Large Scale Content-Based Video Retrieval with LIvRE

Gabriel de Oliveira Barra
University of Barcelona
Email: gabriel.deoliveira@ub.edu

Mathias Lux
Klagenfurt University
Email: mlux@itec.aau.at

Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Email: xavier.giro@upc.edu

***Abstract*—The fast growth of video data requires robust, efficient, and scalable systems to allow for indexing and retrieval. These systems must be accessible from lightweight, portable and usable interfaces to help users in management and search of video content. This demo paper presents LIvRE, an extension of an existing open source tool for image retrieval to support video indexing. LIvRE consists of three main system components (pre-processing, indexing and retrieval), as well as a scalable and responsive HTML5 user interface accessible from a web browser. LIvRE supports image-based queries, which are efficiently matched with the extracted frames of the indexed videos.**

I. Introduction

Multiple Content-Based Image Retrieval (CBIR) systems are publicly available, all with different characteristics regarding performance and features. Content-Based Video Retrieval (CBVR) systems, however, are scarcer and present additional challenges when compared to their image-oriented counterparts. These factors are primarily related to the temporal information available from a video document and the difficulty to handle big amounts of data [1].

The increasing amount of data generated by ubiquitous video cameras increases the need for tools capable of managing them. Retrieval tasks like finding scenes with certain semantic content, or similar moments to the one depicted by an image are common when dealing with video databases.

In this demo paper we present LIvRE, a novel open source tool capable of indexing video files and retrieving moments based on visual content. LIvRE is an extension of the Lucene Image Retrieval Engine (LIRE) [2], a CBIR library developed in Java, and later extended to Solr [3]. LIRE is the backbone of LIvRE and provides tools to cover the three main components of our CBVR: parsing, indexing and retrieval. LIRE has been extended and adapted for video indexing, and wrapped with a responsive user interface accessible from both mobile and desktop devices, with an embedded video player to interact with the search results. Figure 1 presents the basic functionality of LIvRE. A query image is uploaded or referred with its URL to the web-based LIvRE interface. Results are displayed adapted to the user display, presenting on the upper part a video player to inspect the results, while showing in the bottom part of the interface thumbnails of the video segments that match the query. LIvRE is available online [1] under GPLv3 license, along with the tools and documentation for its deployment and setup as well as a working demo.

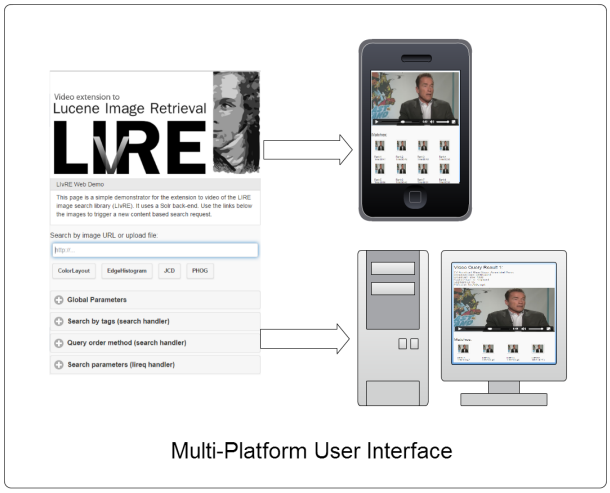

Fig. 1. Overview of the video retrieval procedure and presentation of results.

This paper is structured as follows. Section II overviews other related video search engines. Section III presents how the LIRE image-oriented system was extended to index videos. Section IV tests the performance of LIvRE on the Stanford I2V Light dataset. Finally, Section V contains the conclusions of our work.

II. Related Work

There exist numerous tools that address the challenge of content-based video retrieval with a large variety of approaches and techniques [4].

Content-based video retrieval finds its origins in works such as QBIC [5] or VideoQ [6], which allowed retrieving videos based on the visual similarity of its content when compared with a visual query. Another early work by Jain et al [7] provided support for video queries by subsampling them and comparing the extracted frames with those from the indexed videos.

[1] http://nospotfer.github.io/livre/

978-1-4673-8695-1/16/$31.00

The problem of video retrieval has been addressed in many cases by enriching the low level descriptors extracted from the content with more semantic tags provided by automatic concept detectors. This, combined with contextual metadata associated to the video clips, has resulted in several systems addressing video retrieval as a multimodal. IBM has explored different venues in the field of video retrieval with its iMARS system [8], both based on visual features or on automatic concept detectors. MediaMill [9] was a search engine supporting a semantic indexing process based on a lexicon of 491 concept detectors. CuZero [10] was a web-based system with an intuitive user interface, where a square grid allows formulating multimodal queries and exploring the search results in an intuitive way. Another multimodal contribution was the AXES PRO video search system [11], which allowed the user to perform text-based, concept-based or image-based searches and to refine the results using content from Google Images

More recently, some efforts on video retrieval have addressed the hyperlinking challenge. VIREO-VH [12] is an open source video search system that builds connections between videos to facilitate retrieval.

Latest works have focus on providing video search capabilities from mobile devices. This is the case of Liu et al [13], who built a video search engine to be queried with a video sequence recorded from a mobile device. A lightweight signature for the query is generated at the mobile device and progressively transmitted to a progressive search engine on the cloud.

III. System Architecture

LIRE, the base system LIvRE is built upon, provides a lot of common retrieval functionality including parallel indexing, local feature aggregation, hashing, as well as approximate and linear search. LIRE supports multiple global and local features out of the box, to allow for easy comparison of new features to existing and well-known ones. Most notable global ones for this work are the Joint Color Descriptor (JCD), the Pyramid Histogram of Oriented Gradients (PHOG) , and the MPEG-7 descriptors Edge Histogram, Color Layout and Scalable Color.

Local features support in LIRE includes the bag of visual words approach as well as VLAD aggregation of local features for image retrieval. Local features are based on the OpenCV implementations of SIFT and SURF . In addition to that LIRE fully implements the novel SIMPLE [14] descriptor to using global features on local image patches with configurable key point detectors.

For indexing LIRE builds on the Lucene text search engine and supports linear search in RAM and from Lucene indexes, locality sensitive hashing with an implementation of bit sampling.

A. Overall Architecture

LIvRE is has three main components. The first one, the video parsing component, performs the first step by taking a video dataset as input and extracting keyframes as well as documents containing all the image features from the keyframes

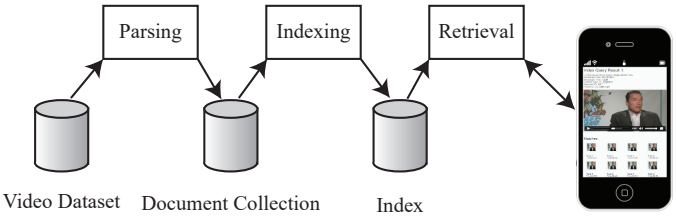

Fig. 2. Main building blocks of the LIvRE CBVR system with: Parsing, Indexing and Retrieval. For the index an Apache Solr server is employed.

for indexing. The second one, the indexing component, then uploads and indexes the documents resulting from parsing. It handles Solr, the search engine, and organizes video segments and keyframes to be shown in result lists. The third and final one, the retrieval component, is integrated with the web-based user interface to query the Solr engine and to present the results to the user in a responsive web application, working on modern mobile as well as desktop browsers. Figure 2 gives an overview on the three main components.

B. The Dataset

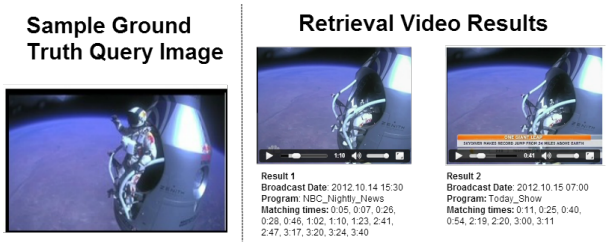

Fig. 3. Query sample from provided ground truth with respective video retrieval results along with the video, matching times and video metadata.

To evaluate LIvRE's performance we used the Stanford I2V Light Dataset [15].The dataset is composed of 23,437 video clips with an average duration of 2.65 min. each, resulting in 1,035 hours of video and 3,808,760 keyframes when using one frame per second for indexing. Each video clip on the dataset corresponds to a single news story, collected from October 2012 until January 2013 from 39 different recurring newscasts in 25 different channels, segmented from a full-length newscast. These story clips are assembled from a coherent collection of successive shots, which cover a single event. Hence, each story clip usually contains tens of shots. These story clips, in the context of news videos, are the equivalent of scenes for general-purpose videos.

The dataset used is accompanied by a set of queries with ground truth annotations, as well as the code and metrics for evaluating results with the ground truth quantitatively. Image queries are collected from news websites and usually depict important events. The ground truth provided consists of 79 topics (queries). A sample query image from the ground truth and retrieval video results from the dataset are shown on Figure 3. For each query image a list of all database clips

matching the topic is provided along with a list of all relevant segments within the clips.

IV. Experimental Results

In order to mimic a broad set of applications, we divide the experiments performed with the dataset in two stages, reflecting the two levels of annotation granularity using the ground truth data and evaluation metrics provided by the Stanford I2V Dataset [15]. The metrics proposed for the data set are mean average precision (map), mean precision at rank 1 (mp@1) and the overlap accurracy using the Jaccard coefficient, measuring how much of the retrieved temporal segment overlaps with the relevant segment. At the first stage, the *scene retrieval*, we evaluate the ranked list of the most likely story clips to contain the query image returned by the system. At the second step, the *temporal refinement*, we evaluate the precision of the specific temporal segments within the clip that contain the query image with a 1 second tolerance. In Table IV we present the evaluation results using features that are known to work well, provided as a baseline. Run time performance results provided by the Solr core, running on a single mid-range server, for 136 non-cached image requests over a population of 3,808,760 indexed video frames report an average time per request of 19,5 seconds and a median request time of 16.3 seconds. Note at this point that the system at hand uses off-the shelf parameters for hashing based search as well as image features and has not been tuned in any way to provide a base line.

Descriptor	map	mp@1	mJaccard
Edge Histogram [16]	0.154	0.372	0.061
JCD [17]	0.174	0.384	0.092
PHOG [18]	0.223	0.450	0.084

TABLE I

Mean average precision (map) and mean precision at 1 (mp@1) obtained at first stage and the mean Jaccard (mJaccard) index obtained at second stage for each of the used descriptors. This table shows evaluation results with off-the-shelve features, provided as a baseline.

V. Conclusions

This paper presents a system extending the LIRE CBIR system to a CBVR system for retrieving shots within large video datasets called LIvRE. The LIvRE system, provided under GPLv3, is the integration of a collection of open-source applications, tools, plugins and modules aimed to work together as an efficient CBVR system. The modular and flexible nature of LIvRE makes it specially interesting to include new tools and descriptors. The scalability of Solr allows for live-indexing and retrieval for a wide range of application possibilities.

Adaptations were done in three main components focusing on the aspects of parsing, indexing and retrieval. Besides the implementation we presented an evaluation using a large video dataset with more than 1000 hours of video. Using the ground truth, we evaluated the performance of the system with the metrics proposed for the dataset for different common image descriptors.

Acknowledgments

This work has been developed in the framework of the project BigGraph TEC2013-43935-R, funded by the Spanish Ministerio de Economia y Competitividad and the European Regional Development Fund (ERDF). The Image Processing Group at the UPC is a SGR14 Consolidated Research Group recognized and sponsored by the Government of Catalonia.

References

[1] S. W. Smoliar and H. Zhang, "Content-based video indexing and retrieval," *IEEE multimedia*, no. 2, pp. 62–72, 1994.

[2] M. Lux, "Lire: Open source image retrieval in java," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 843–846.

[3] M. Lux and G. Macstravic, "The lire request handler: A solr plug-in for large scale content based image retrieval," in *MultiMedia Modeling*, C. Gurrin, F. Hopfgartner, W. Hurst, H. Johansen, H. Lee, and N. OConnor, Eds. Springer, 2014, pp. 374–377.

[4] O. Marques and B. Furht, *Content-based image and video retrieval*. Springer Science & Business Media, 2002, vol. 21.

[5] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic *et al.*, "Query by image and video content: The qbic system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[6] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "Videoq: an automated content based video search system using visual cues," in *Proceedings of the fifth ACM international conference on Multimedia*. ACM, 1997, pp. 313–324.

[7] A. K. Jain, A. Vailaya, and X. Wei, "Query by video clip," *Multimedia systems*, vol. 7, no. 5, pp. 369–384, 1999.

[8] A. Natsev, J. R. Smith, J. Tešić, L. Xie, and R. Yan, "Ibm multimedia analysis and retrieval system," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*. ACM, 2008, pp. 553–554.

[9] M. Worring, C. G. M. Snoek, O. De Rooij, G. P. Nguyen, and A. W. M. Smeulders, "The mediamill semantic video search engine," *IEEE International Conference on Acoustics Speech and Signal Processing ICASSP 2007*, vol. 4, pp. IV–1213 – IV–1216, 2007. [Online]. Available: http://staff.science.uva.nl/∼cgmsnoek/pub/worring-mediamill-icassp2007.pdf

[10] E. Zavesky and S.-F. Chang, "Cuzero: embracing the frontier of interactive visual search for informed users," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 2008, pp. 237–244.

[11] K. McGuinness, N. E. O'Connor, R. Aly, F. De Jong, K. Chatfield, O. M. Parkhi, R. Arandjelovic, A. Zisserman, M. Douze, and C. Schmid, "The axes pro video search system," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 307–308.

[12] S. Tan, C.-W. Ngo, H.-K. Tan, and L. Pang, "Cross media hyperlinking for search topic browsing," *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, p. 243, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2072298.2072331

[13] W. Liu, T. Mei, and Y. Zhang, "Instant mobile video search with layered audio-video indexing and progressive transmission," *Multimedia, IEEE Transactions on*, vol. 16, no. 8, pp. 2242–2255, 2014.

[14] C. Iakovidou, N. Anagnostopoulos, A. C. Kapoutsis, Y. Boutalis, and S. A. Chatzichristofis, "Searching images with mpeg-7 (and mpeg-7-like) powered localized descriptors: the simple answer to effective content based image retrieval," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014, pp. 1–6.

[15] A. Araujo, J. Chaves, D. Chen, R. Angst, and B. Girod, "Stanford I2V: A News Video Dataset for Query-by-Image Experiments," in *Proc. ACM Multimedia Systems*, 2015.

[16] S.-F. Chang, T. Sikora, and A. Purl, "Overview of the mpeg-7 standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 688–695, 2001.

[17] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux, "Selection of the proper compact composite descriptor for improving content based image retrieval," in *The Sixth IASTED International Conference on Signal Processing, Pattern Recognition and Applications SPPRA 2009*, 2009.

[18] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, ser. CIVR '07. New York, NY, USA: ACM, 2007, pp. 401–408. [Online]. Available: http://doi.acm.org/10.1145/1282280.1282340