# Ghent and Cardiff University at the 2012 Placing Task

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

Steven Schockaert
School of Computer Science & Informatics
Cardiff University, UK
s.schockaert@cs.cardiff.ac.uk

Jonathan A. Quinn
School of Computer Science & Informatics
Cardiff University, UK
j.a.quinn@cs.cardiff.ac.uk

Frank C. Langbein
School of Computer Science & Informatics
Cardiff University, UK
F.C.Langbein@cs.cardiff.ac.uk

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

## ABSTRACT

We present the results of our submission to the MediaEval 2012 Placing Task. We used a framework that combines language models and similarity search, which improves our system from last year by using a different feature selection technique, extending our similarity search, tapping into new types of information for videos without any tags and including the use of SIFT features.

## Keywords

Georeferencing, Language models, SIFT features

## 1. INTRODUCTION

We participated in the 2010 and 2011 editions of the Placing Task with a system that combines the use of language models and similarity search. The most important lessons drawn from last year's results were that introducing a prior based on the home location of the user significantly boosts the results and that there is a clear need for a feature selection technique tailored to this task. Also, as this year an even larger part of the test videos contained no tags at all, we experimented with different types of textual information: the home location (as described by the user), the video titles and the descriptions. For a detailed overview of the details of the Placing Task, we refer to [1].

## 2. METHODOLOGY

### *Data aquisition and representation.*

The training data for our system consisted of the same subset of 2 096 712 georeferenced photos we used last year. For run 2, we extracted SIFT features from the training photos, for which we crawled the actual images where they were still available on Flickr. For run 5, a training set of 17 169 341 Flickr photos (with a reported accuracy of 16, meaning street-level precision) was used, being a subset of the crawl we already reported on in [2].

For the runs that allowed the use of gazetteers, we geocoded the home locations from the user profiles (where available), using the Google Geocoding API[1]. We obtained coordinates for 48.9% of the test videos (last year 46.5%).

The locations of the original 2M training photos were clustered into 500, 2500 and 10000 clusters as before; these clusterings will be referred to as $C_{500}$, $C_{2500}$ and $C_{10000}$. Instead of the $\chi^2$ feature selection method, which we used previously, we adopted the geospread measure introduced last year by Hauff [3] to create a ranking of the available features for both the 2M and 17M training sets. For each of the different clusterings, we created the vocabularies (i.e. sets of tags) $V_{500}$, $V_{2500}$ and $V_{10000}$, respectively containing 1.5M, 175K and 125K tags.

### *Estimating locations.*

The approach we use to estimate the location of the test videos consists of two steps. First, given a clustering of the training photos, we use a multinomial Naive Bayes classifier to find the most likely cluster to contain the location of a given test video. Second, within this cluster, we use similarity search to find those training items whose tags best resemble the tags of the test video. The assumption here is that the locations of the most similar training photos are the most plausible locations for the test video. We refer to [2] for more details.

Important parameters in this process are the number of clusters $k$ and the number of features that we retain. Using more clusters means that the Naive Bayes classifier can potentially make a better estimation of the location, reducing the importance of the similarity search step, although more clusters also increases the probability of classification errors. For this reason, we have used an adaptive approach, where we consider a larger number of clusters when a video has sufficiently informative tags. Specifically, if a video contains at least one of the top 125K tags, the clustering $C_{10000}$ is used. Else, if it contains at least one of the top 175K tags, the clustering $C_{2500}$ is used. Else, if it contains at least one of the top 1.5M tags, the clustering $C_{500}$ is used.

If a video does not have any tags from the top 1.5M (because it does not have any tags at all, or only geographically irrelevant tags), we look for other forms of textual information. This is the case for 43.4% of the test data (compared to 16.1% last year). First, we use the textual home location

*MediaEval 2012 Workshop,* October 4-5, 2012, Pisa, Italy

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

of the owner, the video title and description as if they were regular tags, to the extent that this data is available. We only do this for videos that originally did not have any tags. The tokens in these pieces of text are converted to lowercase and concatenations of up to 3 tokens are made (e.g. empirestatebuilding) after which those tokens found in the feature vocabulary $V_k$ are retained. If even with these additional sources of textual information, the video does not have a single tag within the top 1.5M, we default to the coordinates of the home location of the user if it is available (in runs where the use of gazetteers is allowed) and to a system-wide default location of 51.50733460,-0.12768310 (which is the city centre of London and corresponds to estimating the location using a maximum likelihood prior only).

### SIFT features.

In order to compute the similarity between photos in the training set and the test videos, we initially sample the first, middle, and last keyframes of the test videos. We then compute the SIFT features for these keyframes, and for the training photos, using the algorithm described by Lowe [4]. The SIFT feature vectors are then compared using a standard RANSAC similarity measure [5]. The output of this measure is a set of matched features. This measure commonly results in a non-symmetric result, and thus the comparison is done bi-directionally. Thus, for each test video, we have computed the number of intersecting matched features for each keyframe. The maximum of these three values is stored alongside the video.

### Improving similarity search.

Once a cluster has been chosen by the Naive Bayes classifier, we estimate the coordinates by looking at which photos in that cluster are most similar. Last year, we simply used the location $p_{sim}$ of the most similar image in terms of the Jaccard index. As an extension, this year we used one of three possible locations: the location $p_{sim}$ as before, the location $p_{home}$ of the user's home location (if permitted and available), and the location $p_{vis}$ of the most similar photo in terms of SIFT features (if available and the number of matching SIFT features is at least 20). In particular, we choose the location $p \in \{p_{sim}, p_{home}, p_{vis}\}$ minimizing the following expression

$$score(p) = \sum_{s \in \mathcal{S}} dist(p, s) \cdot jaccard(s, x)^{\lambda} \qquad (1)$$

where $\mathcal{S}$ contains the 10 most similar photos from the chosen cluster in terms of the Jaccard index, $dist(p, s)$ is the straight-line distance between $p$ and the location of photo $s$, $jaccard(s, x)$ is the Jaccard similarity between $s$ and the test video $x$ and $\lambda = 5$.

## 3. RESULTS AND DISCUSSION

Table 1 presents the results of the five runs. As can be concluded from the results of runs 1 and 2, using the home location of the user in the prior still makes a difference on this test set. Interesting to note is that runs 2 and 3 produce the same results. The only difference between the submissions is that run 2 uses the SIFT features whereas run 3 does not. This illustrates some of the difficulties we had in combining SIFT features with textual information. Unless there was a very strong visual match between a test video and an image, SIFT features only proved helpful when used in a very cautious way (as explained above). Moreover, in cases where there is a strong match (say, 50 matching features), the video usually contains a landmark. On the development data, the videos for which this was the case also had sufficiently informative tags which allowed us to find accurate coordinates anyway. It is tempting to speculate that the presence of landmarks makes it more likely to have informative tags (viz. the name of the landmark) and thus reduces the need for using visual similarity. Our submission to run 4 consisted of georeferencing every single test video to the system-wide default location in London, mentioned above. Finally, run 5 used only a single clustering, $C_{500}$, in combination with the dataset of 17M photos as training input for the similarity search.

|  | 1km | 10km | 100km | 1000km | 10000km |
|---|---|---|---|---|---|
| run 1 | 459 | 1175 | 1737 | 2422 | 3739 |
| run 2 | 475 | 1240 | 1973 | 2559 | 3763 |
| run 3 | 475 | 1240 | 1973 | 2559 | 3763 |
| run 4 | 4 | 31 | 107 | 887 | 3821 |
| run 5 | 862 | 1432 | 1983 | 2487 | 3753 |

**Table 1: Overview of the results on the test collection of 4182 videos, using textual tags (run 1); using textual tags and a gazetteer (run 3) as well as visual features (run 2); defaulting every estimate to London (run 4); and using tags and a gazetteer on an alternative, larger, training set (run 5)**

Our experiences with this year's task can be summarized as follows. First, given the large number of videos without any tags, it is important to exploit as much of the available information as possible. Using the textual home location, title and description of a video can considerably improve the results, if only for videos for which no (informative) tags are available. Second, while SIFT features may be able to improve the results in some particular cases, their computational cost seems hard to justify for this task. Finally, while moving from $\chi^2$ feature selection to the geospread measure from [3] improved our results, it seems that there is still scope for improving the results by developing feature selection methods tailored to this task.

## 4. REFERENCES

[1] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval2012. In *Working Notes of the MediaEval Workshop*, 2012.

[2] O. Van Laere, S. Schockaert, and B. Dhoedt. Ghent university at the 2011 Placing Task. In *Working Notes of the MediaEval Workshop*, 2011.

[3] Claudia Hauff and Geert-Jan Houben. WISTUD at MediaEval 2011: Placing Task. In *Working Notes of the MediaEval Workshop*, 2011.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, vol. 24, pp. 381–395, 1981.