An Overview of Video Shot Clustering and Summarization Techniques for Mobile Applications

Nicola Adami, Sergio Benini, Riccardo Leonardi and Pierangelo Migliorati
University of Brescia, Department of Electronics for Automation
via Branze 38, 25123, Brescia, Italy
{firstname.lastname}@ing.unibs.it

ABSTRACT

The problem of content characterization of video programmes is of great interest because video appeals to large audiences and its efficient distribution over various networks should contribute to widespread usage of multimedia services. In this paper we analyze several techniques proposed in literature for content characterization of video programmes, including movies and sports, that could be helpful for mobile media consumption. In particular we focus our analysis on shot clustering methods and effective video summarization techniques since, in the current video analysis scenario, they facilitate the access to the content and help in quick understanding of the associated semantics. First we consider the shot clustering techniques based on low-level features, using visual, audio and motion information, even combined in a multi-modal fashion. Then we concentrate on summarization techniques, such as static storyboards, dynamic video skimming and the extraction of sport highlights. Discussed summarization methods can be employed in the development of tools that would be greatly useful to most mobile users: in fact these algorithms automatically shorten the original video while preserving most events by highlighting only the important content. The effectiveness of each approach has been analyzed, showing that it mainly depends on the kind of video programme it relates to, and the type of summary or highlights we are focusing on.

1. INTRODUCTION

As long as we are entering the multimedia era, a large amount of video information has been made available to the normal users. These video-data collections originated from different sources such as digital broadcasting, private collections of home video, stored TV programs or professional video archives, appeal to large audiences so that their efficient distribution over various networks should contribute to widespread usage of multimedia services. As a consequence, the problem of content characterization of video programmes is of great interest because of the strong raise in demand for an efficient retrieval and visualization of the desired piece of information. In particular some specific user needs, such as visualizing sports highlights on mobile devices, or retrieving a particular clip from a movie, or browsing all the scenes with his/her favorite actor from a large digital library, or again having an automatic tool able to organize hours of home videos, are rapidly increasing.

The discipline of video content analysis, studying and developing algorithms that enable automated analysis of large video databases, tries to bridge the gap (see [1]) between these high-level retrieval queries and the analysis of a video, which is still feasible mainly in terms of low-level features. In this scenario the decomposition into shots (*i.e.* the basic video segments filmed in one single camera take) and the consequent key-frame extraction ([21] and [27]) are commonly considered as the prior steps for performing effective content-based indexing, retrieval and summarization. In fact collections of such key-frames extracted from shots, providing a compact representation of a given video sequence, can be used for generating simple static video summaries.

However, a shot separation often leads to a far too fine segmentation: if we consider that there are usually several hundreds of key-frames for an hour long video, it is laborious to check all these images to get a rough idea of the video content. So, building upon this, research efforts (*e.g.* [23] and [2]) are now invested towards grouping shots into more compact structures, by means of their associated low-level visual features, in order to produce more effective video summaries.

In this paper we propose a general overview on recent works regarding video content summarization techniques since the interest in this area is widespread. In fact, with the proliferation of digital video, the process of generating video summaries will become an indispensable component to any practical content management system; once properly realized, a video summary could be displayed on a mobile device without the worry of timing issues. Our analysis starts describing shot clustering methods in order to produce video summaries in the form of static storyboards. Then we focus our attention on dynamic summaries, the so-called video skims, since from the viewpoint of user, a video skim may provide a more attractive choice (since it contains audio and motion information that makes the abstraction more natural, interesting and informative). Finally we do a brief review on the extraction of sport highlights, a high-potential application in the mobile device scenario.

The paper is organized as follows. After the introduction, in Section 2 the previous works on shot clustering are extensively described and briefly discussed. Then Section 3 presents the already proposed schemes regarding video summarization techniques, both static (storyboards) and dynamic (video skims). Then a particular class of video summaries, *i.e.* sport highlights, is investigated and analyzed. Through all the paper, a general discussion on the considered algorithms with a view on performance comparison is given. The paper ends in Section 5 with some concluding remarks.

2. VIDEO SHOT CLUSTERING

Effective segmentation and clustering of video shots are nowadays considered important basic techniques in content-based video analysis and retrieval. In particular recent works have shown how an accurate grouping of similar shots can facilitate the access to video content (as in [45]) and helps in understanding the associated semantics (as in [62] and [16]). Moreover a number of further processing applications, ranging from semantic annotation to video summarization, can largely benefit from effective clustering of similar shots.

Generally speaking, data clustering methods can be classified into two main categories: supervised and unsupervised approaches (the reader can refer to [13] for an accurate overview on data clustering methods). Regarding shots clustering, supervised learning is used for example in [60] and [26]. In these cases, low-level features are extracted from the key-frames belonging to each shot, while the training data are labeled by hand. Specifically, neural network and HMM are used for the statistical training of the classifier. In general, supervised methods are more accurate and efficient than unsupervised methods, but the work of hand labeling requires a lot of time. Moreover, these classifiers can only be applied on the same types of videos and different classifiers should be trained for different video sets.

To overcome these problems, clustering methods based on unsupervised learning have been developed. These clustering methods can be applied directly on data without any hand labeling and, in the case of video data, represent universal solutions for different sets of video programmes. A simple but efficient method is the well known *k-means* clustering algorithm, which for example has been adopted in [15], where a probabilistic hierarchical clustering using Gaussian Mixtures Models is proposed. However the main drawback of such an approach is that the number of the k clusters must be a-priori decided by the user, who sometimes does not share enough knowledge on the clustering data set.

Regarding the issue of shot clustering, in [63] and [47] some approaches based on a time-constrained analysis are presented. In [63] interesting results are obtained clustering shots according to a visual similarity measured by means of color or pixel correlation between key-frames. In [22] instead, the dissimilarity between shots is examined by estimating correlations between key-frame block matching. Building upon this the temporal relations between clusters of shots are then exploited to group contiguous and interconnected shots sharing a common semantic thread into the so called Logical Story Units (which can be considered the best computable approximations to semantic scenes).

Other shot clustering algorithms useful for scene segmentation adopt a short term memory-based model of shot-to-shot *coherence* as in [25] and [52]. Lately, spectral methods (proposed in [39] and [68]) resulted to be effective in capturing perceptual organization features and grouping similar shots into compact structures.

3. VIDEO CONTENT SUMMARIZATION

Among all possible applications of shot clustering, automatic video content summarization has recently attracted numerous attentions due to its commercial potential impact. In general, an extracted video summary should highlight the video content and contain little redundancy, while preserving the balance coverage of the original video.

As described in [38] the techniques for automatic video summarization can be categorized into two major approaches: static storyboard summary (see for example [64], [11], [23] and [6]) and dynamic video skimming (see for example [49], [58], [18], [37], [40], [33], [35], [54] and [34]). In particular, a static summary is a collection of some extracted key-frames of video shots, while a dynamic video skimming is a shorter version of the video composed of a series of selected video clips. On one side, while a static storyboard allows a quick browsing of the content by sacrificing the temporal evolution of the video, a dynamic skim, in contrast, preserves the time-evolving nature of a video by dynamically reproducing certain segments of the content according to a given time length.

A comprehensive review of past video summarization results can be found in [23] and [59], and more specific examples can be retrieved in [11], [12], [17], [23], [53], [27], [21], [44], [20], [66] and [67]. In particular, some of the main ideas and results among the previously published methods are briefly discussed in the next paragraphs, starting from an overview on the existing storyboard summarization techniques.

3.1 Static Video Summarization

In general the solutions to the static summarization problem are typically based on a two step approach: first identifying video shots from the video sequence, and then selecting key-frames according to some criteria from each video shot (as in [21] and [27]). Most key-frames extraction techniques are based mainly on visual information except some approaches like [10], [37], [49] where motion, audio and linguistic information are also incorporated.

Focusing first on methods employing shot clustering techniques, in [67] Zhuang et al. propose an unsupervised method where a video sequence is segmented into video shots by features clustering based on color histogram in the HSV color space. For each video shot, the frame closest to the cluster centroid is chosen as the key-frame for the video shot, regardless of the duration or activity of the shot itself. Similarly Hanjalic et al. propose in [23] a method that divides the video sequence into a certain number of clusters, and find the optimal clustering by cluster-validity analysis. Each cluster is then represented in the video summary by a key-frame, having therefore removed the visual redundancy among frames.

In [38] Ngo et al. present a unified approach for video sum-

marization based on the analysis of video structures and video highlights. In this approach, the video sequence is represented as a complete undirected graph and the normalized cut algorithm is carried out to partition the graph into video clusters. The resulting clusters form a directed temporal graph and a shortest path algorithm is proposed to detect video scenes. The attention values are then computed and attached to the scenes, clusters, shots, and subshots in a temporal graph. As a result, the temporal graph can describe the evolution and perceptual importance of a video.

Other approaches to video static summarization alternative to shot clustering methods include [11] and [53]. In [11] DeMenthon et al. propose an interesting method based on curve simplification. A video sequence is considered as a curve in a high dimensional space, and a video summary is represented by the set of control points on that curve that meets certain constraints and best represents the curve. Sundaram et al. instead, in [53] use the Kolmogorov complexity as a measure of video shot complexity, and computed the video summary according to both video shot complexity and additional semantic information under a constrained optimization formulation.

In [2] a general methodology for automated shot clustering with the purpose of static video summarization is proposed. With respect to many of the methods here presented, this scheme does not require to set the number of clusters in advance, but the final dimension of each cluster is determined by the visual content consistency of its constituent shots. Moreover, instead of obtaining one single summary, the final user can browse multiple summaries organized in a hierarchical scheme at different content granularity.

Many authors (as in [57], [5], [9] and [55]) also try to generate the summary in the form of a video poster, which arranges semantically structured image key-frames in a bi-dimensional plane. As example of this, in [9] Chiu et al. introduce the interesting idea of Stained-Glass visualization. The basic idea of this approach is to find regions of interest in the video and to condense their key-frames into a tightly packed layout by filling the spaces between the packed regions. However, these methods use only low-level features and do not consider the semantic content, and also the time length of the summary and the number of key frames to be displayed can not be changed freely. Moreover, since the generated summary is not semantically structured, users still have to view the whole video to search a specific scene. As a similar form of summaries, Uchihashi et al. in [57] present a method of making video posters in which the key-frame sizes are changed according to an importance measure.

3.2 Dynamic Video Summarization

In addition to the creation of a video poster summary, in [55] a content-based dynamic summarization method for large sports video archives using metadata is also proposed. Specifically, a certain number of video segments are selected according to the significance of play scenes. The quality of a video skim depends on whether the information which a user wants is included in the summary. Therefore, the authors consider creating a video skim which fits the length specified by a user that includes as many important video segments as possible.

To date, if compared with static storyboard summary, there are relatively less works that address dynamic video skimming. Nevertheless, the interest in effective techniques for dynamic video skimming (often directly derived from the obtained static storyboard) is highly in demand. In fact a tool that can automatically shorten the original video while preserving most events by highlighting only the important content would be greatly useful to most users.

Recently, Singular Value Decomposition (SVD) have been proposed as an attractive computational model for video skimming in [18]. However, this approach is computationally intensive since it operates directly on video frames. In [33], a hierarchical tree that consists of events, activities, actions, and shots is constructed to represent the video content. Then a dynamic summary is generated by randomly removing subtrees at different levels to meet the desired output video length. In [54], the rules of cinematic syntax are used to give a syntactical-based reduction schemes for dynamic summarization. Utility functions are derived to maximize the content and coherence of summaries based on the audio-visual information.

In [35] instead, video skimming is achieved by modeling and detecting the motion-attended regions in videos. Specifically, summaries are generated by merging together those video segments that contain high confidence scores in the motion-attended regions. In [34], the attention model proposed in [35] is further generalized by considering the faces and the audio information. One limitation of [35] and [34] is that the structural information such as the inter-shot relationship is not exploited for video skimming. As a result, the dynamic video summary is solely a collection of video highlights that do not take into account the content coverage.

In [49], Smith et al. propose a method for the automatic generation of a video skim, extracting from the video significant information such as audio keywords, specific objects, camera motions and scene breaks. By integrating text, audio, image analysis and language understanding techniques, nevertheless this approach could not generate satisfactory results when speech signals are noisy, which happens frequently in life video recording. As a final example, since it is clear that a dynamic video summary inevitably introduces distortions at the play back time and the amount of summarization distortion is related to the conciseness of the summary, in [32] the skim generation is formulated as a rate-distortion optimization problem.

4. SPORT VIDEO INDEXING AND EXTRACTION OF HIGHLIGHTS

The valuable semantics in a sports video generally occupy only a small portion of the whole content, and the value of sports video drops significantly after a relatively short period of time [7]. The design of efficient automatic techniques suitable to semantically characterize and summarize sports video documents is therefore necessary and very important.

To characterize video documents, a lot of different audio, visual, and textual features have been proposed and discussed in literature [69], [59], [50]. However, if compared to other videos such as news and movies, sports videos have well de-

fined content structure and domain rules. In particular a long sports game is often divided into a few segments. Each segment in turn contains some sub-segments. For example, in american football, a game contains two halves, and each half has two quarters. Within each quarter there are many plays, and each play starts with the formation in which players line up on two sides of the ball. Again, a tennis game is divided into sets, then games and serves. In addition, in sports video, there is a fixed number of cameras in the field that produce unique scenes during each segment. In tennis for example, when a serve starts, the scene is usually switched to the court view. In baseball, each pitch usually starts with a pitching view taken by the camera behind the pitcher. Furthermore, for TV broadcasting, there are commercials or other special information inserted between game sections [70].

Regarding soccer video, the focus was placed initially on shot classification [19] and scene reconstruction [65]. More recently the problems of segmentation and structure analysis have been considered in [61], [60], whereas the automatic extraction of highlights and summaries has been analyzed in [3], [4], [29], [30], [31], [56], [14]. In [56], for example, a method that tries to detect the complete set of semantic events which may happen in a soccer game is presented. This method uses the position information of the player and of the ball during the game as inputs, and therefore needs a quite complex and accurate tracking system to obtain this information.

On the other hand, if we want for example to detect only the goals, we can try to capture the "dynamic" evolution of some low-level features, as suggested in [4], [30], or try to recognize some specific cinematic patterns, as proposed in [14]. In the same way, we are looking to a "dynamic" characteristic if we want to determine automatically the slow-motion replay segments, as suggested in [41].

As far as baseball sequences are concerned, the problem of indexing for video retrieval has been considered in [24], whereas the extraction of highlights is addressed in [46], [8] and [70]. The indexing of F1 car races is considered in [43] and [36], where the proposed approaches use audio, video and textual information. The analysis of tennis videos can be found, for example, in [70] and [42], whereas basketball and football are considered in [71], [48], [51], and [28] respectively.

As a general conclusion, the effectiveness of each approach depends mainly on the kind of sports considered, and from the type of highlights we are interested in.

5. CONCLUSION

In this paper we have analyzed various techniques proposed in literature for the summarization of video content coming from different programmes, that can be useful for mobile service applications. The interest in the area of video content summarization is widespread, since one of the major benefit of digital media has been that the user will be provided with more choices and more interacting viewing experience. However, with the vast amount of data provided through digital channels, it is only through the use of automated content-based analysis that viewers will be given a chance to manipulate content at a much deeper level than that intended by broadcasters, and hence put true meaning into interactivity. Our attention mainly focused on shot clustering methods for generating static storyboards and dynamic summarization techniques for generic videos, including also sport programmes and related highlights generation. The effectiveness of each approach presented mainly depends on the kind of video it relates to, and the type of summary or highlights we are focusing on.

6. REFERENCES

[1] B. Adams, "Where Does Computational Media Aesthetics Fit?," IEEE Multimedia, pp. 18-26, April-June 2003.

[2] S. Benini, A. Bianchetti, R. Leonardi, "Extraction of Significant Video Summaries by Dendrogram Analysis", to appear in Proc. of ICIP'06, Atlanta, GA, October 8-11, 2006.

[3] A. Bonzanini, R. Leonardi, and P. Migliorati, "Semantic video indexing using MPEG motion vectors," in Proc. EUSIPCO'00, pp. 147–150, Sept. 2000, Tampere, Finland.

[4] A. Bonzanini, R. Leonardi, and P. Migliorati, "Event recognition in sport programs using low-level motion indices," in Proc. ICME'01, pp. 920–923, Aug. 2001, Tokyo, Japan.

[5] J. Calic, N. Campbell, "Comic-like Layout of Video Summaries," in Proc. of WIAMIS'06, Seoul, South Korea, April 2006.

[6] H. S. Chang, S. S. Sull, S. U. Lee, "Efficient Video Indexing scheme for Content-Based Retrieval," IEEE Trans. on Circuits and Systems for Video Technol., vol. 9, no. 8, pp. 1269-1279, Dec. 1999.

[7] S.-F. Chang, "The holy grail of content-based media analysis," IEEE Multimedia **9**, pp. 6–10, Apr.-June 2002.

[8] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in Proc. ICIP'02, pp. 609–612, Sept. 2002, Rochester, NY.

[9] P. Chiu, A. Girgensohn, Q. Liu, "Stained-Glass Visualization for Highly Condensed Video Summaries", in Proc. of ICME'04, Taipei, Taiwan, June 2004.

[10] A. Divakaran, R. Radhakrishnan, K. Peker, "Motion activity-based extraction of key-frames from video shots," Proc. of ICIP02, Rochester, NY, Sept. 2002.

[11] D. DeMenthon, V. Kobla, D. Doermann, "Video Summarization by Curve Simplification," in Proc. of CVPR'98, Santa Barbara, CA, 1998.

[12] N. Doulamis, A. Doulamis, Y. Avrithis, S. Kollias, "Video Content Representation Using Optimal Extraction of Frames and Scenes," in Proc. of ICIP'98, Chicago, IL, pp. 875-878, 1998.

[13] R. O. Duda, P. E. Hart, D. G. Stork, "Pattern Classification", Wiley-Interscience, 2^{nd} ed., New York, 2001.

[14] A. Ekin and M. Tekalp, "Automatic soccer video analysis and summarization," in Proc. SST SPIE03, Jan. 2003, CA, USA.

[15] D. Gatica-Perez, A. Loui, M.-T. Sun, "Finding Structure in Home Video by Probabilistic Hierarchical Clustering," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 6, pp. 539-548, June 2003.

[16] D. Gatica-Perez, M.-T. Sun, A. Loui, "Consumer Video Structuring by Probabilistic Merging of Video Segments," in Proc. of ICME'01, Tokyo, Japan, Aug. 2001.

[17] A. Girgenshohn, J. Boreczky, "Time-Constrained Key Frame Selection Technique," in Proc. of IEEE Multimedia Computing and Systems, pp. 756-761, 1999.

[18] Y. H. Gong, X. Liu, "Video Summarization Using Singular Value Decomposition," in Proc. of International Conference on Computer Vision and Pattern Recognition CVPR'00, vol. 2, pp. 174-180, 2000.

[19] Y. Gong, L. Sin, C. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in Proc. ICMCS'95, May 1995, Washington DC,USA.

[20] Y. Gong, X. Liu, "Video Summarization and Retrieval Using Singular Value Decomposition," ACM MM Systems Journal, vol. 9, no. 2, pp. 157-168, Aug 2003.

[21] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved ?," IEEE Trans. on Circ. and Syst. for Video Technol., vol. 12, no. 2, pp. 90-105, Feb. 2002.

[22] A. Hanjalic, R. L. Lagendijk, "Automated High-Level Movie Segmentation for Advanced Video Retrieval Systems," IEEE Trans. on Circuits and Syst. on Video Technol., vol. 9, no. 4, June 1999.

[23] A. Hanjalic, H. J. Zhang, "An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis," IEEE Trans. on Circuits and Syst. for Video Technol., vol. 9, no. 8, pp. 1280-1289, Dec. 1999.

[24] T. Kawashima, K. Takeyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content based video retrieval," in Proc. ICIP'98, pp. 871–874, Oct. 1998, Chicago, IL., USA.

[25] J. R. Kender, B.-L. Yeo, "Video Scene Segmentation via Continuous Video Coherence", in Proc. of CVPR'98, pp. 367-373, Santa Barbara, CA, May 1998.

[26] E. Kijak, L. Oisel, P. Gros, "Hierarchical Structure Analysis of Sport Videos Using HMMs," in Proc. of ICIP'03, Barcelona, Spain, pp. 1025-1028, Sept. 2003.

[27] I. Koprinska, S. Carrato, "Temporal Video Segmentation: a Survey," Signal Processing: Image Commun., vol. 16, pp. 477-500, 2001.

[28] S. Lefevre, B. Maillard, and N. Vincent, "3 classes segmentation for analysis of football audio sequences," in Proc. ICDSP'02, July 2002, Santorin, Grece.

[29] R. Leonardi, P. Migliorati, and M. Prandini, "Modeling of visual features by markov chains for sport content characterization," in Proc. EUSIPCO'02, Sept. 2002, Toulouse, FR.

[30] R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," IEEE Multimedia **9**, pp. 44–51, Apr.-June 2002.

[31] R. Leonardi, P. Migliorati, and M. Prandini, "A markov chain model for semantic indexing of sport program sequences," in Proc. WIAMIS'03, Apr. 2003, London, UK.

[32] Z. Li, G. M. Schuster, A. K. Katsaggelos, "MINMAX Optimal Video Summarization," IEEE Trans. on Circuits and Syst. for Video Technol., vol. 15, no. 10, pp. 1245-1256, Oct. 2005.

[33] R. Lienhart, "Dynamic Video Summarization of Home Video," in Proc. of SPIE'00, vol. 3972, pp. 378-389, San Jose, CA, Jan. 2000.

[34] Y.-F. Ma, L. Lu, H.-J. Zhang, M. Li, "A User Attention Model for Video Summarization," in Proc. of 10^{th} ACM Int. Conf. Multimedia, pp. 533-542, Juan Les Pins, FR, Dec. 2002.

[35] Y.-F. Ma, H.-J. Zhang, "A Model of Motion Attention for Video Skimming," in Proc. of ICIP'02, vol. 1, pp. 129-132, Rochester, NY, Sept. 2002.

[36] V. Mihajlovic and M. Petrovic, "Automatic annotation of formula 1 races for content-based video retrieval," in Tech. report, TR-CTIT-01-41, Dec. 2001.

[37] J. Nam, A. T. Tewfik, "Dynamic Video Summarization and Visualization," in Proc. of 7^{th} ACM Int. Conf. Multimedia, Orlando, Florida, pp. 53-56, Nov. 1999.

[38] C.-W. Ngo, Y.-F. Ma, H.-J. Zhang, "Video Summarization and Scene Detection by Graph Modeling," IEEE Trans. on Circuits and Syst. for Video Technol., vol.15, no.2, pp. 296-305, Feb. 2005.

[39] J-M. Odobez, D. Gatica-Perez, M. Guillemot, "Video Shot Clustering Using Spectral Methods", in Proc. of CBMI'03, Rennes, FR, Sept. 2003.

[40] X. Orriols, X. Binefa, "An EM Algorithm for Video Summarization, Generative Model Approach," in Proc. of Int. Conference on Computer Vision, Vancouver, Canada, vol. 2, pp. 335-342, July 2001.

[41] H. Pan, P. Beek, and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in Proc. ICASSP'01, May 2001, Salt Lake City, USA.

[42] M. Petkovic, W. Jonker, and Z. Zivkovic, "Recognizing strokes in tennis videos using hidden markov models," Marbella, Spain, 2001.

[43] M. Petrovic, V. Mihajlovic, W. Jonker, and S. Djordievic-Kajan, "Multi-modal extraction of highlights from tv formula 1 programs," in Proc. ICME'02, Aug. 2002, Lausanne, Switzerland.

[44] Y. Qi, A. Hauptmann, T. Liu, "Supervised Classification for Video Shot," in Proc. of ICME'03, Baltimore, MD, July 2003.

[45] Y. Rui, T. Huang, "A Unified Framework for Video Browsing and Retrieval," in Image and Video Proc. Handbook, Academic Press, pp. 705-715, 2000.

[46] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in Proc. ACM Multimedia'02, pp. 105–115, 2000, Los Angeles, CA, USA.

[47] E. Sahouria, A. Zakhor, "Content Analysis of Video Using Principal Components," IEEE Trans. on Circuits and Syst. for Video Technol., vol. 9, no. 8, pp. 1290-1298, 1999.

[48] D. Saur, Y. Tan, S. Kulkarni, and P. Ramadge, "Automated analysis and annotation of basketball video," in SPIE Vol. 3022, Sept. 1997.

[49] M. A. Smith, T. Kanade, "Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques," in Proc. of CVPR'97, Puerto Rico, pp. 775-781, June 1997.

[50] C. Snoek and M. Worring, "Multimodal video indexing: a review of the state-of-the-art," in ISIS Technical Report Series, Vol. 2001-20, Dec. 2001.

[51] G. Sudhir, J. lee, and A. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in IEEE Multimedia, 1997.

[52] H. Sundaram, S. F. Chang, "Determining Computable Scenes in Films and their Structures Using Audio-Visual Memory Models," in Proc. of ACM, pp. 95-104, Los Angeles, CA, Nov. 2000.

[53] H. Sundaram, S.-F. Chang, "Constrained Utility Maximization for Generating Visual Skims," in Proc. of IEEE Workshop Content-Based Access of Image and Video Library 2001, Kauai, HI, pp. 124-131, 2001.

[54] H. Sundaram, L. Xie, S.-F. Chang, "A Utility Framework for the Automatic Generation of Audio-Visual Skims," in Proc. of 10^{th} ACM Int. Conf. Multimedia'02, Juan Les Pins, FR, pp. 189-198, 2002.

[55] Y. Takahashi, N. Nitta, N. Babaguchi, "Video Summarization for Large Sports Video Archives," in Proc. of ICME'05, Amsterdam, NL, July 2005.

[56] V. Tovinkere and R. J. Qian, "Detecting semantic events in soccer games: Toward a complete solution," in Proc. ICME'01, pp. 1040–1043, Aug. 2001, Tokyo, Japan.

[57] S. Uchihashi, J. Foote, A. Girgensohn, J. Boreczky, "Video Manga: Generating Semantically Meaningful Video Summaries", in Proc. of ACM Multimedia 1999, Orlando, Florida, pp. 383-392, October, 1999.

[58] N. Vasconcelos, A. Lippman, "A Spatio-Temporal Motion Model for Video Summarization," in Proc. of International Conference on Computer Vision and Pattern Recognition CVPR'98, Santa Barbara, CA, pp. 361-366, June 1998.

[59] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia Content Analysis Using Both Audio and Visual Clues," IEEE Signal Process. Mag., vol. 17, no. 11, pp. 12-36, Nov. 2000.

[60] L. Xie, S.-F. Chang, A. Divakaran, H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Model," in Proc. of ICASSP'02, Orlando, FL, May 2002.

[61] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," in Proc. ICME'01, pp. 928–931, Aug. 2001, Tokyo, Japan.

[62] M. Yeung, B. L. Yeo, B. Liu, "Segmentation of Video by Clustering and Graph Analysis," in Proc. of Computer Vision and Image Understanding, vol. 71, no. 1, pp. 94-109, July 1998.

[63] M. M. Yeung, B.-L. Yeo, "Time-Constrained Clustering for Segmentation of Video Into Story Units," in Proc. of ICPR'96, vol.III-vol.7276, p. 375, Vienna, Austria, Aug. 1996.

[64] M. M. Yeung, B.-L. Yeo, "Video Visualization for Compact Presentation and Fast Browsing of Pictorial Content," IEEE Trans. on Circuits and Syst. for Video Technol., vol. 7, no. 5, pp. 771-785, Oct. 1997.

[65] D. You, B. Yeo, M. Yeung, and G. Liu, "Analysis and presentation of soccer highlights from digital video," in Proc. ACCV'95, Dec. 1995, Singapore.

[66] D. Q. Zhang, C. Y. Lin, S. F. Chang, J. R. Smith, "Semantic Video Clustering Across Sources Using Bipartite Spectral Clustering," in Proc. of ICME'04, Taiwan, June 2004.

[67] Y. Zhuang, Y. Rui, T. S. Huan, S. Mehrotra, "Adaptive Key Frame Extracting Using Unsupervised Clustering," in Proc. of ICIP'98, Chicago, IL, pp. 866-870, 1998.

[68] J. Zhang, L. Sun, S. Yang, Y. Zhong, "Joint Inter and Intra Shot Modeling for Spectral Video Shot Clustering," in Proc. of ICME'05, Amsterdam, NL, July 2005.

[69] T. Zhang and C.-C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," IEEE Trans. on Speech and Audio Processing **9**, pp. 441–457, 2001.

[70] D. Zhong and S.-F. Chang, "Structure analysis of sports video using domain models," in Proc. ICME'01, pp. 920–923, Aug. 2001, Tokyo, Japan.

[71] W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule based video classification system for basketball video indexing," in Proc. ACM Multimedia'02, Dec. 2002, Los Angeles, CA, USA.