RECOD @ Placing Task of MediaEval 2015

Lin Tzy Li[1], Javier A. V. Muñoz[1], Jurandy Almeida[1,2], Rodrigo T. Calumby[1,3], Otávio A. B. Penatti[1,4], Ícaro C. Dourado[1], Keiller Nogueira[6], Pedro R. Mendes Júnior[1], Luís A. M. Pereira[1], Daniel C. G. Pedronette[1,5], Jefersson A. dos Santos[6], Marcos A. Gonçalves[6], 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]Universidade Estadual Paulista (UNESP),
[6]Department of Computer Science, Universidade Federal de Minas Gerais (UFMG) – Brazil
{lintzyli, pedro.mendes, luis.pereira, rtorres}@ic.unicamp.br, jurandy.almeida@unifesp.br, rtcalumby@ecomp.uefs.br, o.penatti@samsung.com, jalvarm.acm@gmail.com, icaro.dourado@students.ic.unicamp.br, daniel@rc.unesp.br, {keiller.nogueira, jefersson, mgoncalv}@dcc.ufmg.br

ABSTRACT

In this work, we describe the approach proposed by the RECOD team for the Placing Task, Locale-based sub-task, at MediaEval 2015. Our approach is based on the use of as much evidence as possible (textual, visual, and/or audio descriptors) to automatically assign geographical locations to images and videos.

1. INTRODUCTION

Geocoding multimedia has gained attention in the latest years given its importance for providing richer services for users such as placing information on maps and providing geographic searches. Since 2011, the Placing Task [2] at MediaEval has been challenging participants to assign the geographical locations to images and videos automatically.

Here we present our approach for the Locale-based subtask. 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. This year, we evaluated new features and a Genetic Programming (GP) [5] approach to multimodal geocoding. GP provides a good framework for modeling optimization problems even when the variables are functions.

Besides combining ranked lists, we also applied combinations of rank aggregation methods by using GP. 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 [7, 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 density analysis [7]. We extracted a large set of features from the data, derived ranked lists, and combined them using rank aggregation methods which in turn are selected and fused by a GP-based framework proposed in [15].

For evaluation purposes in the training phase (as in 2014 [7]) we split the whole training set into two parts: (i) a validation set; and (ii) a sub-training set. The validation set has 4,677 images and 905 videos, while the sub-training set has 12,944 videos and 4,226,559 images.

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

2.1 Features

Textual. The title, description, and tags of photos/videos were concatenated as a single field (here named as fusion). The versions that used only title, description, or tags were also used for the rank aggregation method. The text was stemmed and stopwords were removed. We used BM25, TF-IDF (cosine), information-based similarity (IBSimilarity) and language modelling similarity (LMDirichletSimilarity), which are similarity measures implemented in the Lucene package [8].

Audio/Visual. For visual place recognition of images, we used the provided features: edgehistogram (EHD), scalablecolor (SCD), and tamura. We also extracted BIC [13]. Due to time constraint and feature dimensionality, other planned visual features could not be applied this year. For videos, we used the provided features (all from LIRE [9] and MFCC20 [3]) besides extracting histograms of motion patterns (HMP) [1].

2.2 Rank aggregation, density analysis & 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. Given the ranked lists, we explored two distinct strategies: (i) a rank aggregation approach based on genetic programming (GP-Agg); and (ii) a ranked list density analysis. In addition, we also explored the combination of both strategies.

1. 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 methods [15]. We used this method to combine the textual and visual ranked lists generated for various descriptors.

This method was choosen 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 30th 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 the test set.

Table 1: GP-Agg parameters [15].

Parameter	Value
Number of generations	30
Genetics operators	Reproduction, Mutation, Crossover
Fitness functions	FFP1, WAS*, MAP, NDCG
Rank Agg. methods	CombMAX, CombMIN, CombSUM, CombMED, CombANZ, CombMNZ, RLSim, BordaCount, RRF, MRA

* WAS (Weighted Average Score) as defined in [6].

Among the different fitness functions tested, the best results (more precise) were achivied with the FFP1 as defined in [4]:

$$F_{FFP1} = \sum_{i=1}^{|N|} r(\hat{l}_i) \times k_1 \times \ln^{-1}(i + k_2) \quad (1)$$

where i is the element position after retrieval and $\hat{l}_i$ is the element at position i. $r(\hat{l}_i) \in \{0, 1\}$ is the relevance score assigned to an element, being 1 if the element is relevant and 0 otherwise. $|N|$ is the total number of retrieved elements. k_1, k_2 are scaling factors. Based on [15], we choose $k_1 = 6$ and $k_2 = 1.2$ in our experiments.

2. The ranked list density analysis (RLDA)[1] explores the idea of finding the maximum point in a probability density function (PDF). Firstly, we induce a k-nearest neighbor graph (with $k = 3$), where the graph nodes are defined as being the top-n items of the ranked lists. For each node, we estimate its probability density value by using a Parzen-Window gaussian kernel. This procedure is the same used to find root nodes (nodes with maximum density in a PDF) in the Optimum-path Forest (OPF) clustering algorithm [10]. Finally, to assign a lat/long to a test item, we just verify the lat/long of the graph node (a ranked list item) with the highest density value.

3. OUR SUBMISSIONS & RESULTS

None of our submissions used extra crawled material or gazetteers. Based on parameters of our best results on the evaluation phase, our submissions were configured as shown in Table 2. Runs 1 and 4 were solely based on textual descriptors, while Run 2 was only-visual and Run 3 was a multimodal submission.

From Table 3, our best submission was Run 4, in which we applied the RLDA over the top-5 items of each textual ranked list. We observed on the validation set that the RLDA of the top-n items from aggregated ranked list (visual and textual) seems to improve the results over just taking the first item from a multimodal aggregated ranked list. However, due to the delay in generating ranked lists of the visual features, we did not apply RLDA to the top-n items of the aggregated lists in Run 3 as we had planned initially. The second best submission was achieved by Run 1. Runs 2 and 3 showed that there is a room for improvements in our method based on GP.

[1]Last year [7], we called RLDA as OPF, however this year we renamed it, since we only used the OPF step that finds the most dense point.

Table 2: Runs configurations

	Images			Videos		
Run	**Textual**	**Visual**	**Geocoding**	**Textual**	**Visual/ Audio**	**Geocoding**
1	BM25 + TF-IDF + IBS + LMD	-	GP-Agg	BM25 + TF-IDF + IBS + LMD	-	GP-Agg
2	-	BIC	RLDA (top100)	-	HMP + all LIRE	GP-Agg + RLDA (top100)
3	BM25 + TF-IDF + IBS + LMD	BIC + EHD + SCD + tamura	GP-Agg	BM25 + TF-IDF + IBS + LMD	HMP + all LIRE + MFCC20	GP-Agg
4	TFIDF+ BM25+ IBS+ LMD	-	RLDA (top 5 from each list)	BM25	-	RLDA (top 5)

Table 3: Overall test result: % correct in precision levels.

Km	Run 1	Run 2	Run 3	Run 4
0.001	0.15	0	0.14	0.12
0.01	0.54	0.01	0.53	0.62
0.1	5.49	0.09	5.35	6.44
1	19.75	0.44	19.11	21.74
10	36.60	1.99	35.31	38.38
100	44.89	3.57	43.26	46.91
1000	58.97	20.38	57.67	63.22
10000	91.36	88.13	91.53	94.01
Distance in each quartile (km)				
Q1	1.8967	1239.6011	2.1244	1.4824
Q2	309.8649	5882.7206	394.889	196.0081
Q3	5573.9304	8636.9465	5766.8941	3798.6223

In most of the runs we have dealt with images and video through different settings. For instance, in Run 2 we applied RLDA top-100 of BIC ranked list for images, while for videos we combined other descriptors using GP-Agg followed by RLDA for the GP aggregated list. Thus, in Table 4, we only show the results regarding the videos in the test set. In the validation phase, the geocoding results for videos were relatively better than the ones for images, but it seems that in the test set this tendency was not preserved. For example in Run 4 (Table 4), the rate of correctly geocoded for videos in each precision levels is lower than the overall results for Run 4 (Table 3).

Table 4: Videos test results (% correct)

Km	0.001	0.01	0.1	1	10	100	1000	10000
Run 1	0.08	0.40	5.46	17.62	32.44	40.27	54.13	90.57
Run 2	0.00	0.00	0.01	0.02	0.11	3.80	20.39	91.97
Run 3	0.08	0.37	5.13	16.74	32.10	39.69	53.67	91.50
Run 4	0.06	0.41	5.79	17.89	32.24	39.92	55.68	93.16

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 [11] with a framework for graph-to-vector synthesis [12]; (b) applying results from works that tackle the problem of visual place recognition [14]. Additionally, we plan to devise a GP fitness function that takes advantage of RLDA to geocode, since most of the time RLDA improves geocoding results, besides exploring clustering analysis of the top-n items.

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] J. Choi, C. Hauff, O. V. Laere, and B. Thomee. The Placing Task at MediaEval 2015. In *Working Notes Proc. MediaEval Workshop*, Wurzen, Germany, Sept. 2015.

[3] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, Aug. 1980.

[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, 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.

[7] 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.

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

[9] M. Lux and S. A. Chatzichristofis. Lire: Lucene image retrieval: An extensible java CBIR library. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 1085–1088, New York, NY, USA, 2008.

[10] 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.

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

[12] 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.

[13] 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.

[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, R. d. S. Torres, and M. A. Gonçalves. A soft computing approach for learning to aggregate rankings. In *Proceedings of the 24Th ACM International Conference on Conference on Information and Knowledge Management*, CIKM '15, 2015. (in press).

[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.