CERTH at MediaEval Placing Task 2013

Giorgos Kordopatis-Zilos
Electrical Engineering Dept.
Aristotle University of
Thessaloniki, Greece
gkordopa@auth.gr

Symeon Papadopoulos,
Eleftherios
Spyromitros-Xioufis,
Information Technologies
Institute
CERTH, Thessaloniki, Greece
[papadop,espyromi]@iti.gr

Andreas L. Symeonidis,
Yiannis Kompatsiaris
Information Technologies
Institute
CERTH, Thessaloniki, Greece
[asymeon,ikom]@iti.gr

ABSTRACT

We describe the participation of the CERTH team in the Placing task of MediaEval 2013. We submitted 5 runs on the full test set, two of which are based on tag information, two on visual content, and one uses both tag and visual information. Our best performance (median error 650km) was achieved with the use of tag features.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Miscellaneous

1. INTRODUCTION

The goal of the task is to produce location estimates for a set of 262,000 images using a set of over 8.5 million geotagged images and their metadata for training. One may find more details regarding the challenge and the dataset in [2]. For the tag-based runs, we built upon the scheme of [6], making use of a two-level LDA scheme [1][1] to filter out non-geographic terms. For the visual-based runs, we relied on a simple Nearest Neighbour scheme using SURF+VLAD features [5] and an efficient indexing scheme for very fast retrieval [4]. Our hybrid run combined a tag and a visual run using a simple fall-back scheme. All models were built solely on the training data provided by the organizers (i.e. no external gazetteers or Internet data were used).

2. APPROACHES

2.1 Placing images using tags

The tag-based method relies on an offline analysis, in which a complex geographical-tag model is built from the tags and locations of the approximately 8.5 million images of the training set. The implemented approach comprises three steps.

A. Filtering: In this step, we aim at removing noisy and irrelevant tags from the training data. We, therefore, remove machine-tags from all images and then remove from the training set those images with no tags left. We end up with 7,266,903 images.

B. Spatial clustering and local LDA: Here, we first cluster the training set images based on their location, using k-means on their latitude-longitude values. We opted for $K = 5000$ clusters-areas so that on average each area would contain approximately 1450 images. For each such area, we then apply LDA to derive a *local topic distribution*, using 100 topics and 20 terms per topic. We denote an area j as $A_j = \{u_j, \{w_j\}, \{\tau_{jk}\}\}$, where u_j is the area id, w_j is the set of images belonging to the area, and τ_{jk} is the k-th topic (set of terms) of the local distribution.

C. Creating bag-of-excluded-words (BoEW): In this step, we attempt to create a set of non-geographic tags, i.e. tags that should not be taken into account for geotagging. To this end, we first apply LDA on the whole dataset (*global LDA*) using 500 topics and 50 terms per topic. For each of the resulting topics, we compute its frequency per area[2], thus ending up with a topic-area distribution (histogram). This essentially corresponds to the spatial distribution of the topic. Based on this, we compute its entropy and flag the topics that exceed a threshold of 180 (empirically selected). The terms of these topics form the bag-of-excluded-words. Some example excluded words include the terms *landscape*, *35mm*, *kodak*, *boats*, *christmas*, *sunset*, and *tree*, while some terms that were mistakenly considered as non-geographic include *europe*, *usa*, *atlanticocean* and *newmexico*, most probably due to their large geographic span.

Having created the geographical-tag model, we then proceed with the online location estimation step for each test image T_i. We first filter the tags of the image that are either machine-tags or belong to the BoEW, ending up with the set of $\{t_i\}$ clean tags. We then compute the Jaccard similarity between this set of tags and the set of tags for each topic of each local LDA:

$$s_{ijk} = \frac{|t_i \cap \tau_{jk}|}{|t_i \cup \tau_{jk}|} \qquad (1)$$

In a first variant of the approach (run 1), the image is assigned to the area with the highest Jaccard similarity with any local topic, $area_i = \arg\max_{j,k} s_{ijk}$. In the second variant of the approach (run 3), we first compute the mean Jaccard similarity for each area s_{ij} over all topics with at least one common tag with the test image, and then select the area with the highest mean similarity, $area_i = \arg\max_j s_{ij}$.

Having assigned the test image to an area, we then adopt the location estimation technique of [6]: we first determine the k most similar training images (using Jaccard similarity on the corresponding sets of tags) and use their center-of-

[1] We used the JGibbLDA implementation, available on: `http://jgibblda.sourceforge.net/`.

[2] This is computed by counting the images contained in the area that activate the topic. These are determined by applying a threshold ($= \frac{1}{0.9 \times 500}$) on the image-topic distribution.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

gravity (weighted by the similarity values) as the location estimate for the test image. For test images with no clean tags, we set their location equal to the centroid of the largest area (a kind of maximum likelihood estimation).

2.2 Placing images using visual features

In the offline analysis step, we extract optimized SURF+ VLAD features from each image in the training set (over 8.5 million images) and index the features using the IVFADC scheme proposed in [4]. The SURF+VLAD vectors are based on multiple vocabulary aggregation (four visual vocabularies with $k = 128$ centroids each) and joint dimensionality reduction (to only 48 dimensions for efficiency) with PCA and whitening [3]. The vectors were then quantized using a coarse quantizer with 1024 centroids and applied Product Quantization on the residual vectors using an 8×10 scheme [4], which led to a signature of length 112bits for each image. We made the implementation of the above feature extraction and indexing methods publicly available[3].

For the online location estimation step, we retrieve the top k visually most similar images and use those for the estimate. In the first variant (run 2), $k = 1$ and we simply consider the location of the most similar image as the location of the test image. In the second variant (run 4), $k = 20$ and we apply a simple incremental spatial clustering scheme, in which if the j-th image (out of the k most similar) is within 1km from the closest one of the previous $j - 1$ images, it is assigned to its cluster, otherwise it forms its own cluster. In the end, the largest cluster (or the first in case of equal size) is selected and its centroid is used as a location estimate.

3. RUNS AND RESULTS

As described above, we prepare two tag-based runs, which we will refer to as *tmax* (run 1) and *tmean* (run 3), and two visual runs which we will refer to as *vnn* (run 2) and *vclust* (run 4). A fifth run, referred to as *hyb*, was prepared using a hybrid strategy: if the test image had at least one clean tag associated with it, the *tmax* approach was selected, otherwise the *vnn* was used. All runs were prepared for the full test set of 262,000 images. The tag-based runs took approximately 23 hours to complete (316msec per image), while the visual ones took only 13 hours (179msec per image). These times were recorded on a commodity Quad core@2.40GHz with 8GB RAM with a 1TB 7200rpm hard drive.

Table 1 summarises the obtained results for the full test set. The best performance in terms of median error was attained by *tmax*, closely followed by *hyb*. In terms of accuracy, the *hyb* run performs marginally better than *tmax* in the low ranges (<1km up to <1000km), which is expected since it has a fall-back scheme for the images with no tags. A noteworthy result is the very low performance of visual runs. Potential reasons for this poor performance include the very "cheap" feature extraction and indexing settings (for efficiency reasons) and the nature of the training and test set, i.e. the training set did not contain images that were sufficiently visually similar to those of the test set.

We also computed the selected performance measures for the test subsets proposed by the organizers. Surprisingly, we could not identify a trend with respect to the test set size. For instance, on the small test set (5300), the best median error (by *hyb*) was 1423km, while for the test set of 53000 images, the best median error (by *tmax*) was just 521km. The accuracy measure was more stable, e.g. in the case of *tmax* ranging from 9.62% to 10.95% (for <1km) and from 30.34% to 38.01% (for <10km). Similar fluctuations were noted for all other runs, which indicates that each test set has an inherent degree of placeability.

measure	**tmax**	**vnn**	**tmean**	**vclust**	**hyb**
acc(1km)	10.26	0.60	7.82	0.76	*10.37*
acc(10km)	23.53	0.99	19.86	1.16	*23.70*
acc(100km)	36.27	1.86	31.99	2.04	*36.22*
acc(500km)	47.20	6.49	43.31	6.64	*47.36*
acc(1000km)	53.12	13.43	49.74	13.50	*53.27*
median error	*651*	6715	1028	6691	681

Table 1: Geotagging accuracy (%) for five ranges and median error (in km). Runs *tmax* and *tmean* were based on the approach of subsection 2.1, while *vnn* and *vclust* on the approach of subsection 2.2.

4. FUTURE WORK

In the future, we plan to conduct a more thorough analysis on the different sources of error for the proposed scheme, and extend it to also include additional metadata of the input images, as well as external resources. Regarding the error analysis, we will look into the impact of (a) the number of topics and terms per topic both for the local and the global LDAs, (b) the selection of the entropy threshold, (c) the number K of geographical areas, (d) the type and quality of visual features, (e) the use of better visual matching methods (e.g. geometric verification applied on the list of top k most similar images). Regarding the use of additional metadata, we plan to incorporate the author of an image as an indicator of the image location. Finally, we will consider incorporating resources such as gazetteers, as well as additional geotagged image data collected from the Web with the goal of increasing the visual coverage of the training set.

5. ACKNOWLEDGEMENTS

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] C. Hauff, B. Thomee, and M. Trevisiol. Working Notes for the Placing Task at MediaEval 2013, 2013.

[3] H. Jégou and O. Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *ECCV*, 2012.

[4] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. PAMI*, 33(1):117–128, 2011.

[5] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. An empirical study on the combination of SURF features with VLAD vectors for image search. In *WIAMIS*, 2012.

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

[3] https://github.com/socialsensor/multimedia-indexing