Proc. Fifth IEEE International Conference on Semantic Computing (ICSC 2011), Palo Alto, CA, USA, September 2011, pp. 224-227.

Improving interactive video retrieval by exploiting automatically-extracted video structural semantics

Vasileios Mezaris, Panagiotis Sidiropoulos, Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas
6th Km Charilaou-Thermi Road, Thermi 57001, Greece
Email: {bmezaris, psid, ikom}@iti.gr

***Abstract*—In this work the contribution of automatically-extracted (thus, imperfect) video structural semantics towards improving interactive video retrieval is examined. First, the automatic extraction of video structural semantics, i.e. the decomposition of the video into scenes that correspond to the different sub-stories or high-level events, is performed. Then, these are introduced to the interactive video retrieval paradigm. Finally, their potential contribution is experimentally evaluated. To this end, different members of a family of scene segmentation algorithms are applied to an extensive professional video collection coming from the TRECVID benchmarking activity; subsequently, a large number of user interactions with a retrieval system that exploits these structural semantics is simulated. The experimental results document the contribution of state-of-the-art automatically-extracted video structural semantics to the efficient and effective interactive video retrieval.**

***Keywords*-Video structural semantics; semantic video retrieval; interaction.**

I. Introduction

Semantic video retrieval is a key application in today's networked world. The main related challenge for the industry and the researchers lies in bridging the gap between the possible video content representations, which are typically machine-only-readable (e.g. low-level audio-visual features), unreliable and incomplete (e.g. automatic visual concept detection results, user-assigned tags) or too specific to be meaningful when seen out of context (e.g. tag "Mary"), and the very specific and very diverse at the same time information needs of every possible user. While the research community tries to respond to this challenge at many fronts, e.g. by developing new low-level features and more reliable semantic concept detectors [1], it is the close interaction of the searcher with the video retrieval system that is generally acknowledged as one of the most powerful classes of techniques for facilitating semantic video retrieval [2].

In this work, we go beyond popular techniques for facilitating interactive video retrieval, such as elaborate user interfaces, relevance feedback and affective retrieval, to examine if and to what extent the automatic extraction of (inevitably, imperfect) video structural semantics can contribute to interactive retrieval. The rest of the paper is organized as follows: In section II, some related works are briefly reviewed. This is followed in sections III and IV by an outline of how video structural semantics can be meaningfully introduced to the interactive retrieval paradigm, and how they can be automatically extracted using state-of-the-art techniques. The evaluation setting and results are presented in sections V and VI, and conclusions are drawn in section VII.

II. Related work

Intelligent video retrieval is typically performed at the shot level (e.g. [2]). This is dictated both by the significant variability in the video content of an entire program (e.g. a movie, a documentary), which necessitates separately indexing each elementary temporal segment of it, and the need of users for retrieving only the bits of information that are of interest to them at any given time. In this context, interactive video retrieval is based on providing the user with a set of functionalities for assisting in searching and navigating within a large collection of video shots.

State-of-the-art video search engines integrate a plurality of such functionalities. These include support for different query formulations (e.g. query-by-text, query-by-example), query expansion, relevance feedback, browsers for visualizing the collection or a subset of it according to different criteria (e.g. concept relevance, time), and others. In [3], [4], query interfaces such as the ForkBrowser and CrossBrowser are central to the interactive search system. Their basic building block is the thread, defined as "a linked sequence of shots in a specified order, based upon an aspect of their content". Various threads are defined, such as time threads (:span the temporal similarity between shots), visual threads (:span the visual similarity between shots), etc. The significance of the time thread in particular is emphasized in [4]. In [5], [6], the interactive search system supports a multitude of query formulations, including basic temporal queries such as the presentation of a fixed number of neighboring shots for each specified shot ("side shots") and the shot-segmented view of each entire video. In [7], [8], shot boundaries and ASR transcripts are used for identifying "stories" in the video; these stories provide the basic unit of retrieval during queries. An overview of additional interactive retrieval systems that participated to TRECVID, the de facto standard for video retrieval evaluation, can be found in [2].

Figure 1. Example of simple temporal query (i.e., "show me neighboring shots to the one I have found"), by (a) indiscriminately presenting the N side shots ($N = 3$), and (b), exploiting the scene membership of the query shot to present the same number of most relevant neighboring shots to the user. In both sub-figures, the query shot s_i is shown in a circle, the vertical bars indicate automatically detected scene boundaries, and any neighboring shots that are not returned to the user as a result of the temporal query are shown here shaded.

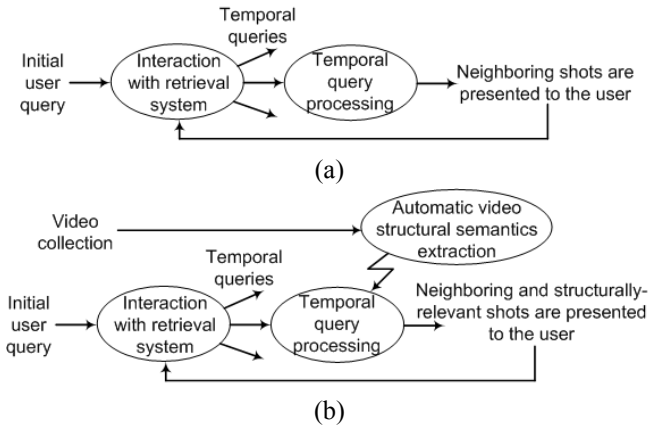

Figure 2. Diagram illustrating (a) a typical sequence of actions in an interactive video retrieval system, and (b) how automatically-extracted video structural semantics can be integrated in interactive retrieval.

Table I
BASIC TEMPORAL QUERIES

(a) Without considering scene boundaries:

query shot $s_i \quad \rightarrow \quad$ show $s_j,\ j \in [i-N,\ i+N],\ N = const$

(b) Based on scene boundary detection (considering a single scene):

query shot $s_i \quad \rightarrow \quad$ show all $s_j \in S_k \mid s_i \in S_k$

(c) Based on scene boundary detection (considering multiple scenes):

query shot $s_i \quad \rightarrow \quad$ show all $s_j \in \{S_{k-X}, ..., S_{k+X}\} \mid s_i \in S_k$ and X is a positive integer

III. VIDEO STRUCTURAL SEMANTICS IN INTERACTIVE RETRIEVAL

Common characteristic of the video search engines reviewed above, as well as of several others, is that they acknowledge the significance of temporal information for interactive retrieval. Specifically, the shots that are temporally close to a correctly retrieved shot s_i are intuitively considered very likely to also be relevant to the query; to this end, basic temporal querying functionalities (such as quickly showing to the searcher the N side shots) are typically implemented. Interestingly, though, the use of temporal information is either governed by ad hoc rules (e.g. N side shots are shown, where N is fixed; see Fig. 1(a)), or not governed at all (e.g. time threads are provided, which are essentially a sequential view of the entire video, shot-by-shot [3], [4]).

We hypothesize in this work that automatically-extracted video structural semantics, i.e. the outcome of algorithms for video segmentation to scenes $S_k,\ k = 1, ..., K$, can in cases such as the above intelligently guide the user in visually inspecting a variable number of temporally neighboring shots that are most likely to also satisfy the query criteria (Fig. 2). Having the possibility to automatically detect scene boundaries, it comes as a natural choice to use this information for guiding the user in visually inspecting shots that have been found to belong to the same scene as the positive result already retrieved, rather than indiscriminately looking at neighboring shots (Fig. 1 and Table I).

It should be emphasized here that a hypothesis such as the above could straightforwardly be accepted only if it involved the use of perfectly accurate structural semantics (such as those generated by manual inspection and structural annotation of the video). However, manual processing of large collections of video for extracting structural semantics is practically infeasible, and the state-of-the-art techniques for performing this task automatically generate results that still deviate considerably from perfection (e.g. [9], [10]). Therefore, it is by no means straightforward to say that video structural semantics extracted automatically by current state-of-the-art techniques are useful in interactive retrieval, nor is it of course possible to quantify their potential contribution without detailed experimentation.

Table II
LIST OF AUTOMATIC SCENE SEGMENTATION ALGORITHM VARIATIONS OF [10] THAT ARE USED IN OUR EXPERIMENTS.

M1 - Using low-level visual features only, optimal parameters
M2 - Using low-level visual features only, parameters favoring over-segmentation
M3 - Using low-level visual features only, parameters favoring under-segmentation
M4 - Combining low-level visual features and concept detector responses, all 101 detectors used
M5 - Combining low-level visual features and concept detector responses, 60 detectors selected according to AP
M6 - Combining low-level visual features and concept detector responses, 50 detectors selected according to ΔAP

IV. AUTOMATIC EXTRACTION OF VIDEO STRUCTURAL SEMANTICS

Early approaches to scene segmentation focused on exploiting just low-level visual or audio features for grouping similar shots into scenes, e.g. [11]. Most recent techniques, e.g. [10], [12], further exploit higher-level information such as visual concept and audio event detection results in order to come to a more accurate extraction of the videos' structural semantics. Specifically, in [10] the possibility of exploiting, for the purpose of video segmentation to scenes, semantic information coming from the analysis of the visual modality, was examined.

For the purpose of the study presented in this work, 6 different variations of the method of [10] were used (Table II). These differ in the information they use as input for extracting the video structural semantics (i.e., low-level visual features only for variations M1 to M3; low-level features and the responses of visual concept detectors ("visual soft semantics") for variations M4 to M6) and in the setting of their parameters (i.e., the shot similarity threshold for variations M1 to M3; the number of considered concept detectors and the strategy for their selection for variations M4 to M6). In general, the variations of the scene segmentation algorithm that take into account visual soft semantics were shown in [10] to produce scene segmentation results that are in better agreement with ground truth scene boundaries. The interested reader is referred to the aforementioned work for further details on these algorithms.

V. EVALUATION SETTING

For the experimental evaluation of the impact of scene segmentation results in interactive retrieval, we used the test portion of the NIST TRECVID[1] 2007 dataset. This is made of approximately 50 hours of professionally-created videos (Dutch TV documentaries), decomposed to 18142 shots. Two classes of queries were defined on this dataset: single-concept queries, and complex queries. Single concept queries refer to queries for shots that depict a particular object or elementary action. The 20 such queries used here correspond to the concepts defined for the 2009 edition of the High-level Feature Extraction Task of TRECVID (e.g. "Cityscape", "People dancing"). Complex queries also refer to particular objects or elementary actions, but tend to introduce additional conditions, such as two objects appearing together. The 24 such queries used here correspond to the queries of the 2007 edition of the Search Task of TRECVID (e.g. "A door being opened", "A person walking or riding a bicycle"). Manually generated ground truth query results are available for both sets of queries: complete ground truth for the single-concept queries and partial for the complex ones. In the latter case, the non-annotated shots are treated as negative samples; this choice derives from the way the pool of ground-truth-annotated shots was formed at NIST.

Two main approaches to showing neighboring shots to the user (i.e., basic temporal queries) were simulated, as shown in Fig. 1 and Table I: (a) indiscriminately presenting the N side shots immediately before and after each positive sample included in the collection, and (b), exploiting the scene membership of the chosen shot to present to the user all shots belonging to the same scene as the given positive sample (or all shots belonging to a number of scenes, in subsequent experiments). These temporal queries were repeated for all positive samples of each original query (overall, 3322 temporal queries in response to single concept queries and 4704 in response to complex queries), and average results are reported. For generating the scene segmentation results, the 6 algorithm variations outlined in section IV were used, and each of them was evaluated separately. For quantifying the results of each experiment, the harmonic mean (F-score) of the widely used precision (P) and recall (R) measures was used; F-score$= \frac{2PR}{(P+R)}$. F-score essentially measures how successful each basic temporal querying strategy is in retrieving additional positive samples of the original query, given that one such positive sample has already been found by the searcher and is used for launching a basic temporal query.

VI. EVALUATION RESULTS AND DISCUSSION

Results from the use of each of the two main temporal query strategies (Table I(a) and (b)) are reported, separately for single-concept queries and complex queries, in Fig. 3(a) and (b). For the single-concept queries (Fig. 3(a)), it can be seen that the F-score attained when exploiting the results of scene segmentation is consistently higher than when the N side shots of query shot s_i are indiscriminately returned. This holds even for scene segmentation methods M2 and M3, which have been selected so as to result in significant over- and under-segmentation, respectively. Considering the remaining segmentation methods (M1, M4-M6), these are shown to present marginal differences among them, and they increase the F-score by over 30% compared to the case where no scene segmentation results are exploited, for the same number of shots returned by the temporal

[1] http://www-nlpir.nist.gov/projects/trecvid/

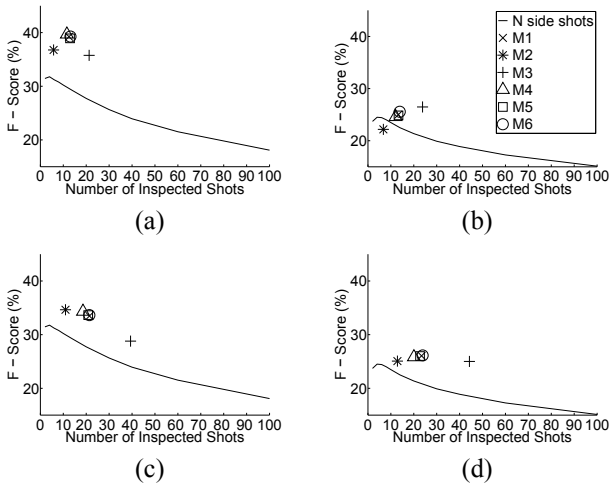

Figure 3. F-score as a function of the number of shots returned by the temporal query (and thus inspected by the user). The solid line corresponds to indiscriminately returning the N side shots immediately before and after each positive sample included in the collection; the other symbols correspond to using the results of the 6 automatic scene segmentation variations listed in Table II for returning: (a), (b) exactly the shots that belong to the same scene as the given positive sample; (c), (d) exactly the shots that belong either to the same scene as the given positive sample or to the scene immediately preceding or following it (therefore, increasing the number of shots returned by the temporal query, compared to the results of sub-figures (a) and (b)). Results are presented for: (a), (c) the single-concept queries; (b), (d) the complex queries.

query. For the complex queries (Fig. 3(b)), the results are in general similar, although the F-score differences are less pronounced. With the exception of the over-segmentation case (M2), exploiting the scene segmentation results leads to higher F-score.

These experiments were subsequently repeated, modifying the scene-based response of the temporal queries as follows:

$$\text{query shot } s_i \;\rightarrow\; \text{show all } s_j \in \{S_{k-1}, S_k, S_{k+1}\} | s_i \in S_k \quad (1)$$

i.e., following the temporal query strategy of Table I(c) and setting $X = 1$. Consequently, all shots belonging to the same scene as the query shot, or to any of the two scenes "bracketing" it [7], are presented to the user as the response of the temporal query. The effect of this is that a larger number of shots is returned to the user, without however compromising the accuracy of scene segmentation (i.e. without enforcing extreme under-segmentation). Results are reported in Fig. 3(c) and (d). We can see that the results for single-concept queries are not significantly affected, in the sense that the gains of exploiting scene segmentation persist. Considering complex queries, though, there are differences: the gains of exploiting scene segmentation become more pronounced, and all 6 segmentation methods are shown to outperform the "show N side shots" strategy. This difference in performance can be attributed to the qualitative differences in the distribution of the positive samples in the dataset for the two classes of queries.

VII. Conclusions

In this work, the use of automatically-extracted video structural semantics for responding to basic temporal queries, which are typically an important part of users' interaction with a video retrieval system, was examined. It was shown that using existing state-of-the-art scene segmentation algorithms to this end can indeed improve the efficiency and effectiveness of interactive retrieval. Future work includes the investigation of whether such an experimental setting can also serve as a method for the objective evaluation of scene segmentation algorithms in large datasets without using ground-truth scene segmentation results.

Acknowledgment

This work was supported by the European Commission under contract FP7-248984 GLOCAL.

References

[1] X. Liu and B. Huet, "Automatic concept detector refinement for large-scale video semantic annotation," in *Proc. Fourth IEEE International Conference on Semantic Computing (ICSC 2010)*, Pittsburgh, PA, USA, September 2010, pp. 97–100.

[2] P. Over, G. Awad, J. Fiscus, M. Michel, A. Smeaton, and W. Kraaij, "TRECVID 2009 - Goals, Tasks, Data, Evaluation Mechanisms and Metrics," in *Proc. TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[3] C. Snoek, K. van de Sande, O. de Rooij, and et. al., "The MediaMill TRECVID 2009 Semantic Video Search Engine," in *Proc. TRECVID 2009 Workshop*, Gaithersburg, MD, USA, November 2009.

[4] O. de Rooij and M. Worring, "Browsing Video Along Multiple Threads," *IEEE Trans. on Multimedia*, vol. 12, no. 2, pp. 121–130, February 2010.

[5] S. Vrochidis, A. Moumtzidou, P. King, A. Dimou, V. Mezaris, and I. Kompatsiaris, "VERGE: A video interactive retrieval engine," in *Proc. 8th International Workshop on Content-Based Multimedia Indexing (CBMI 2010)*, Grenoble, France, June 2010.

[6] A. Moumtzidou, A. Dimou, N. Gkalelis, S. Vrochidis, V. Mezaris, and I. Kompatsiaris, "ITI-CERTH participation to TRECVID 2010," in *Proc. TRECVID 2010 Workshop*, Gaithersburg, MD, USA, November 2010.

[7] J. Pickens, J. Adcock, M. Cooper, and A. Girgensohn, "FXPAL Interactive Search Experiments for TRECVID 2008," in *Proc. TRECVID 2008 Workshop*, Gaithersburg, MD, USA, November 2008.

[8] J. Adcock, M. Cooper, and J. Pickens, "Experiments in Interactive Video Search by Addition and Subtraction," in *Proc. ACM Int. Conf. on Image and Video Retrieval*, Niagara Falls, Ontario, Canada, July 2008, pp. 465–474.

[9] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "On the use of audio events for improving video scene segmentation," in *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Desenzano del Garda, Italy, April 2010.

[10] V. Mezaris, P. Sidiropoulos, A. Dimou, and I. Kompatsiaris, "On the use of visual soft semantics for video temporal decomposition to scenes," in *Proc. Fourth IEEE International Conference on Semantic Computing (ICSC 2010)*, Pittsburgh, PA, USA, September 2010, pp. 141–148.

[11] M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 71, pp. 94–109, July 1998.

[12] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso, "Temporal video segmentation to scenes using high-level audiovisual features," *IEEE Trans. on Circuits and Systems for Video Technology, accepted for publication*, 2011.