Imcube @ MediaEval 2015 Placing Task: A Hierarchical Approach for Geo-referencing Large-Scale Datasets

Pascal Kelm, Sebastian Schmiedeke, and Lutz Goldmann
Imcube Labs GmbH
Berlin, Germany
{kelm, schmiedeke, goldmann}@imcube.de

ABSTRACT

This paper describes Imcube's geo-referencing approach, experiments, and results at the MediaEval 2015 Placing Task benchmark. This task requires to develop techniques to automatically annotate Flickr photos and videos with their geolocation (latitude and longitude) in two individual subtasks. A hierarchical approach combining textual, visual and optional routing information is employed. The results show that for 24% of the images (local-based task) and for 96% of the images (mobile-based task) the error of the estimated location is below city level (10 km).

1. INTRODUCTION

The MediaEval Placing Task 2015 [1] requires that participants use systems to automatically estimate the location of Flickr photos and videos using any or all of metadata, visual/ audio content, and/or user information.

This year the task introduces two new sub-tasks: The *locale-based* sub-task addresses the prediction of missing locations for individual images in an entity-centred way by choosing a location from a given ground truth hierarchy. The *mobile-based* sub-task addresses predicting missing locations within a sequence of photos shot by a travelling photographer.

Similar to the Placing Task 2014, the training set (4,672,382 photos & 22,767 videos) and test sets (931,573 photos & 18,316 videos) were sampled from the YFCC100M [6] data set. One important difference to the past editions is that this year the distances between the predicted and the ground truth geographic coordinates are evaluated using Karney's formula [2], which is based on the assumption that the shape of the Earth is an oblate spheroid.

In this paper, we present an approach that combines different textual and visual descriptors by applying a hierarchical scheme to merge information obtained from several ranked lists.

2. SYSTEM DESCRIPTION

This section describes the different methods created to solve the challenges of the locale-based and mobile-based sub-tasks.

2.1 Local-based Sub-task

The proposed approach is composed of different steps: (i) hierarchical clustering of the provided training set by latitude and longitude, (ii) visual and textual feature extraction, (iii) generation of ranked lists, (iv) re-ranking and (v) estimation of the location for each test item.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

The hierarchy provided contains 221,458 leaf nodes (locations) that are spread across 253 countries. Below the second level (Country $\succ$ State) we segment the states into 360×180 regions according to the meridians and parallels. We also apply a smaller grid of segments with half the spatial dimensions to increase the accuracy and to minimize the computational cost. Each geo-referenced training image is assigned to its corresponding grid cell at the lowest level [3]. For each layer of the hierarchy a ranking model is used to iteratively assign a test image to the most likely spatial segment.

Due to the large size of the dataset and the limited processing time, we did not apply a hierarchical language model approach with multiple modalities [3] but adopted a textual re-ranking model. The vocabulary of the spatial locations includes stemmed[1] words from the tags, titles and descriptions. The text similarity function used is BM25 [4] as implemented by Lucene[2]. The best results for textual similarity computations were achieved with a training set composed of both image and video meta data, regardless of the kind of test query.

The visual similarity relies on a wide spectrum of visual features to describe the color and texture characteristics of the video key frames and photos. These image descriptions are pooled for each leaf node in the different hierarchy level using the mean and median value of each descriptor. A kd-tree that contains all appropriate segments is built for each descriptor in each leaf node. This procedure speeds up the following search because only a portion of data is needed to be computed for nearest neighbour search.

Starting at the top of the hierarchy the nodes of the current level are ranked according to their distance to the test image. The overall distance is obtained by fusing the textual and the visual distances using weighted summation. These weights differ in both fusion experiments as described in results section. Then the node with the lowest distance becomes the most likely location at the given level of granularity. By iteratively traversing the hierarchy the method determines the leaf node that has the highest similarity to the test image and returns the corresponding geolocation.

2.2 Mobile-based Sub-task

For this task, we pursue a similar approach as described in section 2.1 but without the hierarchical layer model and with additional routing information.

We use OpenStreetMap[3] to find the shortest route between two photos that have associated geographic coordinates. Tracks with a distance smaller than 2 km are routed by pedestrian navigation, larger tracks are routed by car navigation, respectively. The results

[1] http://tartarus.org/martin/PorterStemmer/index.html
[2] http://lucene.apache.org/core/
[3] http://www.openstreetmap.org/

Table 1: Results locale-based sub-task.

Distance	Textual		Visual		Fusion1		Fusion2	
	#Items	Percentage	#Items	Percentage	#Items	Percentage	#Items	Percentage
0.001 km	678	0.07 %	0	0.00 %	277	0.03 %	1669	**0.18 %**
0.01 km	1906	0.20 %	3	0.00 %	1030	0.11 %	4549	**0.48 %**
0.1 km	17437	1.84 %	10	0.00 %	11372	1.20 %	31980	**3.37 %**
1 km	81274	8.56 %	150	0.02 %	53172	5.60 %	117491	**12.37 %**
10 km	200103	21.07 %	1676	0.18 %	133321	14.04 %	224080	**23.59 %**
100 km	352851	37.15 %	5121	0.54 %	275985	29.05 %	353357	**37.20 %**
1000 km	658519	69.33 %	52002	5.47 %	634327	66.78 %	658519	**69.33 %**
10000 km	927620	97.66 %	708993	74.64 %	927121	97.60 %	927620	**97.66 %**

Table 2: Results mobile-based sub-task.

Distance	Routing		Visual		Weighted Visual		Textual	
	#Items	Percentage	#Items	Percentage	#Items	Percentage	#Items	Percentage
0.001 km	2	0.02 %	4	0.04 %	3	0.03 %	20	**0.21 %**
0.01 km	81	0.84 %	138	1.43 %	128	1.32 %	244	**2.52 %**
0.1 km	1593	16.47 %	1949	20.14 %	1957	**20.23 %**	1952	20.18 %
1 km	6501	67.19 %	7014	72.50 %	7026	**72.62 %**	6959	71.93 %
10 km	9171	94.79 %	9274	95.86 %	9276	95.88 %	9280	**95.92 %**
100 km	9659	99.83 %	9671	**99.96 %**	9670	99.95 %	9670	99.95 %
1000 km	9674	99.99 %	9675	100.00 %	9675	100.00 %	9675	100.00 %
10000 km	9675	100.00 %	9675	100.00 %	9675	100.00 %	9675	100.00 %

of the *routing* run are predicted linearly in travel time to be a location on these tracks. For test images, which do not have both chronological neighbours, the neighbouring route segment is extrapolating while considering their distance in time. The other runs use additionally textual and visual features to determine the most similar image along the track.

The visual similarity is determined as described in [5]. Densely sampled local features (pairwise averaged DCT coefficients) are represented as a histogram quantised by vector quantisation (a clusterless bag-of-visual-words approach) [5]. As similarity metric between training images and the image to be geo-tagged, histogram intersection of their BoW representation is applied. The two visual runs differ in the assignment of coordinates: the *visual* run assigns the coordinates of the visually most similar image from the training data, the *weighted visual* run calculates the coordinates as the centroid of all training images weighted by their visual similarity.

The *textual* run uses the same textual similarity as the location task, but the training images are restricted to be located within a corridor of 0.001 degree along the estimated routes.

3. RESULTS

3.1 Local-based Sub-task

Table 1 shows the accuracies of selected error margins for the different textual and visual runs. Based on the experience from the previous years we expect the *textual* run to perform better than the *visual* run due to the visual ambiguity at coarser levels. The results clearly show that the visual only approach has low accuracy in all error margins when compared to the textual only approach. For combining the textual and visual information we have tested two different fusion models. We design a set of two fusion experiments to combine textual and visual features. Our first fusion model (*Fusion1*) combines the estimations of textual and visual models equally on each hierarchy level. The second fusion model (*Fusion2*) only combines these estimations on the finest three hierarchy levels. On the coarsest hierarchy levels, only the estimation of the textual model is used. This combination results in more accurate results, since visual feature are not able to solve ambiguities in large scale (i.e. , most cityscapes look similar). The results show that fusing visual and textual information on finer levels (Fusion1) improves the performance for error margins between range between 10 m and 100 km.

3.2 Mobile-based Sub-task

Table 2 shows the results obtained with our four runs (routing, visual, weighted visual, and textual) described in section 2.2. Generally, the use of textual and visual features improves the location performance compared to only using interpolation along routes. Since most of the routes are quite short, improvements are mainly made at smaller error margins. For the margin of error below 1 m the *textual* approach outperforms all other approaches by a factor of 10. The different runs reach a similar effectiveness with increased error margins. The weighted visual similarity approach (*Weighted Visual*) predicts slightly more accurate locations within a range of 100 m to 1 km. For error margins above 10 km all runs produce similar results. A closer look at the location errors of the individual images shows that the textual and visual approaches perform very differently which suggests that a suitable fusion approach may further improve the results.

4. CONCLUSION

The results of the local-based sub-task show that the best performance can be achieved with a multimodal fusion approach that uses textual information on coarser levels and the combination of visual and textual information in finer ones. The results of the mobile-based sub-task show that the use of visual and textual information beside routing information improves the location estimation. The low correlation of the localization errors of the different approaches suggests that more advanced fusion approaches will lead to better results. Another interesting direction to improve the accuracy of the visual approach for both sub-tasks is by using local features to distinct landmarks and points of interest.

5. REFERENCES

[1] J. Choi, C. Hauff, O. V. Laere, and B. Thomee. The placing task at mediaeval 2015. *MediaEval 2015 Workshop*, 2015.

[2] C. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013.

[3] P. Kelm, S. Schmiedeke, J. Choi, G. Friedland, V. N. Ekambaram, K. Ramchandran, and T. Sikora. A novel fusion method for integrating multiple modalities and knowledge for multimodal location estimation. In *Proceedings of the 2Nd ACM International Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '13, pages 7–12, New York, NY, USA, 2013. ACM.

[4] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3 '95*, pages 109–126, 1995.

[5] S. Schmiedeke, P. Kelm, and T. Sikora. Dct-based features for categorisation of social media in compressed domain. In *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, pages 295–300, 2013.

[6] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.