UEC, Tokyo at MediaEval 2013 Retrieving Diverse Social Images Task

Keiji Yanai and Do Hang Nga
The University of Electro-Communications, Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, JAPAN
yanai@cs.uec.ac.jp, dohang@mm.cs.uec.ac.jp

ABSTRACT

In this paper, we describe our method and results for the MediaEval 2013 Retrieving Diverse Social Images Task. To accomplish the task objective, we adopt VisualRank [5] and Ranking with Sink Points [2], which are common methods to select representative and diverse photos. To obtain an affinity matrix for both ranking methods, we used only the officially-provided features including visual features and tag features. We submitted three required runs including only visual feature run, only textual feature run and textual-visual fused feature run.

1. INTRODUCTION

In this paper, we describe our method and results for the MediaEval 2013 Retrieving Diverse Social Images Task [4]. The objective of this task is to select relevant and diverse photos from the given photos regarding the specific locations. To do that, we adopt VisualRank [5] and Ranking with Sink Points [2]. The reason why we adopted these method is that we had used these methods for ranking geotagged photos [6]. First we calculate a similarity matrix using the given features, and we apply VisualRank to select the most representative photo. Then we re-rank the remaining photos by Ranking with Sink Points after removing the first-ranked photo. We repeat re-ranking by Ranking with Sink Points and removing the first-ranked photos until 50 photos are selected.

To obtain a similarity matrix for both ranking methods, we used only the officially-provided features including visual features and tag features. We submitted three required runs including only visual feature run, only textual feature run and textual-visual fused feature run, which are the minimum requirements to participate this task.

2. RANKING METHOD

To obtain representative and diverse photos in the upper rank, we adopt VisualRank [5] and Ranking with Sink Points [2]. In this section, we explain both methods and features briefly.

2.1 VisualRank

VisualRank is an image ranking method based on PageRank [1]. PageRank calculates ranking of Web pages using hyper-link structure of the Web. The rank values are estimated as the steady state distribution of the random-walk Markov-chain probabilistic model.

VisualRank uses a similarity matrix of images instead of hyper-link structure. Eq.(1) represents an equation to compute VisualRank.

$$\boldsymbol{r_{i+1}} = \alpha S \boldsymbol{r_i} + (1-\alpha)\boldsymbol{p}, \quad (0 \leq \alpha \leq 1) \tag{1}$$

S is the column-normalized similarity matrix of images, $\boldsymbol{p}$ is a damping vector, $\boldsymbol{r}$ is the ranking vector each element of which represents a ranking score of each image, and α plays a role to control the extent of effect of $\boldsymbol{p}$. The final value of $\boldsymbol{r}$ is estimated by updating $\boldsymbol{r}$ iteratively with Eq.(1). Because S is column-normalized and the sum of elements of $\boldsymbol{p}$ is 1, the sum of ranking vector $\boldsymbol{r}$ does not change. Although $\boldsymbol{p}$ is set as a uniform vector in VisualRank as well as normal PageRank, it is known that $\boldsymbol{p}$ can plays a bias vector which affects the final value of $\boldsymbol{r}$ [3].

2.2 Ranking with Sink Points

Because VisualRank is a ranking method considering only representativeness of items, higher ranks are sometimes occupied with items which are similar to each other. This is, VisualRank cannot accomplish ranking considering diversity of items. Therefore, we adopt Ranking with Sink Points [2], which can be regarded as an extension of PageRank [1] to make obtained ranking relevant and diverse.

To address the diversity in ranking, the concept of sink points is useful. The sink points are data objects whose ranking scores are fixed at zero during the ranking process. Hence, the sink points will never spread any ranking score to their neighbors. Intuitively, we can imagine the sink points as the "black holes" on the ranking manifold, where ranking scores spreading to them will be absorbed and no ranking scores would escape from them.

First we apply VisualRank to select the most representative photo with the obtained affinity matrix. Then we re-rank the remaining photos by Ranking with Sink Points as shown in Eq.(2), after removing the first-ranked photo as "a sink point". We repeat re-ranking by Ranking with Sink Points and removing the first-ranked photos until 50 photos are selected as following:

$$\boldsymbol{r_{i+1}} = \alpha S I_i \boldsymbol{r_i} + (1-\alpha)\boldsymbol{p} \tag{2}$$

I_i is an indicator matrix which is a diagonal matrix with its $(i,i)-$ element equal to 0 if $x_i \in X_s$ and 1 otherwise. X_s is a set of "sink points".

Note that α is set as 0.85 in the experiments.

2.3 Visual Features

We used the ten kinds of visual features officially provided by the task organizers such as Global Histogram of Oriented

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

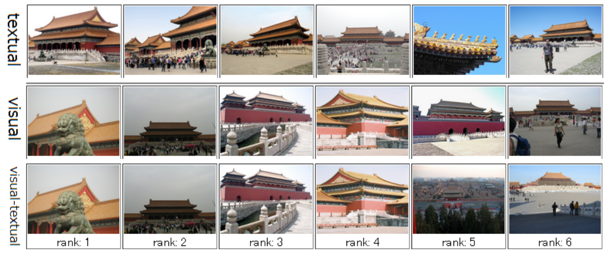

Figure 1: An example of the ranking by the three kinds of features: "The Gate of Forbidden City in China".

Gradient and Color Moments on HSV Color Space. The detail on official visual features is explained in [4].

With histogram intersection, we calculate similarities for each of visual features. Finally we construct an affinity matrix by averaging similarity on ten kinds of visual features.

2.4 Textual Features

We use social TF-IDF weights provided by the task organizers. We extract bag-of-words vectors from Flickr metadata with social TF-IDF weights for all the given images. We calculate an affinity matrix with cosine similarity between bag-of-words vectors within each place.

To obtain an affinity matrix for the visual-textural-fused runs, we simply averaged both visual-feature-based affinity matrix and textual-feature-based affinity matrix.

3. EXPERIMENTAL RESULTS

Tables 1 and 2 show the evaluated results of our three submission runs by experts and crowds, respectively. Note that the results by experts is based on evaluation for the entire dataset of 346 locations, while the results by the crowds is based on evaluation for only 50 locations in the dataset and are obtained by averaging evaluations by three crowd persons.

Basically, the results by only visual were better than the results by only textual and the results by visual-textual, although the difference were not so large.

We show the top six photos of an successful example by the proposed method with three kinds of features: textual, visual and visual-textual features in Figure 1. These photos represents "The Gate of Forbidden City in Beijing, China." In this example, the photos selected by the visual-textual feature is more representative and diverse than the photos selected by the only textual or only visual features. This indicates that our proposed methods works successfully.

In the case of the above example, most of the photos included in the given photo set are relevant and only a few noise photos are included. However, given photo sets of some landmark include many noise photos. In such case, the proposed methods sometimes failed to select relevant photos and selected noise photos in the upper ranking. Therefore, removal of noise photos is one of our important future works.

4. CONCLUSIONS

We tackled MediaEval 2013 Retrieving Diverse Social Images Task with VisualRank [5] and Ranking with Sink Points [2].

Table 1: Results evaluated by experts for the entire test set of 346 locations.

Runs	P@5	**P@10**	P@20	P@30	P@40	P@50
Only visual	0.7164	**0.7056**	0.7092	0.7076	0.6948	0.6752
Only textual	0.7082	**0.6863**	0.6845	0.6904	0.6841	0.6667
Visual-textual	0.7135	**0.7155**	0.7063	0.7026	0.6934	0.6723

Runs	CR@5	**CR@10**	CR@20	CR@30	CR@40	CR@50
Only visual	0.2233	**0.3633**	0.5448	0.6743	0.7572	0.8154
Only textual	0.213	**0.3579**	0.5515	0.6706	0.7549	0.8094
Visual-textual	0.2258	**0.3621**	0.5414	0.6642	0.7427	0.8015

Runs	F1@5	**F1@10**	F1@20	F1@30	F1@40	F1@50
Only visual	0.3288	**0.4617**	0.5926	0.6618	0.6936	0.7068
Only textual	0.318	**0.4531**	0.5869	0.6544	0.689	0.7001
Visual-textual	0.3303	**0.4614**	0.5879	0.6545	0.6869	0.6995

Table 2: Results evaluated by crowd (the average of GT1, GT2 and GT3) for a subset of 50 locations from the test set.

Runs	P@5	**P@10**	P@20	P@30	P@40	P@50
Only visual	0.7061	**0.6959**	0.6857	0.6878	0.6847	0.6845
Only textual	0.6857	**0.6673**	0.6847	0.6966	0.6964	0.6865
Visual-textual	0.6367	**0.6531**	0.6653	0.6823	0.6765	0.6747

Runs	CR@5	**CR@10**	CR@20	CR@30	CR@40	CR@50
Only visual	0.5947	**0.7198**	0.8070	0.8803	0.9153	0.9394
Only textual	0.5875	**0.7331**	0.8429	0.9118	0.9355	0.9449
Visual-textual	0.5573	**0.6824**	0.8050	0.8829	0.9223	0.9447

Runs	F1@5	**F1@10**	F1@20	F1@30	F1@40	F1@50
Only visual	0.5915	**0.6657**	0.7052	0.7435	0.7586	0.7675
Only textual	0.5818	**0.6659**	0.7366	0.7669	0.7770	0.7735
Visual-textual	0.5441	**0.6261**	0.6971	0.7446	0.7578	0.7638

Unfortunately, due to time limitation, we had to give up using some useful additional data including the training dataset, GPS data coordinates and Wikipedia photos. In fact, if you had enough time, we should have used the training data for estimating optimal parameters such as α in the VisualRank formulation and a mixing weight of visual similarity and textual similarity.

5. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the Seventh International World Wide Web Conference*, 1998.

[2] X.-Q. Cheng, P. Du, J. Guo, X. Zhu, and Y. Chen. Ranking on data manifold with sink points. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):177–191, 2013.

[3] T. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE trans. on Knowledge and Date Engneering*, 15(4):784–796, 2003.

[4] B. Ionescu, M. Menendez, H. Muller, and A. Popescu. Retrieving diverse social images at mediaeval 2013: Objectives, dataset and evaluation. In *MediaEval 2013 Workshop, CEUR-WS.org, ISSN: 1613-0073*, Barcelona, Spain, October 18-19 2013.

[5] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1870–1890, 2008.

[6] H. Kawakubo and K. Yanai. Geovisualrank: A ranking method of geotagged images considering visual similarity and geo-location proximity. In *Proc. of the ACM International World Wide Web Conference*, 2011.