



ELSEVIER

Available at

www.ElsevierComputerScience.com

POWERED BY SCIENCE @ DIRECT®

Computer Vision
and Image
Understanding

Computer Vision and Image Understanding xxx (2003) xxx–xxx

www.elsevier.com/locate/cviu

Editorial Introduction

Video retrieval and summarization

This year, it is anticipated that 25% of the population of the wealthy countries will have a digital television camera at their disposal. The combined capacity to generate bits from these devices is astronomical. In addition, the growth in computer speed, disc capacity, and, most of all, the rapid growth of the Internet and WWW will make this information accessible worldwide.

The immediate question is what to do with all the information. One could store the digital video information on tapes, CD-ROMs, DVDs, or any such device but the level of access would be less than the well-known shoe boxes filled with tapes, old photographs, and letters. We need to ensure that the techniques for organizing video stay in tune with the tremendous amounts of information. So, with video on demand about to arrive, there is an urgent need for effective video retrieval and summarization methods.

Creating access to still images had appeared to be a hard problem. It requires hard work, precise modeling, the inclusion of considerable amounts of a priori knowledge, and solid experimentation to analyze the contents of a photograph. Even though video tends to be much larger than images, it can be argued that the access to video is a simpler problem than access to still images. First of all, video comes in color and color provides easy clues to object geometry, position of the light, and identification of objects by pixel patterns, only at the expense of having to handle three times more data than black and white. And, video comes as a sequence, so what moves together most likely forms an entity in real life, so segmentation of video is intrinsically simpler than of a still image, again at the expense of only more data to handle.

That does not mean progress will come for free. Moving from images to video adds several orders of complexity to the retrieval problem due to indexing, analysis, and browsing over the inherently temporal aspect of video. For example, the user can pose a similarity based query of “Find a video scene similar to this one.” Responding to such a query requires representations of the image and of the temporal aspects of the video scene. Furthermore, higher level representations which reflect the structure of the constituent video shots or semantic temporal information such as gestures could also aid in retrieving the right video scene.

A consequence of the growing consumer demand for visual information is that sophisticated technology is needed for representing, modeling, indexing, and retrieving multimedia data. In particular, we need robust techniques to index/retrieve and compress visual information, new scalable browsing algorithms allowing access to

very large databases of images and videos, and semantic visual interfaces integrating the above components into a single concept of content-based video indexing and retrieval (CBVIR) systems.

1. Content-based video indexing and retrieval

There are four main processes involved in content-based video indexing and retrieval [2,4,7]: video content analysis, video structure parsing, summarization or abstraction, and indexing. Each process poses many challenges. In what follows, we briefly review these challenging research issues.

1.1. Video content analysis

The main problem in video content analysis is that we cannot easily map extractable visual features (such as color, texture, shape, structure, layout, and motion) into semantic concepts (such as indoor and outdoor, people, or car-racing scenes). Although visual content is a major source of information in a video, valuable information is also carried in other media components, such as text (superimposed on the images, or included as closed captions), audio, and speech that accompany the pictorial component. A combined and cooperative analysis of these components would be far more effective in characterizing the video for both consumer and professional applications. The Informedia system [6] or AT&T's Pictorial Transcripts [13] system are examples of such approaches.

1.2. Video structure parsing

An important step in the process of video structure parsing is that of segmenting the video into individual scenes. From a narrative point of view, a scene consists of a series of consecutive shots grouped together because they were filmed in the same location or because they share some thematic content. The process of detecting these video scenes is analogous to paragraphing in text document parsing, but it requires a higher level of content analysis. In contrast, shots are actual physical basic layers in video, whose boundaries are determined by editing points or where the camera switches on or off. Fortunately, analogous to words or sentences in text documents, shots are a good choice as the basic unit for video content indexing, and they provide the basis for constructing a table of contents for video. Shot boundary detection algorithms that rely only on visual information contained in the video frames can segment the video into frames with similar visual contents [5]. Grouping the shots into semantically meaningful segments such as stories, however, usually is not possible without incorporating information from the other components of the video. Multimodal processing algorithms involving the processing of not only the video frames, but also the text, audio, and speech components that accompany them have proven effective in achieving this goal [7].

1.3. Video summarization

Video summarization is the process of creating a presentation of visual information about the structure of video, which should be much shorter than the original video. This abstraction process is similar to extraction of keywords or summaries in text document processing. That is, we need to extract a subset of video data from the original video such as keyframes or highlights as entries for shots, scenes, or stories. Abstraction is especially important given the vast amount of data even for a video of a few minutes' duration. The result forms the basis not only for video content representation but also for content-based video browsing. Combining the structure information extracted from video parsing and the keyframes extracted in video abstraction, we can build a visual table of contents for a video.

1.4. Video indexing

The structural and content attributes found in content analysis, video parsing, and abstraction processes, or the attributes that are entered manually, are often referred to as metadata. Based on these attributes, we can build video indices and the table of contents through, for instance, a clustering process that classifies sequences or shots into different visual categories or an indexing structure. As in many other information systems, we need schemes and tools to use the indices and content metadata to query, search, and browse large video databases. Researchers have developed numerous schemes and tools for video indexing and query. However, robust and effective tools tested by thorough experimental evaluation with large data sets are still lacking. Therefore, in the majority of cases, retrieving or searching video databases by keywords or phrases will be the mode of operation. In some cases, we can retrieve with reasonable performance by content similarity defined by low-level visual features of, for instance, keyframes, and example-based queries.

2. Papers in the special issue

The seven papers selected for the *Video Retrieval and Summarization* special issue of *Computer Vision and Image Understanding* follow roughly the structure presented in the previous section. They all have a video content analysis process in which features are being extracted, followed by the structure analysis for object extraction [8,10], parsing based on the geometry of the scene [1,3], video segment boundary extraction and refinement based on multi-level feature selection [9], or scene segmentation based on local feature extraction [12]. Finally, video abstractions are constructed for scene matching and retrieval [9,12], event detection [1,3], object-based analysis and interpretation [8,10]. One exception from the above structure is the paper by Pickering and R uger [11] which provides a comprehensive survey of key frames based retrieval techniques for video.

- A very interesting approach for matching shots which are images of the same 3D location in a film is taken by Schaffalitzky and Zisserman [12]. This is a difficult problem because the camera viewpoint may change substantially between shots, with consequent changes in imaged appearance of the scene due to foreshortening, scale change, partial occlusion, and lighting changes. They present excellent results of matching shots for a number of very different scene types extracted from two entire commercial films.
- A semantic annotation framework for soccer videos is introduced in the paper by Assfalg et al. [1]. Their goal is to achieve highlights detection such as ball motion, the playfield zone, the position of the players, and color of players' uniforms, which exploit visual cues that are estimated from the video stream. The presented results show that all the principal highlights of a soccer game can be detected with great accuracy.
- The paper by Denman et al. [3] presents several tools for content analysis of snooker videos. The first tool is a new feature for parsing a sequence based on geometry without the need for deriving 3D information. The second tool allows events to be detected where an event is characterized by an object leaving the scene of a particular location. The final tool is a mechanism for summarizing motion in a shot for use in a content-based summary. The authors show that by exploiting context, a convincing summary can be made for snooker footage.
- Luo et al. [10] propose a novel scheme for object-based video analysis and interpretation of human motion in sports video sequences based on automatic video object extraction, video object abstraction, and semantic event modeling. Their experiments show that the object-based approach is effective for a complete characterization of video sequences, which includes both macro-grained and fine-grained semantics contained in the video sequences.
- The paper by Liu et al. [8] presents a novel content-based 3D motion retrieval algorithm. The authors construct a motion index tree based on a hierarchical motion description which serves as a classifier to determine the similar motions to the query sample. The experiments demonstrate the effectiveness of the proposed 3D motion algorithm retrieval algorithm.
- The problem of feature selection for video retrieval is investigated by Liu and Kender [9]. Their system uses machine learning techniques to automatically construct a hierarchy of small subsets of features that are progressively more useful for indexing. This hierarchy allows video segments to be segmented and categorized simultaneously in a coarse-fine manner that efficiently and progressively detects and refines their temporal boundaries. The results are demonstrated for a 75-min instructional video and a 30-min baseball video.
- An extensive evaluation of several keyframe-based retrieval techniques for video is presented by Pickering and R uger [11]. The authors perform a very good systematic analysis and discussion of their results. It is interesting to see from their comparison that k -NN performs so well in most tasks compared to boosting.

3. Concluding remarks

Video analysis and understanding is an emerging research area that has received growing attention in the research community over the past decade. Though modeling and indexing techniques for content-based image indexing and retrieval domain have reached reasonable maturity [14], content-based techniques for video data, particularly those employing spatio-temporal concepts, are at the infancy stage. Content representation through low-level features has been addressed fairly, and there is a growing trend towards bridging the semantic gap. Monomodal approaches have proven successful to a certain level, and more efforts are being put for fusion of multiple media. As visual databases grow bigger with advancements in visual media creation, compaction, and sharing, there is a growing need for storage-efficient and scalable search systems.

Acknowledgments

We thank the referees for their help.

References

- [1] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, W. Nunziati, Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding* 91(3) (2003).
- [2] R. Bole, B. Yeo, M. Yeung, Video query: research directions, *IBM Journal of Research and Development* 42 (2) (1998) 233–252.
- [3] H. Denman, N. Rea, A. Kokaram, Content-based analysis for video from snooker broadcasts, *Computer Vision and Image Understanding* 91(3) (2003).
- [4] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, A. Zakhor, Applications of video-content analysis and retrieval, *IEEE Multimedia* 9 (3) (2002) 42–55.
- [5] A. Hanjalic, Shot-boundary detection: unraveled and resolved?, *IEEE Transactions on Circuits and Systems for Video Technology* 12 (2) (2002) 90–105.
- [6] A. Hauptmann, T.D. Ng, R. Baron, W. Lin, Chen M., M. Derthick, M. Christel, R. Jin, R. Yan, Video classification and retrieval with the Infromedia digital video library system, *Text Retrieval Conference (TREC02)*, 2002.
- [7] M. Lew, N. Sebe, P. Gardner, Video indexing and understanding, in: M. Lew (Ed.), *Principles of Visual Information Retrieval*, Springer, Berlin, 2001, pp. 163–196.
- [8] F. Liu, Y. Zhuang, F. Wu, Y. Pan, 3D motion retrieval with motion index tree, *Computer Vision and Image Understanding* 91(3) (2003).
- [9] Y. Liu, J.R. Kender, Fast video segment retrieval by sort-merge feature selection, boundary refinement, and lazy evaluation, *Computer Vision and Image Understanding* 91(3) (2003).
- [10] Y. Luo, T.-D. Wu, J.-N. Hwang, Object-based analysis and interpretation of human motion in sports video sequences by Dynamic Bayesian Networks, *Computer Vision and Image Understanding* 91(3) (2003).
- [11] M. Pickering, S. R uger, Evaluation of key frame based retrieval techniques for video, *Computer Vision and Image Understanding* 91(3) (2003).
- [12] F. Schaffalitzky, A. Zisserman, Automated location matching in movies, *Computer Vision and Image Understanding* 91(3) (2003).

- [13] B. Shahraray, Multimedia information retrieval using pictorial transcripts, in: B. Furth (Ed.), Handbook of Multimedia Computing, CRC Press, Boca Raton, FL, 1999, pp. 345–359.
- [14] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.

Nicu Sebe

*Faculty of Science, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
E-mail address: nicu@science.uva.nl*

Michael S. Lew

*LIACS Media Lab, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
E-mail address: mlew@liacs.nl*

Arnold W.M. Smeulders

*Faculty of Science, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
E-mail address: smeulders@science.uva.nl*