# Diversification for Content-Based Visual Information Retrieval System for Video

Zsombor Paróczi, Bálint Fodor, Gábor Szűcs

The number of user-generated content has grown exponentially due to the media recording and sharing methods becoming easier and cheaper for everyone. In recent years computer vision and data mining communities devoted significant attention to user submitted media analysis and retrieval, this field is usually referred to as content-based visual information retrieval (CBVIR). Due to the huge number of media content, without an efficient search system we could not find the relevant information. Existing retrieval technologies focus on the precision of the results that often provides the user with redundant answer list with near duplicated items. The users would like to retrieve not only relevant items, but also diverse results.

Our work focuses on a potential tourist, who tries to find more information about a famous place, only knowing the name of the place. The literature calls this type of attitude the surfer, who has a moderate goal, has a starting point but may want slightly different outcome in the same context. The goal besides finding the relevant items is to filter the duplicated or similar content. In this special case a diverse result for a location name based search should include shots from various weather conditions and seasons, day/night shots from the same place. We only focus on search results of YouTube, but the same method should work on any video based service.

CBVIR systems usually have three distinct steps: (i) content retrieval, (ii) data organization and indexing and (iii) data-driven applications [2]. In the first, the data is extracted from the source, in our case we used a self written video extractor which helped us to overcome the YouTube location based suggestion system. In this preprocessing phase every video was cut into multiple shots along the time domain. Every shot is intended to contain a coherent scene of the video, this gave us ability for fine-grained analysis of the video content. The dataset we worked on was cut into shots manually to ensure a quality input for the following phases. The raw data is obviously too big for direct analysis, so in the data organization and indexing step the extracted data is transformed to a smaller, but still distinct and representative dataset - in our case we created feature vectors for each shot, using a wide variety of video and image processing algorithms (including edge detection, color histograms, histograms of gradients). In the data-driven application step our goal was to reorder the existing results, we used a combination of machine learning and data mining algorithms.

The relevance of a shot was determined by a pre-trained neural network classifier. The training data was the 20% of the whole dataset. For improving the early diversity of the resulting shot order, first we clustered the initial ordering. The new result list was formed by picking relevant shots from each cluster. While the relevant and diverse shots are preferred to take the first places of the new ordering the algorithm tries to preserve the initial ordering as much as possible.

Our evaluation considers both the relevancy and the diversity factors, this is why the CBVIR algorithms are either evaluated by these two factors separately or using some kind of jointly optimization metric [1]. In our evaluation we used average precision at N seconds (P@N), cluster recall at N seconds (CR@N) (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 at N seconds (F1@N).

We used the same dataset, including the shot annotations in our previous work [3]. In that we used a smaller feature vector, and the reordering was done using K-means algorithm and cluster diameters for the reordering. In that work the evaluation was done in a shot based scale, we managed to improve the original reordering by 9-11% (measured at the 10th shot, 20th shot

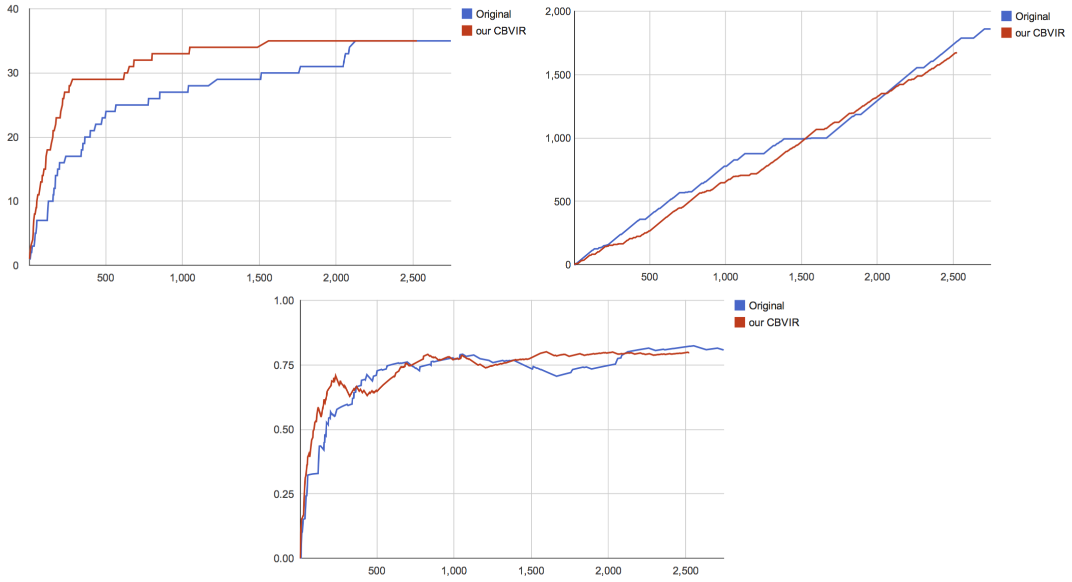

Figure 7: Results from left to right: CR@N, P@N, F1@N

and 50th shot). In this paper we extended the feature vector set with more metrics, and used neural network with support vector clustering. Due to the shot length variation the evaluation was done using the time dimension in seconds, which gives us a much more precise result.

Figure 7 shows the diversity, the relevance and the F1-measure for the "Acropolis of Athens" search result as a function of the time in seconds. The YouTube ordering (Original) is compared to the re-ranked ordering (Our CBVIR). The evaluation shows that our approach improves the original ordering by 14-18% at the first result shots.

### Acknowledgments

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

### References

[1] Thomas Deselaers, Tobias Gass, Philippe Dreuw, and Hermann Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 39:1–39:8, New York, NY, USA, 2009. ACM.

[2] Shi-Min Hu, Tao Chen, Kun Xu, Ming-Ming Cheng, and RalphR. Martin. Internet visual media processing: a survey with graphics and vision applications. *The Visual Computer*, 29(5):393–405, 2013.

[3] Zsombor Paróczi and Bálint Fodor. Video content retrieval diversification system. In *Tavaszi Szél Konferencia*, march 2014.