In-depth Exploration of Geotagging Performance using Sampling Strategies on YFCC100M

Giorgos Kordopatis-Zilos
Information Technologies
Institute (ITI), CERTH
Thermi-Thessaloniki, Greece
georgekordopatis@iti.gr

Symeon Papadopoulos
Information Technologies
Institute (ITI), CERTH
Thermi-Thessaloniki, Greece
papadop@iti.gr

Yiannis Kompatsiaris
Information Technologies
Institute (ITI), CERTH
Thermi-Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

Evaluating multimedia analysis and retrieval systems is a highly challenging task, of which the outcomes can be highly volatile depending on the selected test collection. In this paper, we focus on the problem of multimedia geotagging, i.e. estimating the geographical location of a media item based on its content and metadata, in order to showcase that very different evaluation outcomes may be obtained depending on the test collection at hand. To alleviate this problem, we propose an evaluation methodology based on an array of sampling strategies over a reference test collection, and a way of quantifying and summarizing the volatility of performance measurements. We report experimental results on the MediaEval 2015 Placing Task dataset, and demonstrate that the proposed methodology could help capture the performance of geotagging systems in a comprehensive manner that is complementary to existing evaluation approaches.

CCS Concepts

•**Information systems → Information retrieval; Test collections;** *Web searching and information discovery;*

Keywords

geotagging; social media; evaluation

1. INTRODUCTION

Measuring the performance of multimedia analysis and retrieval approaches and comparing among different systems that try to solve the same problem is a highly challenging task. When carried out without proper care, such comparisons can lead to misleading conclusions and ultimately to wrong decisions regarding the selection of one approach over another in a given setting. While the ultimate and most reliable test for a system is its use and evaluation in actual operational conditions (i.e. live evaluation), this is often not feasible, too costly or very risky. As a result, performance evaluation is typically carried out with the help of test datasets, which in the case of multimedia systems typically comprise static collections of multimedia content.

Among others, there are two major issues when using a test collection to evaluate a multimedia system: a) compiling the collection may introduce certain bias in the evaluation, i.e. comprise specific kinds of content more than others, and hence favour systems that are better in analyzing the kinds of content that are dominant in the collection; b) generating a single or a few metrics to summarize the performance of a system over the whole test collection may not be sufficient to offer a nuanced understanding of the system's behaviour and performance, when presented with different kinds of content. In this paper, we propose an evaluation framework with a view to addressing the second of those issues, and focus on the problem of multimedia geotagging, i.e. estimating the geographic location of a multimedia item based on its content and metadata, to showcase the problem and the benefits of the proposed framework.

To motivate the problem at hand, let us assume that we are interested in evaluating the performance of a multimedia geotagging system in terms of *Precision at 1km range*[1], and to do this, we generate a test collection of one million images (with known geographical location, and accompanied by text metadata). Assume that out of the set of one million images, 700 thousands are located in the US, while the rest of the images are scattered around the world. It is then easy to imagine that a multimedia tagging system, which has been tuned for locations in the US will be evaluated more favourably compared to one that has been tuned for locations in Europe (a kind of *evaluation bias*). In another hypothetical scenario, assume that out of the test set, half of the images depict non-geographic scenes (e.g. cat and dog close-ups) and the accompanying text metadata are equally uninformative, e.g. tags such as `cute`, `puppy`, etc. (a kind of *evaluation noise*). In such scenarios, **producing a single performance score** (P@1km in this case) for a geotagging system **will likely give a misleading impression of the actual system performance**.

Although careful design of the test collection may mitigate or even eliminate problematic cases such as the ones described above, current practices in building test collections often necessitate the use of automation in most parts of the collection building process with a view to achieve large scale. In such scenarios, it is hard to end up with a test collection that does not suffer from problems such as evaluation bias

MMCommons'16, October 16 2016, Amsterdam, Netherlands

ISBN 978-1-4503-4515-6/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2983554.2983558

[1] This is a commonly used evaluation measure for multimedia geotagging systems, often abbreviated as P@1km, and defined in more detail in the next section.

and noise. As a result, simple schemes based on a single evaluation measure are not sufficient to capture the performance of systems under test. To address this limitation, we propose a **new evaluation framework based on the concepts of sampling strategies and performance volatility**. The proposed framework is particularly designed for the problem of multimedia geotagging, but could be adapted for different multimedia analysis and retrieval problems.

To demonstrate the value of the proposed framework in generating insights into the performance of geotagging systems, we apply it to evaluate a recently proposed method, using the MediaEval 2015 Placing Task [2] collection as benchmark, and derive very rich conclusions in addition to the ones that were possible following the official evaluation protocol established by the task organizers. The code and the generated test samples are publicly available[2].

2. BACKGROUND

2.1 Geotagging

Geotagging multimedia content is a challenging task, which is defined as the estimation of the geographic location that is depicted by a multimedia content item. Surveys of early research on geotagging and related approaches are presented in [8] and [16]. Geotagging approaches may be classified based on the modality used as input (text, visual, hybrid).

Visual: One of the first approaches based on visual content was presented by Hays et al. in [5], who utilized a collection of geotagged Flickr images as the *background collection*: for a *query image* the nearest neighbors (in terms of visual features) are retrieved from this collection and their locations are used to estimate the location of the query image. Recently, Weyand et al. [15] adopted a supervised learning approach to tackle the problem: they divided the earth's surface in multi-scale cells and trained a Deep Convolutional Neural Network (DCNN) using millions of geotagged images to classify a query image to one of these cells.

Text: A popular text-based geotagging approach relies on a geographical Language Model (LM) generated from the textual annotations that accompany user-generated geotagged images. The LM aims at linking text content to specific locations. One of the earliest works was proposed by Serdyukov et al. [10], which used a predefined grid of cells and calculated the prior probabilities for image tags. More recently, Van Laere et al. [14] built a LM by clustering a large set of geotagged images and then used the χ^2 feature selection criterion to create a vocabulary for every cluster. They introduced a similarity search technique, using Jaccard similarity. In an extension [13], they proposed different term selection techniques, utilizing kernel density estimation and Ripley's K statistic, to further improve geotagging performance.

Hybrid: Hybrid approaches combine visual features of images and their associated text metadata. Crandall et al. [4] combined image content and textual metadata at two levels of granularity, at a city level (≈100km) and at landmark level (≈100m). They trained classifiers in a relatively small set of landmarks and for a fixed set of cities. Trevisiol et al. [12] processed the textual information of a set of videos to determine their geo-relevance and to find frequent matching items. In case of lack of such information, they resorted to visual features for predicting the respective locations.

2.2 Geotagging using Language Models

Here, we briefly present the approach that will be used in the experimental section of the paper for showcasing the proposed evaluation framework. The approach was originally presented in [6]. The approach is based on a probabilistic LM, which is constructed using a training set (D_{tr}) of Flickr images as input. In particular, the tags, titles and descriptions of images are first processed into *terms* (keywords) and then used to construct the LM.

2.2.1 Language Model

Initially, the earth surface is divided in a set C of (nearly) rectangular cells of size 0.01° latitude/longitude (approximately $1km^2$ size near the equator). We construct a Language Model (LM), following an approach that is similar to [9], by creating a term-cell probability map based on the user count of each term in each cell. In particular, the term-cell probability is calculated as $p(t|c) = N_u/N_t$, where N_u is the number of users in D_{tr} that used the term t inside cell c, and N_t is the total count of different users that used the term t in all cells. Note that a user can be counted in N_t more than once. Finally, the most likely cell (mlc) is computed from the summation of the respective term-cell probabilities based on Equation 1.

$$mlc_j = \underset{c_i \in C}{\arg\max} \sum_{k=1}^{N} p(t_k|c_i) \qquad (1)$$

where, mlc_j is the most likely cell for item j, N is the total number of terms for j and $p(t_k|c_i)$ is the term-cell probability for term t_k in cell $c_i \in C$. As a result, the centre of the estimated mlc may be considered as a prior location estimation for query image j.

2.2.2 Feature Selection

To increase the robustness of the model and reduce its size, feature selection is performed using the *locality* [7] of terms as a feature selection criterion. Locality captures the geographicity of terms based on the number of different individuals that used the same term in a given location. In particular, it is computed based on Equation 2.

$$l(t) = N_t * \frac{\sum_{c \in C} \sum_{u \in U_{t,c}} |\{u' | u' \in U_{t,c}, u' \neq u\}|}{N_t^2}, \qquad (2)$$

where $l(t)$ is the locality score of term t, N_t is the total occurrences of t, C denotes all cells and $U_{t,c}$ denotes the set of users that used term t inside cell c. Only terms with non-zero locality scores are further considered by the approach.

2.2.3 Feature Weighting

Since the locality score is sensitive to term frequency, we consider it inappropriate for directly weighting terms. Alternatively, having computed the locality scores for every term, we sort them based on their scores and calculate their weights using their position in the distribution.

$$w_l = \frac{|T| - (j-1)}{|T|} \qquad (3)$$

where, w_l is the weight value of the term t on the j-th position in the distribution and $|T|$ is the total number of unique

[2] https://github.com/MKLab-ITI/multimedia-geotagging, generated sample collections can be found under `samples`, while the implementation of the sampling strategies under `src/main/java/gr/iti/mklab/mmcomms16`.

terms in the LM. This weighting approach returns values in the range $(0, 1]$.

Additionally, to capture the ambiguity of the terms, we employ the spatial entropy weighting function [6]. Spatial entropy for each term is calculated based on its probabilities over cells based on Equation 4.

$$se(t) = -\sum_{c_i \in C} p(t|c_i) \log p(t|c_i) \tag{4}$$

where $se(t)$ is the spatial entropy value of term t, and $p(t|c_i)$ is the term-cell probability of t in cell $c_i \in C$. The spatial entropy weights are generated using a Gaussian kernel over the spatial entropy values and then normalizing them with the maximum value as in Equation 5.

$$w_{se} = \frac{N(se(t), \mu, \sigma)}{max_T(N(se(t), \mu, \sigma))} \tag{5}$$

where N is the Gaussian function, and parameters μ, σ are the mean value and the variance of the entropy distribution, respectively, and are estimated from D_{tr}.

The two weights are combined using the simple linear scheme $\omega * w_{se} + (1 - \omega) * w_l$, setting ω to 0.2 through empirical assessment on a sample of 10K images. After the calculation of term weight, the term-cell probabilities in the LM scheme are multiplied with the corresponding weight.

2.2.4 *Estimation Refinement*

To ensure more accurate location prediction in finer granularities, we built an additional LM using a finer grid (cell side length of 0.001°). Having computed the *mlc* for both the coarse (0.01×0.01) and fine granularity (0.001×0.001), we apply the following estimation refinement technique: we first select the most appropriate granularity (if the *mlc* of the finer grid falls within the *mlc* of the coarse grid, then we select the former, otherwise we opt for the latter), and then produce the location estimate based on the center-of-gravity of the k most textually similar images inside the selected *mlc* $(k = 5)$, by employing Similarity Search as in [14]. The textual similarity is computed using the Jaccard similarity of the corresponding sets of terms.

2.3 MediaEval Placing Task

MediaEval is an annual benchmarking initiative that includes a number of tasks in the area of multimedia analysis and retrieval. Within its context, the Placing Task is dedicated to the problem of multimedia geotagging. Participants are required to estimate locations (in terms of latitude and longitude) of items in a provided test collection, and they are also provided with a collection to use for training. The task participants are asked to submit up to five runs, among which one should be purely text-based and another one purely visual-based. For the other three runs, participants are allowed to utilize gazetteers, external data or any additional information, but not re-crawl the test images. In terms of evaluation, the submitted runs where benchmarked based on their precision in different ranges and their median error, both of which are described in Section 3.1.

Every year the volume and the origin of the released dataset are determined by the organizers of the task. In the last two editions of the task [3, 2], the released datasets were subsets of the YFCC100M [11].

3. EVALUATION FRAMEWORK

3.1 Overview

The proposed evaluation framework employs an array of sampling strategies in order to analyze the performance of a test geotagging system on different subsets of a *reference test collection*, denoted as D_{ref}, focusing on the effect of each sampling strategy on the measured performance. Each sampling strategy is formulated as a sampling function $f : D \rightarrow D_{test}$, where the test collections D_{test} is the resulting collection of items after the application of the sampling function f on collection D.

For the evaluation of geotagging performance on a collection of images D, two measures are used: precision and median distance error. Precision is defined as the percentage of test items, for which the distance between the estimated and true location is less than R and is referred to as *Precision at range* R and denoted as $P@R$ (e.g. P@1km). For this study, P@1km is considered the most appropriate instance of $P@R$. Median distance error is defined as the median of estimation errors across all query items in the collection D in terms of the distance between the predicted and the actual location. Finally, we define a volatility score that captures the variation between the performance measures in the reference collection D_{ref} and the sampled collection D_{test}. The volatility is computed based on Equation 6.

$$s = \frac{p(D_{test}) - p(D_{ref})}{p(D_{ref})} \cdot 100 \tag{6}$$

where $p(D)$ is the performance score achieved by the test system on collection D. Note that the performance score may exceed 100 in cases of large differences in the performance between the reference collection and the sampled collection. Also, in cases where the performance on the sampled collection is worse compared to the one on reference collection, the volatility score is negative. For the present paper, the corresponding volatility scores are denoted as s_p and s_m, referring to volatility with respect to P@1km and median distance error respectively.

Once the performance score and its volatility are computed according to the different sampling strategies presented in Section 3.2, the overall geotagging performance of a system is summarized with the help of a spider-plot. This can be also used to compare different systems by overlaying the respective plots.

Discussion: The proposed approach bears some similarity to the concept of cross-validation, which is widely established among machine learning practitioners as a means of obtaining a more reliable performance estimate for a machine learning algorithm by averaging performance over different splits of a collection into training and test. In contrast, the proposed framework aims at gaining a more nuanced understanding of a geotagging system by measuring its performance over a variety of subsets that comprise multimedia content with specific characteristics.

3.2 Sampling Strategies

3.2.1 *Geographical Uniform Sampling*

An important factor that has great impact on measuring the performance of geotagging systems is the distribution of test images across the globe. Usually, the total amount of images in a geographically discrete area is proportional to its *popularity*. More precisely, the areas that cover popular places, such as tourist attractions or big cities, tend to have considerably more images in comparison to the rest of the world and hence dominate the performance measurement.

In order to compensate this effect and have a uniform representation of every place on the planet, we apply a geographical uniform sampling strategy. For this sampling strategy, the earth surface is divided in cells of size 0.1° ×0.1° (≈ 10 km ×10 km), which roughly correspond to the extent of a city. Then, the total number of items in every cell is counted and the median value of items per cell is determined. Subsequently, we randomly select a number of items from every cell, equal to the median value[3], and create a collection of items that are almost uniformly distributed across the surface of the earth. In that way, every location has approximately equal impact on the geotagging performance of a system. The sampling function can be expressed by means of Equation 7.

$$f = \{i | c_i \in C, |c_i|_s \leq \text{median}_C(|c|)\} \tag{7}$$

where c_i is the cell of item i, C is the set of all cells in D_{ref}, $|c_i|_s$ is the number of selected items in c_i in the sampled collection D_s, and $\text{median}_C(|c|)$ is the median number of items in the set of cells C in D_{ref}.

Note that the random sampling from the set of cells' items may lead to slightly different results for different runs of the same sampling strategy.

3.2.2 User Uniform Sampling

The highly skewed distribution of user contributions is another factor significantly affecting the measurement of geotagging performance. Users that post a lot more images and videos than the "average" user in the dataset have greater effect on the evaluation outcome. In particular, the way a user annotates his/her content, i.e. select their tags, may considerably affect geotagging performance. Hence, if a geotagging system is tuned to the annotation style of a few high-contributing users, it is expected to achieve considerable gains in performance that could be misleading.

To alleviate this problem, we apply a sampling strategy similar to the previous one. The collection generated by this sampling strategy is formed by randomly selecting only one item from each user. Similar to the previous sampling strategy, the results between different experiments may vary, because of the random selection of users' items. Equation 8 provides the formulation of this sampling strategy.

$$f = \{i | u_i \in U, |u_i|_s = 1\} \tag{8}$$

where u_i is the contributor (user) of item i, U is the set of users in D_{ref} and $|u|_s$ is the number of selected items in D_s from user u.

3.2.3 Text-based Sampling

For this sampling strategy, we distinguish the multimedia items based on the number of terms contained in the accompanying text (i.e. tags and title). One may expect that tags with very few terms will be harder to geotag and vice versa. Hence, it is interesting to explore geotagging performance for different sets of items, i.e. items described by a few terms versus items described by numerous terms.

To this end, we first determine the median number of terms per item and exclude all images that have less terms than this threshold. This sampling strategy is expressed by Equation 9

$$f = \{i | t_i \subset T, |t_i| \geq \text{median}_D(|t|)\} \tag{9}$$

where t_i is the set of terms of image i, T is the set of all terms in D_{ref}, $|t_i|$ is the total number of terms of image i and $\text{median}_D(|t|)$ is the median number of terms in the reference collection.

3.2.4 Text Diversity Sampling

Another sampling strategy that we devised aims at creating test samples with high diversity in terms of text annotations. This is achieved by grouping all images with similar textual content in a single bucket and using only one sample from each bucket.

For the needs of this approach, we employ the Min-Hash [1] technique to quickly estimate pairs of items with highly similar textual content. The features used are the individual terms, without any further pre-processing. Initially, we extract the set of all terms accompanying the multimedia items. Then, we create a binary term occurrence vector. Each dimension of the feature vector corresponds to a specific term. If a term is associated with an item, then the respective position in its vector is set to one. After extracting the binary feature vector, we perform hashing using the Min-Hash technique to generate a binary signature per vector. Finally, all items with the same signature are grouped together in the same bucket. The similarity estimation error used for the Min-Hash technique is 0.1.

The test collection generated by this sampling procedure is composed by one random sample image from each bucket. In this scheme, the random sampling does not have any noteworthy impact on the evaluation results, since the textual content of the items in the same bucket are almost identical. Equation 10 expresses the text diversity sampling strategy.

$$f = \{i | b_i \in B, |b_i|_s = 1\} \tag{10}$$

where, b_i is the bucket of item i, B is the set of buckets in D_{ref} and $|b_i|_s$ is the number of selected images in the sampled collection from bucket b_i.

3.2.5 Geographically Focused Sampling

We also devised a geographically focused sampling. This is implemented by classifying images based on the country/continent where they are located and generates a discrete collection for every individual country/continent.

For the needs of this sampling strategy, we utilize the *places* metadata that accompany the YFCC100M dataset and contains country and continent information about each item. Consequently, discrete collections of images can be generated based on the country/continent they belong. With this sampling strategy, we can evaluate the performance of the approach for every country/continent and record the best and worst results. The sampling function of this strategy is expressed in Equation 11.

$$f = \{i | p_i \in P\} \tag{11}$$

where p_i is the place name (i.e. country or continent) of item i and P is the set of all place names.

[3] In case a cell contains less items than the median value, then all of them are added to the collection.

3.2.6 Ambiguity-based Sampling

This sampling strategy aims at distinguishing between items located in places with ambiguous versus non-ambiguous names. Names that are used to refer to more than one cities are considered to be ambiguous. However, in many occasions there is one dominant city that is most commonly referred.

To proceed with this strategy, for every city name, we calculate its *place entropy* by utilizing the *places* metadata of the YFCC100M. For every city name, we count the frequency of the different place codes that emerge in the test collection. Eventually, the place probabilities are calculated for city names with more than one codes based on the following: $p(q|n) = N_q/N$, where N_q is the times that place code q is found in the test collection for city name n, and N is the total count of the particular city name. Having calculated the place code probabilities, we can compute the place entropy according to Equation 12.

$$pe(n) = -\sum_{q \in Q} p(q_i|n) \log p(q_i|n) \tag{12}$$

where $pe(t)$ is the place entropy value of a city with name n, $p(q_i|n)$ is the code probability of code q and $|Q|$ is the total number of different codes corresponding to a city name. For example, if an item is tagged with the term `London`, then it is most likely that this item is located in London, UK, even though there are at least seven other cities across the world that have the same name. In this example, `London` is expected to have low entropy (i.e. low ambiguity).

To form the sampled collection, we calculate the median place entropy and all items that are associated with a city name with entropy that is higher than the median form the ambiguous collection D_A. The mathematical definition of ambiguity-based sampling is given by Equation 13.

$$f_{D_A} = \{i|p_i \in D_A\} or f_{\overline{D_A}} = \{i|p_i \notin D_A\} \tag{13}$$

where p_i is the place name (i.e. city) of item i and D_A is the set of the ambiguous city names. We can either choose the ambiguous or the non-ambiguous collection for testing.

3.2.7 Visual Sampling

To take into account the depicted content of a multimedia collection, multimedia items are sampled based on their visual content. To this end, we use the *autotags* metadata of the YFCC100M that provide the visual concepts for every image in the dataset. Given these visual concepts, different test collections are built, one for each different visual concept (Equation 14).

With this sampling strategy, we evaluate the performance of the approach on the collection of every visual concept and record the best and worst performances. Furthermore, we manually identify all the visual concepts that are associated with buildings in order to create a collection of images that display buildings (expecting that these images will contain much more geographic information). This sampling strategy is expressed by Equation 15.

$$f = \{i|v_{ij} \in V, \forall j \in |v_i|\} \tag{14}$$

$$f = \{i|v_{ij} \in V, \exists j : v_{ij} \in S\} \tag{15}$$

where v_{ij} is the j-th visual concept of item i, V is the set of visual concepts and S is the set of visual concepts associated with buildings.

4. EXPERIMENTS

The reference collection for all experiments reported here is the test collection that was released by the organizers of the MediaEval 2015 Placing Task (PT) [2]. This contains 949,889 images from YFCC100M. Furthermore, to evaluate the sampling strategies we tested the approach of Section 2.2 using four different set-ups that vary in terms of the use of the additional steps described Sections 2.2.2-2.2.4 and the volume of items used for training: a) `Basic-PT`, plain LM-based location estimation (i.e. only the step described in Section 2.2.1) using the training set distributed by the PT organizers; b) `Full-PT`, full location estimation method (all steps of Section 2.2) using the training set released by PT organizers; c) `Basic-Y`, plain LM-based location estimation using the whole YFCC100M dataset (excluding content coming from users included in the test set); d) `Full-Y`, full location estimation method using the whole YFCC100M (again, excluding users that occur in the test set). The PT training set contains ≈ 4.7M media items, whereas the YFCC100M consists of ≈ 40M items (after removing the ones coming from users that also occur in the test set).

4.1 Baseline Performance

The performance of the different set-ups of the approach are presented in Table 1 in terms of precision at three different ranges (100m, 1km, 10km) and median distance error. As expected, the best results in the reference collection are reported by the set-up using all proposed refinement steps and the YFCC100M dataset for training. Additionally, it is noteworthy that the performance of `Full-PT` is highly competitive with `Full-Y`, given that it uses only 10% of items for training. Both set-ups that use the plain LM-based approach (`Basic-PT/Y`) perform considerably worse.

Table 1: Geotagging precision (%) and median distance error (km) of the four set-ups of the approach on the MediaEval 2015 Placing Task test set.

	`Basic-PT`	`Full-PT`	`Basic-Y`	`Full-Y`
P@100m	0.64	6.42	0.69	**7.72**
P@1km	21.78	24.61	23.43	**27.36**
P@10km	37.67	43.68	39.69	**46.75**
m. error	342	57	240	**22**

4.2 Evaluation using Sampling Strategies

4.2.1 Geographical Uniform Sampling

Figure 1 illustrates the initial distribution of items in the test set. The dots represent the cells that the surface has been divided in, and their color corresponds to the number of items per cell. Cells with > 100 images are displayed in brown red, and as the number of items diminishes the color gradually becomes blue. Geographical uniform sampling is performed by randomly selecting three items from each cell (which is equal to the median number of items per cell).

Table 2 presents the evaluation results using the geographical uniform sampling. It is evident that this strategy has significant impact on the geotagging problem since the performance of all set-ups dropped sharply in comparison to the reference collection. Volatility scores for P@1km fluctuate between -51.3 to -58.7. For median distance error, volatility is much higher, reaching 1163% for `Full-Y`. It is noteworthy that `Basic-Y` reports better P@1km than `Full-PT` in con-

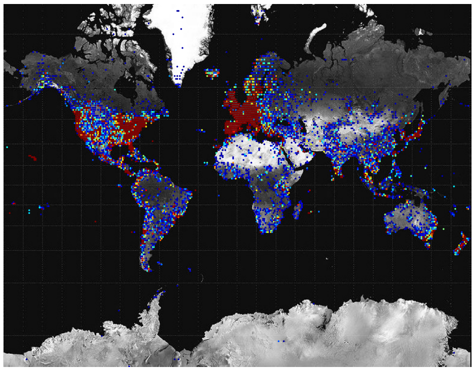


Figure 1: Geographical distribution of items in the reference PT test set.

Table 2: Evaluation of four set-ups using geographical uniform sampling.

	Basic-PT	Full-PT	Basic-Y	Full-Y
P@1km	9.82	10.16	11.33	13.32
m. error	657	399	522	278
volatility s_p	-54.9	-58.7	-51.6	-51.3
volatility s_m	92	600	117	1163

trast to the reference collection, which indicates that a large training set is beneficial to ensure satisfactory performance across a geographically balanced test set.

4.2.2 *User Uniform Sampling*

The results of employing user uniform sampling are summarized in Table 3. This sampling strategy equally affects P@1km for all set-ups, leading to slightly worse results. Instead, it has considerable negative impact on the median distance error, which is more pronounced for the refined (`Full`) set-ups. This indicates that those set-ups may have benefited from being tuned to geotag more accurately the images and videos of frequently contributing users.

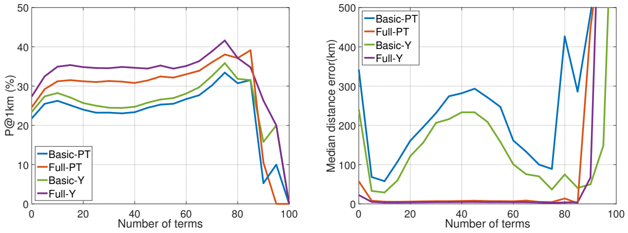
Table 3: Evaluation using user uniform sampling.

	Basic-PT	Full-PT	Basic-Y	Full-Y
P@1km	19.32	21.68	20.63	24.04
m. error	479	186	522	105
volatility s_p	-11.3	-11.9	-12.0	-12.1
volatility s_m	40	226	118	377

4.2.3 *Text-based Sampling*

Figures 2(a) and 2(b) depict the performance of the approach relative to the number of terms per item in terms of P@1km and median distance error, respectively. Overall, the geotagging performance of the approach improves as the test set is increasingly composed of items annotated with many terms. The basic set-ups are more sensitive to this sampling strategy as their performance deteriorates as the number of terms per item increases from 10 to 45. The performance of all set-ups reaches a peak for 75 number of terms per item, and then drops sharply potentially indicating that items with more terms are tagged inappropriately or in a manner that is confusing geotagging systems.

Furthermore, the median number of terms per item in D_{ref} was determined to be equal to seven. Table 4 presents the corresponding results when sampling items with a number of terms equal to or larger than the median. This sampling was found to lead to a collection that is easier to geotag, which is expected given the larger amount of available text information.



(a) (b)

Figure 2: (a) P@1km - number of terms/item, (b) median error - number of terms/item.

Table 4: Evaluation using text-based sampling.

	Basic-PT	Full-PT	Basic-Y	Full-Y
P@1km	26.13	30.33	28.08	33.81
m. error	52.4	6.4	26	4.2
volatility s_p	20.0	23.2	19.8	23.6
volatility s_m	-84.7	-88.8	-89.2	-80.9

4.2.4 *Text Diversity Sampling*

As explained in Section 3.2.4, the Min-Hash technique was used to group images into buckets based on their text similarity. The total number of generated buckets was 478,817. The sampled test collection was formed by randomly selecting one item per bucket. The corresponding evaluation results are presented in Table 6. In terms of P@1km, all set-ups exhibit similar behaviour, overall benefiting as a result of the employed sampling strategy.

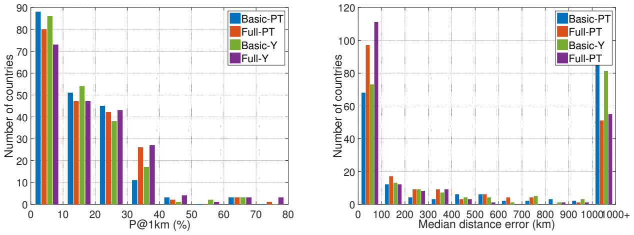
Table 5: Evaluation using text diversity sampling.

	Basic-PT	Full-PT	Basic-Y	Full-Y
P@1km	27.31	31.13	29.37	34.68
m. error	35	5.9	17	3.9
volatility s_p	25.4	26.5	25.4	26.8
volatility s_m	-89.8	-89.6	-92.9	-82.3

4.2.5 *Geographically Focused Sampling*

Here, we study the performance of the approach at country/continent level. Figures 3(a) and 3(b) illustrate the histograms of P@1km and median error over all (201) countries. It is evident that for a large number of countries, geotagging performance is very low (P@1km<10%, median error>1000km), and `Basic` set-ups seem to suffer most from this issue. However, there are also several countries for which geotagging performance is very high, with `Full` set-ups being associated with such high scores.

Further results are provided in Table 6, which presents the performance of `Full-Y` for different continents, and for the top and bottom three countries (ranked by P@1km). The best geotagging performance is recorded in Europe. Note that in America and Australia, even though the P@1km is in a medium range, the median distance error is very high. A possible explanation for this is the fact that many cities in these continents share the same name with European big cities (which actually motivated the sampling strategy in section 3.2.6), hence leading to a number of cases of very



(a) (b)

Figure 3: Histograms of (a) P@1km and (b) median error, by number of countries.

high median error. Finally, the countries with the best results are small European countries (where even the name of the country may lead to an estimate with relatively low median error). Instead, the lowest scores are recorded in large countries (`Pakistan`, `Paraguay`, `Angola`), where there is relatively scarce data for training.

Table 6: Geotagging performance per continent and for sample countries.

Continent	P@1km	m. error	occur.
America	22.93	271	436,010
Europe	35.63	3.49	375,879
Asia	19.39	51	100,857
Australia	21.24	1341	21,798
Africa	13.78	216	14,722

Country	P@1km	m. error	occur.
Liechtenstein	72.04	0.46	93
San Marino	71.79	0.61	39
Vatican City	70.40	51	527
Pakistan	2.37	3909	295
Paraguay	1.35	7496	148
Angola	0.0	6002	49

4.2.6 *Ambiguity-based Sampling*

Figure 4 depicts the P@10km (which is considered the most appropriate range to evaluate geotagging performance at city scale) in relation to the place entropy of city names. One may observe a negative correlation between place entropy and geotagging precision, i.e. city names that have low entropy (less ambiguous) tend to be geotagged with higher accuracy. Another observation is that city names with high place entropy have greater variance in their estimations. The median value of place entropy is highlighted with a red line and is equal to 0.5. The total amount of city names that are considered ambiguous (i.e. have entropy >0.5) is equal to 1579 and the portion of items taken in these cities corresponds to approximately 10% of the reference collection.

Table 7 presents the geotagging results in the *ambiguous* collection (i.e. items tagged with city names that are ambiguous given the above definition). As expected, the volatility scores for this collection are very high, with a notable case on the `Full-Y` set-up, in which median error volatility reaches 2268%. Volatility in terms of P@1km is moderate ($\approx$ 25%). Such results indicate that it may be worth investing in geotagging approaches that perform location disambiguation.

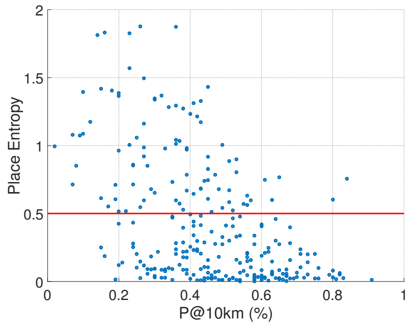


Figure 4: Place entropy - P@10km (%) for city names with more than 100 occurrences.

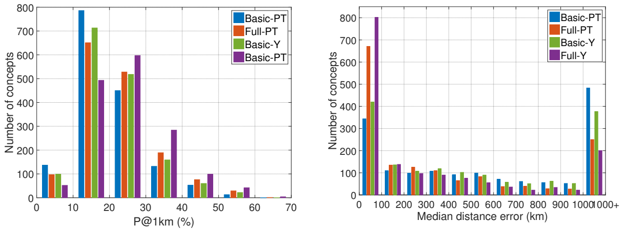
Table 7: Evaluation on ambiguity-based sampling.

measures	`Basic-PT`	`Full-PT`	`Basic-Y`	`Full-Y`
P@1km	16.27	17.98	17.87	20.77
m. error	940	562	770	521
volatility s_p	-25.3	-26.9	-23.7	-24.1
volatility s_m	175	886	220	2268

4.2.7 *Visual Sampling*

Figures 5(a) and 5(b) illustrate the histograms of P@1km and median error over the visual concepts depicted by the items. A large number of visual concepts (especially for the `Full` set-ups) are associated with median errors in the 0-100km range, while there is also a considerable number of visual concepts, for which the median error is very high (>1000km). The differences with respect to the number of these concepts between `Basic` and `Full` are noteworthy.

Table 8 presents a few examples of visual concepts, for which the sampling strategy had significant impact. The best results are achieved for concepts associated with buildings and landmarks (e.g. `capitol`, `coliseum` and `cathedral`). On the other side of the spectrum, we find visual concepts with virtually no geographical information (e.g. `frying pan`, `kitten`, `highchair`). To further explore this observation, we manually constructed a set of 120 visual concepts that are related to buildings and then applied sampling on the reference collection using this set of concepts. The evaluation results on this collection are presented in Table 9. As expected, geotagging performance is significantly improved, especially in terms of median error where even in the basic runs it is <10km. Such a sampling strategy is therefore considered as a valid means of generating test collections for visual-only geotagging methods.



(a) (b)

Figure 5: Histograms of (a) P@1km and (b) median error, by number of visual concepts.

Table 8: Top and bottom three visual concepts in terms of geotagging precision.

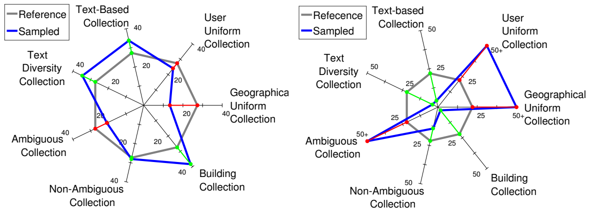
Concept	P@1km	m. error	occur.
capitol	60.44	0.48	2040
coliseum	59.54	0.54	5450
cathedral	59.06	0.57	1630
frying pan	6.61	2125	1013
kitten	5.47	3087	2595
highchair	3.38	3229	1213

Table 9: Evaluation using visual sampling.

	`Basic-PT`	`Full-PT`	`Basic-Y`	`Full-Y`
P@1km	31.76	35.30	33.92	38.35
m. error	8.6	3.6	5.5	2.6
volatility s_p	45.9	43.4	44.8	40.1
volatility s_m	-97.5	-93.7	-97.7	-88.2

4.2.8 *Summarization of Evaluation*

Figures 6(a) and 6(b) employ spider plots to summarize the performance of the `Full-Y` set-up for all sampling strategies that generate a single collection of items. The strategies, for which the overall geotagging performance improved include the text-based and text diversity sampling, the non-ambiguous sampling and the building sampling. On the other hand, the uniform sampling strategies lead to collections that are more challenging for geotagging systems.



(a) (b)

Figure 6: Evaluation summary of (a) P@1km and (b) median error, for the `Full-Y` set-up. Volatility is drawn in green and red on performance improvement and deterioration respectively.

5. CONCLUSIONS

The paper presented a novel evaluation framework for multimedia geotagging based on a number of sampling strategies that aim at generating test collections with specific features that are expected to challenge different aspects of the tested geotagging systems. Comprehensive experiments on a state-of-the-art text-based approach offered several insights into its performance and demonstrated the value of the framework as a complementary means of evaluation. In other words, geotagging systems should not only be benchmarked based on their absolute performance in terms of median error or P@1km, but also based on their resilience with respect to the sampling strategies.

To conclude, the evaluation framework is not inextricably linked to the evaluation of geotagging systems; hence, future work could focus on adapting the proposed framework to other multimedia problems, with concept detection and multimedia retrieval, being obvious candidates.

6. ACKNOWLEDGMENTS

This work is supported by the REVEAL and InVID projects, partially funded by the European Commission under contract numbers 610928 and 687786.

7. REFERENCES

[1] A. Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997.

[2] J. Choi, C. Hauff, O. V. Laere, and B. Thomee. The placing task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, 2015.*, 2015.

[3] J. Choi et al. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of GeoMM Workshop*, pages 27–31, 2014.

[4] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770. ACM, 2009.

[5] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, pages 1–8. IEEE, 2008.

[6] G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. Geotagging social media content with a refined language modelling approach. In *Proceedings of PAISI 2015*, pages 21–40, 2015.

[7] G. Kordopatis-Zilos, A. Popescu, S. Papadopoulos, and Y. Kompatsiaris. CERTH/CEA LIST at mediaeval placing task 2015. In *MediaEval*, 2015.

[8] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision: a survey. *MTAP*, 51(1):187–211, 2011.

[9] A. Popescu. CEA LIST's participation at mediaeval 2013 placing task. In *MediaEval*, 2013.

[10] P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd ACM SIGIR conference*, pages 484–491. ACM, 2009.

[11] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, Jan. 2016.

[12] M. Trevisiol, H. Jégou, J. Delhumeau, and G. Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In *Proceedings of the 3rd ICMR*, pages 1–8. ACM, 2013.

[13] O. Van Laere, J. Quinn, S. Schockaert, and B. Dhoedt. Spatially aware term selection for geotagging. *IEEE transactions on Knowledge and Data Engineering*, 26(1):221–234, 2014.

[14] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. ICMR '11, pages 48:1–48:8, New York, NY, USA, 2011. ACM.

[15] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. *arXiv preprint arXiv:1602.05314*, 2016.

[16] Y.-T. Zheng, Z.-J. Zha, and T.-S. Chua. Research and applications on georeferenced multimedia: a survey. *MTAP*, 51(1):77–98, 2011.