Proc. 20th Int. Conf. on MultiMedia Modeling (MMM'14), Dublin, Ireland, LNCS vol. 8326, pp. 411-414, January 2014.
The final publication is available at link.springer.com

VERGE: An Interactive Search Engine for Browsing Video Collections

Anastasia Moumtzidou[1], Konstantinos Avgerinakis[1], Evlampios Apostolidis[1], Vera Aleksic[2], Fotini Markatopoulou[1], Christina Papagiannopoulou[1], Stefanos Vrochidis[1], Vasileios Mezaris[1], Reinhard Busch[2], Ioannis Kompatsiaris[1]

[1]Information Technologies Institute/Centre for Research and Technology Hellas,
6th Km. Xarilaou - Thermi Road, 57001 Thermi-Thessaloniki, Greece
{moumtzid, koafgeri, apostolid, markatopoulou, cppapagi, stefanos, bmezaris, ikom}@iti.gr
[2]Linguatec Sprachtechnologien GmbH, Gottfried-Keller-Str. 12, 81245 München
{v.aleksic, r.busch}@linguatec.de

Abstract. This paper presents VERGE interactive video retrieval engine, which is capable of searching and browsing video content. The system integrates several content-based analysis and retrieval modules such as video shot segmentation and scene detection, concept detection, clustering and visual similarity search into a user friendly interface that supports the user in browsing through the collection, in order to retrieve the desired clip.

1 Introduction

This paper describes VERGE interactive video search engine[1], which is capable of retrieving and browsing video by integrating different indexing and retrieval modules. VERGE supports the Known Item Search task, which requires the incorporation of techniques for browsing and navigation within a video collection. VERGE was evaluated with participation in workshops and showcases such as TRECVID and VideOlympics, where it was shown to significantly improve user search experience over single or fewer search modalities. Specifically, VERGE demonstrated the best results in the interactive known item search of TRECVID 2011 by achieving a Mean Inverted Rank of 0.56, while the concept detectors of VERGE achieved good balance between detection accuracy (e.g. 15.8% MXinfAP for TRECVID 2013) and low computational complexity.

The proposed version of VERGE aims at participating to the KIS task of the Video Browser Showdown (VBS) Competition [1]. In this context, VERGE supports interactive searching of a known video clip in a large video collection by incorporating content-based analysis and interactive retrieval techniques.

In the next sections we present the content-based analysis and retrieval techniques supported by VERGE, as well as the interaction with the user.

[1] More information and demos of VERGE are available at: http://mklab.iti.gr/verge/

Proc. 20th Int. Conf. on MultiMedia Modeling (MMM'14), Dublin, Ireland, LNCS vol. 8326, pp. 411-414, January 2014.
The final publication is available at link.springer.com

2 Video Retrieval System

VERGE is a retrieval system, which combines advanced retrieval functionalities with a user-friendly interface. The following basic modules are integrated: a) Shot and Scene Segmentation; b) Textual Information Processing Module; c) Visual Similarity Search; d) High Level Concept Detection; e) Clustering.

The **shot segmentation module** performs shot segmentation by extracting visual features, namely color coherence, Macbeth color histogram and luminance center of gravity, and forming a corresponding feature vector per frame [2]. Then, given a pair of frames, the distances between their vectors are computed, composing distance vectors that are finally evaluated using one or more SVM classifiers, resulting to the detection of both abrupt and gradual transitions between the shots of the video.

Scene segmentation is based on the previous analysis and groups shots into sets that correspond to individual scenes of the video. The algorithm [3] introduces and combines two extensions of the Scene Transition Graph (STG); the first one aims to reduce the computational cost by considering shot linking transitivity, while the second one constructs a probabilistic framework towards combining multiple STGs.

The **textual information processing module** applies Automatic Speech Recognition (ASR) on videos. We employ the VPE (Voice Pro Enterprise) framework, which is based on RWTH-ASR technology [4]. Finally, each shot is described by a set of words, which are used to create a taxonomy to facilitate browsing of the collection.

The **visual similarity search module** performs content-based retrieval based on global and local information. To deal with global information, MPEG-7 descriptors are extracted from each keyframe and they are concatenated into a single feature vector. Efficient retrieval is achieved by employing the r-tree indexing structure. In the case of local information SURF features are extracted. We apply two Bag of Visual Words techniques for representing and retrieving images efficiently. On the first, we calculate visual vocabularies via hierarchical k-means clustering [5], while on the second, we follow K-Means clustering and VLAD encoding for representing images.

The **high level concept retrieval module** indexes the video shots based on 346 high level concepts (e.g. water, aircraft). For each keyframe we employ up to 25 feature extraction procedures [6]. For learning these concepts, a bag of linear Support Vector Machines (LSVM) is trained for each feature extraction procedure and each concept. A sampling strategy is applied to partition the dataset into 5 subclasses and for each subset a LSVM is trained. During the classification phase, a new unlabeled video shot is given to the trained LSVMs, each of them returns the degree of confidence that the concept is depicted in the image, and late fusion is used for combining these scores.

Finally, the **clustering module** incorporates an agglomerative hierarchical clustering process [7], which provides a hierarchical view of the keyframes. In addition to the feature vectors used as input to the high level concept retrieval module, we extract vectors consisting of the responses of the trained concept detectors for each video shot. The clustering algorithm is then applied to these representations in order to group the keyframes into clusters, each of which consists of keyframes having visually or semantically similar content.

Proc. 20th Int. Conf. on MultiMedia Modeling (MMM'14), Dublin, Ireland, LNCS vol. 8326, pp. 411-414, January 2014.
The final publication is available at link.springer.com

VERGE is built on Apache server, PHP, JavaScript and mySQL database. Besides the aforementioned basic modules, VERGE integrates the following complementary functionalities: a) basic temporal queries, b) shot storage structure and c) history bin.

3 Interaction Modes

The aforementioned modules aid the user to interact with the system through a user-friendly interface (Figure 1), in order to discover the desired video clip during known item search tasks. The interface comprises of three main components: a) the central component, b) the left side and c) the upper panel.

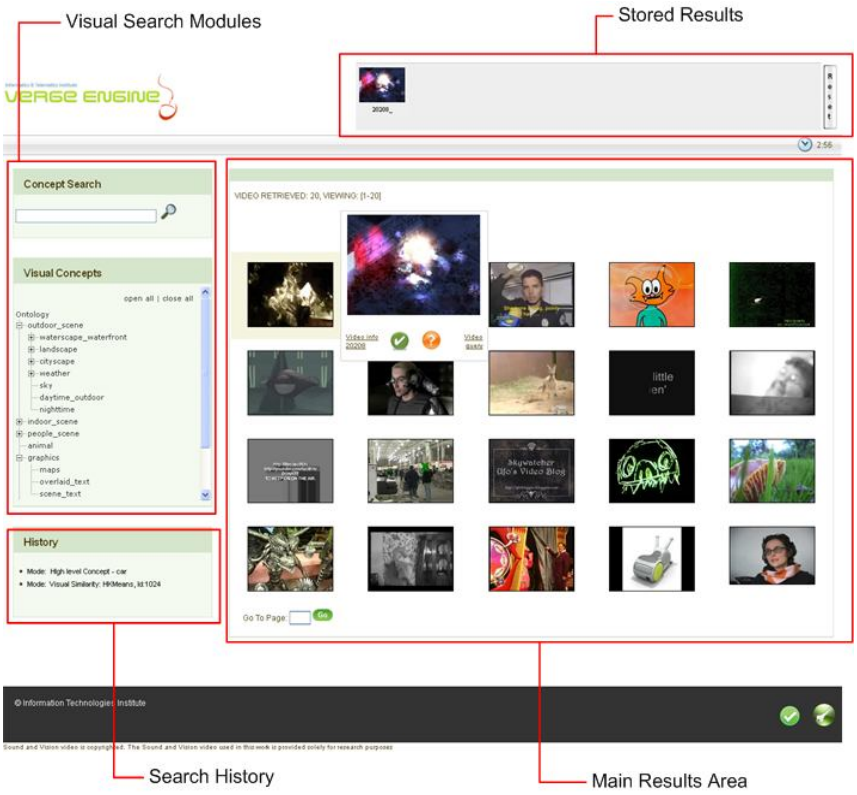

Fig. 1. Screenshot of VERGE video retrieval engine

The central component of the interface includes a shot-based representation of the video in a grid-like interface. In this grid each video shot is visualized by a representative key frame. When the user hovers over an image, a pop-up frame appears that contains a larger preview of the image in order to allow for better inspection of its content, as well as several links that support:

- browsing the temporally adjacent shots and all shots of the specific video

Proc. 20th Int. Conf. on MultiMedia Modeling (MMM'14), Dublin, Ireland, LNCS vol. 8326, pp. 411-414, January 2014.
The final publication is available at link.springer.com

- registering an image as relevant for a specific topic or query
- searching for visually similar images to the given (query) image

On the left side of the interface, the search history, as well as additional search and browsing options are displayed. The history module automatically records all searching actions done by the user, while the search and browsing options include the taxonomy based on the ASR transcriptions, the high level visual concepts and the hierarchical clustering. Using the aforementioned functionalities, the user can browse the dataset at shot and scene level taking also into account ASR and concept taxonomies.

Finally, the upper panel is a storage structure that mimics the functionality of the shopping cart found in electronic commerce sites and holds the shots selected by the user throughout the session.

Acknowledgements This work was partially supported by the European Commission under contracts FP7-287911 LinkedTV, FP7-318101 MediaMixer and FP7-610411 MULTISENSOR.

References

1. Schoeffmann, K., Bailer, W.: Video Browser Showdown. ACM SIGMultimedia Records, vol. 4, no. 2, pp. 1-2 (2012)
2. Tsamoura, E., Mezaris, V., Kompatsiaris, I.: Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. IEEE International Conference on Image Processing, Workshop on Multimedia Information Retrieval (ICIP-MIR 2008), pp. 45-48, San Diego, CA, USA (2008)
3. Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: Temporal video segmentation to scenes using high-level audiovisual features. IEEE Trans. on Circuits and Systems for Video Technology, vol. 21, no. 8, pp. 1163-1177 (2011)
4. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., Ney, H.: The RWTH Aachen University Open Source Speech Recognition System. In: Interspeech, pp. 2111-2114, Brighton, UK (2009)
5. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), vol. 2 (2006)
6. Moumtzidou, A., Gkalelis, N., Sidiropoulos, P., Dimopoulos, M., Nikolopoulos, S., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: ITI-CERTH participation to TRECVID 2012. Proc. TRECVID 2012 Workshop, Gaithersburg, MD, USA (2012)
7. Johnson, S.C.: Hierarchical Clustering Schemes. Psychometrika, vol. 2, pp. 241-254 (1967)