BMEMTM at MediaEval 2013 Retrieving Diverse Social Images Task: Analysis of Text and Visual Information

Gábor Szűcs
Inter-University Centre for
Telecommunications and Informatics,
H-4028 Kassai út 26., Debrecen,
Hungary
szucs@tmit.bme.hu

Zsombor Paróczi
Dept. of Telecommunications and
Media Informatics, BME, Budapest,
Hungary
paroczi@tmit.bme.hu

Dániel Máté Vincz
Dept. of Telecommunications and
Media Informatics, BME, Budapest,
Hungary
dani.vincz@gmail.com

ABSTRACT

In this paper, the possibilities of using visual and textual information are investigated to improve the ranking of photos from Flickr about famous places. We have elaborated improved textual features based on standard ones and visual features e.g. face feature for measure the relative face area on the images. These heuristic features have been used for the solution in the MediaEval 2013 Retrieving Diverse Social Images Task to rerank social photos based on two evaluation metrics, the precision and the diversity.

1. INTRODUCTION

In the MediaEval 2013 Retrieving Diverse Social Images Task [1] retrieved social photos should be reranked. In the use case of the task a potential tourist tries to find more information about famous place, because he/she has only a vague idea about the location, knowing the name of the place. In this task many photos retrieved from Flickr with rank information have been available. These results have been noisy and redundant, so the aim was to refine these results by providing a ranked list of up to 50 photos that are considered to be both relevant and diverse representations of the query.

2. OUR CONTRIBUTION

Firstly we have thought that the supervised learning of machine learning would have been useful, but investigating the locations it can be stated that albums are very different. Because of large difference (statues, buildings, squares, famous ship, etc.) there was a little chance to learn. Thus instead of machine learning we have used statistics, heuristic for solving the task.

2.1 Visual models

The contest organizer has made many visual descriptors available, but we have introduced an additional descriptor, the FACE feature based on the OpenCV's implementation of Haar Like Feature detection [2]. The FACE feature is the ratio of the calculated area occupied by the possible face regions on an image and whole image area. Based on the descriptors we have developed several algorithms for the task. Each algorithm uses the original order of the images (Flickr order), and reorder them.

noface: Using FACE descriptor the algorithm filters out images containing faces, and it takes back them into the end of the queue.

hiercn<N>: Using the CN (Global Color Naming Histogram) descriptor this algorithm creates N clusters by a simple hierarchic clustering using Euclidean distance function in 11 dimension of the descriptor in order to get better diversity. The algorithm takes back every image until the first N images are in different clusters. So if the original order was A1, A2, B1, C4, C5, D2 (where letters represent clusters, numbers represent index in the cluster), than the reordered list will be A1, B1, C4, D5, A2, C5.

facehiercn<N>: This algorithm takes advantage of both the noface and the hiercn<N> by executing them after each other (in the order of 1. face, 2. hiercn<N>)

clustercm<N>: Using the CM (Global Color Moments on HSV Color Space) descriptors this algorithm creates N clusters by hierarchic clustering based on a special distance function, and the ordering is the same as in the hiercn<N> algorithm. The distance function between image *i* and *j* is:

$$d(i,j)=\sum_{k=1}^{3}CM_k(i,j)+\frac{\sum_{k=4}^{6}CM_k(i,j)}{\alpha}+\frac{\sum_{k=7}^{9}CM_k(i,j)}{\alpha^2} \quad (1)$$

where the first 3 CM values are the means, then standard deviations, finally the last 3 CM values are the second momentums, α is tuning parameter (in our experiences 30 was the best value based on the development set), furthermore

$$CM_k(i,j)=\left|CM_k(i)-CM_k(j)\right| \quad (2)$$

clustermodcm<N>: This is a modified version of the clustercm<N> algorithm, which takes back certain images (as punishment) by only 3 places in the queue, therefore the similar images can be too close.

We have tested these algorithms on the development set, and the results can be seen on Figure 1. The baseline is the original Flickr result, and the facehiercn20 algorithm was the best at F1@10 metric; that is why we have chosen this for visual-only run (run1).

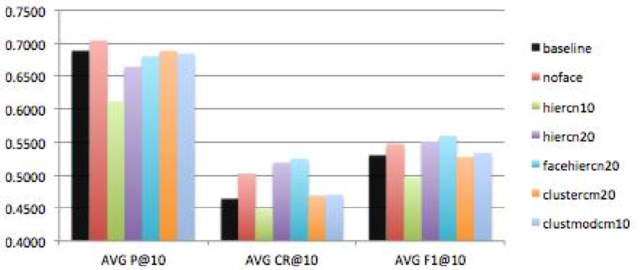

Figure 1. Comparing visual algorithms on the development set

2.2 Textual models

Firstly we have separated the textual task (for run2) into four subtasks: 1: Improving the provided textual models (probabilistic, TF-IDF, Social TF-IDF). 2: Assigning score values to each image for each provided textual model and for each improved model (so an image will possess 6 score values). 3: Calculating the rank

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

score of each image based on the weights of the textual models. 4: Calculating the new order of images for each location. More detailed explanation of subtasks is described below.

1. Rewarding the keywords which appear more often related to one location may lead to a better result, but a keyword is sometimes nested in the tag of the image, e.g. basilica can be found in 'basilicadisantamariadellasalute', 'thebasilicaofstmaryof-health'. In order to handle this problem a developed algorithm has split the tags without spaces into list of keywords (using the estimated position of the spaces, as results of an inference algorithm), and then it has assigned new values to the keywords according to the number of their appearance.

2. Our method calculates an average value for every image based on the number of keywords belonging to the image and the values assigned to those keywords according to all six different textual descriptors – i.e. probabilistic, TF-IDF, social TF-IDF models and the improved versions of these. Then the method calculates a score value for every image (according to each textual model), which is going to be the sum of the maximum value from all the keywords related to the image and the logarithm of the previously calculated average value.

3. We assign weights to the 6 textual models, and our method calculates the weighted average score (final score) for each image.

4. A higher final score means a better final rank position, thus the new ranks (improved order) can be produced for the images.

We executed many test cases with various weights assigned to both the original and the improved textual models and we found, that the best result is in P@10 the improved TF-IDF weighting model, however in case of CR@10 and F1@10 using only the improved probabilistic model led to the best results.

2.3 Combination of visual and textual models

Our text based approach ignores the original ordering of the images and our visual based solution only modifies a predefined order, so it seemed natural to combine them. At the combination the text algorithm was the first phase, then using the ordered result the visual algorithm was the second phase. Our results on the development set have indicated, that this combination is better (at least in the CR@10 metric) than the original two solutions.

2.4 Human-based approach

We have implemented a helping tool for humans, by which the user is able to sort the images into clusters and to store the binary decision about the relevance of each image. After the human's work a developed algorithm has determined the order of the images by the following way: in a cycle the most relevant image in each non-empty cluster is selected (and removed from the cluster) and ordered based on Flickr rank. This cycle is repeated, and the process terminates after the last image.

We have not enough time to survey the Internet, thus the human-based run (run 4) and the general run (run 5, where everything allowed including using data from external sources) were the same in our contribution, so the results of them were the same.

3. RESULTS

Evaluation metrics include precision at top 10 results (P@10), cluster recall (CR@10) (measure of how many of the existing clusters are represented in the final refinement, so this is the diversity) and harmonic mean of them, the F1-measure (F1@10).

Table 1. and table 2. present the results achieved using the expert ground truth and using the crowd-sourcing ground truth, respectively. The expert evaluation is conducted on the entire test set of 396 locations, and the crowd-sourcing evaluation is based on a selection of 50 locations. The crowd-sourcing relevance ground truth was determined after a majority voting scheme, and the crowd-sourcing diversity ground truth was provided with 3 different annotations (*a* column in the Table 2.).

Table 1. Results achieved using the expert ground truth.

	run1	run2	run3	run4, 5
P@10	0.7389	0.8056	0.6754	0.8936
CR@10	0.4076	0.3859	0.3709	0.2963
F1@10	0.5066	0.4979	0.461	0.4115

Table 2. Results at the crowd-sourcing ground truth

	a	run1	run2	run3	run4, 5
P@10		0.6857	0.7653	0.6469	0.8163
CR@10	1	0.8067	0.7731	0.8217	0.6519
	2	0.7371	0.713	0.7814	0.5753
	3	0.6469	0.6219	0.6399	0.4922
F1@10	1	0.7075	0.7429	0.6981	0.6798
	2	0.6825	0.707	0.6711	0.6278
	3	0.6266	0.6472	0.6098	0.5734

4. CONCLUSIONS

The *only visual* (run1) results correspond with our expectations, because the value of F1@10 was 0.559 for the development set. The *text + visual* (run3) result is a big surprise for the team (BMEMTM), because our development set results indicated, that the combined algorithm will be better at least at the CR@10 metric, but both the *only visual* and the *only text* methods reached better results.

In future works we will try to use the remaining visual descriptors as teaching dataset for an SVM classifier, a learning based algorithm (maybe combined with the facehiercn<N> algorithm) may achieve better results in the CR metric. The *text + visual* algorithm should be reevaluated to find out the mistake.

5. ACKNOWLEDGMENTS

The publication was supported by the TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project. Project has been supported by the European Union, co-financed by the European Social Fund.

6. REFERENCES

[1] Ionescu, B., Menéndez, M., Müller, H. and Popescu, A. 2013. Retrieving Diverse Social Images at MediaEval Objectives, Dataset and Evaluation, *MediaEval 2013 Workshop*, ISSN: 1613-0073, 18-19 October 2013, Barcelona, Spain.

[2] Viola, P., and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition. CVPR 2001. Proceedings of the IEEE Computer Society Conference on*. Vol. 1, pp. I-511-I-518.