Combining Multi-Resolution Evidence for Georeferencing Flickr Images

Olivier Van Laere[1], Steven Schockaert[2], and Bart Dhoedt[1]

[1] Department of Information Technology, Ghent University, IBBT, Belgium
{olivier.vanlaere,bart.dhoedt}@intec.ugent.be
[2] Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium
steven.schockaert@ugent.be

Abstract. We explore the task of determining the geographic location of photos on Flickr, using combined evidence from Naive Bayes classifiers that are trained at different spatial resolutions. In particular, we estimate the location of Flickr photos, based on their tags, at four different scales, ranging from a city-level granularity to fine-grained intra-city areas. Using Dempster-Shafer's evidence theory, we combine the output of the different classifiers into a single mass assignment. We demonstrate experimentally that the induced belief and plausibility measures are useful to determine whether there is sufficient evidence to classify the photo at a given granularity. Thus an adaptive method is obtained, by which photos are georeferenced at the most appropriate resolution.

1 Introduction

An increasing number of web systems allow users to organize and share resources, such as photos, videos, or scientific papers. The predominant way of organizing such resources is by the use of short textual descriptions called tags. These tags are added by users in an uncontrolled way, without the need for any semantic resources. Nonetheless, due to the wide availability of such tags, statistically analyzing tag distributions has proven to be a successful way of obtaining (shallow) semantic information in an automated way [13].

Considering photo sharing websites such as Flickr[3] or Panoramio[4], the most important kind of metadata is arguably the location where a photo was taken. Accordingly, these websites typically allow to attach explicit geographical coordinates to a photo, in addition to tag-based descriptions of its content. This is important for at least two reasons. First, it allows to put photos on a map, providing an interesting addition to e.g. Google Maps[5], and allowing users to quickly retrieve photos that were taken in a particular region [1, 4, 11]. Second, by looking at correlations between tag occurrences and locations, it becomes possible to find approximate boundaries of geographic regions [8]. This is particularly important for vernacular (i.e. informal) place names, which have no

[3] http://www.flickr.com/
[4] http://www.panoramio.com/
[5] http://maps.google.com

official boundaries that could be retrieved from gazetteers or other geographic resources. As a consequence, a thorough analysis of Flickr tags may result in rich geographic models that can be used to support geographically informed web search engines.

The question remains of how the location of a photo could be acquired. Certain cameras have a built-in GPS, in which case the exact coordinates are obtained automatically. In most cases, however, users need to manually specify where a photo was taken. Because this puts an extra burden on the user, without any immediate benefit, only a small minority of the users go through this step. Another approach is taken by Suggestify[6], a web application which allows users to suggest the location of a photo of another Flickr user, which she can then choose to accept or refuse. Nonetheless, for the vast majority of all photos on Flickr and Panoramio, the location is not known.

To solve this problem, we may attempt to derive the approximate location of a photo automatically, by comparing its tags with the tags of photos whose exact location is known [4, 11, 14]. These approximate locations may be sufficient to put the photo on a map (at a certain resolution), or to help determine the approximate boundaries of a vernacular region. Alternatively, by establishing the area in which the photo was taken, we may also assist users that are willing to manually specify the exact location, by centering a map on the area that was found, and zooming in at an appropriate level. In each case, it is important not only to find a location that is approximately correct, but also to provide a reliable estimate of how accurate that location is (e.g. street-level, neighborhood-level, city-level, regional level, ...). For some photos, we can easily find a very precise location (e.g. a photo tagged "Eiffel tower"), while for other photos we cannot even indicate in which country it was taken, by only looking at its tags (e.g. a photo tagged "birthday party"). Thus it is of interest to study adaptive techniques that provide reliable estimates at an appropriate resolution, or admit that no reliable location could be established.

To georeference Flickr photos (i.e. to assign a location), in previous work [14] we have proposed to discretize space by clustering the photos from some training set, and then train a Naive Bayes classifier to find the most appropriate cluster for previously unseen photos. Different resolutions can then be considered by repeating the whole process for more or less fine-grained clusterings, i.e. by adopting a larger or smaller number of clusters. Thus we obtain a series of different classifiers, operating at different levels of resolution. In this way, for a given photo, the most appropriate resolution can be chosen by looking at the confidence each of the classifiers has in its respective outcome.

In this paper, we look at how the results of these different classifiers can be combined to find the most appropriate location and resolution. In particular, we experimentally investigate the use of Dempster-Shafer theory, which naturally allows to combine evidence from sources that operate at different levels of granularity. Our hypothesis in using Dempster-Shafer theory is that agreement between the classifiers is a strong indicator of the correctness of the location that

[6] `http://suggestify.appspot.com/`

was found. For instance, if classifier C_1 finds locations at the neighborhood-level and C_2 at the city level, and the neighborhood that was found by C_1 is not from the city that was found by C_2, our confidence that the neighborhood is correct should be low, regardless of the confidence of classifier C_1 in its choice.

The remainder of this paper is structured as follows. Section 2 summarizes our methodology in obtaining training and test data from Flickr. We also explain how we have clustered the images in the training set, and which preprocessing techniques were applied. Then in Section 3 we discuss the details of our proposed method. We briefly recall how a Naive Bayes classifier can be trained to find plausible areas where a photo might be located, at a fixed resolution. Subsequently we provide details on how Dempster-Shafer theory is applied to combine the classifiers that were trained at different resolutions. In Section 4, we present our experimental results, demonstrating substantial improvements over a baseline system. Finally, related work is discussed in Section 5.

2 Methodology

To obtain suitable training and test data, we composed a list of 55 large European cities. These cities were selected by intersecting the set of the 100 most densely populated European cities[7] with the set of the 160 most important European cities for tourism[8]. This choice was motivated by the intuition that a high population should ensure that allocating photos to locations is non-trivial (as opposed to villages where all activity is centered around a small area), while tourist activity should ensure that a sufficient number of photos is available on Flickr. For each georeferenced photo in these cities, we collected the corresponding tags and coordinates using the Flickr API, leading to a total of 3738072 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). From our initial set of photos, we removed those photos whose coordinates had an accuracy of 13 or less, to ensure that all coordinates were meaningful w.r.t. within-city location. Furthermore, we removed photos whose tag set and user name was identical to a photo that is already in our collection (to reduce the impact of bulk uploads [11]). After these two filtering steps, a set of 1029761 photos remained from 54 cities (no photos from Bremen had coordinates whose accuracy was above 13), which was split into 686193 photos for training ($\approx 66\%$) and 343568 photos for testing ($\approx 33\%$). In separating training data from test data, we ensured that all photos from the same user were either in the training set, or in the test set (to avoid an unfair exploitation of user-specific tags).

We then divided the 54 remaining cities into a set of disjoint areas that will serve as classification labels. The areas themselves were obtained by clustering the locations of the photos in the training set using the k-medoids algorithm with geodesic distance. Below, we consider four different resolutions, corresponding to

[7] `http://www.nga.mil`
[8] `http://www.visiteuropeancities.info`

the city level (in which case there are 54 areas, each corresponding to an entire city), as well as the result of clustering all photos in 250, 500, or 1000 clusters. In each case, we chose the number of clusters per city proportional to the number of georeferenced photos we had available for that city (in the training set), with the exception that every city should contain at least one cluster centre. As a result, cities for which we had only few georeferenced photos were divided in areas of a larger scale. This conforms to our intuition that we should try to be precise in estimating the location of a photo only when sufficient information is available for making that decision. In addition, whenever the number of photos in a given cluster dropped below 50, after an iteration of the k-medoids algorithm, that cluster was eliminated and the associated photos added to the nearest remaining cluster. The actual number of areas after the clustering algorithm had converged was respectively 54, 217, 401 and 677.

For efficiency, and to increase the robustness of the approach, we removed all tags that were used by 2 users or less. Next, we applied χ^2 feature selection to eliminate tags that are not indicative of a particular area. In particular, the vocabulary V that was used for classification was obtained by taking for each area a those 25 tags whose χ^2 value was highest. This led to a total number of 1269, 4701, 8452 and 13727 distinct tags, respectively in the case where the initial number of clusters k was 54, 250, 500 and 1000.

3 Georeferencing Images

3.1 Naive Bayes Classification

Let $\mathcal{A}$ be a set of (disjoint) areas, obtained by clustering the locations of the images in our training set. 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. Given a previously unseen image 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. In [14], we proposed a (multinomial) Naive Bayes classifier to this end, which has the advantage of being simple, efficient, and robust. An additional advantage, which will be crucial for combining classifiers that operate at different resolutions, is the fact that Naive Bayes produces probabilities, in contrast to e.g. support vector machines. Specifically, we assume that an image x is represented as its set of tags. Using Bayes' rule, and assuming that occurrences of different tags are independent, the probability $P(a|x)$ that image x was taken in area a is proportional to

$$P(a|x) \propto P(a) \cdot \prod_{t \in x} P(t|a) \tag{1}$$

Using a multinomial language model with Laplace smoothing [18], the probability $P(t|a)$ is estimated as

$$P(t|a) = \frac{N_t + 1}{\left(\sum_{y \in X_a} |y|\right) + |V|}$$

where N_t is the number of images in area a containing tag t, $\sum_{y \in X_a} |y|$ is the total number of tag occurrences over all images in area a, and V is the vocabulary, as before. Note that this technique of estimating $P(t|a)$ originates from Laplace's rule of succession. The maximum likelihood estimation $\frac{N_t}{\sum_{y \in X_a} |y|}$ would not be useful here, as it would imply $P(a|x) = 0$ as soon as x has one tag which does not occur with any image of the training set that is located in area a. The prior probability $P(a)$ of area a, on the other hand, can reliably be estimated using the maximum likelihood method:

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

Finally note that the actual value of $P(a|x)$, for all $a \in \mathcal{A}$, is found from (1) after normalization.

3.2 Combining classifiers using Dempster-Shafer theory

Motivation The fact that areas are spatially distributed should intuitively help to assign photos to areas more accurately. For example, assume that $\mathcal{A} = \{a, b, c, d\}$ and that the Naive Bayes classifier finds for a given photo x that that $P(a|x) = 0.3$, $P(b|x) = 0.25$, $P(c|x) = 0.25$ and $P(d|x) = 0.2$. Now assume furthermore that b, c, and d are adjacent neighborhoods, while a is located in a different city. Then in fact, the correct location is more likely to be near areas $\{b, c, d\}$ than near a. Naive Bayes in its basic form ignores this information and simply treats areas as abstract classes. To make Naive Bayes more spatially-aware, we propose to apply the approach outlined in Section 3.1 at multiple resolutions and combine the results. A classifier working at a higher resolution will then hopefully find the region containing regions b, c, d to be more likely than the region containing a. Based on the agreement between fine-grained classifiers and coarse-grained classifiers, we may then try to find the most appropriate resolution for a given photo: in cases of disagreement, coarser results are preferred, while in cases of strong agreement, fine-grained results may be better suited.

Specifically, let $\{\mathcal{A}_1, ..., \mathcal{A}_k\}$ be different clusterings of the cities of interest into disjoint areas, where $\mathcal{A}_1$ corresponds to the finest clustering and $\mathcal{A}_k$ corresponds to the coarsest clustering, i.e. $|\mathcal{A}_1| > |\mathcal{A}_2| > ... > |\mathcal{A}_k|$. Furthermore let C_i be a classifier that was trained to find the area from $\mathcal{A}_i$ in which a given photo was taken. With each area in $\mathcal{A}_i$, we can now associate a set of areas from the finest level $\mathcal{A}_1$. In particular, for $a \in \mathcal{A}_i$, we let $areas(a)$ denote the set of areas from $\mathcal{A}_1$ that overlap with area a. In this way, classifications at coarser resolutions can be seen as incomplete classifications at the finest resolution. For instance, if classifier $\mathcal{A}_k$ suggests that a is the most plausible area, we can take this as evidence that the correct area, at the finest level, is among those of the set $area(a)$. Such incomplete conclusions are naturally represented in the theory of evidence that was proposed by Dempster and Shafer [5, 12]. In Dempster-Shafer theory, evidence is encoded by a probability distribution on the power set of the universe. This probability distribution is called a belief function, or

mass assignment, to distinguish it from probability distributions on the universe itself.

Obtaining mass assignments In Dempster-Shafer theory, a mass assignment m in the universe U maps any subset of U to a value in $[0, 1]$ such that $\sum_{X \subseteq U} m(X) = 1$ and $m(\emptyset) = 0$. Intuitively, $m(X)$ represents the amount of evidence that the correct value is among those in X. Subsets X such that $m(X) > 0$ are called focal elements. If all focal elements are disjoint, then $m(X)$ can be interpreted as the probability that the correct area is among those in X. In general, two measures of uncertainty are typically defined in Dempster-Shafer theory, for any $X \subseteq U$:

$$Bel(X) = \sum_{Y \subseteq X} m(Y) \qquad\qquad Pl(X) = \sum_{Y \cap X \neq \emptyset} m(Y)$$

The degree of belief $Bel(X)$ can be interpreted as a lower bound on the probability that X contains the correct value, while the degree of plausibility $Pl(X)$ is an upper bound for this probability.

In the context of this paper, the universe will always be the set of areas (clusters) in the most fine-grained clustering, viz. the set $\mathcal{A}_1$. Let $p_i(a)$ be the probability that classifier C_i has assigned to area $a \in \mathcal{A}_i$ for the photo under consideration. Intuitively, we can take this information as evidence that the correct area, among the fine-grained areas in $\mathcal{A}_1$, is among those that overlap with a, i.e. among those in $areas(a)$. This idea leads to the following mass assignment corresponding to classifier C_i ($X \subseteq \mathcal{A}_1$):

$$m_i(X) = \begin{cases} p_i(a) & \text{if } X = areas(a) \text{ for some } a \in A_i \\ \sum_{a \in (\mathcal{A}_i \setminus A_i)} p_i(a) & \text{if } X = \mathcal{A}_1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $A_i \subseteq \mathcal{A}_i$ is the set of areas that are most likely according to classifier C_i. In principle, we may take $A_i = \mathcal{A}_i$ but there are at least two reasons for taking A_i to be a much smaller set of areas. The mass assigned to the universe $\mathcal{A}_1$ corresponds to a degree of ignorance, i.e. we only put belief in the most plausible areas of each classification, and admit that we are ignorant about the correct area when it turns out that none of the most plausible areas is correct. The underlying motivation is that Naive Bayes can be useful to find which are the most likely areas, but that the probability estimates for the remaining areas are not meaningful. Moreover, restricting attention to a relatively small subset of areas A_i is a prerequisite for obtaining a sufficiently scalable method. In our experiments, the set A_i was constructed by adding areas in decreasing order of likelihood (according to C_i), until $\sum_{a \in A} p_i(a) \geq 0.95$.

Note that alternatively, we could also assign the mass $\sum_{a \in \mathcal{A}_i \setminus A_i} p_i(a)$ to $\mathcal{A}_i \setminus A_i$ instead of $\mathcal{A}_i$; we do not consider this possibility, however, in the remainder of this paper.

Combining mass assignments An important advantage of using Dempster-Shafer theory in this context is that it allows to combine evidence from different sources. In particular, for two mass assignments m and m' in the universe $\mathcal{A}_1$, the joint mass assignment $m \oplus m'$ is defined using Dempster's rule of combination as

$$(m \oplus m')(\emptyset) = 0 \tag{3}$$

$$(m \oplus m')(X) = \frac{\sum_{Y \cap Z = X} m(Y) \cdot m'(Z)}{1 - \sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z)} \tag{4}$$

for any subset $\emptyset \subset X \subseteq \mathcal{A}_1$, and provided that $\sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z) < 1$. It can be shown that this combination rule is associative. By treating the classifiers $C_1, ..., C_k$ as independent sources, we obtain the following mass assignment:

$$m = m_1 \oplus m_2 \oplus ... \oplus m_k \tag{5}$$

Note that the assumption that classifiers $C_1, ..., C_k$ are independent sources is a simplification, as they have essentially been trained on the same data. However, as different classifiers operate at different resolutions, implying among others that different tags have been retained by the χ^2 method in each case, this simplification appears to be reasonable.

The combination rule (3)–(4) is the combination rule proposed by Dempster. It is not entirely uncontroversial, however, and in particular when the degree of conflict $\sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z)$ is close to 1, it is reputed to provide counterintuitive results [17]. As an alternative, Yager [16] proposed the following rule for combining k mass assignments in a universe U ($X \subset U$):

$$m(X) = \sum_{\bigcap_i Y_i = X} m_1(Y_1) \cdot ... \cdot m_k(Y_k) \tag{6}$$

$$m(U) = m_1(U) \cdot ... \cdot m_k(U) + \sum_{\bigcap_i Y_i = \emptyset} m_1(Y_1) \cdot ... \cdot m_k(Y_k) \tag{7}$$

$$m(\emptyset) = 0 \tag{8}$$

Clearly, Yager's combination rule only differs from the one proposed by Dempster in what happens with the mass $\sum_{Y \cap Z = \emptyset} m(Y) \cdot m'(Z)$ that would normally be assigned to the empty set. While Dempster's rule distributes this mass over all focal elements, leading to an associative operator, in Yager's rule this mass is assigned to the universe U. As such, Yager's rule can be considered more cautious as the degree of ignorance increases when different sources are in conflict with each other.

4 Experimental Results

In this section, we present the results of a number of experiments which we have carried out to compare the performance of the Dempster-Shafer based

approach with a baseline that uses the probabilities from Naive Bayes in a more straightforward way.

In a first experiment, we have verified whether we could improve the accuracy of Naive Bayes by using the combined mass assignment defined by (3)–(4). In particular, the task consists of choosing one area from the clustering $\mathcal{A}_i$ at a given resolution ($k = 54, 250, 500, 1000$), and we have compared the following three methods:

Probability Choose the cluster for which the highest probability was found using Naive Bayes.

Plausibility Choose the cluster a for which the value of $Pl(areas(a))$ is maximal.

Belief Choose the cluster a for which the value of $Bel(areas(a))$ is maximal.

The result is summarized in Table 1. The evaluation metric that was used is accuracy, i.e. the percentage of photos in the test set for which the correct area was found. Clearly, for higher values of k a lower accuracy is generally obtained, as there are more areas to choose from, and less information is available for each area. In approximately 87% of the cases, a photo can be assigned to the correct city, while the correct area at the finest level can only be found in about 40% of the cases. Comparing the different methods, we find that except for $k = 500$, using *belief* leads to slightly better performance than using *probability*. *Plausibility*, on the other hand, leads to worse performance than *probability*, except for the case $k = 54$. This latter fact is not surprising as for any area a which represents an entire city, the only focal elements that overlap with this city will actually be contained in the city, hence $Bel(areas(a)) = Pl(areas(a))$. Overall, in this task, the Dempster-Shafer approach does not allow to substantially improve over the standard Naive Bayes approach.

	54	250	500	1000
Probability	0.8694	0.5137	0.4622	0.4126
Plausibility	0.8729	0.4756	0.3838	0.4134
Belief	0.8729	0.5211	0.4457	0.4151

Table 1. Comparing the use of probability, plausibility and belief for finding the area in which a photo was taken (accuracy).

A second experiment was targeted at evaluating the behavior of the Dempster-Shafer approach when it comes to finding the right resolution for a given photo. Here the task is as follows. For a given photo, choose the most appropriate value of k (54, 250, 500 or 1000) and choose an area from the corresponding clustering. Accuracy is defined as the percentage of cases in which the true location of the photo was within the area that was chosen. However, the idea is that as often as possible, areas should be chosen from the more fine-grained clusterings, while accuracy will clearly be higher when selecting areas from the more coarse-grained

clusterings. In addition to accuracy, it is therefore important to compare the average size of the areas that are returned by different methods.

As clusters are simply defined as sets of photos (as opposed to e.g. polygons) we have measured the size of an area (cluster) in terms of the distance between the centroid of that cluster and the remaining photos of the cluster. In particular, for a given area a, represented by the set of photos X_a, the centroid c_a of a is the most central photo, i.e.:

$$c_a = \arg\min_{x \in X_a} \sum_{y \in X_a} d(x, y)$$

where $d(x, y)$ is the geodesic distance between the locations of photos x and y. To measure the size $size(a)$ of area a, we have used the median value of the set $\{d(x, c_a) | x \in X_a\}$. The size intuitively corresponds to the radius of area a, if we think of this area as a circle. Note that the median is used, rather than the maximum or average, because several areas contain outliers, i.e. photos that are not close to any other photos, and that are added to the cluster centre that happens to be closest. The median is more robust against such outliers than the maximum or average, and thus appears to be better suited as an evaluation measure. As an evaluation criterium, in addition to accuracy, we consider the average of $size(a)$ over all areas a that were chosen by a particular method. Ideally, methods should exhibit a high accuracy and a small average size.

The methods that have been compared all follow the same basic strategy. First, the areas from the clustering corresponding to $k = 1000$ are ranked. For the top-ranked area a_1, it is checked whether sufficient support is available. If this is the case, area a_1 is returned as the chosen area. If not, the process is repeated for the clustering at level $k = 500$ and, if necessary for $k = 250$. If there is insufficient support for the top-ranked area a_3 at level $k = 250$, the best area at level $k = 54$ is always chosen. Thus our method is parametrized by a ranking function, a way of measuring support, and a threshold value. The threshold value will be used to control the trade-off between accuracy and average size. As for the remaining two parameters, we have compared the following configurations:

Probability Areas are ranked according to the probability that was assigned to them by the Naive Bayes classifier. This probability value also serves as a measure of support.

Plausibility Areas a are ranked according to the plausibility degree $Pl(areas(a))$. This degree also serves as a measure of support.

Belief Areas a are ranked according to the belief degree $Bel(areas(a))$. This degree also serves as a measure of support.

Hybrid Areas a are ranked according to the plausibility degree $Pl(areas(a))$. Support is measured as $Bel(areas(a))$.

The result is depicted in Figure 1. This figure was obtained by varying the value of the threshold from 0.01 to 0.99 in steps of 0.01. This led, for each of the four methods, to 99 data points, each of which corresponds to an *(accuracy,average size)* pair. After interpolation of these 99 data points, the graphs in Figure 1

were obtained. Clearly, the three methods based on the combined mass assignment perform substantially better than the method based on the probabilities of the Naive Bayes classifier. For instance, to obtain an accuracy of 75%, using the method *probability* we need to accept an average cluster size of about 2.15 km, whereas this is around 1.5 for the other methods. At lower accuracy levels, the difference becomes somewhat less pronounced, e.g. an accuracy of 50% corresponds to an average cluster size of about 1.15 using the *probability* method and a size of about 1 km using the other methods.

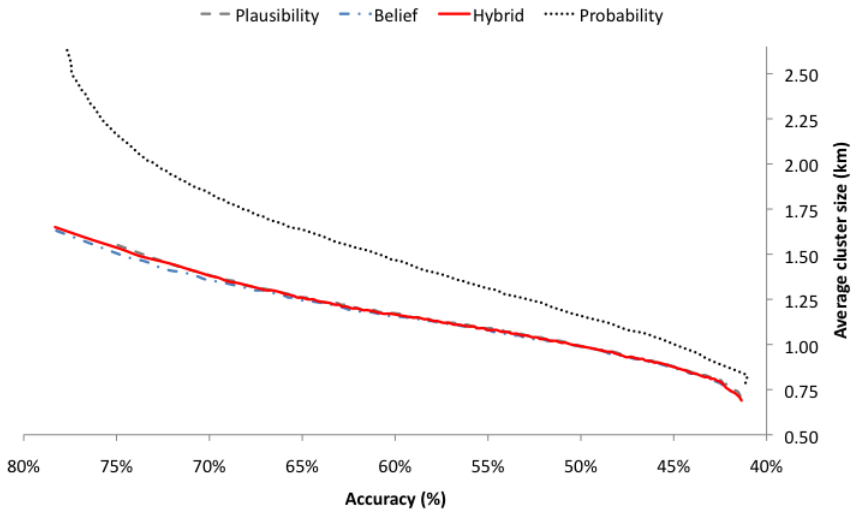

Fig. 1. Comparison of the trade-off between accuracy and average cluster size for four different methods.

Figure 1 is based on the combined mass assignment (5) obtained using Dempster's rule. As this task is essentially about deciding whether there is enough support to assign a photo to a cluster at a particular level, Yager's rule may be more suitable. Indeed, Dempster's rule ignores any conflict among the different levels by normalizing the masses. When using Yager's rule, on the other hand, whenever there is conflict, the degrees of belief and plausibility will have lower values. This will lead to photos being assigned to areas of coarser clusterings. In Figure 2, the result of the *hybrid* method is depicted when using either the combined mass assignment (5) based on Dempster's rule or the combined mass assignment (6)–(8) based on Yager's rule. The most important conclusion is that the graph corresponding to Yager's rule is to the left of the graph corresponding to Dempster's rule. This is indeed in accordance with the cautious nature of the method: photos tend to be assigned to coarser levels, leading to higher accuracy at the cost of a higher average size. For the accuracies attained by both methods,

i.e. the accuracies in the interval $[0.47, 0.78]$, Yager's rule and Dempster's rule perform comparably, with Dempster's rule performing slightly better.

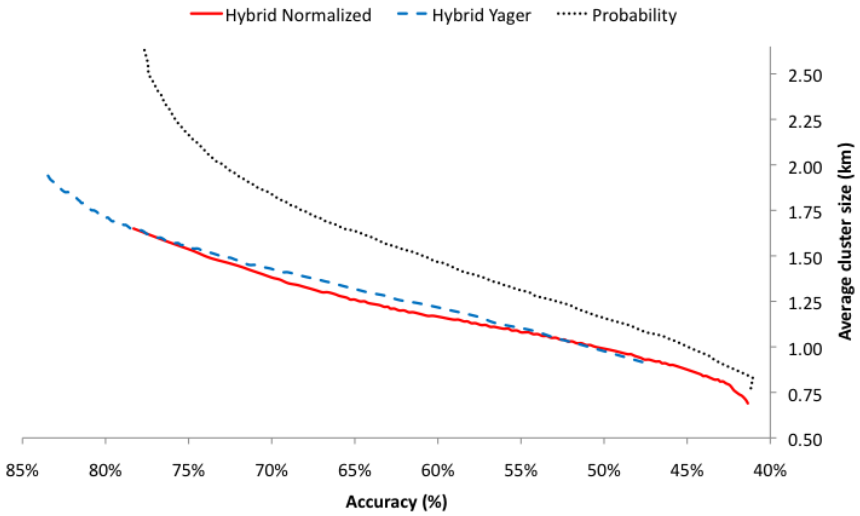

Fig. 2. Comparison the performance of Dempster's and Yager's rule of combination.

5 Related Work

Some authors have already studied the task of georeferencing photos based on clustering. One such approach is presented in [4], where target locations are determined using mean shift clustering, a non-parametric clustering technique from the field of image segmentation. The advantage of this clustering method is that an optimal number of clusters is determined automatically, requiring only an estimate of the scale of interest. Specifically, to find good locations, the difference is calculated between the density of photos at a given location and a weighted mean of the densities in the area surrounding that location. 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 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 [7], 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. Using k-means to spatially cluster geotagged Flickr images has been proposed in [1], where the clusters are used to find representative textual descriptions of each area. The goal is to visualize these textual descriptions on a map, to assist users in finding images of interest.

The idea that when georeferencing images, the spatial distribution of the classes (areas) could be utilized to improve accuracy has already been suggested in [11]. Their starting point is that typically not only the correct area will receive a high probability, but also the areas surrounding the correct area. Indeed, the expected distribution of tags in these areas will typically be quite similar. Hence, if some area a receives a high score, and all of the areas surrounding a also receive a relatively high score, we can be more confident in a being approximately correct than when all the areas surrounding a receive a low score. Motivated by this intuition, [11] proposes to smooth $P(a|x)$ as follows (using a uniform prior):

$$P^*(a|x) \propto \alpha P(x|a) + (1-\alpha) \cdot \sum_{b \in neigh_d(a)} \frac{P(x|b)}{(2d+1)^2 - 1}$$

where $d > 0$ and $neigh_d(a)$ is the set of all areas that are within distance d of a.

Some Flickr tags are intuitively more important than others in determining the location of a photo. Toponyms in particular are by definition indicative of geographic location. One way of recognizing toponyms is by looking for so-called comma-groups. These are groups of words that are comma-separated, e.g. *San Francisco, California, USA*. In this example, there is a clear relationship between the comma-separated values, as San Francisco is a city, located in the state of California, which is in turn one of the states of the USA. As a result, resolution of the toponyms represented by this group reveals an unambiguous geographical reference. Resolution of such comma-groups has been studied by Lieberman in [9]. In [8], Hollenstein studied the way people tag images in order to discover how people refer to a location. She found that the city toponym was by far the most essential reference to a specific location. This is in accordance with our results, where we have also found classification accuracies to be particularly high for the city level. It was furthermore shown in [8] that the average user has a distinct idea of specific places, their location and extent. Despite this tagging behaviour, Hollenstein concluded that the data available in the Flickr database meets the requirements to generate spatial footprints at a sub-city level.

Various authors have investigated the use of Dempster-Shafer theory for combining the results of different classifiers [2, 6, 10, 15]. However, the aim of using Dempster-Shafer theory in this context is quite different from our aim in this paper. Specifically, these methods mainly use Dempster-Shafer theory for its ability to represent partial ignorance. For instance, if a given classifier assigns a probability p_i to each class c_i, a belief function may be constructed by choosing $m(\{c_i\}) = f_i$ for some $f_i < p_i$, and $m(C) = 1 - \sum_i f_i$, for $C = \{c_1, ..., c_n\}$ the set of all classes. The value $1 - \sum_i f_i$ can then intuitively be interpreted in terms of confidence in the associated classifier. Note also that all focal elements are

then either singletons or the universe, which makes Dempster-Shafer theory sufficiently scalable to deal with large numbers of classes, although sometimes focal elements of the form $C \setminus \{c_i\}$ are also used. In [3], Dempster-Shafer theory is used for retrieving images of people, combining evidence from a face recognition module and a classifier based on textual descriptions; again only singletons and the entire universe are considered as focal elements.

6 Conclusions

We have studied the problem of finding the geographic location of a photo, particularly emphasizing the importance of determining the appropriate resolution for any given photo. While the precise location of some photos can easily be established, for other photos we can only hope to find a rough idea of where it was taken. Our basic approach consists of clustering the part of geographic space that is of interest, and use standard machine learning techniques (viz. Naive Bayes) to find the cluster which is most likely to contain the correct location of a photo. By varying the number of clusters, different classifiers are obtained which operate at different resolutions.

While adaptive methods, heuristically choosing the most appropriate resolution, can be obtained by straighforwardly analyzing the outputs of these different classifiers, a significant gain in performance is obtained by first combining these outputs using Dempster-Shafer's evidence theory. Experimental results have indicated that the belief and plausibility measures induced by the resulting mass assignment are particularly suitable for determining whether sufficient support is available to classify a photo at a given resolution.

Acknowledgments

Steven Schockaert was funded as a postdoctoral fellow of the Research Foundation – Flanders (FWO).

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. A. Al-Ani and M. Deriche. A new technique for combining multiple classifiers using the dempster-shafer theory of evidence. *J. Artif. Int. Res.*, 17(1):333–361, 2002.
3. Y. Aslandogan and C. Yu. Multiple evidence combination in image retrieval: Diogenes searches for people on the Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 88–95, 2000.
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, 2009.

5. A. Dempster. A Generalization of Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247, 1968.
6. T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE transactions on systems, man, and cybernetics*, 25(5):804–813, 1995.
7. J. H. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2008.
8. L. Hollenstein. Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zurich, 2008.
9. M. D. Lieberman, H. Samet, and J. Sankaranayananan. Geotagging: using proximity, sibling, and prominence clues to understand comma groups. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, 2010.
10. G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.
11. P. Serdyukov, V. Murdock, and R. van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491, 2009.
12. G. Shafer. *A mathematical theory of evidence*. Princeton university press Princeton, NJ, 1976.
13. J. Tang, H.-f. Leung, Q. Luo, D. Chen, and J. Gong. Towards ontology learning from folksonomies. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 2089–2094, 2009.
14. O. Van Laere, S. Schockaert, and B. Dhoedt. Towards automated georeferencing of flickr photos. In *GIR '10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, 2010.
15. L. Xu and C. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.
16. R. R. Yager. On the dempster-shafer framework and new combination rules. *Information Sciences*, 41(2):93 – 137, 1987.
17. L. A. Zadeh. A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination. *AI Mag.*, 7(2):85–90, 1986.
18. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.