A Multimedia Interactive Search Engine based on Graph-based and Non-linear Multimodal Fusion

Anastasia Moumtzidou, Ilias Gialampoukidis, Theodoros Mironidis, Dimitris Liparas,
Stefanos Vrochidis and Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research and Technology Hellas
Emails: {moumtzid, heliasgj, mironidis, dliparas, stefanos, ikom}@iti.gr

***Abstract*—This paper presents an interactive multimedia search engine, which is capable of searching into multimedia collections by fusing textual and visual information. Apart from multimedia search, the engine is able to perform text search and image retrieval independently using both high-level and low-level information. The images of the multimedia collection are organized by color, offering fast browsing in the image collection.**

***Index Terms*—Multimedia retrieval, Non-linear fusion, Unsupervised multimodal fusion, Search engine, Multimedia, Colormap, GUI**

I. Introduction

This paper describes the VERGE interactive multimedia search engine, which is capable of retrieving and browsing multimedia collections. Contrary to the previous versions of VERGE [1], which perform video retrieval using video shots, the present system fuses textual metadata and visual information. The engine integrates a multimodal fusion technique that considers high-level textual and both high- and low-level visual information. Multimedia retrieval engines have been supported by tree-based structures [2], or multilayer exploration structures [3] for scalability purposes. However, our multimedia retrieval module filters out non-relevant objects, with respect to the textual modality and performs multimodal fusion of all modalities (textual and visual) on the filtered documents, in order to retrieve the most relevant objects, in response to a multimodal query. The images of the multimedia collection are organized by color, so as to ensure efficient and fast access to the image collection, associated to the multimedia collection. All images are indexed using visual concepts and query-based search is supported. Metadata of each image are indexed and text search is possible.

II. Multimedia Retrieval System

The VERGE interactive retrieval system[1] combines advanced retrieval functionalities (Fig. 1) with a user-friendly interface (Fig. 2 and Fig. 3), and supports the submission of queries and the accumulation of relevant retrieval results. The system integrates a fusion module that is able to retrieve multimodal documents and supports (independently) text-based search and image retrieval, using high-level concepts. The aforementioned capabilities allow the user to search through a collection of multimodal documents. Moreover, the system incorporates an Image ColorMap Module that clusters the images according to their color and allows for presenting all images of the collection in a single page. The VERGE architecture is shown in Fig. 1, with multimodal fusion of low- and high-level visual and textual information, color-based clustering, served by the VERGE Graphical User Interface (GUI). The overall system is novel, since it integrates the fusion of multiple modalities [4], in a hybrid graph-based and non-linear way [5], with several functionalities (eg. multimedia retrieval, image retrieval, search by visual or textual concept, etc.) already presented in [1], but in a unified user interface.

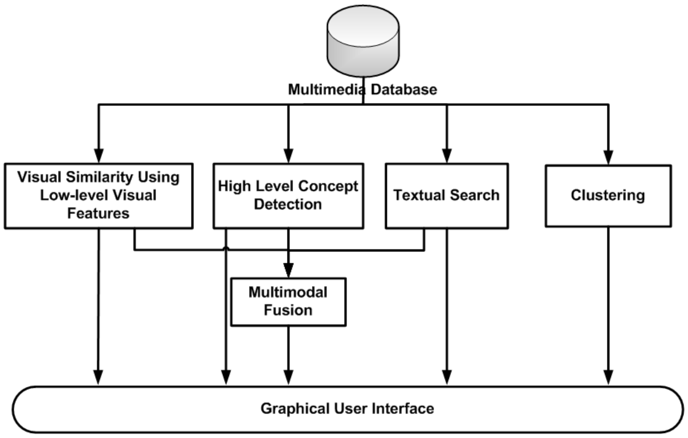


Fig. 1: VERGE architecture.

In the following, the multimedia retrieval module is discussed, along with the utilized visual and textual features.

A. Multimedia Retrieval Module based on Multimodal Fusion

In this module it is employed an extended version of the state-of-the-art unsupervised multimedia retrieval framework [6], which fuses two modalities, aiming at scalable and efficient multimedia retrieval. However, our approach fuses textual concepts, visual concepts and visual descriptors, using a hybrid graph-based and non-linear multimodal fusion method

[1] http://mklab-services.iti.gr/verge/

978-1-4673-8695-1/16/$31.00 ©2016 IEEE

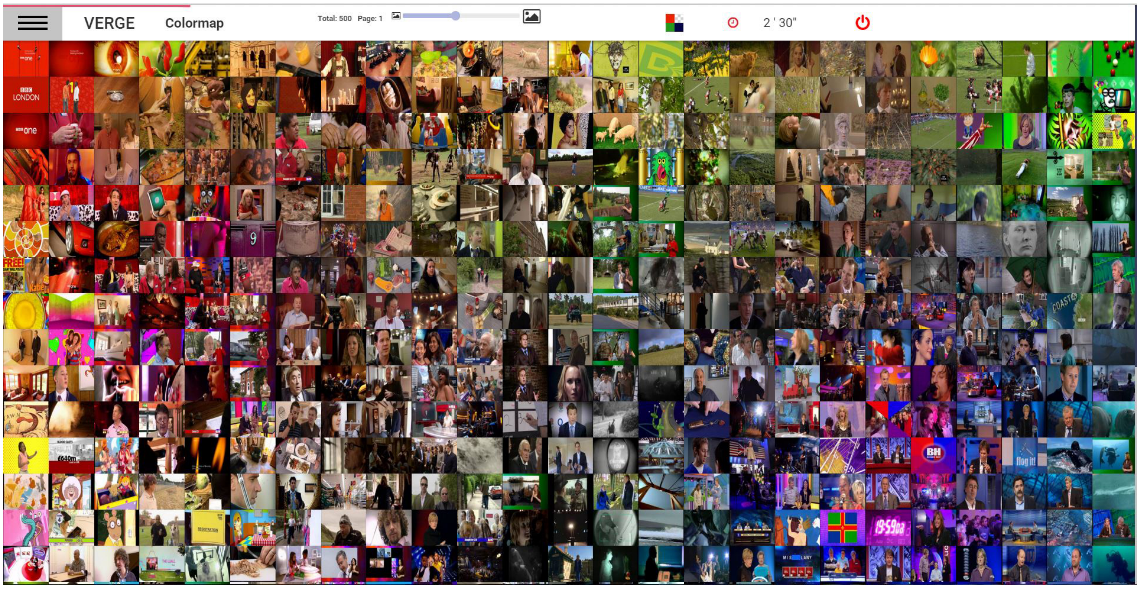


Fig. 2: Screenshot of VERGE ColorMap.

[5]. The multimodal fusion method of the VERGE system is based on the construction of a uniform multimodal contextual similarity matrix and the non-linear combination of relevance scores from query-based similarity vectors.

Multimodal fusion is performed on semantically filtered multimodal objects, utilizing the textual modality [6], and therefore, the overall memory and computational complexity of the multimedia retrieval module is reduced [4].

In brief the multimedia retrieval module [5], constructs one similarity matrix per modality and one similarity vector (query based) per modality, given M modalities and a query, but only for the results of a text-based search, assuming that text description is the main semantic source of information [6]. A graph-based fusion of multiple modalities [4] is combined with all similarity vectors in a non-linear way [5], which in general may fuse multiple modalities. In this context, we employ $M = 3$ modalities, namely visual features (RGB-SIFT), locally aggregated into one vector representation using VLAD encoding (Section II.B.1), text description (Section II.C), 346 high-level visual concepts (Section II.B.2), and textual high-level concepts, which are DBpedia[2] entities.

B. *Visual Features and Image Retrieval*

This part of the multimedia retrieval module performs content-based retrieval based on visual low-level and high-level information.

1) Low-level Visual Features: Each image is described using the RGB-SIFT feature, which is extracted in more than one square regions at different scale levels. Then, the descriptors are compacted using PCA and are aggregated using the VLAD encoding [7]. Eventually, these VLAD vectors are compressed using a modification of the random projection matrix [8].

For the Nearest Neighbor search, we create an Asymmetric Distance Computation index for the database vectors and then, we compute the K-Nearest Neighbors from the query image. Specifically, we have used the IVFADC [9], which is the non-exhaustive search variant of Asymmetric Distance Computation. IVFADC combines ADC with an inverted file to restrict the search to a subset of vectors and thus it speeds up the searching procedure.

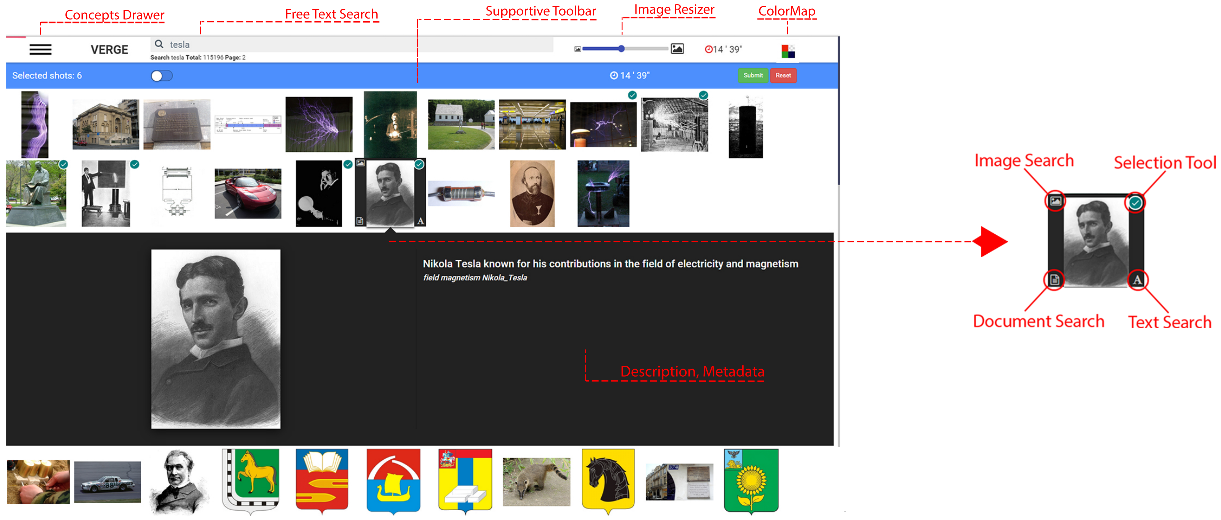
Finally, a web service is implemented in order to accelerate the querying process. In order to query the indexing structures, a two-step procedure is realized that involves: a) the loading of the index on the RAM memory, and b) the querying to the index. The loading of the index is a rather time-consuming since it requires more than two minutes for a database of size 400K. Therefore, in order to eliminate the time required for the index loading, web services are created that load these indexing structures on the RAM memory at the beginning of all runs. Then, querying of the structures is realized, which requires eventually less than a second. Therefore, the web service is created in order to achieve low response time to the user queries and thus improve her satisfaction of the system.

2) High-level Visual Features: Independent visual concept detectors are built for the detection of 346 high level concepts studied in TRECVID SIN task such as water, aircraft. The VLAD vectors of Section II.B.1 are served as input to Logistic Regression (LR) classifiers. These classifiers are trained per concept, and their output is combined by means of late fusion (averaging) [10].

C. *Textual Features and Text Search*

This part of the multimedia retrieval module utilizes the text information related to each image of the collection and the corresponding textual concepts (DBpedia concepts) [11]. The indexing of the text information is realized by Apache

[2] http://dbpedia-spotlight.github.io/demo/



VERGE

tesla

Search tesla Total: 115196 Page: 2

14 ' 39"

Selected shots: 6

14 ' 39"

Nikola Tesla known for his contributions in the field of electricity and magnetism

field magnetism Nikola_Tesla

Fig. 3: Screenshot of VERGE multimedia retrieval engine and Image options appearing over an image.

Lucene search platform, which allows full-text search and enables fast retrieval as well easy formulation of complicated queries. Apache Lucene is a high-performance search engine that allows full-text search. It allows submitting several type of queries such as phrase queries, wildcard queries, proximity queries, range queries. All these types are built-in the engine. For all textual concepts, Lucene indexing provides the similarity score between any two text documents, where Lucene-based similarity[3] scores are obtained by “Lucene’s Practical Scoring Function”.

D. ColorMap clustering module

The ColorMap (Fig. 2) clusters all images into color classes using Self Organizing Maps for image visualization and offers fast browsing in the multimedia database [12], [13]. Using Self Organizing Maps [12], all images are clustered, hence, all images of the collection are organized by color. The collection of images is represented in the GUI as a color-map, using the most representative image of all color classes. The ColorMap module restricts the image collection into color clusters of size 50-100 images, hence it is possible to quickly find an image, given its color cluster.

III. VERGE Interface

The modules described in Section II are incorporated into a unified user-friendly interface (Fig. 3) in order to aid the user search in a fast and effective way. The system architecture consists of 3 main components: MongoDB for data storage, a RESTful API and the front-end. MongoDB is reliable and fast enough while it can easily scale for use cases that demand a lot of data storage. The RESTful API has been developed in PHP. It supports Server-side paging, sorting, filtering and searching functionalities. These aforementioned components enable asynchronous data calls from the front-end (GUI).

The main results area extends to the 90% of the screen and the video search toolbar is fixed to the top covering about 10% of screen height. All the other components are collapsible. The interface comprises of three main components: a) the central component, b) the left side and c) the top search toolbar. The central component of the interface includes search results in a grid-like interface. When hovering over an image, four options appear (Fig. 3) that allow text search on the images’ description, document-based search using the multimedia retrieval module and visual search. On the left side of the interface reside the search history and additional search and browsing options that include a high level visual concepts and the ColorMap tool.

IV. Interaction Modes and Results

The aforementioned modules can aid the user interact with the system in order to discover the desired image in the collection. The user can browse the dataset by taking into account concept taxonomies and color. In addition, she can apply visual and textual search to retrieve similar images or multimodal search in order to combine all available modalities. Finally, the user can store the desirable images in a basket structure.

In order to demonstrate the practicality of the aforementioned modules, a specific usage scenario is presented, supposing that a user is interested in finding shots, which contain instances of “male color portrait”. In Fig. 4, we demonstrate the results for the textual query “portrait”. The user can start either using the high level concept “male person” or with doing a simple text search using the keyword portrait. Then from the results appearing, the user selects the images, illustrating the desired object, which in this case is the colored portraits of men, and use either multimodal search or visual search to retrieve similar images.

[3] https://lucene.apache.org/core/3_0_3/api/core/org/apache/lucene/search

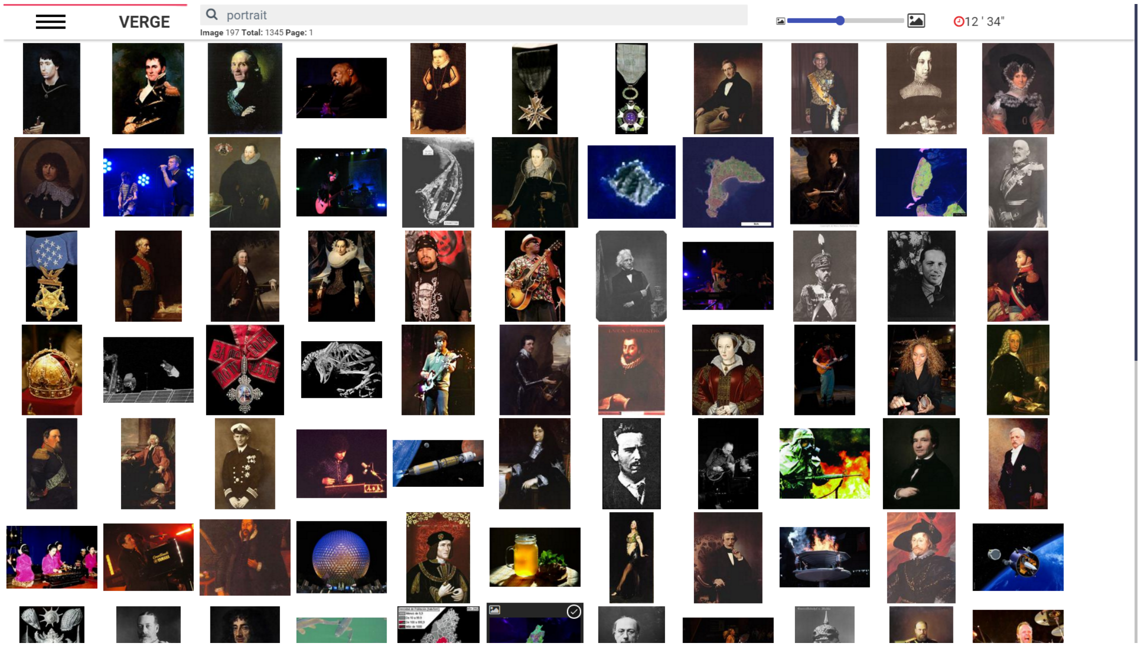


Fig. 4: VERGE instances related to “male color portrait”.

V. Conclusion

VERGE multimedia search engine allows for retrieving and browsing large multimedia collections. Low- and high-level visual and textual descriptors are extracted and enrich the multimedia collection, offering fast and efficient retrieval with respect to a given modality. Moreover, VERGE’s multimodal fusion allows for combining all modalities for retrieving relevant-to-a-query multimodal objects, where queries are served as an image collection. Finally, all images of the multimedia databased are organized by color, and therefore, allows for browsing in VERGEs image collections efficiently.

In the future, we plan to extend the present VERGE system towards the retrieval of textual documents, images and video scenes, utilizing also features from deep convolutional neural networks. This direction requires the efficient but scalable multimedia retrieval method from multiple modalities.

Acknowledgments

This work was partially supported by the European Commission by the projects MULTISENSOR (FP7-610411), HOMER (FP7-312883) and KRISTINA (H2020-645012).

References

[1] A. Moumtzidou, T. Mironidis, E. Apostolidis, F. Markatopoulou, A. Ioannidou, I. Gialampoukidis, K. Avgerinakis, S. Vrochidis, V. Mezaris, I. Kompatsiaris *et al.*, “Verge: A multimodal interactive search engine for video browsing and retrieval,” in *MultiMedia Modeling*. Springer, 2016, pp. 394–399.

[2] C. Beecks, M. S. Uysal, P. Driessen, and T. Seidl, “Content-based exploration of multimedia databases,” in *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*. IEEE, 2013, pp. 59–64.

[3] J. Moško, J. Lokoč, T. Grošup, P. Čech, T. Skopal, and J. Lánskỳ, “Mles: Multilayer exploration structure for multimedia exploration,” in *New Trends in Databases and Information Systems*. Springer, 2015, pp. 135–144.

[4] I. Gialampoukidis, A. Moumtzidou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, “Retrieval of multimedia objects by fusing multiple modalities,” in *ACM International Conference on Multimedia Retrieval (ICMR)*, 2016.

[5] I. Gialampoukidis, A. Moumtzidou, D. Liparas, S. Vrochidis, and I. Kompatsiaris, “A hybrid graph-based and non-linear late fusion approach for multimedia retrieval,” in *Content-based Multimedia Indexing (CBMI), 14th International Workshop on*, 2016.

[6] J. Ah-Pine, G. Csurka, and S. Clinchant, “Unsupervised visual and textual information fusion in cbmir using graph-based methods,” *ACM Transactions on Information Systems (TOIS)*, vol. 33, no. 2, p. 9, 2015.

[7] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3304–3311.

[8] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[9] H. Jegou, M. Douze, and C. Schmid, “Product quantization for nearest neighbor search,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 117–128, 2011.

[10] N. Gkalelis, F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, K. Avgerinakis, N. Pittaras, S. Vrochidis, V. Mezaris, I. Kompatsiaris, and I. Patras, “Iti-certh participation to trecvid 2014,” in *Proceedings of the TRECVID Workshop*, 2014.

[11] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, “Improving efficiency and accuracy in multilingual entity extraction,” in *Proceedings of the 9th International Conference on Semantic Systems*. ACM, 2013, pp. 121–124.

[12] T. Kohonen and P. Somervuo, “Self-organizing maps of symbol strings,” *Neurocomputing*, vol. 21, no. 1, pp. 19–30, 1998.

[13] K. U. Barthel, N. Hezel, and R. Mackowiak, “Imagemap-visually browsing millions of images,” in *MultiMedia Modeling*. Springer, 2015, pp. 287–290.