Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11296. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05716-9_53

VERGE IN VBS 2019

Stelios Andreadis[1], Anastasia Moumtzidou[1], Damianos Galanopoulos[1], Foteini Markatopoulou[1], Konstantinos Apostolidis[1], Thanassis Mavropoulos[1], Ilias Gialampoukidis[1], Stefanos Vrochidis[1], Vasileios Mezaris[1], Ioannis Kompatsiaris[1], and Ioannis Patras[2]

[1] Information Technologies Institute/Centre for Research & Technology Hellas, Thessaloniki, Greece
{andreadisst, moumtzid, dgalanop, markatopoulou, kapost, mavrathan, heliasgj, 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, an interactive video retrieval engine that enables browsing and searching into video content. The system implements various retrieval modalities, such as visual or textual search, concept detection and clustering, as well as a multimodal fusion and a reranking capability. All results are displayed in a graphical user interface in an efficient and friendly manner.

1 Introduction

VERGE interactive video search engine integrates a multitude of indexing and retrieval modules, aiming to provide efficient browsing and search capabilities inside video collections. During the last decade, VERGE has participated in several video retrieval related conferences and showcases, with the most recent being TRECVID [1] and Video Browser Showdown (VBS) [2]. The system is adapted to support Known Item Search (KIS), Instance Search (INS) and Ad-Hoc Video Search tasks (AVS). Inspired by other participations in VBS, i.e. VIREO [3], HTW [4], and SIRET [5], a novel graphical user interface (GUI) was introduced in 2018, transitioning from a multi- to a single-page website with a dashboard menu. This year, VERGE incorporates notable improvements regarding the user experience and new approaches to previously ineffective techniques, e.g. video clustering and fusion. Following the example of SIRET, winner of the last VBS, an advanced keyword search is also investigated. Section 2 describes the current version of the search modules and Section 3 illustrates the GUI.

2 Video Retrieval System

VERGE serves as a video retrieval system with straightforward browsing in a friendly environment and diverse search functionalities, which can be combined either with fusion or reranking the relevant results. So far the following indexing and retrieval modules have been integrated: a) Visual Similarity Search;

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11296. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05716-9_53

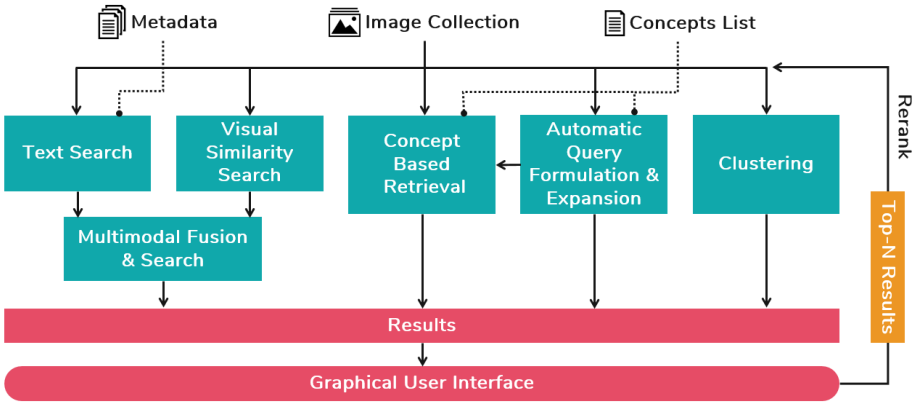

Fig. 1. VERGE System Architecture.

b) Concept-Based Retrieval; c) Automatic Query Formulation and Expansion; d) Clustering; e) Text-Based Search; and f) Multimodal Fusion. Utilizing these modules, the user is able to search through a collection of images or videos and the retrieved results are displayed on a GUI. Furthermore, the system allows the user to rerank the top-N results using alternative modalities. The general architecture of the VERGE system is depicted in Figure 1.

2.1 Visual Similarity Search

This module performs content-based retrieval using deep convolutional neural networks (DCNNs). Compared to last year's implementation, which utilized as global keyframe representation the output of the last pooling layer of the GoogleNet trained on 5055 ImageNet concepts, a more complicated feature representation will be evaluated. Specifically, the output of different layers from different networks such as the AlexNet, the GoogLeNet, the ResNet, the VGG Net and the CaffeNet will be concatenated into a single feature vector. In the sequel, an IVFADC index database vector will be created for fast indexing and K-Nearest Neighbors will be computed for the query image [6].

2.2 Concept-Based Retrieval

This module indexes each shot that is part of the VBS video collection using a pool of 1000 ImageNet concepts, 345 TRECVID SIN concepts, 500 event-related concepts, and 205 place-related concepts [7]. To obtain scores regarding the 1000 ImageNet concepts, we applied five pre-trained ImageNet deep convolutional neural networks on the shot's keyframes. For each of the 1000 concepts, the average (arithmetic mean) of the output of these networks was used as the final score for the given concept. To obtain scores for the 345 TRECVID SIN concepts, we used the deep learning framework of [8]. This deep architecture learns to

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11296. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05716-9_53

annotate video frames not just by looking at each concept separately, but also by explicitly making use of concept relations that exist at two different levels: the (lower) visual representation level, and the (higher) semantic correlations level. Training of the network for the 345 concepts was performed using the 600-hours training portion of the TRECVID-SIN 2013 dataset [9]. Finally, to obtain scores for the event- and place-related concepts we applied the publicly available DCNNs fine-tuned on the EventNet [10] and Places [11] datasets, respectively.

2.3 Automatic Query Formulation and Expansion

This module splits the input query into smaller and more meaningful textual parts, and uses them to translate the original query into a set of high-level visual concepts C_Q (those listed in Section 2.2). For this, the concept-based query representation procedure proposed in [12] is adopted. Specifically, a set of steps is defined in order to decompose the query and associate each part of it with visual concepts. In the first step, we search for one or more concepts that are semantically similar to the entire query, using the Explicit Semantic Analysis (ESA) measure. If such concepts are found (according to a threshold θ) we assume that the entire query is well described by them; the selected concept(s) is (are) added in the initially empty set C_Q, and no further action is taken. Otherwise, in a second step we examine if any of the concepts in our concept pool appears in the query by string matching, and if so these concepts are added in C_Q. In the third step the original query is transformed into a set of elementary "subqueries", and for that we search for Noun Phrases in the query. Then in the fourth step, concepts that are semantically similar to any of the "subqueries", are added into the set C_Q. Otherwise, if C_Q is still empty, in a final step the original query and all the "subqueries" are used as input to the zero-example event detection pipeline [13], which attempts to find the concepts that are most closely related to them. The outcome of this procedure is a set of concepts C_Q describing the input query.

2.4 Clustering

Two clustering approaches are applied for the effective visualization of the dataset:

ColorMap clustering: Video keyframes are clustered by color using three MPEG-7 descriptors related to color, i.e. Color Layout, Color Structure, Scalable Color. These descriptors are extracted for each video frame, and then each frame is mapped to a color of the palette by using either euclidean distance or k-means.

Video clustering: Videos are clustered by using the visual concepts of their keyframes and their metadata. Specifically, for each video we retrieve the top-N ranked concepts among its keyframes, assuming that these concepts describe accurately the video, since most times a certain topic is presented in their short length. Regarding the video metadata, they are processed by stemming and removing stopwords and punctuation. Thus, each video is represented as a text vector and then Latent Dirichlet Allocation is applied on the vectors in order to identify a predefined number of topics inside the video collection.

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11296. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05716-9_53

2.5 Text-Based Search

This year we evaluate a more sophisticated approach that entails the exploitation of online lexical resources to leverage semantic features. Online databases like WordNet or DBpedia use an elaborate system of interlinked metadata that enables semantically similar words to be interconnected. The process of moving from terms literally occurring in treated text to their semantically corresponding concepts in semantic resources is called conceptualization [14]. Capitalizing on this feature we manage to retrieve words like "client" when the available terms only include "customer". Naturally, not all terms have respective equivalents in these external semantic resources, so only specific concepts can be mapped to the original list of available terms.

2.6 Multimodal Fusion and Search

This module fuses the results of two or more search modules, such as the visual descriptors of Section 2.1, the concepts of Section 2.2 and the color features of Section 2.4. Similar shots are retrieved by performing center-to-center comparisons among video shots by using the selected modules. It is possible to describe the query with more than one features (e.g. a shot, a color and/or some concepts) and in that case one feature, considered as dominant, returns the top-N relevant shots, while the others rerank the initial list by using a non-linear graph-based fusion method [15]. Eventually, on the top-N retrieved shots, reranking is performed, taking into account the adjacent keyframes of the top retrieved shots.

3 VERGE User Interface and Interaction Modes

The VERGE user interface (Fig. 2) is designed in a modern style and aims to be a competitive tool for image or video retrieval. For this year the target is to improve the user experience in terms of responsiveness and intuitiveness.

The GUI[a] can be divided into three main parts: a vertical dashboard-like menu on the left, a results panel that spans to the majority of the screen and a filmstrip on the bottom. From top to bottom, the menu includes a countdown timer that shows the remaining time for submission during the contest, a slider to adjust the image size, a back button to restore results from previous search queries and a switch button to select whether new results will be retrieved or returned results will be reranked by a different approach. The various search options are presented to the user in a compact manner, in the form of sliding boxes. In detail, *Search in Metadata* is a text input field that looks for the typed keywords into the video metadata, e.g. their title or description, (Section 2.5), *Search for Concepts* transforms a natural language sentence to suggested concepts (Section 2.3), while *Concepts* is the entire list, along with the options of auto-complete search and multiple selection (Section 2.2). Moreover, *Events* is a set of predefined queries that combine persons, objects, locations, and activities,

[a] http://mklab-services.iti.gr/vbs2018

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11296. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05716-9_53

Fig. 2. Screenshot of the VERGE web application.

Video Similarity provides a grouping of videos that are most similar and *Colors* offer a palette in order to retrieve images of a specific shade. The outcome of every search module is displayed on the central and largest panel of the interface either as single shots or group of shots (video) in a grid view, sorted by retrieval scores. Hovering over an image allows users to run the *Visual Similarity Search* (Section 2.1) or to submit the shot to the contest. Clicking on the image updates the bottom filmstrip with the complete scene where this frame belongs to, showing the related shots in a chronological order.

To illustrate the capabilities of the VERGE system, we propose a simple usage scenario. Supposing that we are looking for an image of *a man next to his red car*, we can initiate the retrieval procedure by converting this sentence to concepts, utilizing the corresponding module. The suggested concepts are "Adult Male Human" and "Car", so we can combine these two to get the first results. In order to detect a red car, we are able to rerank the retrieved shots by the color red from the palette. Alternatively, we can search for the word "vehicle" into the videos' metadata and when a shot of a red car appears, we exploit the visual similarity modality to find more related images.

4 Future Work

Future work spans in two directions: on the one hand, we intend to improve the existing search modules as far as it concerns their performance and their responsiveness; on the other hand, we would like to explore novel techniques for video content retrieval, e.g. color sketching.

Acknowledgements This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-779962 V4Design, H2020-700024 TENSOR, H2020-693092 MOVING and H2020-687786 InVID.

References

1. G. Awad, A. Butt, J. Fiscus, et al. Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA, 2017.
2. C. Cobârzan, K. Schoeffmann, W. Bailer, et al. Interactive video search tools: a detailed analysis of the video browser showdown 2015. *Multimedia Tools and Applications*, 76(4):5539–5571, Feb 2017.
3. Phuong Anh Nguyen, Yi-Jie Lu, Hao Zhang, and Chong-Wah Ngo. Enhanced vireo kis at vbs 2018. In *International Conference on Multimedia Modeling*, pages 407–412. Springer, 2018.
4. Kai Uwe Barthel, Nico Hezel, and Klaus Jung. Fusing keyword search and visual exploration for untagged videos. In *International Conference on Multimedia Modeling*, pages 413–418. Springer, 2018.
5. Jakub Lokoč, Gregor Kovalčík, and Tomáš Souček. Revisiting siret video retrieval tool. In *International Conference on Multimedia Modeling*, pages 419–424. Springer, 2018.
6. H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):117–128, January 2011.
7. F. Markatopoulou, A. Moumtzidou, D. Galanopoulos, et al. ITI-CERTH participation in TRECVID 2017. In *Proc. TRECVID 2017 Workshop*, USA, 2017.
8. F. Markatopoulou, V. Mezaris, and I. Patras. Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. *IEEE Trans. on Circuits and Systems for Video Technology*, 2018.
9. P. Over et al. TRECVID 2013 - an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. TRECVID 2013 Workshop*, USA, 2013.
10. Y. Guangnan, Yitong L., Hongliang X., et al. Eventnet: A large scale structured concept library for complex event detection in video. In *Proc. ACM Multimedia Conference (ACM MM)*, 2015.
11. B. Zhou, A. Lapedriza, J. Xiao, et al. Learning deep features for scene recognition using places database. In *Proc. NIPS*, pages 487–495, 2014.
12. F. Markatopoulou, D. Galanopoulos, V. Mezaris, and I. Patras. Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 407–411. ACM, 2017.
13. D. Galanopoulos, F Markatopoulou, V. Mezaris, and I. Patras. Concept language models and event-based concept number selection for zero-example event detection. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 397–401. ACM, 2017.
14. S. Albitar, S. Fournier, and B. Espinasse. The impact of conceptualization on text classification. In *International Conference on Web Information Systems Engineering*, pages 326–339. Springer, 2012.
15. 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 *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, June 2016.