RECOD @ Placing Task of MediaEval 2016: A Ranking Fusion Approach for Geographic-Location Prediction of Multimedia Objects

Javier A. V. Muñoz[1], Lin Tzy Li[1], Ícaro C. Dourado[1], Keiller Nogueira[5], Samuel G. Fadel[1],
Otávio A. B. Penatti[1,4], Jurandy Almeida[1,2], Luís A. M. Pereira[1],
Rodrigo T. Calumby[1,3], Jefersson A. dos Santos[5], Ricardo da S. Torres[1]
[1]RECOD Lab, Institute of Computing, University of Campinas (UNICAMP),
[2]GIBIS Lab, Institute of Science and Technology, Federal University of São Paulo (UNIFESP), [3]University of Feira de Santana,
[4]SAMSUNG Research Institute Brazil, [5]Department of Computer Science, Universidade Federal de Minas Gerais (UFMG) – Brazil
jalvarm.acm@gmail.com, {lintzyli, samuel.fadel, luis.pereira, rtorres}@ic.unicamp.br, icaro.dourado@students.ic.unicamp.br,
o.penatti@samsung.com, jurandy.almeida@unifesp.br, rtcalumby@ecomp.uefs.br, {keiller.nogueira, jefersson}@dcc.ufmg.br

ABSTRACT

We describe the approach proposed by the RECOD team for the estimation-based sub-task of Placing Task at MediaEval 2016. Our approach uses genetic programming (GP) to combine ranked lists defined in terms of textual and visual descriptors to automatically assign geographic locations to images and videos.

1. INTRODUCTION

By having multimedia content annotated with geographic information, we can provide richer services for users such as placing information on maps and providing geographic searches. Since 2011, the Placing Task [3] at MediaEval has been challenging participants to assign the geographical locations to images and videos automatically.

Here we present our approach for the estimation-based subtask of the Placing Task 2016. It combines textual, audio, and/or visual descriptors by applying rank aggregation and ranked list density analysis to combine multimodal information encoded in ranked lists. We evaluated new features and a genetic programming (GP) [5] approach for multimodal geocoding. GP provides a good framework for modeling optimization problems even when the variables are functions. We applied combinations of rank aggregation methods defined by a GP framework. The idea is to automatically select a set of suitable features and rank aggregation functions that yield the best result according to a given fitness function. Previous works [8, 16] have shown that combining rank aggregated lists and rank aggregation functions [15] yields very effective results.

2. PROPOSED APPROACH

Our approach estimates location based on rank aggregation of a multitude of ranked lists and their top-K density analysis [8]. We extracted a large set of features from the data, derived their ranked lists, and combined them using rank aggregation methods which in turn are selected and fused by the GP-based framework proposed in [15] (GP-Agg).

For evaluation purposes in the training phase (as in 2015 [6]) we split the whole training set into two parts: (i) a validation set; and (ii) a sub-training set. The validation set has 4,674 images and 903 videos, while the sub-training set has 12,935 videos and 4,188,484 images.

2.1 Features

Textual. The title, description, and tags of photos/videos were concatenated as a single field. The text was stemmed and stopwords were removed. We used BM25, TF-IDF (cosine), information-based similarity (IBSimilarity - IBS) and language modelling similarity (LMDirichletSimilarity - LMD), which are similarity measures implemented in the Lucene package [9].

Audio/Visual. For visual place recognition of images, we used the provided features: edgehistogram (EHD), scalable-color (SCD), GIST (static feature), cedd, col, jhist, and tamura. We also extracted BIC [12] and deep-learning based features (GoogleNet) [13]. For video data, due to time and infrastructure constraints for extracting features for new videos in test set, we were only able to use features of histograms of motion patterns (HMP) [1].

2.2 GP-based Rank aggregation & Geocoding

We used the full training set as geo-profiles and each test item was compared to the whole training set for each feature independently. For a given test item, a ranked list for each feature was generated. Then, these ranked lists were aggregated through the GP-Agg framework [15]. Given the improvements obtained in the last year by applying the ranked list density analysis (RLDA) over the final combined ranked list [6], we explored the idea of including this RLDA function into the GP-Agg framework: both in the fitness function evaluation and in the tree structure of GP's individuals (as an unary and binary operator). In this way, the GP-Agg framework was able to apply the RLDA density function in previous steps of the combination, which improved the results. Including the RLDA density function in the set of rank aggregation functions turns it in the unique function that uses geo-localization in the combination, whereas the other classic approaches only use similarity or rank position.

The GP-Agg method uses genetic programming to combine a set of methods for rank aggregation in an agglomerative way, in order to improve the results of the isolated

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

methods [15]. We used this method to combine the textual and visual ranked lists generated for various descriptors. This method was chosen because in [15] the authors showed that GP-Agg produced better or equal results than the best supervised technique in a wide range of rank aggregation techniques (supervised and unsupervised). Moreover, it required a reasonable time for training (a couple of hours), and it was relatively fast to apply the best individual (discovered function) on the test set.

The GP-Agg method was trained using 400 queries from the validation set (randomly chosen) and their ranked lists. We stopped the evolution process at the 20th generation. We used the fitness function, genetic operators, and rank aggregation techniques that yielded the best results in [15]. The GP-Agg parameters are shown in Table 1.

For the training phase of GP-Agg, an element of a ranked list was considered relevant if it is located no farther than 1 km from the ground truth location of the query element. The best individuals discovered in the training phase were applied to combine the ranked lists of test set. The predicted lat/long for an test-set element is obtained by picking the lat/long of the first element of its respective combined ranked list (which could be the single result of RLDA).

Table 1: GP-Agg parameters [15].

Parameter	Value
Number of generations	20
Genetics operators	Reproduction, Mutation, Crossover
Fitness functions	FFP1, WAS*, MAP, NDCG
Rank Agg. methods	CombMAX, CombMIN, CombSUM, CombMED, CombANZ, CombMNZ, RLSim, BordaCount, RRF, MRA, RLDA

* WAS (Weighted Average Score) as defined in [7].

Among the different fitness functions tested, the best results (more precise) were achieved with the WAS [7] and FFP1 [4].

3. OUR SUBMISSIONS & RESULTS

Based on parameters of our best results in the evaluation phase, our submissions were configured as shown in Table 2. For each Run, it shows the combination function applied on the test set, some of them discovered by the GP-Agg framework and others we choose based on experimental results, as it will be explained in next paragraphs. Runs 1 and 4 were based on textual-only descriptors, Run 2 was visual-only, and Run 3 was our multi-modal submission. For textual and multimodal runs, we set the K-top parameter of RLDA at 5, and for the visual ones at 100. No extra crawled material or gazetteers were used in our submissions.

In the case of photos, for Runs 1-3, we used the GP-Agg framework to discover a semi-optimal combination of rank aggregation functions and ranked lists. For the Run 4, we used the configuration with which we got the best results in the past year. Results in Table 3 show slight improvements at including RLDA in GP-Agg framework (Run 1 vs. Run 4).

As shown in Table 3, most of our best results were from Run 1, where GP-Agg applied rank aggregation for textual descriptors. For visual run (Run 2), combining rank aggregation functions and different visual features, including GoogleNet, improved our results over last year's.

The results for videos are presented in Table 4. As in the case of images, the best video results were obtained by applying GP-Agg over textual ranked lists. For Run 1 and Run 3, we combined the ranked lists using individual found by GP-Agg. We were unable to use the GP-Agg for Run 2 (visual) because we had only the HMP descriptor, thus we applied RLDA over it. In Run 4 we used only the best textual descriptor, since the best configuration of past year decreased the precision of video results. We can observe in Table 4 significant improvements in the combination of textual ranked lists through GP-Agg framework over the best textual descriptor (Run 1 vs. Run 4).

Table 2: Configurations of Runs

Run	Combination function
1	**Photo**: RLDA(RLDA(CombSUM(CombMED(BM25, IBS), RRF(IBS, BM25)), RLDA(CombSUM(BM25, BM25), RLDA(LMD))), CombANZ(IBS, BordaCount(CombMNZ(LMD, BM25), CombMNZ(LMD, BM25))))
	Video: CombSUM(BordaCount(RLDA(RLDA(BM25, TF-IDF)), BordaCount(CombMAX(BM25, BM25), CombSUM(BM25, BM25))), CombMNZ(CombMAX(RLDA(BM25, TF-IDF), CombMIN(LMD, IBS)), CombMED(CombMIN(BM25, TF-IDF), TF-IDF)))
2	**Photo**: RRF(RRF(CombMNZ(CombANZ(GoogleNet, jhist), RRF(tamura, cedd)), CombMAX(GoogleNet, tamura)), CombMED(MRA(col, GIST), RLDA(BIC, RLDA(EHD, GoogleNet))))
	Video: RLDA(HMP)
3	**Photo**: RLDA(CombMNZ(CombMED(BM25, CombMNZ(GoogleNet, IBS)), BM25), MRA(CombMNZ(LMD, LMD), CombMNZ(CombSUM(TF-IDF, BM25), TF-IDF)))
	Video: BordaCount(CombANZ(CombMAX(BM25, RLSim(TF-IDF, LMD)), BordaCount(CombANZ(HMP, TF-IDF), CombMED(LMD, BM25))), CombSUM(RLDA(BM25, RLDA(BM25, LMD)), CombMAX(RRF(IBS, TDIDF), RLSim(LMD, LMD))))
4	**Photo**: RLDA(IBS, BM25, LMD, TF-IDF)
	Video: LMD

Table 3: % of photos predicted correctly in test set.

Precision	Run1	Run2	Run3	Run4
10m	0.59	0.09	0.56	0.56
100m	6.07	0.87	5.97	5.94
1km	21.06	2.36	20.83	20.73
10km	38.00	4.47	37.72	37.47
100km	46.23	5.88	46.04	45.71
1000km	59.69	21.46	59.89	59.28
Avg (km)	2872.49	5664.14	2797.58	2919.3
Median (km)	254.67	5821.07	261.66	279.37

Table 4: % of videos predicted correctly in test set.

Precision	Run1	Run2	Run3	Run4
10m	0.45	0.00	0.51	0.37
100m	5.74	0.03	5.82	4.03
1km	18.69	0.15	18.46	13.51
10km	33.57	1.15	33.38	25.76
100km	41.56	2.46	41.20	33.02
1000km	54.51	13.54	54.77	47.67
Avg (km)	3204.8	6085.23	3123.38	3739.79
Median (km)	566.96	6085.63	571.24	1236.51

In both cases, for photos and videos, results obtained show no gain in the combination of textual and visual information (Run 3) through GP-Agg. It is explained due to the fact that the visual ranked list has significantly lower precision than textual ranked lists, and it is hard to find complementary between these types of lists by just applying classical rank aggregation methods.

4. FUTURE WORK

We plan to evaluate more textual and visual descriptors and give them as input to GP-Agg to select descriptors and rank aggregation methods. For example: (a) a textual descriptor that combines graph representation [10] with a framework for graph-to-vector synthesis [11]; (b) applying results from works that tackle the problem of visual place recognition [14] and of geolocation with Convolutional Neural Networks [2, 17]; (c) extracting visual features using GoogleNet and BIC for video frames.

Acknowledgments

We thank FAPESP, CNPq, CAPES, and Samsung.

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] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Computer Vision and Pattern Recognition (CVPR)*.

[3] J. Choi, C. Hauff, O. V. Laere, and B. Thomee. The Placing Task at MediaEval 2016. In *Working Notes Proc. MediaEval Workshop*, Hilversum, Netherlands, Oct. 2016.

[4] W. Fan, E. A. Fox, P. Pathak, and H. Wu. The effects of fitness functions on genetic programming-based ranking discovery for web search. *Journal of the American Society for Information Science and Technology*, 55(7):628–636, 2004.

[5] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[6] L. T. Li, J. A. Muñoz, J. Almeida, R. T. Calumby, O. A. Penatti, Í. C. Dourado, K. Nogueira, P. R. M. Júnior, L. A. Pereira, D. C. Pedronette, et al. Recod@ placing task of mediaeval 2015. In *Working Notes Proc. MediaEval Workshop*, volume 15, page 2.

[7] 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. http://dx.doi.org/10.1007/s11042-013-1588-4.

[8] L. T. Li, O. A. B. Penatti, J. Almeida, G. Chiachia, R. T. Calumby, P. R. M. Júnio, D. C. G. Pedronette, and R. da S. Torres. Multimedia geocoding: The RECOD 2014 approach. In *Working Notes Proc. MediaEval Workshop*, volume 1263, page 2, 2014.

[9] A. Lucene. Apache Lucene Core. Web Site. `http://lucene.apache.org/core/`. As of Sept. 2015.

[10] A. Schenker, H. Bunke, M. Last, and A. Kandel. *Graph-Theoretic Techniques for Web Content Mining*. World Scientific Publishing Co., Inc., NJ, USA, 2005.

[11] F. B. Silva, S. Tabbone, and R. d. S. Torres. BoG: A New Approach for Graph Matching. In *ICPR*, pages 82–87. IEEE, Aug. 2014.

[12] R. d. O. Stehling, M. Nascimento, and A. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, CIKM '02, pages 102–109, 2002.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '15, pages 1808–1817, 2015.

[15] J. A. Vargas Muñoz, R. da Silva Torres, and M. A. Gonçalves. A soft computing approach for learning to aggregate rankings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 83–92, New York, NY, USA, 2015. ACM.

[16] M. N. Volkovs and R. S. Zemel. CRF framework for supervised preference aggregation. In *Proceedings of the 22Nd ACM International Conference on Conference on Information; Knowledge Management*, CIKM '13, pages 89–98, New York, NY, USA, 2013.

[17] T. Weyand, I. Kostrikov, and J. Philbin. Planet - photo geolocation with convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016.