Multimedia Geocoding: The RECOD 2014 Approach

Lin Tzy Li[1], Otávio A. B. Penatti[1,2], Jurandy Almeida[1,3], Giovani Chiachia[1], Rodrigo T. Calumby[1,4], Pedro R. Mendes Júnior[1], Daniel C. G. Pedronette[1,5], Ricardo da S. Torres[1]

[1]RECOD Lab, Institute of Computing, University of Campinas (UNICAMP), Campinas, SP – Brazil, 13083-852
[2]Advanced Technologies, SAMSUNG Research Institute, Campinas, SP – Brazil, 13097-160
[3]Institute of Science and Technology, Federal University of São Paulo (UNIFESP), Sao José dos Campos, SP – Brazil, 12247-014
[4]Dept. of Exact Sciences, University of Feira de Santana (UEFS), Feira de Santana, BA – Brazil, 44036-900
[5]Dept. of Stat., Applied Math. and Computing, Universidade Estadual Paulista (UNESP), Rio Claro, SP – Brazil, 13506-900
{lintzyli, chiachia, pedro.mendes, rtorres}@ic.unicamp.br, o.penatti@samsung.com,
jurandy.almeida@unifesp.br, rtcalumby@ecomp.uefs.br, daniel@rc.unesp.br

ABSTRACT

This work describes the approach proposed by the RECOD team for the Placing Task of MediaEval 2014. This task requires the definition of automatic schemes to assign geographical locations to images and videos. Our approach is based on the use of as much evidences as possible (textual, visual, and/or audio descriptors) to geocode a given image/video. We estimate the location of test items by clustering the geographic coordinates of top-ranked items in one or more ranked lists defined in terms of different criteria.

1. INTRODUCTION

Geocoding multimedia material has gained greater attention in the latest years given its importance for providing richer services for users, like placing information on maps or providing geographic searches. The Placing Task at MediaEval 2014 [2] challenges participants to assign geographical locations to images and videos automatically.

In this paper, we present our approach that combines different textual, audio, and/or visual descriptors uniformly by applying a clustering scheme to merge information defined by several ranked lists.

2. PROPOSED APPROACH

The approach used is composed of five steps: (i) image/video feature extraction, (ii) generation of ranked lists, (iii) re-ranking, (iv) clustering by lat/long of the top-ranked items (considering one or multiple ranked lists), and (v) assigning to the test item the lat/long of the sample with the highest density value.

For evaluation purposes in the training phase, we created a *validation set* sampling 5,000 images and 1,000 videos from the development/training set. This set was created as follows. First, each item in the development set was assigned to a fixed cell of 1-by-1 degree based on its latitude and longitude. Then, the resulting grid was summarized by the number of photos (density) in each cell. Next, we randomly picked up images/videos from each cell considering their proportional distribution over the original dataset. To keep the validation step with similar characteristics to the real development and testing sets, items from users who have image/video selected for the validation set were removed from the new training set, creating a subset from the original full development set with 4,485,331 images and 14,115 videos. Therefore, to evaluate our strategies before conducting the final runs, we used the validation set with the partial training set created as described above.

2.1 Features

Textual. From textual metadata, the title, description, and tags of photos/videos were concatenated as one field to compute text similarities between the test and training items. The text was stemmed and stopwords were removed. The text similarity functions used were BM25 and TF-IDF as implemented by the Lucene.[1] The best results for textual similarity computation used a training set composed of both image and video metadata, regardless the kind of test query.

Audio/Visual. Videos and images were handled differently. For images, we used the provided CEDD, Gabor, and FTCH and extracted additional features: OverFeat[2] and BIC [3]. Before extracting additional features from images, we resized them to at most 100k pixels. For videos, we used the provided features: GIST (static feature) and MFCC (audio feature), besides extracting HMP motion feature [1].

2.2 Re-ranking, clustering & geocoding

We first used the full development set as geo-profiles and each test item was compared to the whole development set for each feature independently. For a given test item, a ranked list for each feature was produced. Given the ranked lists, we explored two strategies:

1. Re-ranking items using the RL-Sim algorithm [5]. It relies on contextual information encoded in the similarity between ranked lists. This method exploits the fact that if two images are similar, their ranked lists should be similar as well. Therefore, a contextual distance measure is defined based on the similarity of ranked lists. As the top-n positions hold more relevant items, we focus on them to define the final list considering m input ranked lists (m features).

We were able to apply the re-ranking algorithm (using the top $n = 15$ items of the original ranked lists) only to the video dataset, due to its small size and to the number of required inputs for the algorithm.

2. Clustering lat/long points derived from the top-n items of ranked lists. Input lists of the clustering method were

[1]`http://lucene.apache.org/core/` (as of 10/2014).
[2]`https://github.com/sermanet/OverFeat` (as of 09/2014).

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Runs configuration.

	Photos/Images (500,000)			**Videos** (10,000)		
Run	**Textual**	**Visual**	**Geocoding**	**Textual**	**Visual/ Audio**	**Geocoding**
1	BM25 & TF-IDF	-	OPF (top10)	BM25 & TF-IDF	-	OPF (top10)
2	-	BIC & OverFeat	OPF (top100, 50 per list)	-	Re-ranking (HMP & GIST & MFCC)	OPF (top100)
3	BM25 & TF-IDF	BIC & OverFeat	OPF (top30, 7 per list)	BM25	Re-ranking (HMP & GIST & MFCC)	OPF (top10, 5 per list)
4	BM25	-	OPF (top5)	BM25	-	OPF (top5)
5	-	OverFeat	OPF (top100)	-	HMP	OPF (top100)

Table 2: Overall effectiveness (test set).

Precision/Run	**1**	**2***	**3**	**4**	**5***
10m	0.55%	0.09%	**0.59%**	0.52%	0.10%
100m	6.06%	0.78%	**6.26%**	5.77%	0.78%
1km	21.04%	1.86%	**21.15%**	20.52%	1.89%
10km	**37.59%**	4.02%	37.50%	37.00%	3.98%
100km	46.14%	5.91%	46.03%	**45.39%**	5.88%
1,000km	**61.69%**	21.39%	61.41%	60.52%	20.91%
5,000km	**76.76%**	45.08%	75.07%	76.02%	45.17%

* no metadata use

defined for a single feature (i.e., one list only), for the result of the re-ranking of m features, or from a set of m independent lists associated with m different features. We used Optimum-Path Forest (OPF) [6] for clustering the input list(s) related to a given test sample. OPF created a graph as follows: for each item s, a node was defined; each node s was then linked to its k nearest neighbors ($k = 3$ was used in all the cases). Then, each item/node in the graph received a density value according to the formula proposed by Rocha et al. [6]. The lat/long of the test item were inherited from the graph's sample/node with highest density value. When using m ranked lists generated for m different descriptors, we combined the top $\lfloor \frac{n}{m} \rfloor$ items for each ranked list to create the graph.

3. OUR SUBMISSIONS & RESULTS

None of our submissions used extra crawled material or gazetteers. Based on configuration from our best results on evaluation set, our submission was set as shown in Table 1.

For test items that had no lat/long estimation (because of missing/empty features), we randomly selected an item from the development set to assign its latitude and longitude to the test item. For runs of textual feature only (Runs 1 & 4), those represented 1.07% of the test items, while for visual only runs (2 & 5) they were 0.02% and 0.03% respectively. For multimodal run (3) there were only 2 cases. We have also noted that 0.58% of the test images were the *unavailable message* of Flickr, warning that the item was unavailable.

As we can observe in Table 2, the test run combining textual and visual information (Run 3) yields the best results for lower precision radii (10 m, 100 m, and 1 km), while using only textual descriptors via OPF clustering (Run 1) produces better from 10 km precision level on.

For non-textual runs (Run 2 and Run 5), at precision level up to 1 km the results using only one visual feature (Run 5) are slightly higher (0.01, 0.00, and 0.03 percentage point) than combining different features (Run 2). The opposite is true when we observe results from 10 km on. It seems that there were some disagreement between the two combined visual features that were accommodated by the geocoding method applied, which affected the results precision.

Table 3: 10,000 test videos results (visual only).

Run	**10m**	**100m**	**1km**	**10km**	**100km**	**1000km**	**5000km**
2	0.02%	0.04%	0.18%	1.28%	2.49%	10.38%	33.29%
5	0.01%	0.03%	0.14%	1.33%	3.10%	13.93%	43.65%

During the validation stage of the OPF clustering, we have noticed that when textual features are used, the number of top-n items to be clustered should be lower than when using only visual features. Otherwise, the textual results were degraded when more points are considered in the clustering process. For example, Run 4 (textual) result was derived from top-5 point clustering, while Run 5 (visual) was based on lat/long from top-100 items.

Comparing the results using re-ranking to combine visual features of videos (Run 2) with just HMP feature (Run 5), the test results showed that up to 1km precision the fusion by re-ranking (Run 2) improved the results over using just one feature (Run 5), but for larger radii it is the other way around, as shown in Table 3. Considering that we aim to geocode items as precisely as possible, re-ranking and clustering strategies have shown promising results.

4. CONCLUSIONS

In this work, we explored re-ranking and clustering approaches to geocode multimedia items based on the similarity of ranked lists. We observed that geocoding results were influenced by the number of top-n items of a ranked list used to cluster or re-rank. It seems that textual features require less top items than visual descriptors.

As future work, we plan to explore further configuration and approaches using different clustering and re-ranking strategies. We also plan to combine the strategies used this year with rank aggregation methods [4].

Acknowledgments

We thank FAPESP (#2013/08645-0 and #2013/11359-0), CNPq (306580/2012-8 and 484254/2012-0), CAPES, Samsung, and Placing Task organizers.

5. REFERENCES

[1] J. Almeida, N. J. Leite, and R. da Silva Torres. Comparison of video sequences with histograms of motion patterns. In *ICIP*, pages 3673–3676, 2011.

[2] J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, and D. Poland. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *ACM GeoMM*, 2014.

[3] R. de O. Stehling, M. A. Nascimento, and A. X. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *CIKM*, pages 102–109, 2002.

[4] L. T. Li, D. C. G. Pedronette, J. Almeida, O. A. B. Penatti, R. T. Calumby, and R. da Silva Torres. A rank aggregation framework for video multimodal geocoding. *Mult. Tools and App.*, pages 1–37, 2013.

[5] D. C. G. Pedronette and R. da Silva Torres. Image re-ranking and rank aggregation based on similarity of ranked lists. *PR*, 46(8):2350–2360, 2013.

[6] L. M. Rocha, F. A. M. Cappabianco, and A. X. Falcão. Data clustering as an optimum-path forest problem with applications in image analysis. *Int J Imag Syst Tech*, 19(2):50–68, 2009.