Ghent University at the 2011 Placing Task

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 the results of a system that georeferences Flickr videos using a combination of language models and similarity search. The system extends our approach from last year by using language models with a more adaptive granularity, and by taking into account the home location of the user.

Keywords

Georeferencing, Language models, Dempster-Shafer theory

1. INTRODUCTION

The Placing Task requires participants to estimate the geographical coordinates of a video, based on the visual and auditory features of the video, textual tags that have been assigned to it by its owner, context information about the owner, etc. Training data consists of a portion of the georeferenced photos on Flickr. For a detailed description of this task, we refer to [2]. Participants were allowed to submit five runs, which differ in the kind of meta-data and external resources that are allowed.

We participated in the 2010 Placing Task with a system based on a two-step approach [6]. In the first step, language models are used to determine the area which is most likely to contain the location of a previously unseen video. The second step determines the location of the most similar photo within the chosen area and uses its location as the prediction. An important lesson drawn from last year's participation was that the chosen granularity of the areas in the first step crucially influences the performance, and that moreover this optimal granularity varies greatly across different test videos. Therefore, this year we have experimented with two methods to determine a suitable granularity. As a second extension, this year we have included the possibility of using the home location of the user, which is available in textual form for a majority of all test videos.

2. METHODOLOGY

A total number of 3 185 258 georeferenced photos from Flickr were provided as training data by the task organizers. As last year, photos that have been uploaded on the same day by the same user with identical tags are treated as duplicates, to reduce the impact of bulk uploads, after which 2 096 712 photos remained. For run 5, a larger training set was used, crawled using the Flickr API, consisting of 11 770 000 photos with the highest level of location accuracy (i.e. level 16). We ensured not to crawl any videos and thus any possible items from the test set.

In both cases, the locations of the photos in the training set were clustered using agglomerative hierarchical clustering, from which flat clusterings into 500, 2500, 5000 and 7500 clusters have been obtained; these clusterings will be referred to as C_{500}, C_{2500}, C_{5000} and C_{7500} respectively. For each cluster within these four clusterings, the most relevant tags are determined using χ^2 feature selection, leading to the vocabularies (i.e. sets of tags) V_{500}, V_{2500}, V_{5000} and V_{7500}.

Finding the most likely area.

To determine the probability $P(a|x)$ that a video x was taken in area $a \in C_k$, a unigram language modeling approach is used (except for run 4, which does not permit the use of textual tags), whereby [3]

$$P(a|x) \propto \left(\prod_{t \in tags_k(x)} P(t|a) \right) \cdot P(a) \qquad (1)$$

where $tags_k(x)$ is the set of tags from V_k that have been assigned to video x. The probability $P(t|a)$ is estimated using Bayesian smoothing (see [6] for more details). Different to our system of last year, we estimate the prior probability $P(a)$ using the home location of the owner of video x, in those runs where the use of gazetteer look-up was allowed, and for those videos where a textual home location was available and georeferencing did not fail. Specifically, we take

$$P(a) \propto \left(\frac{1}{d(p_a, p_{home})} \right)^{\theta} \qquad (2)$$

where d refers to geodesic distance, p_a are the coordinates of the most central photo of area a (i.e. the medoid of the locations of the photos from the training data located in area a) and p_{home} are the coordinates obtained from the textual home location using the Google Geocoding API[1]. The parameter θ was set to 0.75 in our experiments. If coordinates of the home location cannot be obtained, $P(a)$ is estimated as the percentage of all photos from the training

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

[1] http://code.google.com/apis/maps/documentation/geocoding/

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

data that are contained in area a, i.e.

$$P(a) = \frac{|a|}{\sum_{a \in C_k} |a|} \quad (3)$$

identifying a with the set of photos from area a in the training data. In run 1, where a textual home location may be available, but gazetteer look-up is not allowed, (3) can be refined by looking at tags from the vocabulary V_k that appear in it:

$$P(a) \propto \left(\prod_{t \in homeTags(x) \cap tags_k(x)} P(t|a)^{\mu} \right) \cdot \frac{|a|}{\sum_{a \in C_k} |a|} \quad (4)$$

where μ was set to 0.45 in the experiments.

Determining the level of granularity.

The language modeling approach to georeferencing requires an appropriate level of granularity to be determined: for videos with more informative tags, it is beneficial to consider a finer-grained clustering. As a baseline technique for selecting the optimal value of k, we check the number of tags a video x has in common with the different vocabularies. If $tags_{7500}(x) \cap V_{7500} \geq t_{7500}$, with t_{7500} an appropriate threshold value, $k = 7500$ is chosen. Otherwise, if $tags_{5000}(x) \cap V_{5000} \geq t_{5000}$ we select $k = 5000$, etc. For run 1 and 2 the threshold values where chosen as $t_{500} = 1$ and $t_{2500} = t_{5000} = t_{7500} = 2$. For run 3, on the other hand, we set $t_{500} = t_{2500} = t_{5000} = t_{7500} = 1$. Run 4 is not based on language models. For run 5, we used a technique based on Dempster-Shafer theory which was proposed in [5]. Intuitively, this approach combines the probability distributions obtained at each of the granularity levels into a single structure, called a belief function, and then determines the most likely area at the most appropriate level of granularity[2]. While this approach allows for a better informed decision, it requires language model probabilities to be calibrated, which necessitates the use of a sufficiently large development set which is disjoint from the training set. Initial experiments revealed that the training set provided by the task organizers was not sufficiently large to allow for both accurate training and accurate calibration. Therefore this technique was only applied in run 5, using 10.7M photos for training and 1.07M photos for calibration.

Determining the location.

Once a suitable value of k has been chosen, the area a from C_k that maximizes (1) is determined. Subsequently the photo from area a (in the training data) which is most similar to the video x is determined, and its location is used as the prediction for the location of x. Similarity is determined by comparing the tags assigned to each photo with the tags assigned to x using Jaccard similarity (without feature selection).

As a fall-back strategy, if no tags have been assigned to x at all, the home location of x is used as the prediction (in those runs where the use of a gazetteer is allowed). If no home location is available, we use the location of the photo which is visually most similar to video x. To measure visual similarity, a photo is compared against the key frames of video x that were provided by the task organizers. Visual features were extracted using the Color and Edge Directivity Descriptor (cedd) of the LIRE tool [1] . When different key frames of the video yield conflicting predictions (i.e. when they are most similar to different photos), the (keyframe,photo) pair which provided the highest degree of similarity is used.

[2]Specifically, the most likely area was determined using the pignistic probability decision rule [4], choosing the granularity level as the most fine-grained level for which pignistic probability was above the threshold of 0.6.

	1km	10km	100km	1000km	10000km
run 1	1245	2386	3340	4010	5207
run 2	1294	2753	3883	4578	5232
run 3	1263	2665	3759	4499	5231
run 4	2	6	49	624	4332
run 5	2567	3528	4109	4672	5263

Table 1: Overview of the results on the test collection of 5347 videos, using textual tags and visual features (run 1); using textual tags, gazetteer services and visual features (runs 2 and 3); using only visual features (run 4); and using tags, gazetteers and visual features on an extended training set with the Dempster-Shafer approach (run 5).

3. RESULTS AND DISCUSSION

The results of the five runs are provided in Table 1. In particular, the table shows how many of the 5347 videos in the test collection were localized within 1km, 10km, 100km, 1000km and 10000km of the correct location.

As can been concluded by comparing the results of runs 1 and 2, using the geocoded home location is really boosting the results. Also, determining a good threshold value to fall back to a coarser clustering can impact the results, as is demonstrated in run 3 which only differs from run 2 in its choice of the threshold values t_{500}, t_{2500}, t_{5000} and t_{7500}. Run 4 is a baseline run which only uses visual features. Unsurprisingly, run 5, which is based on a larger training set, yielded the best results. As further experiments have indicated, however, this increased performance is not only due to the larger training set, but also to the use of Dempster-Shafer theory to combine the different granularity levels.

4. REFERENCES

[1] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java CBIR library. In *Proc. ACM Multimedia*, pages 1085–1088, 2008.

[2] A. Rae, V. Murdock, P. Serdyukov, and P. Kelm. Working Notes for the Placing Task at MediaEval2011. In *Working Notes of the MediaEval Workshop*, 2011.

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

[4] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In *Proc. UAI*, pages 29–40, 1990.

[5] O. Van Laere, S. Schockaert, and B. Dhoedt. Combining multi-resolution evidence for georeferencing Flickr images. In *Proc. SUM*, pages 347–360. 2010.

[6] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proc. ACM ICMR*, 2011.