Finding locations of Flickr resources using language models and similarity search

Olivier Van Laere
Department of Information Technology, IBBT
Ghent University, Belgium
olivier.vanlaere@ugent.be

Steven Schockaert*
Dept. of Applied Mathematics and Computer Science
Ghent University, Belgium
steven.schockaert@ugent.be

Bart Dhoedt
Department of Information Technology, IBBT
Ghent University, Belgium
bart.dhoedt@ugent.be

ABSTRACT

We present a two-step approach to estimate where a given photo or video was taken, using only the tags that a user has assigned to it. In the first step, a language modeling approach is adopted to find the area which most likely contains the geographic location of the resource. In the subsequent second step, a precise location is determined within the area that was found to be most plausible. The main idea of this step is to compare the multimedia object under consideration with resources from the training set, for which the exact coordinates are known, and which were taken in that area. Our final estimation is then determined as a function of the coordinates of the most similar among these resources. Experimental results show this two-step approach to improve substantially over either language models or similarity search alone.

Categories and Subject Descriptors

H.4 [**INFORMATION SYSTEMS APPLICATIONS**]: Miscellaneous; H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital libraries

General Terms

Experimentation, Measurement

Keywords

Geographic information retrieval, Language models, Semi-structured knowledge, Geo-annotation

1. INTRODUCTION

Web 2.0 systems such as Flickr bring structure in collections of shared multimedia objects by taking advantage of both structured and unstructured forms of metadata. Unstructured metadata is mainly available in the form of tags, i.e. short (but otherwise unconstrained) textual descriptions that are provided by users, although in the case of Flickr, only owners can add tags. Such tags help users to organize the resources they find interesting or to otherwise facilitate retrieval of such resources (by themselves or by others) in the future [2]. In the case of photos and videos, most of the structured metadata is provided automatically by the camera, without any involvement of the user. These types of metadata usually include the type of camera, the settings that were used (e.g. aperture, focal distance, etc.) and the time and date. In a limited number of cases, cameras also provide geographic coordinates, using a built-in or externally connected GPS device. Flickr additionally offers the possibility of manually indicating on a map where a photo was taken.

The availability of location metadata is important for at least two reasons. First, it allows users to easily retrieve photos or videos that were taken at a particular location, e.g. by explicitly supporting spatial constraints in queries [13], or by displaying the resources on a map which users can explore [14]. Second, by analyzing the correlation between geographic location and the occurrence of certain tags, we may discover geographic knowledge beyond what is usually described in gazetteers [5, 7]. As a result of these considerations, and given that only a small fraction of Flickr resources are currently geo-annotated, there has been a recent interest in techniques that could automatically estimate the geographic location of photos and videos [4]. More generally, there seems to be a trend towards leveraging user-contributed, unstructured information to structured, semantic annotations, e.g. automatically completing Wikipedia infoboxes [16] or building ontologies from user tags [12].

Several kinds of information are available to estimate the geographic location of a photo or video: visual features, user profiles, and tags. Visual features may be useful to recognize certain types of landmarks, or to differentiate photo or videos that were taken e.g. at the beach from resources taken in a city center. In general, however, visual information alone is not likely to be sufficient for determining a specific location. Similarly, user profiles may be useful to introduce a bias (e.g. users are more likely to take photos closer to the place where they live), but they do not contain sufficient information to pinpoint where a photo or video was taken. In this paper, we exclusively focus on the third type of available information, viz. the tags associated with a resource. Indeed, before the value of visual features or user

*Postdoctoral Fellow of the Research Foundation – Flanders (FWO).

ICMR '11, April 17-20, Trento, Italy

profiles for this task can be assessed, in our opinion, a more thorough understanding is needed of the kind of geographic information that can be extracted from tags.

To estimate the location of a multimedia object based on its tags, three natural strategies present themselves. First, we may use gazetteers to find the locations of those tags that correspond to toponyms. Although intuitive, this strategy has proven to be particularly challenging in practice, among others due to the fact that no capitalization occurs in tags, making it difficult to identify the toponyms (e.g. *nice* vs. *Nice*), as well as due to the high ambiguity of toponyms and the limited amount of context information that is available for disambiguation. Second, we may interpret the problem of georeferencing as a classification problem, by partitioning the locations on earth into a finite number of areas. Standard language modeling approaches can then be used to determine the most likely area for a given resource, represented as its set of tags. This method eliminates the problem of determining which tags are toponyms, or any form of (explicit) disambiguation. A drawback, however, is that it results in an entire area, rather than a precise coordinate. The more areas in the partition, the more fine-grained our conclusion will be, but, the higher the chances of classification error become. Third, we may resort to similarity search, and estimate the location of a given resource as a weighted average of the locations of the most similar objects in our training set, e.g. using a form of similarity that is based on the overlap between tag sets. In this case, we do obtain precise coordinates, but the performance of the method may be limited by the fact that it treats spatially relevant tags in the same way as others. For instance, a resource tagged with *paris,bridge* will be considered as similar to a resource tagged with *london,bridge* as to a resource tagged with *paris,cathedral*. In this paper, we propose to combine the best of the latter two strategies: first use a classifier to find the most likely area in which a photo or video was taken, and then use similarity search to find the most likely location within that area.

We have participated in the Placing Task of the 2010 MediaEval benchmarking initiative [4] using a system based on this two-step approach. Our system came out best, localizing about 44% of the videos in the test collection within 1km of their true location. In this paper, we present the details of our system, and we analyze which aspects are responsible for its performance, focusing on two crucial points. First, we stress the importance of combining classification (e.g. using language models) with interpolation (e.g. using similarity search), revealing that neither method alone is capable of producing equally good results. Second, we analyze the influence of user-specific tags. In particular, in case of the Placing Task, it turns out that most of the users that own a video from the test collection also own one or more photos in the training data: among the 4576 test videos with at least one tag, 923 different users appear of whom 873 own at least one photo in the training set. We analyze to what extent the availability of such previous geo-annotations by the same user influences the performance of the system.

The paper is structured as follows. First, we detail the nature of the data sets that have been used, as well as the preprocessing methods that were applied. The subsequent two sections individually consider the two strategies that lie at the basis of our system: finding the most plausible area, using a standard language modeling approach, and finding the most likely location within that area, using similarity search.

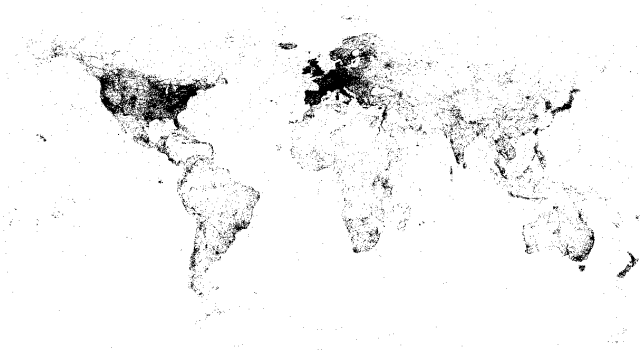

Figure 1: Plot of all the photos in the training set

Next, in Section 5 we explain how these two methods can be combined, and show that this combination performs better than the two components on which it is based. Finally, we provide an overview of related work and conclude.

2. DATA ACQUISITION AND PREPROCESSING

As training data, we used a collection of 8 685 711 photos, containing the 3 185 258 georeferenced Flickr photos that were provided to participants of the Placing Task, together with an additional crawl of 5 500 368 georeferenced Flickr photos. In addition to the coordinates themselves, Flickr provides information about the accuracy of coordinates as a number between 1 (world-level) and 16 (street level). When crawling the additional data, we only crawled Flickr photos having an accuracy of at least 12, to ensure that all coordinates were meaningful w.r.t. within-city location. Once retrieved, photos that did not contain tags or valid coordinates were removed from the collection. Next, we ensured that at most one photo was retained in the collection with a given tag set and user name, in order to reduce the impact of bulk uploads [14]. Once filtered, the remaining dataset contained 3 271 022 photos. A visual representation of this dataset is shown in Figure 1.

The test videos provided for the Placing Task contain videos that are part of bulk uploads, in the sense that some videos were uploaded around the same time as some photos in the training set by the same user, often resulting in a very high similarity between the tag set of the corresponding videos and photos. To avoid any undesirable effects of bulk uploads in our evaluation, we crawled a collection of 10 000 Flickr videos that have been uploaded later than the most recent photo from the training set. We furthermore restricted ourselves to videos with an accuracy level of 16, to ensure that our gold standard was as accurate as possible. This data set was then split into 7 400 videos that are owned by a user who also has at least one photo in our training set, and 2 600 videos by users who do not appear in the training set.

Next, the locations of the photos in the training set were clustered in a set of disjoint areas $\mathcal{A}$ using the k-medoids algorithm with geodesic distance, considering a varying number of clusters k. We consider ten different resolutions and thus ten different sets of areas $\mathcal{A}_k$. The datasets were clustered into 50, 500, 2 500, 5 000, 7 500, 10 000, 12 500, 15

000, 17 500 and 20 000 disjoint areas respectively.

Subsequently, a vocabulary V consisting of 'interesting' tags is compiled, which are tags that are likely to be indicative of geographic location. We used χ^2 feature selection to determine for each area in $\mathcal{A}$ the m most important tags. Let $\mathcal{A}$ be the set of areas that is obtained after clustering. Then for each area a in $\mathcal{A}$ and each tag t occurring in photos from a, the χ^2 statistic is given by:

$$\chi^2(a,t) = \frac{(O_{ta} - E_{ta})^2}{E_{ta}} + \frac{(O_{t\bar{a}} - E_{t\bar{a}})^2}{E_{t\bar{a}}} + \frac{(O_{\bar{t}a} - E_{\bar{t}a})^2}{E_{\bar{t}a}} + \frac{(O_{\bar{t}\bar{a}} - E_{\bar{t}\bar{a}})^2}{E_{\bar{t}\bar{a}}}$$

where O_{ta} is the number of photos in area a in which tag t occurs, $O_{t\bar{a}}$ is the number of photos outside area a in which tag t occurs, $O_{\bar{t}a}$ is the number of photos in area a in which tag t does not occur, and $O_{\bar{t}\bar{a}}$ is the number of photos outside area a in which tag t does not occur. Furthermore, E_{ta} is the number of occurrences of tag t in photos of area a that could be expected if occurrence of t were independent of the location in area a, i.e. $E_{ta} = N \cdot P(t) \cdot P(a)$ with N the total number of photos, $P(t)$ the percentage of photos containing tag t and $P(a)$ the percentage of photos that are located in area a, i.e.:

$$P(a) = \frac{|X_a|}{\sum_{b \in \mathcal{A}} |X_b|} \quad (1)$$

where, for each area $a \in \mathcal{A}$, we write X_a to denote the set of images from our training set that were taken in area a. Similarly, $E_{t\bar{a}} = N \cdot P(t) \cdot (1 - P(a))$, $E_{\bar{t}a} = N \cdot (1 - P(t)) \cdot P(a)$, $E_{\bar{t}\bar{a}} = N \cdot (1 - P(t)) \cdot (1 - P(a))$. The vocabulary V was then obtained by taking for each area a, the m tags with highest χ^2 value. In the default configuration of our system, the m values are 640 000 for the coarsest clustering, 6 400, 256, 64, 28, 16, 10, 7, 5 for the intermediate resolutions and 4 for the finest clustering level. This choice of features ensures that the language models, introduced next, require approximately the same amount[1] of space for each clustering level. In Section 6, we will analyze the impact of the choice of the m values.

3. LANGUAGE MODELS

3.1 Outline

Given a previously unseen resource x, we try to determine in which area x was most likely taken by comparing its tags with those of the images in the training set. Specifically, using standard generative unigram language modeling, the probability of area a, given the tags that are available for resource x is given by

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a) \quad (2)$$

where we identify the resource x with its set of tags. The prior probability $P(a)$ of area a can be estimated using maximum likelihood, as in (1), which means that in absence of other information, resources are assigned to the area containing the largest number of photos from the training set.

[1]Space requirements increase quadratically with the number of clusters.

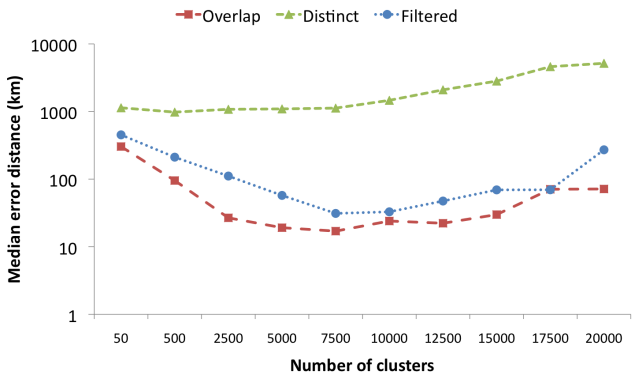

Figure 2: Median error between the medoid of the found cluster and the true location of the videos in the test set.

To obtain a reliable estimate of $P(t|a)$, some form of smoothing is needed, to avoid a zero probability when x is associated with a tag that does not occur with any of the photos in area a from the training set. We have experimented with Laplace smoothing, Jelinek-Mercer smoothing, and Bayesian smoothing with Dirichlet priors, the latter yielding the best results in general (with Jelink-Mercer producing similar results). These findings conform to experimental results in other areas of information retrieval [17], and to earlier work on georeferencing Flickr photos [14]. Specifically, using Bayesian smoothing with Dirichlet priors, we take:

$$P(t|a) = \frac{O_{ta} + \mu \left(\frac{\sum_{a' \in \mathcal{A}} O_{ta'}}{\sum_{a' \in \mathcal{A}} \sum_{t' \in V} O_{t'a'}} \right)}{(\sum_{t' \in V} O_{t'a}) + \mu}$$

where, as before, we write O_{ta} for the number of occurrences of term t in area a, and V is the vocabulary (after feature selection). The parameter μ takes a value in $]0, +\infty[$ and was set to 1750, although good results were found for a large range of values. The area a_x assigned to resource x is then the area maximizing the right-hand side of (2):

$$a_x = \arg\max_{a \in \mathcal{A}} P(a) \cdot \prod_{t \in x} P(t|a) \quad (3)$$

Thus an area is found which is assumed to contain the true location of x. It may be useful to convert this area to a precise location, e.g. for comparison with other methods. To this end, an area a can be represented as its medoid $med(a)$:

$$med(a) = \arg\min_{x \in a} \sum_{y \in a} d(x,y) \quad (4)$$

where $d(x,y)$ represents the geodesic distance. Note that the medoid is the most central element of an area, rather than its center-of-gravity. The latter is avoided here because it is too sensitive to outliers.

3.2 Experimental results

Whether or not (4) provides a good estimation depends on the number of clusters that are considered. If this number is too small, the clusters will be too coarse, and the medoid will not be a good approximation of the true location in general. If this number is too large, however, the chances of classification error increase. Thus there is a trade-off to be

found, as can clearly be seen in Figure 2. This figure depicts the median error that was obtained for a variety of cluster sizes, i.e. the median of the geodesic distance between the medoid of the cluster that was found by (3) and the true location. The figure reports the results of three experimental set-ups: one experiment considers the 7 400 videos whose owner appears in our training set (*Overlap*), one experiment considers the results for these same videos when the photos from these video owners have been excluded from the training set (*Filtered*), and one experiment considers the 2 600 videos whose owners are distinct from the owners of the photos in the (complete) training set (*Distinct*).

Regarding the influence of previously geo-annotated photos by the same user, the bad performance of the *Distinct* experiment is particularly noticeable. Closer inspection of the results has revealed that the bad results are to a large extent due to the fact that the videos in the corresponding test set have less (and less informative) tags. For instance, while the average number of tags per video is 9.39 for the *Overlap* experiment, we only have 5.92 tags on average for the videos of the *Distinct* experiment. We may speculate that users owning a larger number of resources tend to put more effort in accurately tagging these resources. As the users of the videos in the *Overlap* experiment own photos as well as videos, they are more likely to belong to this latter category. The *Filtered* experiment confirms this intuition, showing that the mere lack of geo-annotated objects by the same user has a much milder impact, although the optimal median error is still worse by almost a factor two. This suggests that the number of (good) tags has a much stronger influence than the presence or absence of geo-annotated objects by the same user. To test this hypothesis, we have separately evaluated those videos that contain a given number of tags, starting from a combined test set containing all 10 000 videos. The results, which are shown in Figure 3, clearly show that videos with more tags also tend to contain more descriptive tags and can therefore be more accurately localized. However, the results for videos with more than 10 tags are, somewhat surprisingly, worse than those for videos with 6 to 10 tags. This appears to be due to the fact that among the videos with more than 10 tags, many contain tags that have not been manually added, e.g. *taxonomy:phylum=chordata*. In particular, we found that 9.25% of all tag occurrences contain a ':' in the [11,75] category, as opposed to 0.45% in the [6,10] category. Clearly, the assumption that the number of tags provides an indication of how much effort the user has spent to describe the video no longer applies when tags are added automatically. Figure 4 shows that the same conclusions can be drawn, when restricted to the videos from the *Distinct* set-up, providing evidence that it is indeed the lack of appropriate tags that cause the overall results of the *Distinct* and *Overlap* configurations to be so different.

For the *Overlap* experiment, the optimal median error of 17.02 km is obtained when using 7 500 clusters, for the *Distinct* experiment, the optimal median error of 979.86 km is obtained when using 500 clusters, and for the *Filtered* experiment, we again need 7 500 clusters to obtain the optimal median error of 31.10 km. The lower optimal number of clusters in the case of *Distinct* suggests that the less informative the tags of a given video are, the coarser the clustering should ideally be. This is also confirmed by the results in Figure 3 which show the optimal number of clusters to be

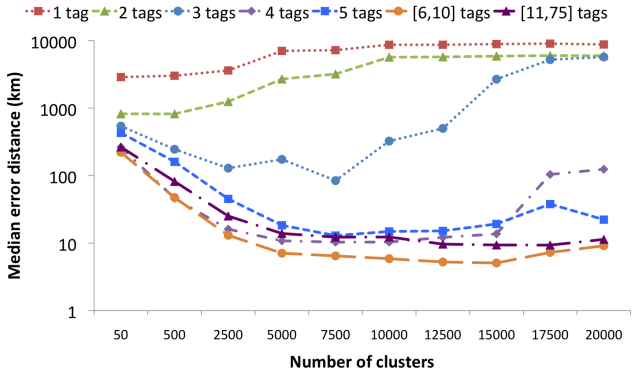

Figure 3: Median error between the medoid of the found cluster and the true location, each time using all test videos containing a given number of tags.

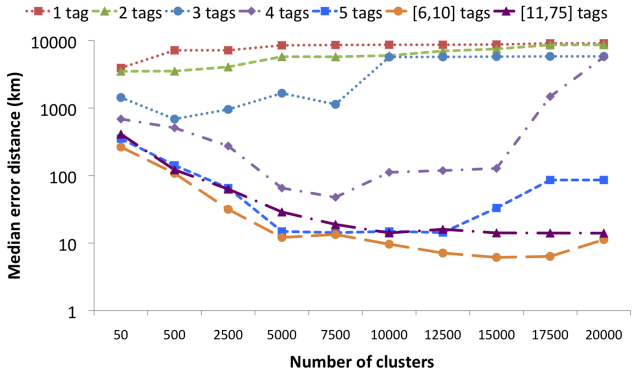

Figure 4: Median error between the medoid of the found cluster and the true location, using only the test videos from the *Distinct* set-up containing a given number of tags.

50 for photos with 1 tag (2876.46 km), 500 for photos with 2 tags (820.61 km), 7 500 for photos with 3 (84.33 km), 4 (10.32 km), or 5 (12.92 km) tags, 15 000 for photos with 6 to 10 tags (5.07 km), and 17 500 for photos with more than 10 tags (9.33 km).

4. SIMILARITY SEARCH

4.1 Outline

Rather than converting the problem at hand to a classification problem, a more direct strategy to find the location of a photo or video x consists of identifying the photos from the training set that are most similar to x, and estimate the location of x by averaging these locations. Specifically, let $y_1, ..., y_k$ be the k most similar photos from our training set. We then propose to estimate the location of x as a weighted center-of-gravity of the locations of $y_1, ..., y_k$:

$$loc(x) = \frac{1}{k} \sum_{i=1}^{k} sim(x, y_i)^{\alpha} \cdot loc(y_i) \qquad (5)$$

where the parameter $\alpha \in]0, +\infty[$ determines how strongly the result is influenced by the most similar photos only. The

similarity $sim(x, y_i)$ between resources x and y_i was quantified using the Jaccard measure:

$$s_{jacc}(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

where we identify a resource with its set of tags *without feature selection*. In principle, Jaccard similarity may be combined with other types of similarity, e.g. based on visual features.

In (5), locations are assumed to be represented as Cartesian (x, y, z) coordinates rather than as (lat, lon) pairs[2]. In practice, we thus need to convert the (lat_i, lon_i) coordinates of each photo y_i to its Cartesian coordinates:

$$x_i = cos(lat_i) \cdot cos(lon_i)$$
$$y_i = cos(lat_i) \cdot sin(lon_i)$$
$$z_i = sin(lat_i)$$

Subsequently, the right-hand side of (5) is evaluated, yielding a point (x^*, y^*, z^*), which is usually not on the surface of the earth. Unless this point is exactly the center of the earth, its latitude lat^* and longitude lon^* can be determined:

$$lat^* = atan2(z^*, \sqrt{x^{*2} + y^{*2}})$$
$$lon^* = atan2(y^*, x^*)$$

In addition to the choice of the parameter α, the performance of (5) depends on the set of resources R_x that is considered when determining the k most similar photos $y_1, ..., y_k$. In principle, we could take R_x to be the entire training set. However, we also experiment with putting a threshold on the similarity with x, considering in R_x only those resources that are sufficiently similar. This restriction is motivated by the fact that center-of-gravity methods are sensitive to outliers. Note that using medoids to alleviate the influence of outliers is not appropriate when the number of points is small. Also note that as a result of this restriction, sometimes less than k similar photos may be used. In each case R_x will contain the most similar photo, even if its similarity is below the threshold. Other photos are added only if they are sufficiently similar.

4.2 Experimental results

Three parameters influence the performance of the similarity search: the threshold on the similarity with the object to be classified, the number k of similar photos to consider, and the exponent α in (5). Table 1 displays the result for different choices of the threshold on similarity, and different values of k, in case of the *Overlap* configuration. Regarding the similarity threshold, we find that a small threshold of 0.05 slightly improves the results for the smaller values of k. Indeed, the smaller the value of k, the more the result may be influenced by outliers, and the more important it thus becomes to avoid them. Surprisingly, small values of k appear to be better than larger values, although the optimal choice $k = 2$ is substantially better than $k = 1$.

Tables 2 illustrates the influence of varying the exponent α in (5), where we take the similarity threshold fixed at 0.05. Choosing a higher value of α essentially serves the same purpose as choosing a higher similarity threshold, i.e. reducing the impact of potential outliers on the result. We can observe that higher values of α tend to produce better results. Again the choice of $k = 2$ turns out to be optimal.

In general, it seems that similarity search performs a lot worse than the language models, yielding an optimal error of 340.69 km, as opposed to 17.02 km in the case of language models. Similar effects are witnessed for the *Distinct* and *Filtered* configurations (not shown), where we respectively find an optimal error of 1302.95 km (instead of 979.86 km) and 578.22 km (instead of 31.10 km). However, as we will see in the next section, when combined with the language models, similarity search may be of great value.

[2]See `http://www.geomidpoint.com/calculation.html` for an explanation of this coordinate transformation, and a comparison with alternative methods to calculate "average locations".

Table 1: Influence of the similarity threshold on the median error distance for the *Overlap* configuration (using an exponent α of 1).

threshold	0	0.05	0.10	0.15	0.20
1	2528	2528	2528	2528	2528
2	477	424	477	773	1150
3	685	604	662	880	1150
4	748	741	773	899	1181
5	790	821	835	952	1242
6	824	799	837	954	1238
7	808	823	850	961	1247
8	843	829	856	980	1246
9	855	856	871	971	1242
10	863	868	872	968	1243

(k)

Table 2: Influence of the exponent α on the median error distance for the *Overlap* configuration (using a similarity threshold of 0.05).

α	1	25	50	75	100
1	2528	2528	2528	2528	2528
2	424	343	341	341	341
3	604	435	413	411	410
4	741	417	383	370	370
5	821	410	368	350	349
6	799	419	399	395	393
7	823	422	400	395	395
8	829	440	427	420	419
9	856	459	450	441	440
10	868	475	459	451	449

(k)

5. A HYBRID APPROACH

5.1 Outline

The two methods that have been presented in the previous sections can be combined in a natural way: first an area is determined using the language modeling approach from Section 3 and then the similarity based method from Section 4 is applied, but restricted to the photos in the found area. When no photo in the clustering satisfies the chosen similarity threshold, the medoid of the found cluster can be used instead. Thus, we may take advantage of the language modeling's ability to implicitly discriminate between occurrences of more and less important tags, while keeping the advantage of the similarity search that a precise coordinate is obtained.

A second extension is related to choosing the right number of clusters. In particular, when we discover that a given re-

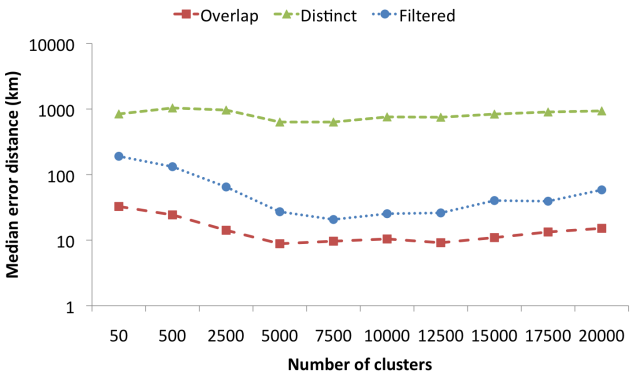

Figure 5: Median error obtained using the hybrid method with $k = 1$ and without a similarity threshold.

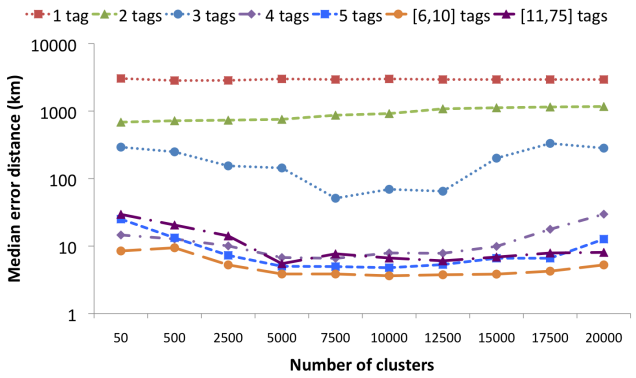

Figure 6: Median error between the medoid of the found cluster and the true location, each time using all test videos containing a given number of tags.

Table 3: Number of the test videos for which the location that was found is within a given distance of the true location.

	1km	5km	10km	50km	100km
Overlap (7 400)	2135	3362	3773	4500	4694
Distinct (2 600)	465	803	903	1066	1128
Filtered (7 400)	1428	2770	3248	4012	4265

source has no tag in common with the vocabulary of the chosen clustering, we fall-back to the next (coarser) clustering.[3] In this way, if a resource contains no tags that are indicative of a precise location (e.g. *eiffeltower*) but does contain some tags that define a larger-scale area (e.g. *france*), it may not have any tags in common with the vocabulary of the finest clusterings, but after falling-back to a coarser clustering, a suitable area can still be determined.

5.2 Experimental results

Figure 5 shows the median distance that is obtained when language models are combined with similarity search. Interestingly, we find that choosing $k = 1$ with similarity threshold 0 (shown in Figure 5) performs slightly better than choosing $k = 2$ with similarity threshold 0.05 (not shown), despite that the latter configuration is clearly better when similarity search is applied alone. This can be explained by the fact that within a cluster, all photos are relatively close to each other anyway, hence the problem of outliers is alleviated. As a result, the positive effect of filtering photos that are not sufficiently similar becomes counter-productive. A more detailed analysis of the results is presented in Table 3. For all three set-ups, a marked improvement is witnessed over the results of the language models from Section 3, the optimal results now being attained for 5 000 clusters in the case of the *Overlap* (8.82 km) and *Distinct* (633.36 km) set-ups, and for 7 500 clusters in the case of the *Filtered* (20.64 km) set-up.

Figures 6 and 7 provide a more detailed picture of the performance of our method, considering all test videos and

[3]Recall that in absence of any tags, without fall-back, the prior probabilities determine to which cluster a resource is assigned, hence the cluster containing the largest number of resources will be chosen.

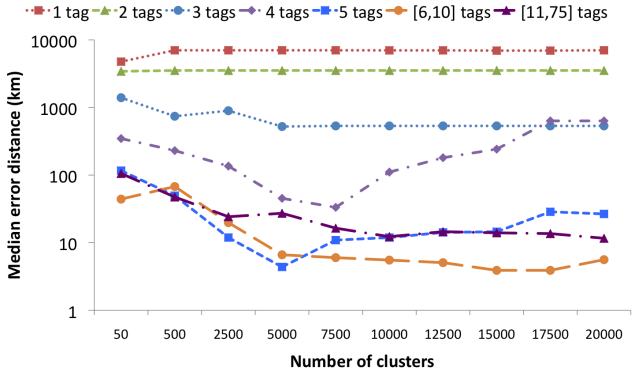

Figure 7: Median error between the medoid of the found cluster and the true location, using only the test videos from the *Distinct* set-up containing a given number of tags.

those from the *Distinct* set-up respectively. As in Section 3, we find that the bad performance in the *Distinct* set-up can be attributed to the fewer number of videos with sufficient tags. In particular, if we only consider those videos with 6 to 10 tags (21.77% of the test videos), a median distance of 3.90 km is attained when using either 15 000 or 17 500 clusters. In case of the *Overlap* experiment (not shown), the median distance in the [6,10] range (23.19% of the test videos) is only slightly better, with 3.54 km being attained when using 10 000 clusters. These results indicate that rather precise coordinates can be found for most videos, provided that a sufficient number of (manually chosen) tags are available.

Finally we analyze the impact of feature selection. The purpose of feature selection is to eliminate all tags that are not spatially relevant, before the language models are built. This may be useful not only for speeding up calculations, but also to improve classification accuracy. Figures 8, 9 and 10 display how choosing a different number of features impacts the median error distance. The results for all features refers to the set of features that have been used in the experiments throughout the paper, applying χ^2 feature selection as explained in Section 2. The other results show what happens when only the best 25%, 50% and 75% of these features (according to the χ^2 statistic) are retained. The main ob-

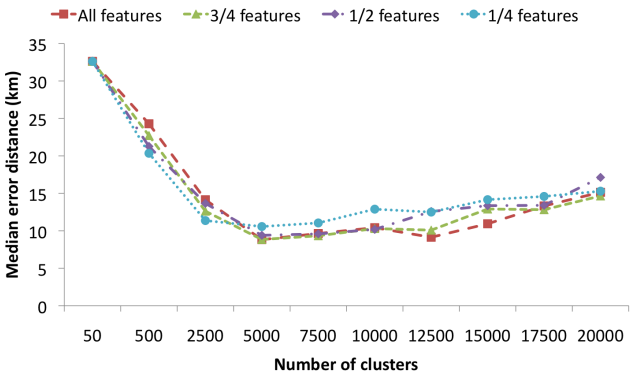

Figure 8: Impact of the amount of feature selection, in case of the *Overlap* set-up

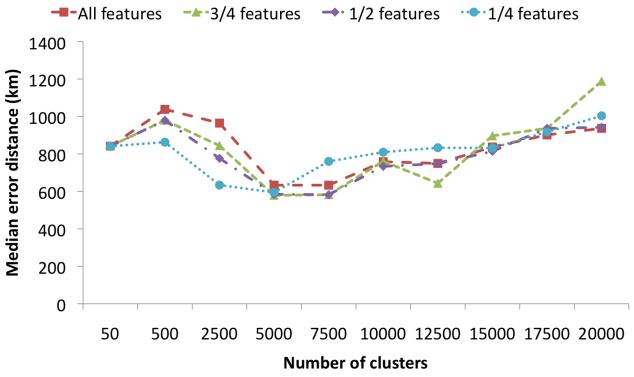

Figure 9: Impact of the amount of feature selection, in case of the *Distinct* set-up

servation is that the optimal value is quite robust w.r.t. the number of selected features. Only when a suboptimal number of clusters is chosen we find some differences, favoring fewer features for the coarser clusterings.

6. RELATED WORK

The related work falls in two categories: finding the geographic scope of resources, and using it when it is available.

Finding locations of tagged photos.

The task of deriving geographic coordinates for multimedia objects has recently gained in popularity. A recent benchmark evaluation of this task was carried out at MediaEval 2010 [4], where an earlier version of our system was shown to substantially outperform all other approaches. This result confirms and strengthens earlier support for using language models in this task [14].

Most existing approaches are based on clustering, in one way or another, to convert the task into a classification problem. For instance, in [3] target locations are determined using mean shift clustering, a non-parametric clustering technique from the field of image segmentation. To assign locations to new images, both visual (keypoints) and textual (tags) features were used. Experiments were carried out on a sample of over 30 million images, using both Bayesian classifiers and linear support vector machines, with slightly

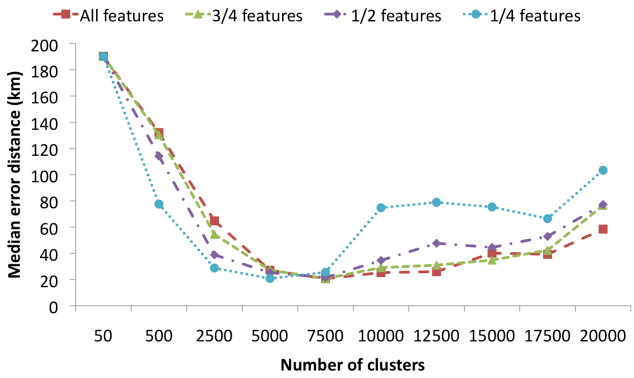

Figure 10: Impact of the amount of feature selection, in case of the *Filtered* set-up

better results for the latter. Two different resolutions were considered corresponding to approximately 100 km (finding the correct metropolitan area) and 100 m (finding the correct landmark). It was found that visual features, when combined with textual features, substantially improve accuracy in the case of landmarks. In [6], an approach is presented which is based purely on visual features. For each new photo, the 120 most similar photos with known coordinates are determined. This weighted set of 120 locations is then interpreted as an estimate of a probability distribution, whose mode is determined using mean-shift clustering. The resulting value is used as prediction of the image's location.

Next, [14] investigates the idea that when georeferencing images, the spatial distribution of the classes (areas) could be utilized to improve accuracy. Their starting point is that typically, not only the correct area will receive a high probability, but also the areas surrounding the correct area. An appropriate adaptation of the standard language modeling approach is shown to yield a small, but statistically significant improvement.

Using locations of tagged photos.

When available, the coordinates of a photo may be useful for a variety of purposes. In [1], for instance, coordinates of tagged photos are used to find representative textual descriptions of different areas of the world. These descriptions are then put on a map to assist users in finding images that were taken in a given location of interest. The approach is based on spatially clustering a set of geotagged Flickr images, using k-means, and then relying on (an adaptation of) tf-idf weighting to find the most prominent tags of a given area. Similarly, [9] looks at the problem of suggesting useful tags, based on available coordinates. Some authors have looked at using geographic information to help diversify image retrieval results [8, 10].

Geotagged photos are also useful from a geographic perspective, to better understand how people refer to places, and overcome the limitations and/or costs of existing mapping techniques [5]. For instance, by analyzing the tags of georeferenced photos, Hollenstein [7] found that the city toponym was by far the most essential reference type for specific locations. Moreover, [7] provides evidence suggesting that the average user has a rather distinct idea of specific places, their location and extent. Despite this tagging be-

haviour, Hollenstein concluded that the data available in the Flickr database meets the requirements to generate spatial footprints at a sub-city level. Finding such footprints for non-administrative regions (i.e. regions without officially defined boundaries) using georeferenced resources has also been addressed in [13] and [15]. Another problem of interest is the automated discovery of which names (or tags) correspond to places. Especially for vernacular place names, which typically do not appear in gazetteers, collaborative tagging-based systems may be a rich source of information. In [11], methods based on burst-analysis are proposed for extracting place names from Flickr.

7. CONCLUDING REMARKS

We have advocated a two-step approach for georeferencing tagged multimedia objects. In the first step, the task of finding suitable geographic coordinates is treated as a classification problem, where the classes are areas that have been obtained by clustering the locations of the objects in the training set. Once the most likely area has been identified, we determine a precise location by interpolating the locations of the most similar objects, in that area. Experimental results confirm the usefulness of this hybrid methodology. We have also analysed the influence of previously geo-annotated resources by the same user, and found that, while the availability of such resources in the training set positively influences the performance, the difference in performance all but disappears if a sufficient number of tags is available for that resource.

We have experimented with several gazetteers (Geonames, DBpedia, and the US and world sets of USGS/NGA), but have not been able to improve our results. On the other hand, preliminary analyses that use an oracle for disambiguating toponyms show that using gazetteers together with our current method has the potential of reducing the median distance considerably. It thus remains unclear whether (or how) such resources could be useful for this task. In addition to gazetteers, other types of information could be taken into account, which we have not examined, including visual features and information about the profile and social network of the corresponding user.

Acknowledgments.

We thank Johannes Deleu for interesting discussions on the use of gazetteers.

8. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 1–10, 2007.

[2] 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*, pages 971–980, 2007.

[3] 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, 2009.

[4] M. L. et al. Automatic tagging and geotagging in video collections and communities. In *Proc. ACM ICMR*, 2011.

[5] M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.

[6] J. H. Hays and A. A. Efros. IM2GPS: Estimating geographic information from a single image. In *Proc. Computer Vision and Pattern Recognition*, 2008.

[7] L. Hollenstein and R. Purves. Exploring place through user-generated content: Using Flickr to describe city cores. *Journal of Spatial Information Science*, 1(1):21–48, 2010.

[8] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proceeding of the 17th International Conference on World Wide Web*, pages 297–306, 2008.

[9] E. Moxley, J. Kleban, and B. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from Flickr. In *Proceeding of the 1st ACM International Conf. on Multimedia Information Retrieval*, pages 24–30, 2008.

[10] A. Popescu and I. Kanellos. Creating visual summaries for geographic regions. In *IR+SN Workshop (at ECIR)*, 2009.

[11] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web*, 3(1):1–30, 2009.

[12] P. Schmitz. Inducing ontology from Flickr tags. In *Proceedings of the Collaborative Web Tagging Workshop*, pages 210–214, 2006.

[13] S. Schockaert and M. De Cock. Neighborhood restrictions in geographic IR. In *Proc. ACM SIGIR*, pages 167–174, 2007.

[14] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *Proc. ACM SIGIR*, pages 484–491, 2009.

[15] F. Wilske. Approximation of neighborhood boundaries using collaborative tagging systems. In *Proceedings of the GI-Days*, pages 179–187, 2008.

[16] F. Wu and D. Weld. Automatically refining the Wikipedia infobox ontology. In *Proceeding of the 17th International Conference on World Wide Web*, pages 635–644, 2008.

[17] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.