g. [1, 8]).

To support the mentioned research directions, a dataset of geotagged photos with the associated metadata is presented in this paper. The dataset is obtained from Flickr by em-

[1] download here: https://drive.google.com/folderview?id=0B-mRR4rjwHPOQUJ1dOx5aHVHVWM&usp=sharing

[2] www.flickr.com

[3] http://latimesblogs.latimes.com/technology/2011/08/flickr-reaches-6-billion-photos-uploaded.html

[4] Usually the neighborhood consists of a collection of images which are visually similar to the input image.

ploying a crawling strategy based on the small-world phenomena [14] and Flickr friendship's graph to ensure the spatial representativeness of the collected data. To improve the quality of the associated user-tags, a cleaning procedure is applied to remove noisy tags. Furthermore, to achieve efficient retrieval, the dataset is indexed based on the geographical information using the quad-tree data structure.

The rest of the paper is organized as follows. In the next section, geo-based crawling techniques as well as a subset the most used geotagged image datasets are reviewed. Our data crawling strategy, the tag cleaning approach, the applied spatial indexing method as well as diverse statistics on the created dataset are presented in section 3. The work is then concluded in section 4.

2. BACKGROUND

Creating photo datasets with the associated metadata from community contributed photos is an essential component of several research activities which aim at extracting new information from user-collective knowledge. In addition to the commonly available image metadata, such as user-tags and image titles, several efforts have been made to provide information about the location of image capture. This became feasible according to the increasing number of geotagged images shared on the web. Before we present our contribution in this regard, we discuss different strategies for creating image datasets based on geographical information and provide a compact report of the available datasets.

2.1 Geo-based Data Crawling

Crawling image data from online collections has been the subject of several research efforts. The authors in [12] propose an approach to crawl geotagged photos based on keyword search. For this purpose, photo sharing services are first queried using keywords (e.g. city names). Next, all geotagged images annotated with that keywords are retrieved. The datasets presented in [8, 11, 18, 9, 22] have been created by using the geographic query feature provided by Flickr API. The quires are built based on the geographic boundaries of specific cities or urban centers. A first effort to build world-scale photo dataset was introduced in [16]. For this purpose, the authors divide the world map into a grid of overlapping tiles. After that, the boundaries of each tile are used to query Flickr. A world-scale photo dataset is also presented in [3]. The authors propose a crawling strategy which aims at gathering photos from Flickr, so that the real spatial distribution of the data is preserved. That means, the density of photos collected from a given place should reflect the popularity of that place among photographers. The crawling method starts by randomly selecting a photo identifier from the pool of Flickr photo identifiers. Next, the uploader of that photo is identified and the corresponding geotagged photos are downloaded with the associated metadata. Additional photos are then acquired by traversing the friendship graph of the initial user to identify new users and downloading the corresponding geotagged photos. To crawl more data, the complete process is repeated by selecting a new photo identifier.

2.2 Geotagged Photo Datasets

In the recent years, a number of photo datasets which provide location information (explicitly or implicitly) have been made available for research purposes. For the Photo Annotation and Retrieval Task, *ImageCLEF* initiative[5] provides a dataset based on *MIRFlickr* [10]. It contains 1 million Flickr images with a subset of 25,000 manually annotated photos. *MIRFlickr* provides different kinds of metadata about the downloaded images, such as the EXIF files and the associated user-tags. However, by investigating the EXIF descriptors, we found out that location information are either missing or inaccurate for a large part of the photos in the dataset. *NUS-Wide* is another dataset based on Flickr [2]. It consists of 269,648 images with the associated user-tags as well as six types of low-level image features. Additionally, the dataset provide a ground-truth for 81 concepts. However, only a small part of the photos in the dataset are geotagged (around 50,000). Additional dataset of about 1 million photos was introduced in [11]. The data were crawled from Flickr and correspond to 22 *European cities*. The dataset was extended in [18] to 40 world cities with a total of about 2,23 million images. However, these datasets provide only the photos without the associated metadata. The authors of [22] provide a script for a dataset called *Paris500k*. The dataset contains more than 500 thousands photos taken in the city of Paris. A further dataset with a main focus on reverse geotagging is presented by the *MediaEval* benchmarking initiative[6]. The dataset, named *MediaEval Placing Task 2013 Data Set* [7] contains around nine million geotagged images crawled from Flickr. User tags are also provided, however, in their raw "noisy" form. Additionally, the authors did not give any information on the applied crawling strategy and the spatial representativeness of the data.

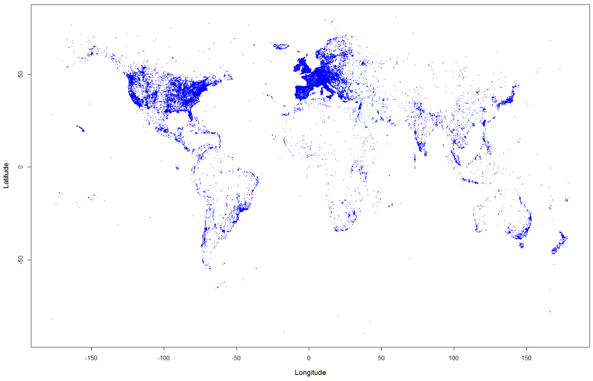


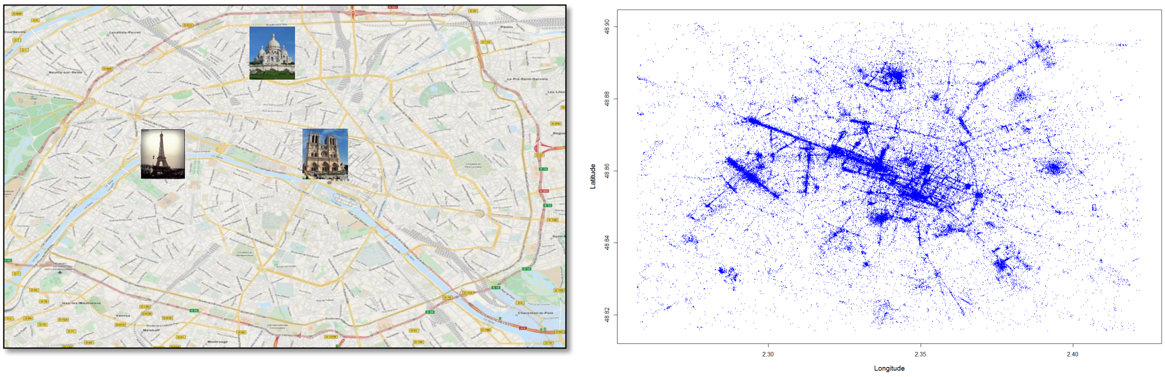
Figure 1: The geographical coordinates (latitude vs. longitude) of a sample of 300,000 images from our dataset

3. OUR DATASET

To ensure the quality of our dataset we defined the following criteria. First, the dataset should be big enough to cover the whole world map. Additionally, the data should be spatially representative. That means, the density of the data corresponding to a certain location should reflect the popularity of that location among photographers. Another aspect is the quality of the provided metadata. An important resource for metadata is user-tags. However, user-tags are inherently

[5] http://www.imageclef.org

[6] http://www.multimediaeval.org/about/



a) Paris city map with famous landmarks

b) Approximation of Paris city map using the geotags of images taken in Paris

Figure 2: Photo density in the city of Paris

noisy [13]. Therefore, they must undergo a cleaning procedure before they can be used by further applications.

The phases of creating our dataset according to the mentioned criteria are discuss in detail in the next subsections.

3.1 Spatial Representativeness

To fulfill the requirement of spatial representativeness, we followed a data crawling strategy based on Flickr's friendship graph and the principle of small-world [14]. The proposed method is inspired from [3]. However, instead of creating a random sample of photo identifiers, we generate a sample of identifiers corresponding to users resident in different places of the world. We start from an initial set of spatially well-distributed users and traverse their associated friendship graphs to extend the user set. According to the principle of small world, the final user set would contain users who have taken photos covering the whole world map and with a realistic density distribution. To achieve this, we used Flickr API to, first, create a set of Flickr users (the seed set) living in different areas of the world. The users are selected randomly and the seed set are then extended as follows. First, the friendship graph of each user in the seed set is obtained from Flickr. After that, breadth-first search is applied on the graph to acquire additional users. This process is applied recursively on the newly acquired users until a certain number of unique users is reached. Finally, for each user, the corresponding geotagged photos are crawled with the associated metadata.

During the crawling process, only photos which are defined as public by their owners are downloaded. Additionally, we applied two filtering conditions. First, we used the metadata provided by Flickr to discard images with poor geographical accuracy[7]. Second, since many applications require photos of acceptable resolution, photos of resolution below 320 × 240 pixels were also removed.

Figure 1 shows the a scatter plot of the coordinates of a sample of 300,000 photos taken from our dataset. Each image is represented by a point in a two dimensional space of longitude on the x-axis and the latitude on the y-axis. The graphic shows how the coordinates of the crawled images can approximate the world map. Moreover, dark areas indicate densely photographed places. This conforms to several studies on Flickr (e.g. [3]) which shows that certain places in Western Europe and the United States are most popular among photographers.

A closer look on the spatial distribution of the crawled photos is given in Figure 2.b. Photos taken in Paris are represented according to their geographical coordinates in the longitude-latitude space. Dense areas correspond to places which attract photographer at most. Compared to the map of Paris shown in Figure 2.a, we observe dense amounts of photos around touristic attractions, such as the city center, around Eiffel Tower and along the Seine River.

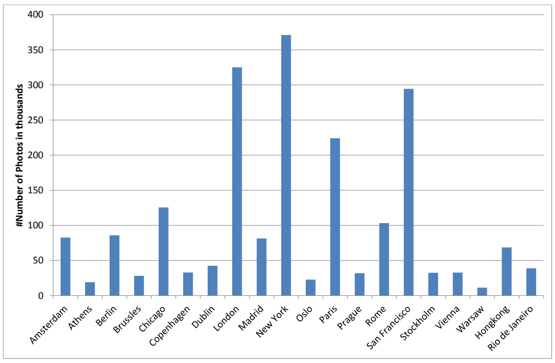


Figure 3: The number of images per city according to our dataset

[7] Flickr defines 16 different accuracy levels for the geographical coordinates of a geotagged image. The highest level 16 indicates that the location is accurate at the street-level, while the lowest value 1 corresponds to world-level. For our dataset, we set the minimum accuracy level for the downloaded images to a city-level (value 11).

We also compared our dataset to the findings of a study conducted by Crandall et al [3] in terms of the most photographed cities. Crandall et al. analyzed a dataset of 35 million Flickr photos and demonstrated that the cities New York, London, San Francisco, and Paris belong to the top photographed cities in the world and in the provided order. The same was observed in our dataset as illustrated in Figure 3.
The final dataset contains a collection of 14,1 million photos with the associated metadata. The photos were contributed by more than 200,000 users in the time period from 14/5/2000 until 01/04/2012. For each photo, the following metadata are provided: the photo identifier, the identifier of the user who uploaded the photo, the title of the photo (if existing), the list of associated user-tags, the location information represented by the longitude and the latitude, the accuracy level of the location information as defined by Flickr, the date of photo capture, the date when the photo was upload to Flickr server, and the information needed to construct the photo URL[8].

3.2 Tag Cleaning

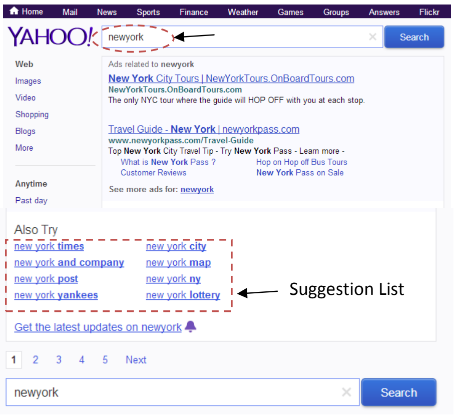
As discussed before, the dataset should also provide clean metadata. User-tags represent a main resource of metadata for describing photo semantic. However, the uncontrolled way of tag creation make tags noisy. In the following, we apply a simple tag cleaning procedures which mainly focus on addressing problems related to the syntax of the tags.

Tag Preprocessing

Before dealing with syntactic problems of user-tags, a filtering step is applied to remove tags corresponding to stop words. For this purpose, we manually identified a list of stop words. This includes non-descriptive tags, such as the words *photo*, *picture* and the like. Another kind of stop words are tags referring to technical terms, such as camera types and camera settings (e.g. canon, longexposure, d40x). Furthermore, tags specific to Flickr, e.g. *flickr.com*, *platinumheartaward*, etc. and other tags referring to dates, web services or photo editing programs are also added to the stop word list. An additional refinement step is to filter tags with low frequency. Usually, tags that are used by a small number of users are noisy since they might be too specific. Accordingly, we eliminated tags which were used by less than 5 users from the dataset. The final dataset contains 415,369 unique tags with a total occurrence of 100,791,616 and an average of 7.14 tags per photo.

Tag Syntactic Cleaning

With respect to the syntax, user-tags suffer from problems such as misspelling and syntactic variations. The latter problem arises because users use different ways to express the same term. For example, different users may annotate photos taken in New York with ”newyork”, ”new-york” or ”new york”. To deal with these problems, we developed an automatic approach based on the correction suggestions provided by Yahoo![9] search engine. For a given tag t, we use it to query Yahoo!. In the case where t is misspelled or consists of combined words, Yahoo provides proposals for related search terms (see Figure 4).



Ads related to **newyork**

New York City Tours | NewYorkTours.OnBoardTours.com
NewYorkTours.OnBoardTours.com
The only NYC tour where the guide will HOP OFF with you at each stop.

Travel Guide - **New York** | newyorkpass.com
www.newyorkpass.com/Travel-Guide
Top **New York** City Travel Tip - Try **New York** Pass - Learn more -
What is **New York** Pass ? Hop on Hop off Bus Tours
Customer Reviews **New York** Pass on Sale

See more ads for: **newyork**

new york **times** new york **city**
new york **and company** new york **map**
new york **post** new york **ny**
new york **yankees** new york **lottery**

Get the latest updates on newyork

Figure 4: Search results for the term ”newyork” according to Yahoo search engine with suggestions for related search terms

We denote the set of all suggestion sets $S = \{S_1, ..., S_n\}$. Each suggestion set $S_i \in S$ consists in turn of one or more words in a specific order $S_i = (w_1, ..., w_k)$. Next, we build the set of unique words $W = \cup_i Si = \{w_1, ...w_m\}$ as the union of all suggestion sets. After that, for each word $w \in W$ we compute the total occurrence of w, denoted as $C(w)$, over all suggestions sets. Finally, a set of terms, denoted as Corr_t, for correcting the input tag t, is determined as follows:

$$\mathrm{Corr}_t = \{w_j | C(w_j) \geqslant \theta\} \qquad (1)$$

In Equation 1, θ is a lower bound for word occurrence and can be set experimentally. We used $\theta = Max(2, 0.8 \times Max(C(w))$, that means, in order for a word to belong to the correction set, it must appear at least in 80% of the suggestions $S_i \in S$ and for more than two times.
After the correction set has been identified, a final correction term is created by determining the right order of the terms in the correction set. To do that, we used a simple technique which determines the order of the words according to their order in the majority of the suggestion sets $S_i \in S$. That is, for two words $w_1, w_2 \in \mathrm{Corr}_t$, if w_1 occurs before w_2 in the majority of the suggestion sets, then w_1 should come before w_2 in the final correction term.
In Figure 4, for example, a correction set for the input tag *newyork* can be built out of the most frequent words in the suggestions list, i.e., $\mathrm{Corr}_{newyork} = \{new, york\}$. As the word *new* occurs before the word *york* in all suggestions, the same order must be followed in the final correction term, i.e., *newyork* have to replaced by *new york*.
Table 1 shows examples of misspelled and multiple-word tags and the automatically identified corrections according to the described algorithm.

3.3 Indexing using Quad-tree

A initial processing step of applications that use geotagged images datasets (e.g search-based image annotation) is to identify images taken in a certain geographical location. To efficiently process geographic queries, the entries of the dataset have to be spatially indexed. For this purpose, we provide an approach for indexing large amounts of data using the quad-tree data structure [6].
Quad-tree is a hierarchical data structure which is based on

[8] `http://www.flickr.com/services/api/misc.urls.html`
[9] `http://www.yahoo.com/`

Original Tag	Corrected Tag
abandoned-building	abandoned buildings
abrahamlincoln	abraham lincoln
portlandmusic	portland music
greatsanddunes national-park	great sand dunes national park
sanpedrolalaguna	san pedro la laguna
enviroment	enviro**n**ment
fr**ei**nd	fr**ie**nd

Table 1: Sample user-tags acquired from Flickr (first column) automatically corrected according to our algorithm (second column)

the principle of recursive decomposition. It is wildly used for indexing two dimensional data, such as geographical coordinates. For this purpose, data points are recursively divided into four regions until a stopping condition is met. This condition is defined in terms of the maximum allowed capacity of a single quad-tree region. With a large number of data points, a direct application of the quad-tree algorithm becomes impractical. Additionally, using a relatively low maximum capacity threshold leads to immense memory requirements due to the high recursion depth[10]. To deal with this problem, we propose a method for distributing the computation of the quad-tree. Initially, we dived the world map into tiles. A tile is created only if there are photos in the dataset taken in the area specified by that tile. After that, dense tiles are further divided into sub-tiles. This process is repeated as long as the number of photos in the tile exceeds a predefined upper bound (Figure 5). In the next step, the quad-tree algorithm is applied on each tile (Figure 6). The final index consists of the boundaries of each tile as well as the corresponding quad-tree regions. The boundaries of a region are defined by the coordinates, i.e., longitude and latitude pairs, of the left bottom and the right top corners of the bounding box, respectively. To allow flexible retrieval, the index also keeps track of the neighborhood information of each quad-tree region. This can be useful when a specific quad-tree region is sparse. Accordingly, additional data points can be efficiently retrieved by extending the result set to data points of neighboring quad-tree regions.

We applied the described approach on our dataset using initial squared tiles of size 10 × 10. After that, the boundaries of each tile (the width and the height) are shrinked to the minimum possible rectangular area which contains the complete set of data points associated with the original tile. Next, tiles containing more than 300,000 photos were further divided. Figure 5 shows the results of this phase. The produced tiles show an approximation of the continents of the world. Additionally, we can see that tiles corresponding to areas of high photo density (e.g. parts of North America and Europe) are further divided into sub-tiles shown as smaller rectangles inside the corresponding tiles. Finally, we applied the quad-tree algorithm on each tile using a maximum capacity threshold of 800 data points per a quad-tree region (Figure 6).

We collected statistics about the generated tiles and the corresponding quad-trees. Indexing the collection of 14,1 million geographical coordinates resulted in 215 tiles with an average of 312 quad-tree regions per tile. Each tile contains about 65,500 data points (images) on average, however, with a large standard deviation of about 122,000. This due to the sharp differences in the density of photos from place to place. In fact, the density of photographed places follow the power law. There are very few places in the world which are frequently photographed, while quit large number of places are photographed much less (see Figure7).

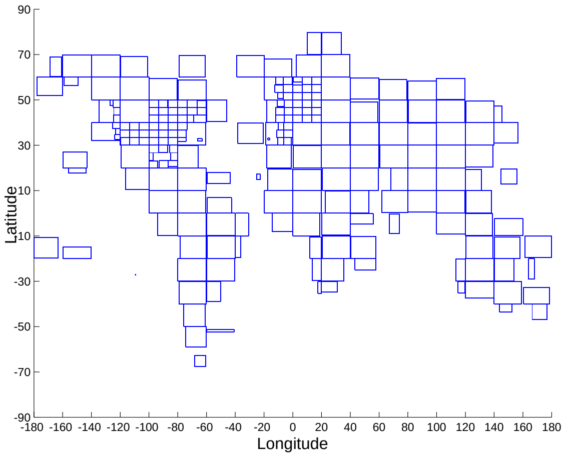


Figure 5: World map divided into tiles according to the photo density as given by our dataset. Dense tile further divided into sub-tiles

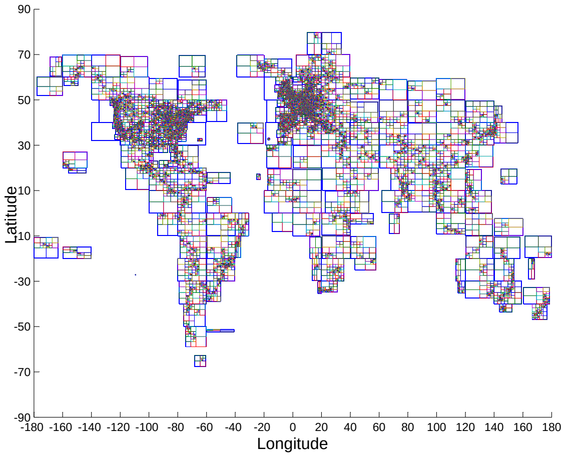


Figure 6: Quad-tree regions for our dataset. The quad-tree algorithm is applied on each tile separately to allow efficient computation

4. SUMMARY

In this paper a dataset of geotagged images on world-wide scale is presented. The dataset contains a snapshot of Flicker of 14,1 million images with the corresponding metadata. The dataset can be used to assist research on automatic image annotation as well as reverse geotagging. The representativeness of the data was achieved through a crawling approach based on Flickr friendship's graph. Additionally,

[10] On a machine with 8GB RAM and using Matlab, the maximum recursion limit of 500 was reached with a relatively small set of 300,000 data points and a maximum capacity of 2,000 data points per quad-tree region

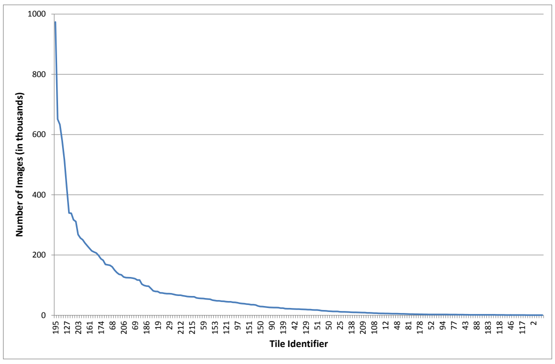


Figure 7: The number of photos per tile according to our dataset

the associated user-tags were cleaned to boost their utility. Finally, efficient retrieval can be performed using the provided spatial index which is based on the quad-tree data structure.

5. REFERENCES

[1] Cao, L., Gao, Y., Liu, Q., and Ji, R. Geographical retagging. In *Advances in Multimedia Modeling*, S. Li, A. Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, and C. Gurrin, Eds., vol. 7733 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 47–57.

[2] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y.-T. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)* (Santorini, Greece., July 8-10, 2009).

[3] Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web* (New York, NY, USA, 2009), WWW '09, ACM, pp. 761–770.

[4] Doherty, A. R., and Smeaton, A. F. Automatically augmenting lifelog events using pervasively generated content from millions of people. *Sensors 10*, 3 (2010), 1423–1446.

[5] Duygulu, P., Barnard, K., de Freitas, J. F., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer VisionâĂŤECCV 2002*. Springer, 2006, pp. 97–112.

[6] Finkel, R., and Bentley, J. Quad trees a data structure for retrieval on composite keys. *Acta Informatica 4*, 1 (1974), 1–9.

[7] Hauff, C., Thomee, B., and Trevisiol, M. Working notes for the placing task at mediaeval 2013. In *MediaEval* (2013).

[8] Hays, J., and Efros, A. A. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2008).

[9] Hollenstein, L., and Purves, R. Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, 1 (2013), 21–48.

[10] Huiskes, M. J., and Lew, M. S. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (2008), ACM, pp. 39–43.

[11] Kalantidis, Y., Tolias, G., Avrithis, Y., Phinikettos, M., Spyrou, E., Mylonas, P., and Kollias, S. Viral: Visual image retrieval and localization. *Multimedia Tools and Applications* (2011).

[12] Kessler, C., Maué, P., Heuer, J. T., and Bartoschek, T. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In *GeoSpatial semantics*. Springer, 2009, pp. 83–102.

[13] Mathes, A. Folksonomies-cooperative classification and communication through shared metadata. *Computer Mediated Communication 47*, 10 (2004).

[14] Milgram, S. The Small World Problem. *Psychology Today 2* (1967), 60–67.

[15] Moxley, E., Kleban, J., and Manjunath, B. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (2008), ACM, pp. 24–30.

[16] Quack, T., Leibe, B., and Van Gool, L. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 international conference on Content-based image and video retrieval* (2008), ACM, pp. 47–56.

[17] Sergieh, H. M., Gianini, G., Döller, M., Kosch, H., Egyed-Zsigmond, E., and Pinon, J.-M. Geo-based automatic image annotation. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (New York, NY, USA, 2012), ICMR '12, ACM, pp. 46:1–46:8.

[18] Tolias, G., and Avrithis, Y. Speeded-up, relaxed spatial matching. In *in Proceedings of International Conference on Computer Vision (ICCV 2011)* (Barcelona, Spain, November 2011).

[19] Torralba, A., Fergus, R., and Freeman, W. T. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 30*, 11 (2008), 1958–1970.

[20] Wang, X.-J., Zhang, L., Jing, F., and Ma, W.-Y. Annosearch: Image auto-annotation by search. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, IEEE, pp. 1483–1490.

[21] Wang, X.-J., Zhang, L., Liu, M., Li, Y., and Ma, W.-Y. Arista-image search to annotation on billions of web photos. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (2010), IEEE, pp. 2987–2994.

[22] Weyand, T., Hosang, J., and Leibe, B. An evaluation of two automatic landmark building discovery algorithms for city reconstruction. In *Trends and Topics in Computer Vision*. Springer, 2012, pp. 310–323.