Inter and Intra-Video Navigation and Retrieval in Mobile Terminals

Andrei Bursuc, Titus Zaharia
Institut Télécom; Télécom SudParis; ARTEMIS Dept.
UMR CNRS 8145 MAP 5
9, rue Charles Fourier, 91011 Evry Cedex
{Andrei.Bursuc, Titus.Zaharia}@it-sudparis.eu

Françoise Prêteux
Mines ParisTech
60, Boulevard Saint-Michel 75272 Paris Cedex, France

Francoise.Preteux@mines-paristech.fr

Abstract—**This paper introduces a novel on-line video browsing and retrieval platform, so-called OVIDIUS (*On-line VIDeo Indexing Universal System*). In contrast with traditional and commercial main stream video retrieval platforms, where video content is treated in a more or less monolithic manner (i.e., with global descriptions associated with the whole document), the proposed approach makes it possible to browse and access video content in a finer, per-segment basis. The hierarchical metadata structure exploits the MPEG-7 approach for structural description of video content. The MPEG-7 description schemes have been here enriched with both semantic and content-based metadata. The developed approach shows all its pertinence within a multi-terminal context and in particular for video access from mobile devices. The platform has been recently (February, 2010) validated within the framework of the Médi@TIC French national project.**

Keywords - video indexing, video search engines, user interfaces, MPEG-7 standard, visual descriptors, description schemes.

I. Introduction

Over the last ten years, mobile devices have been undergoing a booming prosperity in our everyday life. A broader variety of handheld devices with audio/video playback functionalities are available in the market with a reasonable price. The huge steps forward made by third generation communication networks enable telecom operators to provide better mobile multimedia services, such as smoother streaming time and higher quality of video resolution.

A. Context and objectives

Currently we are witnessing a proliferation of powerful mobile phones capable of fast connections and high-fidelity multimedia rendering; the so-called “smartphones”. According to the consultancy company Nielsen, the amount of smartphone subscribers in the United States has augmented from 14% at the end of 2008 to 29% at the beginning of 2010. The same study predicts that by the end of 2011 in the U.S. there will be more smartphones than feature phones [1].

Meanwhile, average mobile users seem to take more advantage of the new features provided by the phone, such as the faster internet connection. Thus, the use of Wi-Fi increases 10 times from 5% for feature phone owners to 50% for smartphone users [1]. A recent study made by Nielsen illustrates that in the U.S. active mobile video users grew by 57% from the fourth quarter of 2008 to the fourth quarter of 2009, from 11.2 million to 17.6 million [2]. This means that mobile subscribers are also developing a growing appetite for online video content.

However, the functionalities of such devices are constrained by their size and computational/memory capacities. Existing studies have put an emphasis around the consumer and his behavior, aiming to build user-driven interfaces and applications [3]. In this context, an important technological challenge concerns the issue of accessing/retrieving video content from mobile devices. The critical point relates to the high complexity of video content, in terms of amount of heterogeneous information included. In order to tackle the issue of complexity, appropriated presentation and search engines, as well as novel interaction modalities need to be developed. In addition, it is necessary to ensure a personalized access to *segments* of interest, defined as parts of an audio-visual document. User should have the possibility of rapidly browsing the content and identify/access the segments of interest.

Let us analyze how such aspects are treated in the state of the art.

B. Related work

Our work draws upon research in several areas concerning multimedia content management: content-based image and video retrieval, multimedia content management, user feedback management and distributed content sharing.

One of the most active areas over the last decade has been the content-based image retrieval area. Within this context, let us first mention the Multimodal Automatic Mobile Indexing (MAMI) system, which allows users to annotate and search for digital photos via speech input combined with time, date and location information [4]. Jesus *et al.* used geographical queries to retrieve personal pictures when visiting points of interest [5]. Zhu *et al.* integrate the features mentioned above in their user-centric system, called iScope [6]. iScope uses multi-modality clustering of both content and context information for efficient image management and search, and online learning techniques for predicting images of interest, while supporting distributed content-based search among networked devices.

Kim *et al.* have used CBIR for visual-content recommendation [7] while Yeh *et al.* have used mobile images for content-based object search [8]. CLOVER searches sketches or photos of leaf images on a server starting from a mobile phone [9]. Photo-to-Search performs queries directly on the web using images taken with a mobile device [10].

The issue of incompatibility between video resolution and certain mobile phones is addressed by the researchers from Zhejiang University [11]. User feedback is used in order to update the metadata for video clips, including the acceptable resolution. In addition, redundant versions with lower resolution are generated for video clips by estimating their popularities, and sent to users with lower resolution directly to save the downsizing transcoding process.

The Multimedia Content Creation Platform (MMCP) [12] takes full advantage of the context information provided by the mobile device each time a new picture or video is created and added immediately as a metadata (date, time, location, etc.). The platform can be used both for generating new contact and retrieve content as well.

The system proposed in [13] allows users to retrieve video content starting from a mobile photo uploaded by the user. The search engine in the background returns videos containing key frames that are similar with the uploaded picture. An automatic key frame extractor and the Contrast Context Histogram (CCH) [14] have been used for the elaboration of the system.

Let us also mention the innovative approach proposed by Miller *et al.* Their mobile media content browser has been developed and placed on an iPhone device. MiniDiver [15] is based on four user interfaces serving mobile context sensitive video. After selecting the desired video or program via a simple interface, the user can view the content from one or two camera perspectives simultaneously. In the case of live broadcasts, users can have the moving objects (*e.g.,* hockey players) highlighted and can choose the favorite camera angles.

The existing approaches offer interesting preliminary solutions to the issue of mobile video access. However, most of them are basically mobile versions of their desktop-based counterpart and hence, do not take into account the specificity of mobile devices, environments and usages/services. In addition, they massively focus on textual queries, while an efficient search process should consider rich and multimodal queries, combining text, image, audio and video features.

Another drawback of such approaches comes from the fact that videos are considered in a monolithic manner, without taking into account the intrinsic spatio-temporal structure of the video content. However a common video document may include a huge amount of heterogeneous information that needs to be identified, described and accessed independently.

Creating dedicated tools for query formulation, metadata driven visualization / navigation and ergonomic user interaction in both fixed and mobile environments is still a challenge that needs to be addressed and solved.

C. Contributions

The analysis of the state of the art shows that currently no system fully exploits the intrinsic spatio-temporal structure of video documents. The OVIDIUS (*On-line VIDeo Indexing Universal System*) platform proposed in this paper aims at solving such limitations, ensuring all the interaction and navigation capabilities needed to access video content at a fine level of granularity, from fixed or mobile devices.

The strong points of the proposed system are the following :

- Modular and distributed architecture, achieved with the help of web services, which makes it possible to easily upgrade the system in order to keep pace with inherent future technological advances,
- Fine granularity access to video content, based on the MPEG-7 structural approach for video content description [16]
- Core interoperability achieved with open MPEG-7 standard technologies,
- Enrichment of MPEG-7 structural description schemes with semantic and content-based descriptors,
- Advanced interaction functionalities integrating browsing, search, hierarchical navigation and visualization capabilities,
- Support of both textual, content-based and hybrid queries,
- Compatibility with a vast variety of platforms.

The rest of the paper is organized as follows. Section 2 presents the description framework that we have adopted for our platform. Section 3 gives an overview of the proposed OVIDIUS system and presents the audio-video analysis techniques that have been considered in the metadata extraction engine. Finally, Section 4 concludes the paper and opens perspectives of future work.

II. MPEG-7 STRUCTURAL VIDEO CONTENT DESCRIPTION

The MPEG-7 structural approach for video description [16] is based on an abstract class (description scheme) of *Segment*. An MPEG-7 Segment represents an arbitrary part of a video and includes generic descriptors (*e.g.*, textual annotations, keywords, temporal localization elements for specifying the starting and the ending time stamps of a segment, etc.). Starting from this abstract structure, which cannot be instantiated, a set of media-specific segments is derived, by applying an inheritance mechanism. In our implementation, we have considered the following MPEG-7 segments: AudioVisualSegment DS, AudioSegment DS, Video DS, StillRegionDS, and MovingRegion DS.

Each segment can include dedicated descriptors, adapted to each segment type. Examples are color features for still region/video segment, audio features for audio segment, motion parameters for MovingRegion DS, etc.

A second MPEG-7 mechanism exploited is the Segment Decompostion DS, which allows the partition of a segment into sub-segments. Applied recursively, this mechanism makes it possible to represent a video as a hierarchical tree structure made of scenes, shots, transcriptions segments, and key-images.

The adopted MPEG-7 language for specifying all these descriptors and descriptions schemes is XML schema. This choice facilitates the parsing and interpretation of the descriptions, since various XML utilities are available and can be directly used (*e.g.*, Xerces parser).

Disposing of metadata is a first and essential step in the indexing process. However, appropriate video visualization and interaction capabilities need to be elaborated in order to efficiently exploit such a description. The OVIDIUS user interface integrates all the necessary interaction and navigation capabilities, as described in the following section.

III. OVIDIUS PLATFORM: SYSTEM OVERVIEW

Figure 1 illustrates the distributed architecture adopted in the OVIDIUS platform. It includes a content management module (*i.e.*, storage and editing of content and metadata), a metadata extraction engine, a MPEG-7 search engine and a web interface which can be remotely accessed from mobile and fixed environments.

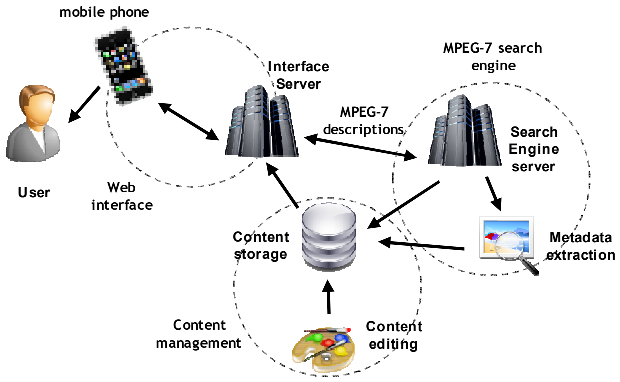


Figure 1. Overview of the OVIDIUS platform.

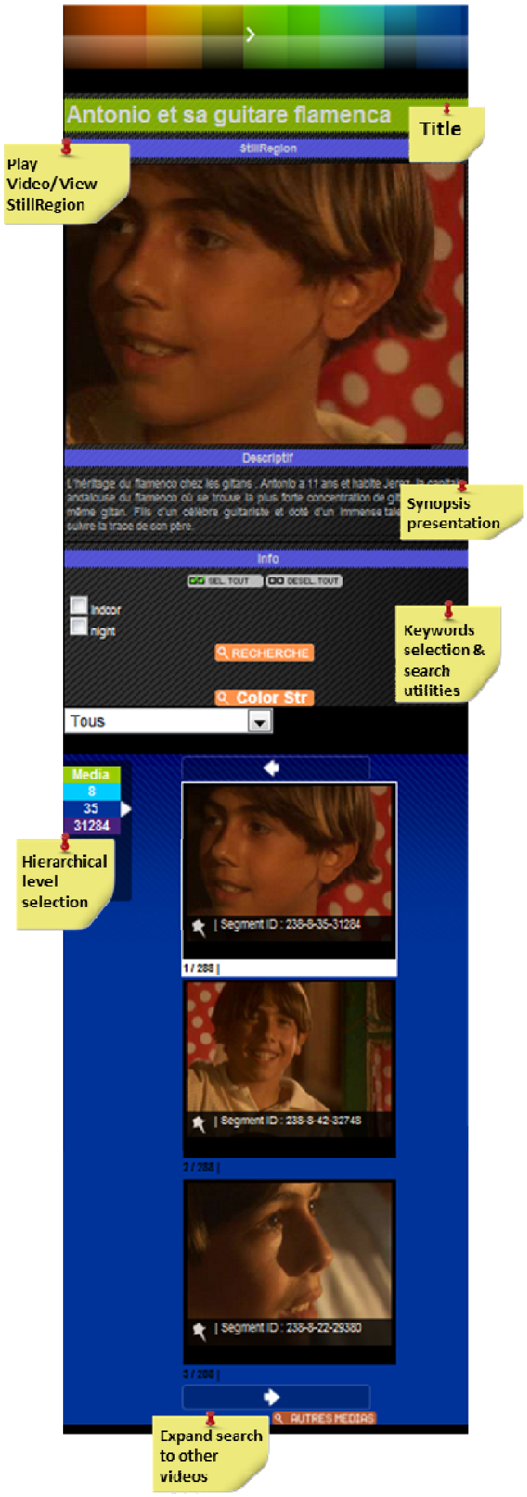
All the components of the system can be distributed on various servers. The communication between them is performed by using web services with http requests. Such a modular approach facilitates the future extension of the platform as well as the replacement of individual components (*e.g.*, video players, search engine…).

A. OVIDIUS user interface

In order to ensure a large interoperability, the interface has been developed by using HTML, PhP and JavaScript technologies. We have successfully tested our system on iPhone 3G and 3GS devices and Android smartphones as well. Due to its construction, OVIDIUS can be accessed from other types of smartphones with a mobile internet browser JavaScript compatible (*e.g*. Android phones). The interface server automatically detects the operating system of the user terminal and adapts accordingly. The sole aspect that should be taken into consideration when switching to other types of terminals is the issue of video encoding formats.

Figure 2 illustrates the OVIDIUS user interface. The following components are included: selector of the segment type, selector of the hierarchical level of each segment, iconic representation of segments (which are dynamically derived from the media), navigation/browsing buttons, summary and keyword visualization and selection.

ISBN: 978-1-61208-100-7


Antonio et sa guitare flamenca
Title
Play Video/View StillRegion
StillRegion
Descriptif
Synopsis presentation
Info
SEL. TOUT
DESEL. TOUT
□ indoor
□ night
RECHERCHE
Color Str
Keywords selection & search utilities
Tous
Media
8
35
31284
Hierarchical level selection
Segment ID : 238-8-35-31284
1 / 288
Segment ID : 238-8-42-32748
2 / 288
Segment ID : 238-8-22-29380
3 / 288
AUTRES MEDIAS
Expand search to other videos

L'héritage du flamenco chez les gitans . Antonio a 11 ans et habite Jerez, la capitale andalouse du flamenco où se trouve la plus forte concentration de gi même gitan. Fils d'un célèbre guitariste et doté d'un immense tal suivre la trace de son père.

Figure 2. Components of the OVIDIUS user interface and results for a query by Color Structure descriptor.

One column of iconic preview representations is dedicated to each level of hierarchy of the video. Scenes, shots, and still regions can be browsed and accessed in a hierarchical manner, as MPEG-7 segments. Thus, the user can navigate inside a video both horizontally (through video segments on the same level) and vertically (through different levels of the scene hierarchical tree structure).

The iconic images are dynamically generated at each access with the use of a Java based frame extractor and the MPEG-7 time stamps. Each iconic image provides information about the visual content of the segment (preview image), as well as information regarding the classification and hierarchy (type of segment, segment identifier).

OVIDIUS can exploit arbitrary extraction methods and utilities, provided that the representation of the description is compliant with the MPEG-7 specification.

A specific feature has been added envisioning time and bandwidth efficiency. By taking advantage of the FFMpeg library [17], OVIDIUS can automatically cut a segment from the video at user request. Whenever a user wants to access a certain video segment, this video segment will be automatically cut from the media and sent to user. This way the user will receive the exact requested video segment.

Thanks to the web implementation of the OVIDIUS platform, a PC version is also available, offering the same navigation functionalities, features, speed and user experience (Figure 3).

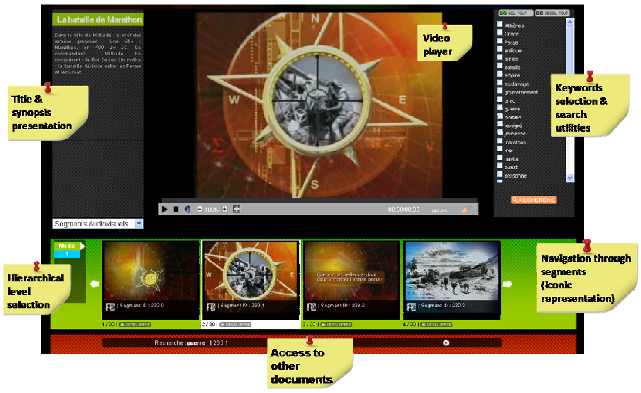

La bataille de Marathon
Video player
Title & synopsis presentation
Keywords selection & search utilities
Segments Audiovisuels
Media
Hierarchical level selection
Navigation through segments (iconic representation)
Access to other documents

Figure 3. PC version of the OVIDIUS user interface.

B. *Metadata extraction engine*

In order to instantiate the MPEG-7 structural video description, we have implemented a video segmentation on three levels, including scenes, shots and key frames.

The first step in the segmentation process was to develop an efficient shot boundary detector. Starting from the techniques and results presented in [18], we have developed a combined two levels method able to detect both CUTs and gradual transitions. Abrupt CUTs are tackled at the first level of detection, based on similarity of color histograms.

In the case where no CUT is detected and the similarity score is above a certain threshold on a group of 10 frames, a second level of detection is applied. Here, we use the graph partition model proposed in [19]. This second step allows the detection of gradual transitions.

For key-frame extraction we have used an adaptation of the algorithm developed by Zhuang et al. [20] which detects multiple frames based on the visual variations in shots.

Finally, the approach described in [21] has been adopted in order to aggregate shots into scenes. The principle consists of constructing a weighted undirected graph so-called shot similarity graph, which involves a similarity measure that combines color and motion features.

The detected elements are represented using MPEG-7 segments (*cf.* §II). To each segment, appropriate descriptors are associated with. For this purpose, the entire set of MPEG-7 visual descriptors [22], related to color, texture and motion features is currently supported, based on an optimized version of the MPEG-7 reference software.

Currently we have tested the performances of the MPEG-7 Color Structure Descriptor in retrieving visually similar video sequences. The experiments have been carried out on the Médi@TIC corpus, kindly provided by LBA (Vodeo.TV) which includes about 15 documentary videos. As shown in Figure 4 the Color Structure Descriptor is very effective in retrieving instances of the same person and same environment within the video corpus.

Textual information which corresponds to the transcription of the audio track available for the Médi@TIC corpus is also included as semantic information.

For all the descriptors considered, dedicated XML schema representations have been elaborated and used to enrich the MPEG-7 schema definition.

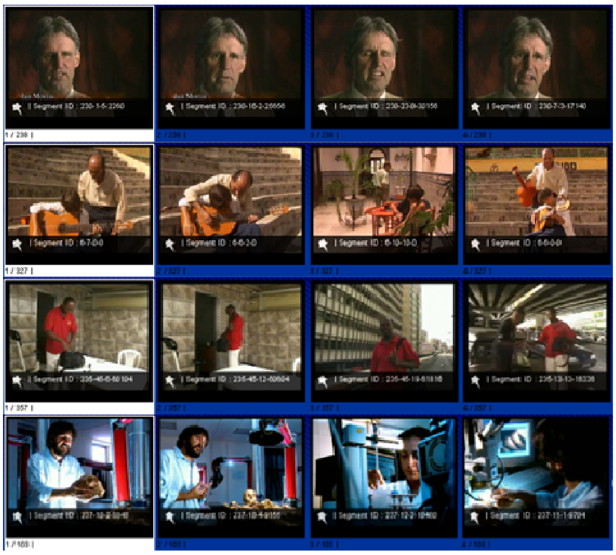


Figure 4. Search results based on the MPEG-7 ColorStructure descriptor with queries by example.

C. *Main functionalities*

Let us summarize the main functionalities available for the OVIDIUS platform:

- browsing of a database of videos with keyword search
- selection of a video and navigation through the hierarchical structure of scenes, shots and key-images, in order to quickly discover the contents
- query formulation,
- search engine capabilities: search by visual features,
- identification of features of interest and search of segments in the whole video database.

Concerning the query formulation, OVIDIUS provides a simple keyword panel, displaying information extracted from the soundtrack of the video. Keyword of interest can be selected through checkboxes and searched within the video segments on the same level. Query results from the same video or from other videos in the data base can be browsed and accessed.

In the near future, we plan to extract also keywords describing the visual content of the video sequence (*e.g.*, indoor, outdoor, night, day, etc.).

IV. CONCLUSIONS AND PERSPECTIVES

In this paper, we have presented the OVIDIUS video indexing platform, which supports hierarchical and multigranular MPEG-7 descriptions while integrating all the necessary interaction, browsing and visualization utilities. The proposed approach makes it possible to discover a video document in a few clicks, and can be accessed from various platforms (*e.g.*, iPhone or PC).

The perspectives of future work concern the extension of the system with new, advanced descriptors and extraction engines. Notably, we will focus on the elaboration of an object detection/recognition/retrieval framework, able to identify and describe user-specific elements of interest.

ACKNOWLEDGMENT

This work has been partially supported by the French national project Medi@TIC, double labeled by the Systematic and CapDigital competitiveness clusters. The OVIDIUS platform has been developed by Institut TELECOM.

REFERENCES

[1] http://blog.nielsen.com/nielsenwire/consumer/smartphones-to-overtake-feature-phones-in-u-s-by-2011/, last accessed April 2010

[2] http://blog.nielsen.com/nielsenwire/online_mobile/three-screen-report-q409/, last accessed April 2010

[3] L. Wang, D. Tjondrongoro, and Y. Liu, "Clustering and visualizing audiovisual dataset on mobile devices in a topic-oriented manner," *Proceedings of the 9th international conference on Advances in visual information systems*, Shanghai, China: Springer-Verlag, 2007, pp. 310-321.

[4] X. Anguera, J. Xu, and N. Oliver, "Multimodal photo annotation and retrieval on a mobile phone," *Proceeding of the 1st ACM international conference on Multimedia information retrieval*, Vancouver, British Columbia, Canada: ACM, 2008, pp. 188-194.

[5] R. Jesus, R. Dias, R. Frias, and N. Correia, "Geographic image retrieval in mobile guides," *Proceedings of the 4th ACM workshop on Geographical information retrieval*, Lisbon, Portugal: ACM, 2007, pp. 37-38.

[6] C. Zhu, K. Li, Q. Lv, L. Shang, and R.P. Dick, "iScope: personalized multi-modality image search for mobile devices," *Proceedings of the 7th international conference on Mobile systems, applications, and services*, Kraków, Poland: ACM, 2009, pp. 277-290.

[7] C. Y. Kim, et al. , "VISCORS: A visual-content recommender for the mobile web," IEEE Intelligent Systems, vol. 19, no. 6, pp. 32-39, 2004.

[8] T. Yeh, K. Grauman, K. Tollmar, and T. Darrell, "A picture is worth a thousand keywords: image-based object search on a mobile platform," *CHI '05 extended abstracts on Human factors in computing systems*, Portland, OR, USA: ACM, 2005, pp. 2025-2028.

[9] S. Kim, Y. Tak, Y. Nam, and E. Hwang, "mCLOVER: mobile content-based leaf image retrieval system," *Proceedings of the 13th annual ACM international conference on Multimedia*, Hilton, Singapore: ACM, 2005, pp. 215-216.

[10] M. Jia, et al., "Photo-to-Search: Using camera phones to inquire of the surrounding world," *in MDM '06: Proceedings of the 7th International Conference on Mobile Data Management,* 2006, pp. 46-46.

[11] Y. Liu, Z. Yang, X. Deng, J. Bu, and C. Chen, "Media Browsing for Mobile Devices Based on Resolution Adaptive Recommendation," *Proceedings of the 2009 WRI International Conference on Communications and Mobile Computing - Volume 03*, IEEE Computer Society, 2009, pp. 285-290.

[12] S. Järvinen, J. Peltola, J. Lahti, and A. Sachinopoulou, "Multimedia service creation platform for mobile experience sharing," *Proceedings of the 8th International Conference on Mobile and Ubiquitous Multimedia*, Cambridge, United Kingdom: ACM, 2009, pp. 1-9.

[13] C. Chen, Y. Wang, H. Wang, and C. Chiu, "Digital Video Retrieval via Mobile Devices," *Proceedings of the 2008 Fourth IEEE International Conference on eScience*, IEEE Computer Society, 2008, pp. 376-377.

[14] C. Huang, C. Chen, and P. Chung, "Contrast Context Histogram - A Discriminating Local Descriptor for Image Matching," *Proceedings of the 18th International Conference on Pattern Recognition - Volume 04*, IEEE Computer Society, 2006, pp. 53-56.

[15] G. Miller, S. Fels, M. Finke, W. Motz, W. Eagleston, and C. Eagleston, "MiniDiver: A Novel Mobile Media Playback Interface for Rich Video Content on an iPhone^{TM}," *Proceedings of the 8th International Conference on Entertainment Computing*, Paris, France: Springer-Verlag, 2009, pp. 98-109.

[16] ISO/ IEC 15938-5:2003, Information technology - MultimediaContent Description. Interface - Part 5: Multimedia Description Schemes. 2003.

[17] http://ffmpeg.org/

[18] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B.Zhang, "A formal study of shot boundary detection," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 2, pp. 168-186, Feb. 2007.

[19] J. Yuan, B. Zhang, and F. Lin, "Graph Partition Model for Robust Temporal Data Segmentation," *Advances in Knowledge Discovery and Data Mining*, 2005, pp. 758-763.

[20] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proceedings 1998 International Conference on Image Processing. ICIP98* (Cat. No.98CB36269), Chicago, IL, USA: , pp. 866-870.

[21] Z. Rasheed, M. Shah, Detection and Representation of Scenes in Videos, IEEE Trans. on Multimedia, 7(6): 1097-1105, Dec. 2005.

[22] ISO/ IEC 15938-3:2003, Information technology - MultimediaContent Description. Interface-Part 3: Visual. 2003.