Multimodal image geocoding: the 2013 RECOD's approach

Lin Tzy Li[1,2], Jurandy Almeida[1,3], Otávio A. B. Penatti[1], Rodrigo T. Calumby[1,4], Daniel C. G. Pedronette[1,5], Marcos A. Gonçalves[6], and Ricardo da S. Torres[1]

[1]RECOD Lab, Institute of Computing, University of Campinas (UNICAMP), Campinas, SP – Brazil, 13083-852

[2]Telecommunications Res. & Dev. Center, CPqD Foundation, Campinas, SP – Brazil, 13086-902

[3]Institute of Science and Technology, Federal University of Sao Paulo (UNIFESP), Sao Jose dos Campos, SP – Brazil, 12231-280

[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

[6]Dept. of Computer Science, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG – Brazil, 31270-010

{lintzyli, jurandy, penatti, rtripodi, rtorres}@ic.unicamp.br, daniel@rc.unesp.br, mgoncalv@dcc.ufmg.br

ABSTRACT

This work describes the approach used by the RECOD team in the MediaEval Placing Task of 2013, in which we were required to develop an automatic scheme to assign geographical locations to images. Our approach is multimodal, considering textual and visual descriptors, which are combined by a rank aggregation strategy. We estimate the location of test images based on the coordinates of top-ranked images in the list of combined results.

1. INTRODUCTION

Geocoding multimedia material has gained great attention in the latest years given the importance of providing richer services for users, like placing information on maps. Image geocoding is the objective of the Placing Task in 2013, *i.e.,* it requires participants to assign geographical locations to images. Details about the Placing task, its dataset, and the evaluation protocol can be found in [1].

In this paper, we present our multimodal approach that combines different textual and visual descriptors uniformly. We combine them using a rank aggregation strategy, previously introduced in [4].

2. PROPOSED APPROACH

We handled the task of automatically assigning a geographical location to images using nearest neighbor searches on aggregated ranked lists, which combine textual and visual features. The strengths of our approach are its simplicity and its power to combine multiple description modalities.

For evaluation purposes in the training phase, we have selected a validation set of 5,000 images from the development set of around 8.5 million images. First, each photo from the development set was assigned to a fixed cell of 1-by-1 degree based on its ground truth latitude and longitude. Then, the resulting grid was summarized by the total of photos (density) in each cell regarding to the dataset size. Finally, the evaluation images (5,000 photos) were randomly picked up from each cell, by taking into account its density.

2.1 Features

Textual

From textual metadata, we used only the photo tags to compute similarities between the images. The tags were stemmed and stopwords were removed. The text similarity functions used were BM25 and TF-IDF, as implemented by the Lucene API.

Visual

Given the large dataset, we had to select carefully the descriptors to be used. Initially, we have evaluated some of the descriptors provided with the dataset, like: color and edge directivity descriptor (CEDD), scalable color (SCD), gabor filter. Using the validation set, we have noticed that the best results were achieved by CEDD. Although SCD has shown the best results in [2], in our validation set, it did not performed well for our geocoding approach.

Additionally to CEDD, we used BIC (border/interior pixel classification). This descriptor was chosen due to its good results in large scale experiments [5]. For this, we downloaded the whole photo dataset, resizing the images to have at most 100 thousand pixels, as suggested by [6] for large scale experiments, and extracted the 128-dimensional BIC feature vector of each image. The Manhattan distance (L1) was used for both BIC and CEDD.

2.2 Rank aggregation

As last year, we used a rank aggregation strategy to combine different descriptors [3]. For this year, due to the size of the development set, we created a ranked list limited to the top 1,000 most similar photos for each test image.

We have used an aggregation function similar to sim_a (numerator is m instead of 2) proposed in [3]. When the intersection of top-1000 lists computed by different features are small, the size of the final aggregated list tends to ($m \times 1000$), being m the number of features combined. We select the top-1000 images that present the highest aggregated score as the output of the rank aggregation step.

2.3 Geocoding

For geocoding the test images, we have used a nearest neighbor approach. We used the development set ($\sim$8.5 million images) as geo-profiles and each test image was compared to the whole development set. For comparing the images, we have used each type of feature independently (textual or visual). For a given test image, the ranked list of each feature is produced. All the lists are then combined by our rank aggregation strategy and the final ranked list is generated. The lat/long of the first image (most similar) in this final list is assigned to the test image.

3. OUR SUBMISSIONS & RESULTS

Submitted runs

Our submissions for this year are:

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

run1: combines 2 textual descriptors: BM25 + TF-IDF;

run2: combines 2 visual descriptors: BIC + CEDD;

run3: one visual descriptor: BIC;

run4: combines 2 textual and 2 visual descriptors: BM25 + TF-IDF + BIC + CEDD;

run5: combines 4 textual descriptors: BM25 + TF-IDF[1].

Runs 1 and 5 used only textual features. Thus, for test images without tags, there was no way to apply our similarity ranked list approach. As post-processing, we randomly selected an item from the development set to transfer its latitude and longitude to the test image.

3.1 Results

Besides the organizers' standard evaluation metric, we also applied the WAS score we proposed in [4]. This evaluation metric gives an overview of a method's performance expressed by a score between [0,1], 0 being very bad and 1 indicating a perfect estimate with a higher weight assigned to more precise results. The WAS takes into account every single result of the whole test set to indicate and summarize the level of precision of an evaluated method as a whole.

Let $d(i)$ be the geographic distance between the predicted and the ground truth location of the image i. The proposed score for the result of a given test image i is defined as: $score(i) = 1 - \frac{\log(1+d(i))}{\log(1+R_{max})}$, where R_{max} is the maximum distance between any two points on the Earth's surface (half of Earth's circumference at the Equator is 20,027.5 km).

Let D be a test dataset with n images whose locations need to be predicted. The overall score for the predictions of a method m is defined as: $WAS(m) = \frac{\sum_{i=1}^{n} score(i)}{n}$.

Table 1: Validation set results.

Precision	Run 1	Run 2	Run 3	Run 4	Run 5
1km	64.56%	16.86%	15.32%	68.82%	64.62%
10km	73.64%	17.68%	16.10%	75.90%	73.60%
100km	77.58%	18.64%	17.04%	78.94%	77.58%
500km	80.20%	22.86%	13.40%	81.10%	80.22%
1000km	82.18%	28.32%	20.12%	82.74%	82.32%
WAS score	**0.7866**	**0.3053**	**0.2889**	**0.8019**	**0.7866**
Distance distribution					
1st Quartile	0.00	698.40	885.30	0.00	0.00
Median	0.03	5,499.40	5,835.80	0.00	0.04

Table 2: Test results using *test3* set (53,000 items).

Precision	Run 1	Run 2	Run 3	Run 4	Run 5
1km	20.14%	0.37%	0.28%	20.11%	18.82%
10km	37.60%	0.80%	0.67%	37.10%	35.93%
100km	47.66%	1.69%	1.51%	46.97%	45.97%
500km	56.62%	6.73%	6.25%	55.83%	55.74%
1000km	63.17%	14.32%	13.78%	62.26%	62.43%
WAS score	**0.5240**	**0.1653**	**0.1623**	**0.5190**	**0.5128**
Distance distribution					
1st Quartile	1.73	1,869.00	1,962.00	1.76	2.05
Median	168.22	6,632.00	6,729.00	196.79	225.67

As we can observe in Table 2, the test runs based solely on textual information yielded the best results (runs 1, 4, and 5), while those based only on visual descriptors presented low accuracy. The possible reason is the semantic gap, as there might be many different places with similar visual appearance, specially in a large dataset like the one used for training. Another potential issue was the large number of ties in the first positions of ranked lists of visual descriptors. Given our 1-nn geocoding approach, this probably degraded our results. However, we can see that by combining BIC+CEDD (run 2) we improve the results of BIC alone (run 3). The combination of textual and visual descriptors (run 4) was slightly worse than the textual descriptors isolated. One possible reason is the large difference between textual and visual results.

Observe that for the test set (Table 2), our results were quite different for our validation set (Table 1), mainly for the visual features. While in the *test3* set, BIC achieved less than 1% in the 1km radius, in the validation set, it presented 15.32%. Because of this, in the validation set, the fusion (run 4) results improved over run 1. The huge difference between validation and test results might be due to a property of the test set not considered when building the validation set: the users who contributed for the photos in the training set are different from those who contributed for the photos in the test set.

Regarding the distribution of test results, for the visual descriptors (runs 2 and 3), the 1st Quartile shows that 25% of the items were geocoded at most 1,900km from the correct location. On the other hand, for the textual descriptors and their combinations (runs 1, 4, and 5), 25% of the items are very close to their correct locations (less than 3km).

4. CONCLUSIONS

Our best results were observed for the methods based only on textual description. For them, we could geocode within 1km radius around 20% of the testing set (test3). Considering visual descriptors, the main challenge this year was the large scale dataset, which poses time and space constraints in the descriptors to be used. Our rank aggregation strategy, for the test set, was only effective for combining textual descriptors. Combining textual and visual descriptors did not improve the results. As future work, we would like to evaluate a more elaborate geocoding approach, similar to the scheme used to create our validation set, for example.

Acknowledgments

We thank the support of FAPESP (2011/11171-5, 2009/10554-8), CNPq (306580/2012-8, 484254/2012-0), CAPES, FAPEMIG, Samsung, ACM SIGIR, and MediaEval organizers.

5. REFERENCES

[1] C. Hauff, B. Thomee, and M. Trevisiol. Working Notes for the Placing Task at MediaEval. In *MediaEval 2013 Workshop*, volume 1043, October 18-19 2013.

[2] P. Kelm, S. Schmiedeke, and T. Sikora. Multimodal geo-tagging in social media websites using hierarchical spatial segmentation. In *International Workshop on Location-Based Social Networks*, pages 32–39, 2012.

[3] L. T. Li, J. Almeida, D. C. G. Pedronette, O. A. B. Penatti, and R. da S. Torres. A multimodal approach for video geocoding. In *Working Notes Proc. MediaEval Workshop*, volume 927, 2012.

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

[5] O. A. B. Penatti, E. Valle, and R. da S. Torres. Comparative study of global color and texture descriptors for web image retrieval. *J. Vis. Comm. and Image Repr.*, 23(2):359–380, 2012.

[6] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, pages 3482–3489, 2012.

[1]Non-English tags were translated to English using the Google Translate service and combined with the original tags.