VERGE IN VBS 2017

Anastasia Moumtzidou [1], Theodoros Mironidis [1], Fotini Markatopoulou [1,2], Stelios Andreadis [1], Ilias Gialampoukidis [1], Damianos Galanopoulos [1], Anastasia Ioannidou [1], Stefanos Vrochidis [1], Vasileios Mezaris [1], Ioannis Kompatsiaris [1], Ioannis Patras [2]

[1]Information Technologies Institute/Centre for Research and Technology Hellas,
6th Km. Charilaou - Thermi Road, 57001 Thermi-Thessaloniki, Greece
{moumtzid, mironidis, markatopoulou, andreadisst, heliasgj, dgalanop, ioananas, stefanos, bmezaris, ikom}@iti.gr
[2]School of Electronic Engineering and Computer Science, QMUL, UK
i.patras@qmul.ac.uk

Abstract. This paper presents VERGE interactive video retrieval engine, which is capable of browsing and searching into video content. The system integrates several content-based analysis and retrieval modules including concept detection, clustering, visual similarity search, object-based search, query analysis and multimodal and temporal fusion.

1 Introduction

VERGE interactive video search engine is capable of retrieving and browsing video collections by integrating multimodal indexing and retrieval modules. VERGE has evolved to support Known Item Search (KIS), Instance Search (INS) and Ad-Hoc Video Search tasks (AVS). The aforementioned tasks require the incorporation of browsing, exploration, or navigation capabilities of the video or image collection.

The VERGE search engine was evaluated by participating in several video retrieval related conferences and showcases such as TRECVID, VideOlympics and Video Browser Showdown (VBS). Specifically, ITI-CERTH participated with consistency in several TRECVID Search tasks including the KIS task and the INS task for consecutive years starting from 2007. Moreover, it has participated in the VideOlympics event, and in VBS competition starting from 2014. The proposed version of VERGE aims at participating to the KIS and AVS tasks of VBS [1].

2 Video Retrieval System

VERGE combines advanced browsing and retrieval functionalities with a user-friendly interface, and supports the submission of queries and the accumulation of relevant results. The following indexing and retrieval modules are integrated in the developed search application: a) Visual Similarity Search; b) Object-based Visual Search; c) High Level Concepts Retrieval; d) Automatic Query Formulation and Expansion; e) ColorMap Clustering; f) CNN-based visualization; and g) Multimodal and

adfa, p. 1, 2011.

Proc. 23rd Int. Conf. on MultiMedia Modeling (MMM'17), Reykjavík, Iceland, Springer LNCS vol. 10133, pp. 486-492, Jan. 2017.

Temporal Fusion and Search. The above modules allow the user to search through a collection of images and/or video keyframes. Figure 1 depicts the general framework.

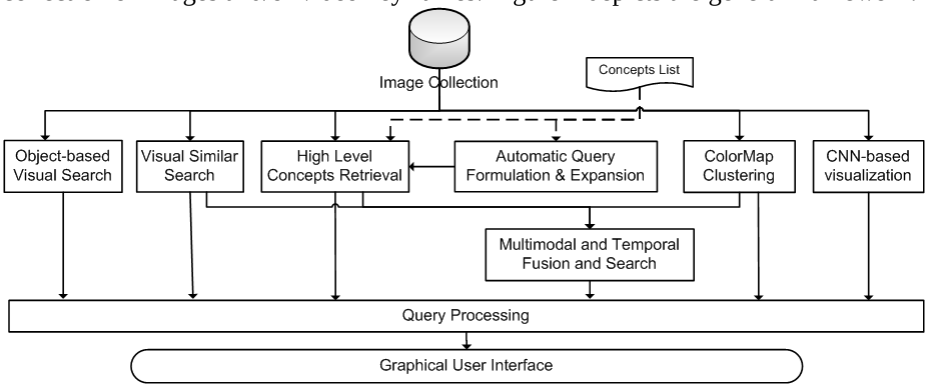

Fig. 1 Framework of VERGE.

2.1 Visual Similarity Search

This module performs content-based retrieval, also known as query by image content, by using deep convolutional neural networks (DCNNs). Specifically, we have trained GoogleNet [2] on 5055 ImageNet concepts, and we used the output of the last pooling layer, with dimension 1024, as a global keyframe representation. In order to achieve fast retrieval of similar images, we constructed an IVFADC index for database vectors and then computed K-Nearest Neighbours from the query file [3].

2.2 Object-based Visual Search

This module performs instance-based object retrieval using two different methods. The first one is based on the Bag-Of-Word model [4]. An inverted index is built for searching the image database BoW vectors, while tf-idf weights and the position of each frame in the retrieved list are used for ranking. The second one relies on Convolutional Neural Networks (CNNs). Several pre-trained CNNs are explored in order to represent each frame with features extracted either from a fully-connected or a convolutional layer. Similarity between the query and the database images is measured based on an appropriate distance. In both methods, the query can be either the keyframe or any cropped part of it.

2.3 High Level Concepts Retrieval

This module indexes the video shots based on 1000 ImageNet and 346 TRECVID SIN high level concepts (e.g. water, aircraft). The rationale for selecting a different network compared to Section 2.1 is that the retrieval accuracy of the 2.4 module that uses these concepts is higher in terms of MAP compared to the complete set of 5055 concepts. To obtain scores regarding the 1000 ImageNet concepts, we applied five pre-trained ImageNet DCNNs on the AVS test keyframes. The output of these net-

Proc. 23rd Int. Conf. on MultiMedia Modeling (MMM'17), Reykjavík, Iceland, Springer LNCS vol. 10133, pp. 486-492, Jan. 2017.

works was averaged in terms of arithmetic mean to obtain a single score for each of the 1000 concepts. To obtain the scores regarding the 346 concepts we fine-tuned (FT) two of the above pre-trained ImageNet networks on the 346 concepts using the TRECVID AVS development dataset [4]; we experimented with many FT strategies and selected the single best performing FT network. We applied it on the AVS development dataset and we used as a feature the output of the last hidden layer to train one Support Vector Machine (SVM) per concept. Then, we applied this FT network on the AVS test keyframes to extract features, and served them as input to the trained SVM classifiers in order to gather scores for each of the 346 concepts. The final step of high-level concepts retrieval was to refine the calculated detection scores by employing the re-ranking method proposed in [5].

2.4 Automatic Query Formulation and Expansion using High Level Concepts

This module formulates and expands an input query in order to translate it into a set of high level concepts. First, we check if the entire query is included in the available pool of high-level concepts. If the query is found, then no further action is necessary. Otherwise, we transform the original query to a set of elementary "subqueries", using Part-of-Speech tagging and a task-specific set of NLP rules. For example, if the original query contains a sequence in the form "Noun – Verb – Noun", this triad is considered to be a "subquery"; the motivation is that such a sequence is much more characteristic of the original query than any of these three words alone would be, and at the same time it is easier to find correspondences between this and the concepts in our pool, compared to doing so for a very long and complex query. Subsequently, we check if any of the "subqueries" are included in our concept pool. Otherwise, the original query and the subqueries are used as input to the zero-example event detection pipeline [6] and the most relevant concepts are identified. In contrast, if at least one of the subqueries is included in the pool, then we select the corresponding concepts and, we use the semantic relatedness measure [7] to select the single most semantically-relevant concept for each of the remaining subqueries. Either way, the results is a set of high-level concepts that are much related and describe well the input query given the relatively limited number of concepts in our pool.

2.5 ColorMap Clustering

Motivated by [8] and [9], video keyframes are clustered by color using Self Organizing Maps (SOM) into color classes. Each color class is represented in the GUI by the most representative image within the color class. All representative images are determined by their distances to the SOM's best matching unit per color class. Each particular image I in the collection is represented in RGB form, and indexed as a tor $(s_R, s_G, s_B)_I$, where s_R, s_G and s_B is the similarity score between the image I and the pure red (s_R), pure green (s_G) and pure blue (s_B) image, respectively. The similarity score is based on pixel-by-pixel comparisons and averaged over all pixels of the image I. In this way, VERGE clusters all images into color classes and offers fast browsing in the collection of video keyframes.

Proc. 23rd Int. Conf. on MultiMedia Modeling (MMM'17), Reykjavík, Iceland, Springer LNCS vol. 10133, pp. 486-492, Jan. 2017.

2.6 CNN-based visualization

CNNs are proposed for the effective visualization of datasets, given that they can be interpreted as gradually transforming the images into a representation in which the classes are separable by a linear classifier. The method tested is the t-SNE method [10] which has shown very satisfactory results. The procedure followed involves taking a set of images and extracting CNN codes. These codes are plugged into t-SNE and a 2-dimensional vector is produced for each image. Finally, the corresponding images are visualized in a grid.

2.7 Multimodal and Temporal Fusion and Search

This module fuses the visual descriptors of Section 2.1, the concepts of Section 2.3 and the color features of Section 2.5. Given a query shot and its central keyframe in the time domain, this module retrieves similar shots by performing center-to-center comparisons among video shots. On the top-k retrieved shots, re-ranking is performed, taking into account the adjacent keyframes of the top-retrieved shots.

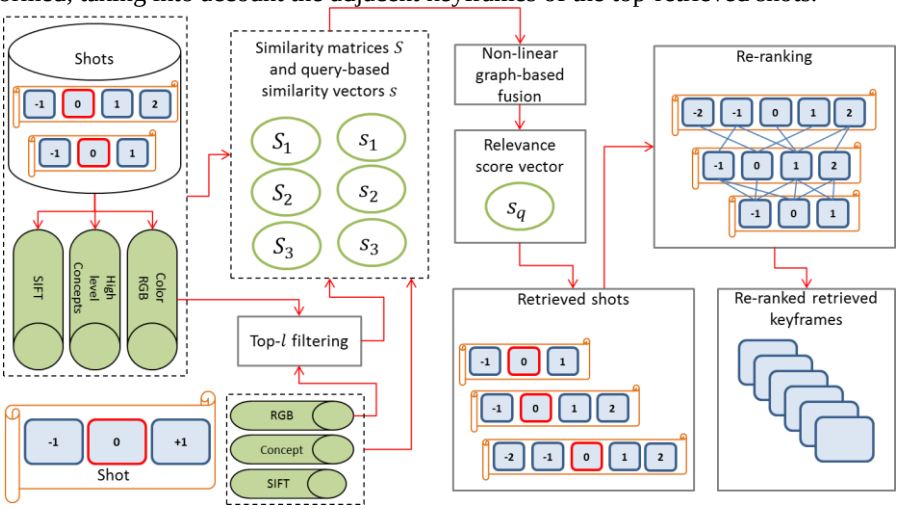

Fig. 2 The VERGE multimodal fusion and search module.

More specifically, as depicted in Fig. 2, given a query shot, its color features, its concepts and its DCNN descriptors are extracted from the central frame, which is marked by zero in the figure. An initial filtering stage keeps only the top-l relevant-to-the-color shots and then computes an $l \times l$ similarity matrix and an $l \times 1$ similarity vector per modality. The similarity matrices and vectors are then fused in a non-linear and graph-based way, following [11], providing a fused relevance score vector s_q for the retrieved shots. Pairwise comparisons on the keyframes of the top-k retrieved shots result to the final list of re-ranked retrieved keyframes. In the initial top-l filtering stage we used the color features as the dominant modality, but this could be set on demand. The final re-ranking stage involves the computation of mutual similarities among not necessarily central keyframes between any two shots.

3 VERGE Interface and Interaction Modes

This year the retrieval utilities incorporated into the system are more than ever, offering a multitude of search options. Thus, the user interface has to give the end user an intuitive and effective way to run queries fast and obtain the best possible results.

A novel component of the interface is the General purpose search input field, common to the user's experience of other search engines. The user can initiate the search procedure by simply describing the shot he/she is looking for, in natural language. The system analyses the text in an intelligent manner, using the *Automatic Query Formulation and Expansion using High-Level Concepts* module, and returns a single or more combined concepts. The user can edit this proposed list by adding or removing concepts and perform a concept-based search. Another novelty is the *Multimodal and Temporal Fusion* that the user can easily invoke by clicking on one or more shots to serve as query, and selecting the dominant modality for each shot.

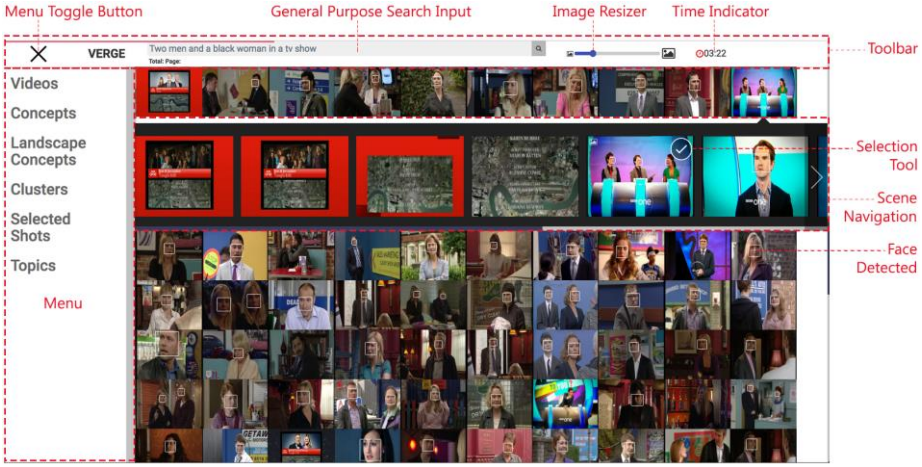

Fig. 3 Screenshot of VERGE video retrieval engine.

Describing the user interface (Fig. 3), there is a toolbar with many useful options on the top. In detail, from left to right, a burger icon opens a toggle menu that contains the different search capabilities, namely the *Concept-* and *Topic-based* search, and the *ColorMap Clustering*. The menu also includes the user's selected shots and the total set of video shots. Next to the application's logo, the General purpose search input field can be seen, followed by the Image Resizer that modifies the amount of results in the viewport, by changing the size of the shots. The last toolbar component applies only to the contest and shows the remaining time for the submission, accompanied by an animated red line on the top of the screen. The central component of the interface includes a shot-based representation of the video results in a grid-like view. Clicking on a shot allows the user to navigate through the whole scene where this frame belongs, displaying the related shots in a chronological order. Moreover, each shot supports tools to run the *Visual Similarity* and the *Object-based Visual* modules. Finally, all selected images are saved in a deposit that can be quickly accessed for further searching or just for the submission.

Proc. 23rd Int. Conf. on MultiMedia Modeling (MMM'17), Reykjavík, Iceland, Springer LNCS vol. 10133, pp. 486-492, Jan. 2017.

To illustrate the functionality of the VERGE interface[1], we describe a simple usage scenario. Supposing that the user is interested in finding a clip of two men and a black woman in a TV show (Fig. 3), he/she can begin with the General purpose search. An appropriate selection of proposed concepts is received (e.g. *Female_Person*, *Two_People*), that the user is able to edit before performing the concept-based search. If a relative image is found during this step, the user can continue with all the above-mentioned retrieval modules to collect more similar images, as well as browse the complete scene to find previous or next shots.

4 Future Work

Future work includes applying multimodal temporal fusion and search on multiple queries. The retrieved results can then be fused and the user will be presented with a single list. Another feature would be to allow the user to create a more complicated query using as base a shot, and describe it by considering multiple modalities.

Acknowledgements This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-687786 InVID, H2020-693092 MOVING, H2020-645012 KRISTINA and H2020-700024 TENSOR.

References

1. Cobârzan, C., et al.: Interactive video search tools: a detailed analysis of the video browser showdown 2015. Multimedia Tools and Applications, pp. 1-33 (2015)
2. Szegedy, C., et al.: Going deeper with convolutions. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9 (2015)
3. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, pp. 117-128 (2011)
4. Markatopoulou, F., et al.: ITI-CERTH participation to TRECVID 2015. In TRECVID 2015 Workshop, Gaithersburg, MD, USA (2015)
5. Safadi B., and Quénot, G.: Re-ranking by local re-scoring for video indexing and retrieval. 20th ACM Int. Conf. on Information and Knowledge Management, pp. 2081–2084 (2011)
6. Tzelepis, C., Galanopoulos, D., Mezaris, V., and Patras, I.: Learning to detect video events from zero or very few video examples. Image and Vision Computing (2015)
7. Gabrilovich, E., and Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In IJCAI, vol. 7, pp.1606–1611 (2007)
8. Barthel, K. U., Hezel, N., and Mackowiak, R.: ImageMap - Visually Browsing Millions of Images. In MultiMedia Modeling, pp. 287-290 (2015)
9. Moumtzidou, A., et al.: A multimedia interactive search engine based on graph-based and non-linear multimodal fusion. In CBMI 2016 International Workshop, pp. 1-4 (2016)
10. t-SNE visualization of CNN codes, http://cs.stanford.edu/people/karpathy/cnnembed/
11. Gialampoukidis, I., et al.: A hybrid graph-based and non-linear late fusion approach for multimedia retrieval. In CBMI 2016 International Workshop, pp. 1-6. IEEE. (2016)

[1] http://mklab-services.iti.gr/vss2016

Proc. 23rd Int. Conf. on MultiMedia Modeling (MMM'17), Reykjavík, Iceland, Springer LNCS vol. 10133, pp. 486-492, Jan. 2017.