An Efficient Visual Search Engine for Cultural Broadcast Archives

Emanuele Caimotti[1], Maurizio Montagnuolo (✉)[2], and Alberto Messina[2]

[1] Politecnico di Torino, Torino, Italy
ema86c@hotmail.it
[2] RAI Radiotelevisione Italiana, Torino, Italy
{maurizio.montagnuolo, alberto.messina}@rai.it

Abstract. In todays digital age, the ability to access, analyze and (re) use large amounts of data is a strategic asset of fundamental importance for the broadcast and media industry. The challenge lies in the ability to search, organize and access multimedia assets in a fast and semantically relevant way. Visual search is the new frontier to achieve these objectives, by allowing users to match image and video contents depicting the same objects, such as buildings, paintings and logos, based on visual similarities and without the need of querying for manually generated metadata. This paper presents the implementation of a Content Retrieval Architecture (CRA) for visual content analysis and search. Experimental results to demonstrate the feasibility of the proposed architecture on cultural TV programme archives are discussed as well.

Keywords: Content-based video retrieval, visual search, broadcast archives

1 Introduction and Related Work

In todays digital age, television content life cycle has a very long span: after being produced and broadcasted, a copy of the content, possibly enriched with metadata, is stored in the archive to be reused when needed or to be published online. Multimedia asset management (MAM) systems provide tools to store and retrieve media files. Pioneer systems used by the industry employed text-based queries to search over textual information and metadata, typically associated to each stored file using either semi-automatic or handmade annotations. While this procedure is still in practise these days, due to its overall reliability and robustness, it presents some critical weaknesses. In fact, metadata extraction is an expensive and time consuming process, which requires human supervision and needs to be done both for audiovisual content that is already produced digitally, as well as for vintage footage that is converted from analog to digital formats. New technologies are needed to increase documentation efficiency, as well as access and (re)use of video archives.

Content-based retrieval (CBR) relies on the idea of indexing and matching image and video contents based on visual characteristics, in addition to manually generated metadata. Many methods have been developed to achieve this goal.

Despite the considerable effort, almost all the available CBR systems still suffer from the semantic gap issue, being based on low-level features, e.g. color, shape and motion, rather than on high level concepts. To overtake this issue, efficient algorithms for object recognition, such as those for key-point feature detectors and descriptors, have been proposed. For this purpose, the Scale-Invariant Feature Transform (SIFT) algorithm is considered a pioneer work [9]. The Moving Picture Experts Group (MPEG) started in 2010 a standardization initiative called Compact Descriptors for Visual Search (CDVS, now ISO/IEC 15938-14) that provides a robust and interoperable technology to create efficient visual search applications in image databases. The core building blocks of CDVS consist in global and local descriptor extractors and compressors based on selected SIFT features [2]. Duan et al [7] provide an overview of the technical features of the related MPEG CDVS standard. MPEG defines also a reference software (Test Model) that implements common visual search operations (pairwise matching and content retrieval) using CDVS technology [3]. In pairwise matching mode two images are compared using both local and global SIFT descriptors and a similarity score is provided. Whereas in content retrieval mode firstly a CDVS database is filled with descriptors of reference images, then a query image is compared with the entire database and an image list is provided. In the end the returned list is sorted by a score based on global descriptors. Recently the interest is moving forward to the video domain. Intuitively, video analysis is a more challenging problem than still images due to temporal and spatial redundancy in video, which increases the amount of data that need to be processed. The LIvRE project [11] represents an interesting attempt at exploring the expansion of Lucene Image Retrieval Engine (LIRE), an open-source Content-Based Image Retrieval system, for video retrieval on large scale video datasets. Furthermore, in order to meet industrial needs, the MPEG CDVA (Compact Descriptors for Video Analysis) Evaluation Framework aims to enable efficient and interoperable design of compact video description technologies for search and retrieval in video sequences [1].

Being a public broadcaster, RAI has the promotion of Italy's historical, artistic and cultural heritage among its mission objectives. For this purpose, several hours of programmes are produced and broadcasted daily, as well as archived for preservation and future access. In order to maximize the reuse of those assets, the ability to efficiently search, organize and access content in a fast and semantic-driven way is an asset of fundamental importance. A novel approach for image retrieval and automatic annotation of cultural heritage images is proposed in [8]. An automatic video analysis and retrieval system for searching in historical collections of broadcasts of the former German Democratic Republic (GDR) is presented in [10]. A comprehensive overview of key issues and research efforts in multimedia analysis for cultural heritage is discussed in [6].

In this paper, an automatic Content Retrieval Architecture (CRA) for video analysis and search is presented. The architecture is designed to meet requirements given by handling large volume of video contents. The paper is organized

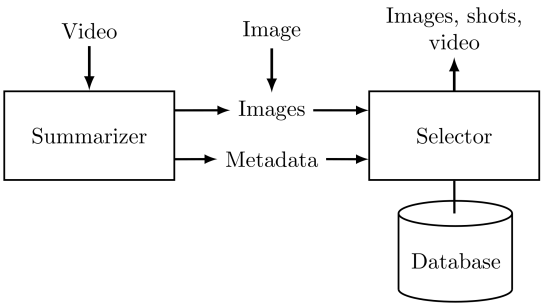


Fig. 1: Content Retrieval Architecture (CRA) overview.

as follows. Section 2 describes the proposed architecture. Section 3 presents preliminary experimental results, and finally, Section 4 concludes the paper.

2 System Architecture

The proposed architecture extracts local and global features from video and performs retrieval operations based on content similarities. It has been designed to strongly reduce video redundancy and concentrate processing resources on informative contents. Its functionalities extend the MPEG CDVS technology to video content keeping the standard interoperability. The architecture is made of three main modules, as illustrated in Fig. 1: (i) The *Summarizer* segments the video in shots and extracts the most representative key-frames; (ii) The *Selector* extracts CDVS descriptors from key-frames, gathers similar shots in clusters and performs key-frame ranking and cluster ranking by relevance; (iii) The *database* (DB) stores information and metadata about video structure, ranking lists and visual descriptors. The architecture can work in two modalities, namely extraction and retrieval. In extraction mode, a reference (input) video is segmented into shots and representative key-frames are selected from each shot. CDVS descriptors are computed from each key-frame and stored in the database for retrieval operations. In retrieval mode a query video is processed in the same way as a reference video. CDVS descriptors are then extracted and matched against those stored in the database. Matching videos are returned as lists of key-frames, shots and videos sorted according to the matching score.

2.1 Summarizer Building Blocks

Fig. 2a shows the Summarizer building blocks. This is the first extraction stage of the CRA architecture. It takes an input video, performs video pre-processing, shot detection, key-frame extraction and provides in output a selection of key-frames grouped by shot. The main goal of the Summarizer is to reduce as much as possible the video redundancy creating a structure based on images (i.e. key-frames) that represents the original video. Shot detection and key-frame extraction make use of video motion information. The input video is re-encoded using

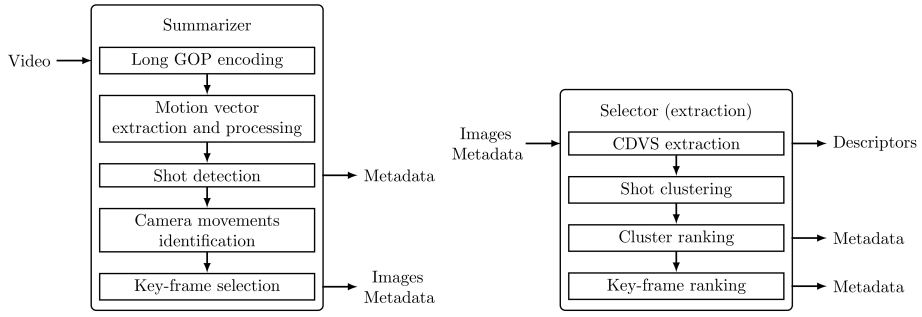
very long GOP (Group Of Pictures) settings and motion vectors are extracted. Then unreliable motion vectors on the edges and within smooth regions are discarded and the remaining vectors are accumulated. Shot transitions are detected when abrupt changes in the motion vectors field happen and the percentage of intra coded motion vectors exceeds a defined threshold τ_{intra}. Each frame is classified as "zooming" for zoom and "dolly" or "panning" for pan, tilt, pedestal and truck accordingly to the global camera movement. Furthermore a two pass filter, based on a temporal window of size $T_w = 5$, evens spurious zooming or panning frames. Next a frame is selected as key-frame when the following conditions are met: (i) It is the first of a shot; (ii) It is the last frame of a zoom; (iii) The distance crossed during a panning exceeds the size of the frame height.

2.2 Selector Building Blocks

Fig. 2b shows the Selector building blocks. This is the second stage in the CRA pipeline and it is responsible for content selection and indexing, starting from a list of video to be processed. Indexing is based on saliency that is the representation of the temporal contents presence. The selector allows for flexibility and scalability of the system since it creates an index structure that allows to retrieve contents at granular level, from single key-frames to whole videos. Processing starts with the extraction of the CDVS descriptors from each key-frame. Then duplicate and near duplicate shots are clustered to reduce redundancy. Two video shots s_a and s_b are merged if at least one key-frame in s_a and one key-frame in s_b matches with a CDVS score grater than a defined threshold θ. After that, each generated cluster c_k is assigned to a score $W_k = \sum_{i=1}^{N} w_i$, where N is the number of key-frames of cluster c_k, and $w_i = |kf_i - kf_{i-1}| + |kf_i - kf_{i+1}|$ counts how many frames are represented by key-frame kf_i. The output of the cluster ranking module is a sorted list of video clusters $\mathcal{C} = (c_1, \ldots, c_K)$ ordered by weight w_k, and counting the representativeness of cluster c_k w.r.t. the analyzed video (see Fig. 3a). All the key-frames within each video cluster are compared to each other in order to select the most representative key-frame for each cluster (see Fig. 3b). In the end the Selector creates a tree structure in the database that is a top-down representation of the video content. From the highest to the lowest aggregation level, the database includes information about shot clusters, video shots and key-frames within a video. Besides this tree structure, the cluster and key-frame ranking lists together with key-frame CDVS descriptors are store in the database as well.

3 Experimental Evaluation

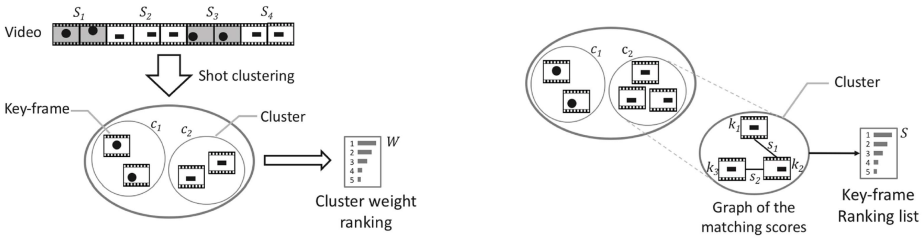
In this section, the datasets for our experiments are first introduced and subsequently more details about our experimental settings are provided. Three datasets have been selected among those provided as part of the CDVA evaluation framework [1]. The Telecom Italia dataset (TI-CTurin180) includes 30 minutes of short videos recorded with mobile phone cameras and showing Turin



(a) Summarizer processing blocks.

(b) Selector processing blocks.

Fig. 2: Detail of the CRA architecture.



(a) Shot clustering and cluster weight ranking creation.

(b) Key-frame graph and key-frame ranking list.

Fig. 3: Video hierarchical segmentation process.

buildings.[3] The David Daniels (DD) dataset is focused on London historical buildings recorded in non-professional way. The "RAI Monuments of Italy" dataset (RAI-IT-MON) includes about 2,000 clips depicting about 200 monuments from all over Italy, mainly acquired from RAI regional newscasts.[4] Furthermore, a new dataset, "RAI Viaggio nella Bellezza" (RAI-BELLEZZE), made of over 20 hours of video material showing monuments, paintings, sculptures and historic locations belonging to the artistic heritage of Italy, was acquired from RAI's cultural TV programmes. A visual example of keyframes extracted from the experimental datasets is shown in Fig. 4. Three types of experiments have been executed aimed to evaluate (i) shot boundaries detection effectiveness, (ii) content compression efficiency and (iii) content retrieval performance.

Shot boundary detection (SBD) has been tested using part of the RAI-IT-MON dataset manually annotated with ground truth. Despite this is not the focus of the system, we performed this test to verify that content retrieval performance were not affected by possible wrong shot segmentations. Considering as true positives the shot boundaries correctly detected, false positives the

[3] https://pacific.tilab.com/www/datasets/ (last accessed July 2017)

[4] The "RAI Monuments of Italy" dataset can be made available on request

Fig. 4: Example of keyframes from the experimental datasets. From top to bottom, left to right: Royal Palace (TI-CTurin180); Santa Giulia church (TI-CTurin180); St. Paul (DD); London Bridge (DD); Florence, Palazzo Vecchio (RAI-IT-MON); Milan Cathedral (RAI-IT-MON); Agrigento, Valley of the Temples (RAI-BELLEZZE); Mantua, Palazzo Te (RAI-BELLEZZE).

Table 1: Content compression performance.

	Compression	**Matching score (avg)**	**Quality (avg)**
Sub-sampler	1:30	7.7	0.23
CRA Summarizer	1:48	6.5	0.63

shot boundaries wrongly detected and false negatives the missing ground truth boundaries, the shot boundaries detector achieved average Precision, Recall and F-measure of 0.86, 0.89 and 0.88, respectively. This is comparable with state of the art, where F-measure ranging from 0.84 and 0.9 is reported [4, 5].

In the second experiment, the content compression provided by the Summarizer with the key-frame selection has been tested using single-shot video, as reported in Table 1. The number of extracted key-frames is compared with the results of a uniform subsampling algorithm (4 frames per shot). Then CDVS Test Model pairwise matching is used to compare a query image with the extracted key-frames and obtain an average score $\overline{s_c}$. The quality of the extracted key-frames is evaluated as $q = \frac{\overline{s_c}}{\overline{\#k}}$, where $\overline{\#k}$ is the average number of extracted key-frames. The efficiency (i.e. compression ratio) of our key-frame selection algorithm is comparable with uniform subsampling. However, the selected key-frames are more representative (i.e. higher quality in Table 1) of the video shots. Furthermore, the loss of CDVS matching accuracy is not significant since empirical studies demonstrated that CDVS matching scores higher than 3 result in near-100% matching precision.

The last experiment is aimed at evaluating the performance of the CRA architecture when used for search and retrieval. Two tests were conducted for video-to-video search and image-to-video search, respectively. Video-to-video search was performed according to the following steps: (i) The datasets have been ran-

Table 2: Video-to-Video retrieval performance. TP = True Positive; FP = False Positive; FN = False Negative; P = Precision; R = Recall

Dataset	**Query (Reference) videos**	**TP**	**FP**	**FN**	**P**	**R**	$\mathbf{F}_{\text{measure}}$
TI-CTurin180	180 (1800)	1790	0	10	1	0.994	0.997
RAI-IT-MON	463 (1700)	963	200	1991	0.828	0.326	0.468
DD	16 (117)	33	1	58	0.97	0.363	0.526

domly split in two parts, the former used as query dataset and the latter as reference dataset; (ii) All the videos have been processed in order to detect key-frames, shots and clusters as previously described; (iii) Reference videos have been stored in the database according to the Extraction mode of the architecture; (iv) Query videos have been matched to the reference database according to the Retrieval mode of the architecture. Results have been collected and analyzed in terms of Precision, Recall and F-measure. Precision and recall are measured considering retrieved videos as true positives when related to the query video and false positives when unrelated to the query. Furthermore expected videos that are not retrieved, are considered as a false negative. Results are reported in Table 2. The TI-CTurin180 dataset got optimal performance in terms of both precision and recall. The peak of false positives in the RAI-IT-MON dataset is due to some elements appearing in most of the videos, such as logos, graphics or studio settings. This behavior might be mitigated applying some pre filtering heuristics to the input data (e.g. frame cropping). Recall significantly drops down for both DD and RAI-IT-MON datasets. However, this issue mainly depends on the datasets themselves rather than on the CRA architecture. In fact, many buildings in these datasets are captured from different sides. The semantic gap between non superimposable view of the same object can not be overcome because of the lack of common elements. This behavior may be mitigated e.g. including more views of the same object in the reference dataset. Furthemore, precision is preserved by all the analyzed dataset. Image-to-video search was performed similarly as for video-to-video search, using the RAI-BELLEZZE as reference dataset and Web images as query items. Images concerning some of the artistic subjects (i.e. monuments, paintings, statues, buildings and archeological sites) depicted in the reference videos have been automatically collected from Google Image search. Achieved precision is 1, meaning that no false positive results were returned by the system. Finally, we examined processing time spent at each pipeline stage using the TI-CTurin180 dataset and an Ubuntu 14.04 LTS virtual machine configured with dual Intel Xeon E5-2690@2.9GHz 16 cores (32 threads, maximum number of running threads limited to 12) and 32GB RAM DDR3L@1.6GHz (maximum memory usage limited to 8GB). The longest operation was performed by the Summarizer ($\sim 15'$). The extraction process took $\sim 2'$. Even if both summarization and extraction are performed only one time for each video and are normally server side jobs, a faster summarization algorithm may be investigated in future developments. Retrieval took $\sim 2''$.

4 Conclusions

This paper presented an end to end video retrieval architecture based on global and local feature descriptors. Due to its flexibility, this architecture may be implemented in different demanding application, from cultural applications on smartphones to professional catalogs management on servers. Experimental results demonstrated the feasibility of the system, in particular when the objective is to achieve high precision, while lower recall is acceptable. Processing times demonstrated that the architecture implementation is compatible with an asymmetric client-server implementation, where the core jobs (summarization and extraction) are performed in the background on server side. Future work will include the analysis of the impact of different video summarization techniques.

Acknowledgments. This work was partly funded by the European Union's Horizon 2020 programme (grant No 731667, MULTIDRONE).

References

1. Evaluation framework for compact descriptors for video analysis - search and retrieval. ISO/IEC JTC1/SC29/WG11/N15338 (2015)
2. Information technology - Multimedia content description interface - Part 13: Compact descriptors for visual search. Tech. rep., ISO/IEC 15938:13 (2015)
3. Information technology - Multimedia content description interface - Part 14: Reference software, conformance and usage guidelines for compact descriptors for visual search. Tech. rep., ISO/IEC 15938:14 (2015)
4. Apostolidis, E.E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp. 6583–6587 (2014)
5. Baraldi, L., Grana, C., Cucchiara, R.: Shot and scene detection via hierarchical clustering for re-using broadcast video. In: 16th Int. Conf. on Computer Analysis of Images and Patterns. pp. 801–811 (2015)
6. Cucchiara, R., Grana, C., Borghesani, D., Agosti, M., Bagdanov, A.D.: Multimedia for Cultural Heritage: Key Issues, pp. 206–216. Springer Berlin Heidelberg (2012)
7. Duan, L.Y., Huang, T., Gao, W.: Overview of the MPEG CDVS Standard. In: Proc. of the 2015 Data Compression Conference. pp. 323–332. DCC '15 (2015)
8. Grana, C., Serra, G., Manfredi, M., Cucchiara, R.: Beyond Bag of Words for Concept Detection and Search of Cultural Heritage Archives. In: Proc. of the 6th Int. Conf. on Similarity Search and Applications. pp. 233–244 (2013)
9. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision 60(2), 91–110 (Nov 2004)
10. Mühling, M., Meister, M., Korfhage, N., Wehling, J., Hörth, A., Ewerth, R., Freisleben, B.: Content-based video retrieval in historical collections of the german broadcasting archive. In: 20th Int. Conf. on Theory and Practice of Digital Libraries. pp. 67–78 (2016)
11. de Oliveira-Barra, G., Lux, M., Giró-i Nieto, X.: Large Scale Content-Based Video Retrieval with LIvRE. In: 14th Int. Workshop on Content-based Multimedia Indexing (CBMI) (2016)