Multimedia Systems 9: 586–593 (2004)
Digital Object Identifier (DOI) 10.1007/s00530-003-0130-3

Multimedia Systems

Multimodal detection of highlights for multimedia content

Serhan Dagtas[1], **Mohamed Abdel-Mottaleb**[2]

[1] Department of Information Science, University of Arkansas, 2800 S. University Ave., Little Rock, AR 72204, USA
[2] Department of Electrical & Computer Engineering, University of Miami, 1251 Memorial Drive, Coral Gables, FL 33146, USA

Abstract. We present a multimedia information analysis framework for content-based browsing of video. Specifically, we develop algorithms for the automated extraction of video highlights in sports video that are based on audio, text, and image features. The extracted annotations are used to build applications for selective browsing of sports videos. Such summarization techniques enable content-based indexing of multimedia documents for efficient storage and retrieval. In addition, in the context of the newly emerging standard MPEG-7, these methods will enable applications that use MPEG-7 descriptions. As this standard provides only the syntax for representing such descriptions and not specific algorithms for extracting them, these algorithms are of great value for establishing MPEG-7 as an accepted standard. We provide experimental results for the proposed algorithms on several hours of sports programs that prove the feasibility of efficient video access techniques in a multimedia environment.

1 Introduction

One of the most important aspects in multimedia data management is to characterize and index the content for efficient access to the desired information. Due to the sheer size of the data involved and its complex semantics, effective content-based analysis methods are crucial for the management of multimedia documents. One such method is the detection of highlights from video streams for efficient user access in consumer applications.

This paper presents several multimedia algorithms that detect highlight segments from broadcast video as part of a broader goal of providing an effective multimedia data management system.

Thanks to countless challenging research topics, multimedia data management has been a fertile ground for many research projects. Initially, the focus was mostly on image retrieval based on features such as color, texture, and shape. Various systems have been developed for content-based image and video retrieval including QBIC [7], Photobook [23], CONIVAS [1], FourEyes [30], Virage [17], ViBE [7], VideoQ [7], VisualSeek [33], Netra [19], and MARS [23]. These systems represent image and video content using a set of low-level attributes such as color, texture, shape, layout, and motion. Retrieval is performed by comparing the features of a query image or video clip with features stored in the database and presenting the user with the content that has the most similar features. Users normally are interested in semantics rather than low-level features, e.g., "show me scenes of goals in a soccer game" rather than "show me scenes that have a certain dominant color". The early approaches to video retrieval used image retrieval of the video keyframes to retrieve video segments. This approach, though practical, has its limitations in describing semantic attributes of the content. Recently, researchers have started to address issues related to semantic-based video retrieval. For this purpose, domain knowledge of the video category is used to enable the automatic extraction of the descriptions. Multiple cues from the different modalities, i.e., audio, images sequences, text, and graphics, can be used to obtain the desired descriptions. This enables a richer analysis that can take into account more than one medium in order to produce better content description.

In this paper, we present algorithms for detecting highlights in sports programs using multimedia cues, namely, audio, text, and visual features. Extracting highlights from video allows users to view the recorded material in a time-saving mode by providing the consumers with the most "interesting" portions of TV programs. While this process may be straightforward, though time consuming, for humans, automatic detection of interesting events is a difficult task that requires sophisticated tools for analysis.

Prior work on video indexing and sports video analysis

Most of the reported work for video indexing focuses on a unimodal approach, using only one of the visual [10,16,17,28,38,40,42,43], audio (www.informatik.uni-mannheim.de/informatik/pi4/projects/MoCA) [6,13,15,24,26,27,30,41], or textual [7,19,44] modalities. Recently, approaches using combined multimodal analysis have been reported [2,4,12,20,25,31,35,37].

Sports video has been a content of choice for many content-based video applications lately due to the popularity of sports

Correspondence to: S. Dagtas (e-mail: sxdagtas@ualr.edu)

programming and the relatively high financial cost of the video content itself. Shot classification of soccer video was done in [16], and rule-based semantic classification was addressed in [33]. Babaguchi et al. [5] explore the advantages of multimodal techniques for retrieval of sports events by combining the textual and color-based image processing techniques. While providing an interesting multimodal technique, their method lacks applicability to general sports programs and does not utilize audio. Assfalg et al. [3] reported another multimodal work that combined graphics and image analysis for annotating sports video that does not use audio analysis, either. In [40], an interesting segment of video is first located by using one feature, and then "interestingness" is validated using another feature. Rui et al. [34] carried out extensive analysis of audio features for detecting baseball highlights, but they do not address other modalities. Clearly, multiple approaches that use different modalities for the same purpose have not been adequately explored for a comparative analysis. Here we use the three modalities, namely, text, audio, and image, for extracting highlights of various sports programs and analyze the outcomes to provide the reader with a perspective of how these modalities may support the task of video summarization.

3 provides an overview of the methods we present in this paper. The first method utilizes audio features to efficiently locate the highlight segments. Then we explore the advantages of multimedia analysis by combining the audio and text features toward the same goal. The third method uses color features for an image-based detection of highlight segments. In the following three sections, we present detailed descriptions of these three methods. Section 5 provides implementation details as well as experimental results, with the last section outlining the major conclusions of our work.

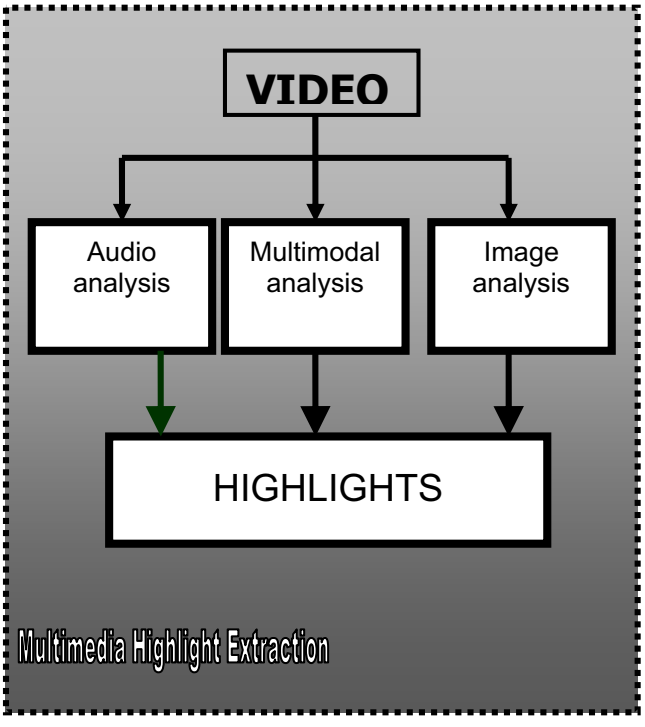

Fig. 1. Overview of the methods presented in this paper

2 Highlight extraction using audio features

Audio signal properties have been widely used in content analysis. Audio features such as zero crossing rate have been used to detect and represent certain characteristics (e.g., speech-music discriminator) [39]. More sophisticated feature extraction and classification methods have also been used quite successfully [22]. These methods generally either run offline or require high computation power for real-time operation. Highlight extraction can take place at either the consumer device or the broadcaster's end. Our algorithms are designed to run on consumer devices in real time, which requires a fair degree of automation as the consumers will demand a minimum level of human intervention. Furthermore, applications running on consumer electronics devices require very efficient algorithms to work with limited processing power. Therefore, techniques that can work with lower sampling rates and computationally inexpensive algorithms are needed on these platforms. This is a key point in the design of the proposed methods in this paper: computational efficiency for feasible porting of the algorithms to the consumer electronics platforms.

We take full advantage of the potential efficiency of audio processing by using an algorithm of linear complexity that can run much faster than real time. In addition, our algorithm can run with sequences sampled at rates much lower than the typical rates. For that, we subsample 44.1-kHz data and work with 441-Hz sequences. This further improves the computational efficiency and makes it possible to implement this method on platforms with significantly less computation power.

The fundamental purpose of this method is to use audio characteristics to detect segments from sports programs where interesting events occur. This typically means increased crowd activity manifested in an increased energy level of audio. However, energy level alone is not sufficient for generating meaningful segments that would make up a reasonable summary or highlights; further processing is needed to group the video segments that have high energy audio into meaningful highlights. The grouping of these segments eliminates the need for boundary detection.

As shown in Algorithm 1 below, we process the audio streams in units of segments 1 s long. We use a form of audio energy that we call **Excitement Level, X(A)**: If A is a segment of the audio signal, the excitement level is computed as $X(A) = Avg(Abs(A))$, where *Abs* is the absolute value function. Then, a sliding window of five units (equivalent to 5 s) is used to compute the average level; in most situations true events last for at least 5 s.

Then, segments with averages above a certain threshold are combined to form a sequence. The threshold value we have used is half of the maximum value over the entire program (approximately 2 h). The measure of importance used in ranking each segment is the summation of all average energy values throughout the entire range. This ensures that energy as well as duration are taken into account in picking that segment. The parameters used in the algorithm such as the 5-s interval or the threshold levels are experimental, and further optimization may be possible.

Algorithm 1: Audio-based highlight extraction

Divide the input audio signal sampled at 441 Hz into N segments: A(k), k=1,…,N

Compute Excitement Level of A: X(k) = Ave(Abs(A(k)))

Compute the average over five seconds, AX :

$$AX(k) = \sum_{m=1}^{5} X(k+m-1)$$

Combine units with AX(k)> Threshold into one segment, S, and compute the signal $Str(S)$:

$$Str(S) = \sum_{l=1}^{M} AX(l)$$

where M is the number of units in segment S

3 Highlight extraction using audio and text features

Textual transcripts of speech carry plenty of semantic information about a TV program. Often, the words and the context are sufficient to detect segments that represent a concept. In soccer programs, for example, important events such as "goal" or "yellow card" can be detected by searching the commentators' speech transcripts. We have observed that textual search of the transcript is a good first step in detecting such events, as in most cases commentators will say an associated word when an event happens in the game. In addition, searching for text is much faster than audio- or image-based searching of the content. However, textual search alone is not sufficient since there are a large number of false positives in the set returned by textual search. We use an audio-based method similar to the one in the previous section to refine the search and hence improve the precision of this algorithm. This also demonstrates the potential advantage of a multimedia method over a monomedia one to achieve the same goal.

For illustration, we employed this method to detect *goals* from soccer programs and *touchdowns* from football programs. We used the closed captions in a football (NFL) game and manually marked the occurrences of these words in the transcript in four soccer games for which closed captions were not available. As shown in Algorithm 2, a textual search is first performed over the transcript of the program to find the segments where the associated word (*goal* or *touchdown*) exists. After this step, the system returns a long list of candidates, i.e., segments of the program that contain the specified keyword. These segments are the 10-s windows around the instance that the word was spoken. Many of these instances, however, do not represent a real event since commentators often use those words in reference to previous games or plays. To eliminate such false positives, we compute the average energy values for the associated segments and mark the ones that exceed the threshold.

4 Highlight extraction using image features

In sports video, interesting events are usually accompanied by a particular camera effect. We have used this observation to identify possible locations of interesting events in soccer matches. In these videos, after an interesting event the camera usually zooms in on the viewers or shows a close-up of the players. When the camera focuses on the viewers, no or little grass can be seen in the field of view.

Figure 3 shows an example event from a soccer game; the frames in the figure are subsampled from the game to scan frames a few seconds before the event up to frames a few seconds after the event. We see in the figure that after the event, i.e., a goal, the camera gives a close-up of the players and there is no grass in the sequence of frames that follow the event.

Algorithm 2: Highlight extraction using audio and text features

Search the transcript and find 10-sec-long segments that contain the searched keyword K

Generate a 10 sec long segment for each occurence of the keyword:

A(k), k=1,…,N

Compute normalized excitement level X,of A:

NX(k)=Ave(Abs(A(k)))/Max(Abs(A))

where Max operation is over the segment, Abs(A)

Select units with

NX(k) > Threshold into the final list

The absence of large grassy areas in the frames that follow the interesting events usually lasts for a few seconds. Our highlight extraction algorithm is based on this observation; it works by detecting sequences of frames in which there are no or small grassy areas. If the sequence of those frames is larger than a predefined threshold in terms of time, this means that the camera has focused for that amount of time on the players or the viewers and hence an interesting event may have occurred. We have introduced a preliminary version of this algorithm in [11].

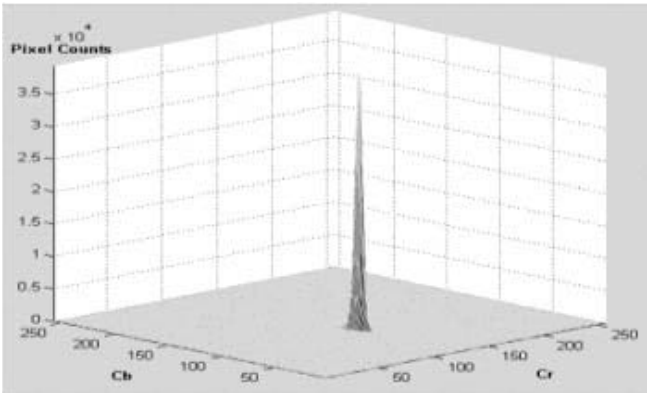

Fig. 2a. Distribution of pixel values for grassy areas

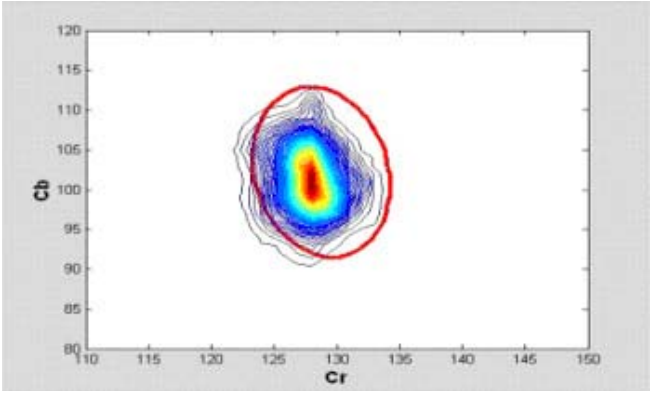

Fig. 2b. Contours of distribution in **a** and ellipse used for the classification

Detection of grassy areas

We use color characteristics to detect the grassy areas. Texture can also be used to distinguish between areas that have similar color to the grass and real grassy areas. We use the **YCbCr** color space and keep the chrominance (i.e., **Cr** and **Cb**) and drop the luminance, **Y**. To detect the grass pixels, we estimate the probability density function (pdf) for the pixel values in grassy areas. This is done by using patches from a set of images in the training phase.

The result is a pdf: ***P(pixel value|grass)***. Figure 2a shows the pdf obtained from a large number of patches. It is clear that the values are concentrated around a mean value and the pdf has the shape of a Gaussian distribution. Figure 2b shows the contours projected from the pdf on the **CbCr** plan and the ellipse that we used for the grass classification. The detection of grass is then accomplished by marking the pixels in a frame that have values inside the ellipse as grass and the ones that have values outside that ellipse as nongrass.

Figure 4 shows the results of grass detection for the frames shown in **Fig. 3**. It is clear from the figure that after the event there is a transition between frames that have large grassy areas to frames that have small grassy areas.

Visual event detection for game highlights

To detect the interesting events, we need to find frame sequences that have transitions from a sequence of frames that include large grassy areas to a sequence of frames that do not include large grassy areas. The first part of the sequence will typically correspond to an event (normal frame), and the second will typically correspond to a break scene, where the camera zooms in on the players or the fans. To find these sequences, we detect the connected components of the detected grassy areas and keep the ones that have a large number of pixels and drop the connected components that have a small number of pixels. For each frame we then calculate the ratio of grass pixels that were not dropped to the number of non-grass pixels. If for a given frame the ratio is high, we mark that frame as type normal; otherwise we mark it as type break. We then look for significant runs of break frames and use these as indications of the endings of interesting events. In our experiments, we used a threshold of 300 break frames to identify an interesting event.

Fig. 3. Subsampled sequence of frames from a video clip

The detected events can then be categorized using the text from closed captions.

Figure 5 shows a snapshot of the system software after the user has selected a specific soccer game to view. The system displays different interesting events that are present in the system, e.g., goals, free kicks, etc. When the user chooses to view the goals, the next snapshot in Fig. 5b shows a keyframe icon for each of the three goals that occurred in the game. By selecting one of the icons the user can view the video segment corresponding to a goal.

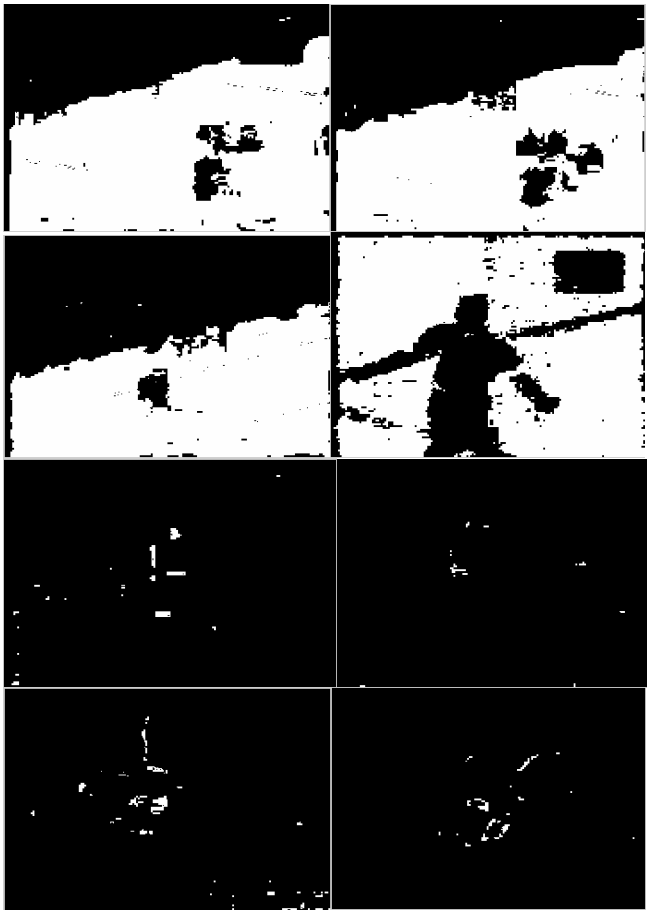

Fig. 4. Detected grassy areas for the sequence of images in Fig. 3

Algorithm 3: Highlight extraction using image features

1. Compute P(pixel-value|grass), the grass pdf function from Γ

Fig. 6. Application for video highlight detection using text and audio

10 min with a notably high (approximately 90%) recall value and about 80% precision. Details of each game are listed in **Table 1**, where we show the results for each game separately. The COR (Correct) column displays the number of actual interesting events (such as free kicks, saves, cards, etc.) out of the 25 displayed segments, and INC (Incorrect) is the total number of incorrect classifications (errors). PRE (Precision) is the ratio of correct detections to all the detected segments. REC (Recall) percentage is more complicated as it is conceptually very difficult to come up with an objective count of such a subjective concept as "interesting event". Nevertheless, the recall measure is important to determine the retrieval rate accurately. We therefore used the number of goals as a measurable event (as it is the most important one), and the REC column reflects the ratio of goals included in the retrieval set of 25 segments for each game. The missed goals (e.g., in games Yug-Slo and Ita-Bel) are also detected, but they are simply ranked lower than the 25th position and therefore not listed in the results set. The DUR (Duration) column is the total length of the highlights in seconds.

Table 1. Results for the audio-based method (Algorithm 1)

GAMES	COR	INC	PRE (%)	REC (%)	DUR (s)
Spa-Yug	21	4	84	100	631
Tur–Bel	21	4	84	100	697
Hol–Fra	20	5	80	100	725
Yug-Slo	18	7	72	50	683
Nor-Spa	18	7	72	100	371
Eng-Rom	19	6	76	100	586
Hol–Ita	19	6	76	100	321
Ita–Bel	21	4	84	67	361
Ave.	19.25	5.75	77	89.6	547

A notable observation is the correlation between the total duration of the 25 highlights and the activity level of the game. Generally, the duration is higher for the games with a high number of goals (e.g., Spa-Yug with seven goals) and lower for the games with few goals (e.g., Hol-Ita with no goals). In addition, as an improvement in the user interface, the duration can be adjusted to a certain value (e.g., 5 min, 10 min, etc.) which gives the user another way to set the desired recall/precision, i.e., trading time with more or fewer highlights.

Table 2 shows the results for the method that uses the combination of audio and text. Accordingly, keyword-only search results in 17.6% precision. Combining text and audio improves the precision to 65.4%. Recall rate was 100% in both cases. All segments are 10 s long. In this table, the TOTAL column indicates the total number of segments returned as a result of the textual search. Numbers in parentheses are the audio results, marked as the positive (selected by the audio processing step) and negative (eliminated by the audio processing step), respectively. COR (Correct) is the total number of correct classifications (both positive and negative), and INC (Incorrect) is the total number of errors. PRE (Precision) values are for text only (left column) and combined (right column) algorithms. REC (Recall) is the ratio of the retrieved events, which shows how many of the searched events (goals or touchdowns) are in the final list.

For the image-based algorithm, our data set was comprised of 45 min of a soccer game. The algorithm returned 21 instances, of which 8 were important events, 7 were normal plays, and 6 were injury or substitute type events. Recall rate in terms of number of goals was 100%. These results demon-

Table 2. Results for audio- and text-based method (Algorithm 2)

GAMES	TOTAL	COR	INC	PRE (%)		REC (%)
Spa-Yug	24 (12+12)	22	2	46	92	71(5/7)
Tur-Bel	7(0+7)	7	0	0	100	0(0/2)
Yug-Slo	14(6+8)	6	8	7	17	17(1/6)
Eng-Rom	16(8+8)	11	5	19	38	60(3/5)
NFL	25(5+20)	24	1	16	80	80(4/5)
Tot./Ave.	86	70	16	18	66	46

strate that the algorithm performs very well in terms of the recall rate, although the precision is not perfect. For the purpose of creating video summary, i.e., to produce a few minutes of summary of a nearly 2-h game, these results are fine. But if the objective is to use these results for retrieval, we can enhance the precision by using the audio cue in addition to the visual cue. One possibility is to detect events based on the visual cues and then only keep the detected events that satisfy the audio algorithm (e.g., Algorithm 1) presented earlier in the paper. Another approach would be to integrate both cues simultaneously using hidden Markov models [20].

6 Conclusion

We have demonstrated how three modalities – text, audio, and image – can be used successfully for extracting video highlights. The methods presented here have the potential to be a powerful feature in consumer digital video devices by providing the essence of recorded programs efficiently. These methods use efficient audio and image processing techniques as well as a combination of text-based search and audio processing for improving the accuracy of the automatically extracted highlights. We have demonstrated that such a combination of multimodal processing of video is a viable and powerful approach for effective and efficient extraction of highlights. Our experiments with several hours of sports programs provided satisfactory results with reasonable 5- to 10-min highlights created for 2-h games. The processing efficiency is due to the combination of computationally efficient algorithms and the domain knowledge for the sports video. In addition, audio processing is performed efficiently thanks to the linear nature of the algorithm and its ability to work with undersampled (e.g., 441 Hz as opposed to 44.1 kHz) audio signal.

The use of domain knowledge has simplified the problem. Instead of trying to solve the general problem of finding interesting events, cues from the different modalities, i.e., visual, audio, and text, have been used to enable the detection of the interesting events. For example, in the case of visual cues, rather than analyzing the segmented images, we have used a simple measure of the grass ratio to detected possible transitions from a play sequence to a break sequence.

Another notable conclusion is that, while the additional use of audio (after text) improves the precision, the recall rate for that technique appears to be relatively low. The textual search step relies on the existence of the associated words (e.g., goals) in the transcript, which is not always guaranteed because of errors in the closed captions. A possible workaround may be the use of grammatical analysis or natural language processing (NLP) techniques to improve the textual phase of the search. Therefore, this technique is better suited for situations where the time instance of the event is more important than providing a full list of the events in demand. The use of visual cues also has great potential, but it should be supported by other modalities for improved precision and boundary determination.

The overall conclusion of our work is that multimedia features (text, audio, and image) can be used effectively for efficient presentation of multimedia content. As would be expected, different modalities have different advantages in certain applications and a careful selection of the right modality and parameters leads to an effective solution for the automated generation of video highlights for efficient storage and retrieval of multimedia documents.

References

1. Abdel-Mottaleb M, Dimitrova N, Desai R, Martino J (1996) CONIVAS: CONtent-based image and video access system. In: Proceedings of ACM Multimedia'96, Boston, 18–22 November 1996, pp 427–428
2. Alatan A, Akansu A, Wolf W (2001) Multimodal dialogue scene detection using Hidden Markov Models for content-based multimedia indexing. Multimedia Tools Appl 14(2):137–151
3. Assfalg J, Bertini M, Colombo C, Del Bimbo A (2002) Semantic annotation of sports videos. IEEE Multimedia 9(2):52–60
4. Babaguchi N, Sasamori S, Kitahashi T, Jain R (1999) Detecting events from continuous media by intermodal collaboration and knowledge use. In: Proceedings of the IEEE international conference on multimedia computing and systems, Florence, Italy, 1–7 June 1999, pp 782–786
5. Babaguchi N, Kawai Y, Kitahashi T (2002) Event based indexing of broadcasted sports video by intermodal collaboration. IEEE Trans Multimedia 4(1):782–786
6. Brown GJ, Cooke M (1994) Computational auditory scene analysis. Comput Speech Lang (8):297–236
7. Brown M, Foote J, Jones G, Sparck-Jones K, Young S (1995) Automatic content-based retrieval of broadcast news. In: Proceedings of ACM Multimedia 1995, San Francisco, 5–9 November 1995, pp 35–43
8. Chang SF, Chen W, Meng HJ, Sundaram H, Zhong D (1997) VideoQ – an automatic content-based video search system using visual cues. In: Proceedings of ACM Multimedia, Seattle, November 1997, pp 313–324
9. Chen J-Y, Taskiran C, Delp EJ, Bouman CA (1998) ViBE: a new paradigm for video database browsing and search. In: Proceedings of the workshop on content-based access of image and video libraries (in conjunction with CVPR'98), Santa Barbara, CA, June 1998, pp 96–100
10. Colombo C, Del Bimbo A, Pala P (1999) Semantics in visual information retrieval. In: Proceedings of IEEE Multimedia, 6(3):38–53
11. Dagtas S, Abdel-Mottaleb M (2001) Extraction of TV highlights using multimedia features. In: Proceedings of the IEEE workshop on multimedia signal processing, Cannes France, 3–5 October 2001, pp 91–96
12. Eickeler S, Muller S (1999) Content-based video indexing of TV broadcast news using hidden Markov models. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, Phoenix, AZ, 15–19 March 1999, pp 2997–3000

13. El-Maleh K, Klein M, Petrucci G, Kabal P (2000) Speech/music discrimination for multimedia applications. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing, Istanbul, Turkey, 5–9 June 2000, pp 2445–2448
14. Flickner M et al (1996) Query by image and video content: the QBIC system. IEEE Comput 28(9):23–32
15. Ghias A, Logan J, Chamberlin D, Smith BC (1995) Query by humming – musical information retrieval in an audio database. In: Proceedings of ACM Multimedia, San Francisco, 5–9 November 1995, pp 231–236
16. Gong Y, Sin LT, Chuan CH, Zhang H, Sakauchi M (1995) Automatic parsing of TV soccer programs. In: Proceedings of the international conference on multimedia computing and systems (ICMCS '99), Washington, DC, 15–18 May 1995, pp 167–174
17. Gunsel B, Ferman M, Tekalp M (1996) Video indexing through integration of syntactic and semantic features. In: Proceedings of the 3rd IEEE workshop on applications of computer vision, Sarasota, FL, 2–4 December 1996, pp 90–95
18. Hampapur A, Gupta A, Horowitz B, Shu CF, Fuller C, Bach J, Gorkani M, Jain R (1997) Virage video engine. In: Proceedings of SPIE: Storage and Retrieval for Image and Video Databases V, San Jose, CA, February 1997, pp 188–197
19. Hauptmann AG, Lee D, Kennedy PE (1999) Topic labeling of multilingual broadcast news in the informedia digital video library. In: Proceedings of the ACM DL/ SIGIR MIDAS Workshop, Berkeley, CA, 14 August 1999, pp 287–288
20. Huang J, Lu Z, Wang Y, Chen Y, Wong EK (1999) Integration of multimodal features for video scene classification based on HMM. In: Proceedings of the *IEEE workshop on multimedia signal processing*, Copenhagen, Denmark, 13–15 September 1999, pp 53–58
21. Ma WY, Manjunath BS (1997) Netra: a toolbox for navigating large image databases. In: Proceedings of the IEEE international conference on image processing, Santa Barbara, CA, October 1997, 1:568–571
22. Martin KD (1999) Sound-source recognition: a theory and computational model. Ph.D. thesis, MIT, Cambridge, MA, June 1999
23. Mehrotra S, Rui Y, Ortega M, Huang TS (1997) Supporting content-based queries over images in MARS. In: Proceedings of the IEEE international conference on multimedia computing and systems, Ontario, Canada, 3–6 June 1997, pp 632–633
24. Minam K, Akutsu A, Hamada H, Tomomura Y (1998) Video handling with music and speech detection. In: Proceedings of IEEE Multimedia 5(3):17–25
25. Naphade MR, Huang TS (2001) A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans Multimedia 3(1):141–151
26. Patel NV, Sethi K (1996) Audio characterization for video indexing. In: Proceedings of SPIE on storage and retrieval for still image and video databases, San Jose, CA, 28 January–2 February 1996, 2670:373–384
27. Patel NV, Sethi K (1997) Video classification using speaker identification. In: Proceedings of IS&T SPIE, Storage and Retrieval for Image and Video Databases IV, San Jose, CA, 8–14 February 1997, pp 218–225
28. Peker AK, Alatan AA, Akansu AN (2000) Low-level motion activity features for semantic characterization of video. In: Proceedings of the IEEE international conference on multimedia and expo, New York, 30 July–2 August 2000, 2:801–804
29. Pentland A, Picard RW, Sclaroff S (1994) Photobook: content-based manipulation of image databases. In: Proceedings of SPIE Storage Retrieval Image Video Databases II, San Jose, CA, 6–10 February 1994, 2185:34–47
30. Pfeiffer S, Fischer S, Effelsberg W (1996) Automatic audio content analysis. In: Proceedings of ACM Multimedia 1996, Boston, 18–22 November 1996, pp 21–30
31. Pfeiffer S, Lienhart R, Effelsberg W (2001) Scene determination based on video and audio features. Multimedia Tools Appl 15(1):59–81
32. Picard RW, Minka TP (1995) Vision texture for annotation. Multimedia Sys 3:3–14
33. Qian R, Tovinkere V (2001) Detecting semantic events in soccer games: towards a complete solution. In: Proceedings of the IEEE international conference on multimedia and expo, Tokyo, 22–26 August 2001, pp 833–836
34. Rui Y, Grupta A, Acero A (2000) Automatically extracting highlights for TV baseball programs. In: Proceedings of ACM Multimedia, Los Angeles October 2000, pp 105–115
35. Satoh S, Nakamura Y, Kanade T (1999) Name-it: naming and detecting faces in news videos. IEEE Multimedia 6(1):22–35
36. Smith JR, Chang SF (1996) Visualseek: a fully automated content-based image query system. In: Proceedings of ACM Multimedia, Boston, November 1996, pp 87–98
37. Smith MA, Kanade T (1997) Video skimming and characterization through the combination of image and language understanding techniques. In: Proceedings of CVPR 1997, San Juan, Puerto Rico, 17–19 June 1997, pp 775–781
38. Sudhir G, Lee JCM, Jain AK (1998) Automatic classification of tennis video for high-level content-based retrieval. In: Proceedings of the IEEE international workshop on content-based access of image and video databases, in conjunction with ICCV'98, Bombay, India, 3 January 1998, pp 81–90
39. Toklu C, Liou S, Das M (2000) Video abstract: a hybrid approach to generate semantically meaningful video summaries. In: Proceedings of the 1st IEEE international conference on multimedia and expo (ICME), New York, 30 July–2 August 2000, 3:1333–1336
40. Truong BT, Venkatesh S, Dorai C (2000) Automatic genre identification for content-based video categorization. In: Proceedings of the IEEE international conference on pattern recognition, Barcelona, Spain, 3–8 September 2000, pp 4230–4233
41. Wold E, Blum T, Keislar D, Wheaton J (1996) Content-based classification, search, and retrieval of audio. In: Proceedings of IEEE Multimedia 3(3):27–36
42. Zhang H, Tan S, Smoliar S, Yihong G (1995) Automatic parsing and indexing of news video. Multimedia Sys 2(6):256–266
43. Zhou W, Vellaikal A, Kuo C (2000) Rule-based video classification system for basketball video indexing. In: Proceedings of ACM Multimedia 2000, Los Angeles, 30 October–4 November 2000, pp 213–216
44. Zhu W, Toklu C, Liou S (2001) Automatic news video segmentation and categorization based on closed-captioned text. In: Proceedings of the IEEE international conference on multimedia and expo, Tokyo, 22–25 August 2001, pp 1036–1039