Imcube @ MediaEval 2015 Retrieving Diverse Social Images Task: Multimodal Filtering and Re-ranking

Sebastian Schmiedeke, Pascal Kelm, and Lutz Goldmann
imcube labs GmbH
Berlin, Germany
{schmiedeke, kelm, goldmann}@imcube.de

ABSTRACT

This paper summarizes the participation of Imcube at the Retrieving Diverse Social Images Task of MediaEval 2015. This task addresses the problem of result diversification in the context of social photo retrieval where the results of a query should contain relevant but diverse items. Therefore, we propose a multi-modal approach for filtering and re-ranking in order to improve the relevancy and diversity of the returned list of ranked images.

1. INTRODUCTION

The Retrieving Diverse Social Images Task of MediaEval 2015 [5] requires participants to develop a system that automatically refines a list of images returned by a Flickr query in such a way that the most relevant and diverse images are returned in a ranked list of up to 50 images.

A photo is considered relevant if it is a common representation of the overall query concept in good visual quality (sharpness, contrast, colours) and without people as main subjects except for queries dealing with people as part of the topic. The results are considered diverse if they depict different visual aspects (time, location, view, style, etc) of the target concept with a certain degree of complementarity.

The refinement and diversification process can be based on the social metadata associated with the collected photos in the data set and/ or on the visual characteristics of the images. Furthermore, the task provides information about user annotation credibility as an automation estimation of the quality of a particular user's tag.

2. SYSTEM DESCRIPTION

In this section, we present our approach that combines textual, visual and credibility information to filter and re-rank the initial results. Our approach consists of two steps – relevancy improvement and diversification – as depicted in Figure 1.

The goal of the first step is to improve the relevancy of the ranked image list by re-ranking the images based on more reliable textual and visual criteria and filtering them in order to remove images which are irrelevant for the given application scenario. The goal of the second step is to improve the diversity of the ranked image list through textual filtering and visual clustering and re-ranking. The individual modules will be described in the following sections.

2.1 Textual relevancy improvement

This step exploits additional information extracted from the corresponding Wikipedia article that is provided together with the query.

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

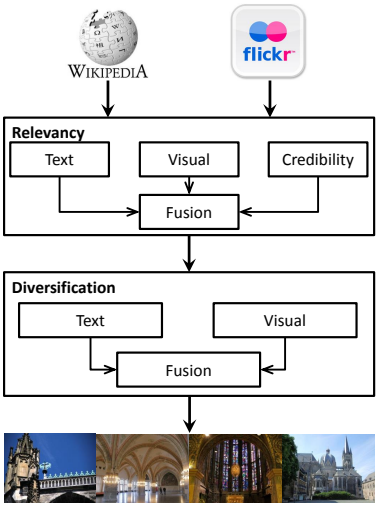

Figure 1: Proposed approach.

To improve the ranking of the images the query is expanded with the most frequent words from the Wikipedia article and the images are re-ranked using a bag-of-words representation. The relevancy is further improved by removing images that do not match the original query and the location information from the Wikipedia article. The location information is extracted by analysing the original query or Wikipedia title (e.g. , "Great Sphinx of Giza") taking into account typical prepositions for locations (e.g. "in", "at", "on", "of", "de"). In the case that no location information can be extracted (e.g. "Niagara Falls") toponyms are not considered for relevancy filtering.

2.2 Visual relevancy improvement

Visual information is also used to improve the relevancy by re-ranking them according to different criteria. For each visual feature a ranked image list is derived based on the computed relevancy scores.

Since images with persons as main subjects are considered irrelevant, we employ a face detector [7] trained for frontal and profile faces to determine the *size of facial regions*. The inverse relative size of the detected faces determines the relevancy. Hence, the smaller the area covered by faces the more relevant is the image.

Additionally, photos taken from the target location but not displaying it are considered irrelevant. We model that relevancy by computing the visual similarity between the retrieved images and the images which are available from the associated Wikipedia article. We use histogram of oriented gradients (*HOG*) features and a *clusterless BoW approach* [6] based on speeded up robust features

Table 1: Evaluation of different runs.

	Average P@20			Average CR@20			Average F1@20		
	one-concept	multi-concept	overall	one-concept	multi-concept	all	one-concept	multi-concept	overall
run1	0.7014	0.6743	0.6878	0.3963	**0.4209**	0.4087	0.4885	**0.5027**	0.4957
run2	**0.7819**	**0.7143**	**0.7478**	**0.4380**	0.3986	**0.4182**	**0.5478**	0.4917	**0.5195**
run3	0.7674	0.6993	0.7331	0.4285	0.4064	0.4174	0.5340	0.4926	0.5132
run5	0.5928	0.6671	0.6302	0.3410	0.3508	0.3460	0.4251	0.4448	0.4350

(SURF) features to generate histograms for each of the images. The similarity between the retrieved images and the wikipedia images is computed through histogram intersection. The retrieved images are re-ranked by considering the maximum score across the set of wikipedia images.

We further incorporate aesthetic aspects to emphasize more visually appealing images, since less blurry and salient images are usually considered more relevant. The *sharpness* is calculated as the ratio of magnitude of image gradients between different blurred versions of the original image. The larger that ratio, the more relevant the image. *Saliency* is measured using a spectral residual approach [4]. Considering the different criteria described above we obtain 5 ranked image lists (Face, HOG, BoW, Sharpness, Saliency) which are fused using weighted rank fusion.

2.3 Credibility based relevance improvement

This step is intended to be the baseline approach for improving the relevance. It re-ranks the image list according to the credibility of the owner of an image. The re-ranking is based on 3 scores which describe the user credibility [5]: the use of correct tags (*visualScore*), specific tags (*tagSpecificity*) and their preference for photographing faces (*faceProportion*). Following the application scenario the combined credibility score for an image is high if the user has a high *visualScore*, a high *tagSpecificity* and a low *faceProbability*.

2.4 Textual diversification

The final image list should not only contain relevant images but also diverse ones, i.e. , depicting different aspects of the topic. With the assumption that images which have an identical textual description often depict very similar content, the images are clustered based on their textual similarity. The ranked image list is then obtained by ranking the clusters in descending order according to their relevancy and iteratively selecting the most relevant image from each cluster.

2.5 Visual diversification

The visual diversification considers multiple visual characteristics including colour (*ColorMoment*), structure (*HOG*, *clusterless BoW approach* [6]) and texture (local binary patterns (*LBP*)). For each feature the normalized distances between the retrieved images are combined using weighted summation and then projected in a lower dimensional space by applying the FastMap [2] algorithm. On the resulting 5-dimensional feature space, kMeans++ clustering [1] is applied.

The number of clusters is estimated by Hartigan's Leader clustering algorithm [3], but the number is restricted to be between 5 and 21. Clusters with a low mean relevancy or clusters containing only a few images are discarded, Since these small clusters are very likely to contain outliers. The remaining clusters are ordered in descending order according to their maximum relevancy and ranked image list is obtained by iteratively selecting the best image from them.

3. EXPERIMENTS & RESULTS

The following experiments were performed based on the system and the individual modules described above following the guidelines of the task.

Run1 is an approach using visual information only (as described in Sec. 2.2 and 2.5). *Run2* is an approach based on purely textual information (as described in Sec. 2.1 and 2.4). *Run3* is an approach based on textual and visual information (as described in Sec. 2.1, 2.2 and Sec. 2.5). *Run5* is an approach using credibility based relevancy and visual diversity (as described in Sec. 2.3 and 2.5).

These experiments are performed on the test set provided which contains 69 one-concept location queries and 70 multi-concept queries related to events.

Table 1 shows the results on the test set for all the runs defined above. Since we want to evaluate our filters for different conditions, scores for the one-concept and multi-concept queries are also provided.

In general, the textual run (*run2*) achieves the best results. It achieves the higher precision ($P@20 = 0.748$) and also a slightly better recall ($CR@20 = 0.418$) compared to the visual run (*run1*). The advantage is more significant for one-concept queries than for multi-concept queries. The textual run fails for queries which main topic is not correlated to a location (e.g. "chinese new year in Beijing" (its main topic is firework), "paragliding in the mountains" or "tropical rain"). For these cases, the visual run reaches considerably higher F1 scores. Generally, the purely visual run achieves a better recall ($CR@20 = 0.4209$) and thus a slightly better F1 metric ($F1@20 = 0.5027$) for multi-concept queries.

Since, the combination of visual and textual features (*run3*) constantly achieves lower scores than the individual modalities, we analyse the cases where improvements were made. For example the previously mentioned query ("chinese new year in Beijing") benefits from visual information with a considerable increase of the F1 measure ($\Delta F1@20 = 0.18$). In comparison to *run2*, *run3* achieves a lower precision ($\Delta P@20 = -0.015$) and a similar recall ($\Delta CR@20 = -0.001$) leading to slightly lower F1 score ($\Delta F1@20 = -0.006$). However, it is interesting to note that the results differ for one-concept and multi-concept queries. The recall of *run3* is actually higher than that of *run2* for multi-concept queries ($\Delta CR@20 = 0.008$) while it is lower for one-concept queries ($\Delta CR@20 = 0.009$).

4. CONCLUSION

The results of the different runs show that overall the best results can be achieved with textual information only and that the fusion of visual and textual information leads to slightly worse results. Analysing the results in more detail shows that visual information provides better results for multi-concept queries and queries where the main topic is not correlated to a location while textual information achieves better performance for one-concept queries. This shows that a more advanced fusion approach for combining textual and visual information may improve the results further.

5. REFERENCES

[1] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[2] C. Faloutsos and K. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 163–174. ACM New York, NY, USA, 1995.

[3] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.

[4] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[5] B. Ionescu, A. L. Gînscă, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving Diverse Social Images at MediaEval 2015: Challenge, Dataset and Evaluation. *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[6] S. Schmiedeke, P. Kelm, and T. Sikora. DCT-based features for categorisation of social media in compressed domain. In *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, pages 295–300, 2013.

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