# TV Content Analysis: Techniques and Applications

Yiannis Kompatsiaris, Bernard Merialdo and Shiguo Lian

June 25, 2011

# Chapter 1

# TV Program Structuring Techniques: A Review

## 1.1 Introduction

The objective of the proposed chapter is to present the problem of TV program structuring, its applications in real-world services, and to provide a panorama of existing techniques.

Due to the high number of television channels and their various audiovisual content, the amount of broadcasted TV programs has significantly grown. Consequently, data organization and tools to efficiently manipulate and manage TV programs are needed. TV channels broadcast audiovisual streams, in a linear way, 24h/24. This makes the content available only at the broadcast time. Services like Catch-up TV and PVRs remove this constraint and allow users to watch previously broadcasted TV programs. To enable these services, TV Programs have to be extracted, stored, indexed and prepared for a later usage. Additional indexing steps are also required. After choosing a program, the user might want to get an overview of the program before watching it. He/she might also want to directly access a specific part of the program, to find a certain moment of interest or to skip a part of the program and go to the following one or even to the final one. These features represent an

alternative for the basic fast forward/backward options. The value of these features was explored by Li et al. [25], where users were provided with such capabilities for a wide variety of video content. The results showed that the users found the ability to browse video content very useful for the reasons of time saving and the feeling of control over what they watched.

Program extraction requires a macro-segmentation of the TV stream. Macro-segmentation algorithms generally rely on detecting "inter-programs", which include commercials, trailers, jingles and credits. These are easier to detect as they have a set of common properties related to their duration, visual and audio content. The idea is to detect these inter-programs and to deduce the long programs as the rest of the stream. Metadata (like EIT or EPG) can be used afterwards to annotate the programs.

As for non-linear TV program browsing and summary generation, the original structure of TV programs has to recovered and all possible moments of interest have to be precisely tagged. A similar option exists on DVDs, where an anchorpoint indicates the different chapters of the video. But in order to obtain such a representation, a human observer is required to watch the entire video and locate the important boundaries. Obviously this could be done in the case of TV programs also, but these pre-processing steps are costly and highly time consuming, especially when dealing with large amount of TV programs, as is the case in real-world services. One challenge is hence to develop content-based automatic tools for TV program structuring. These tools will allow the easy production of information that will give to the users the capability to watch programs ondemand or to watch just the parts of the programs they are interested in. As mentioned earlier, another important field where TV program structuring is extremely useful is video summarization [32, 31, 42]. Video summarization techniques can use the structure of the program and its moments of interest in order to build the best compact overview of the TV program.

A video can be analyzed at different granularity levels. The elementary level is the image/frame, generally used to extract features like color, texture, shape. The next level is represented by shots, basic video units showing a continuous action in both time and space. The shots are separated by editing effects called transitions. A transition can be abrupt, namely *cut*, and groups directly successive shots, or *gradual*, and groups together successive

shots by different editing effects like dissolve, wipe, fade etc. Shot boundary detection, is the process of identifying the transitions, abrupt (cut) or gradual, between the adjacent shots. Shot boundary detection techniques were intensively studied and a lot of methods have been proposed [53]. The shot however is not a relevant level to represent pertinent parts of a program as it usually lasts few seconds and has low semantic content. A video may have hundreds or thousands of shots, which is not practical for human navigation. Therefore, high-level techniques are required to group video shots into a more descriptive segment of the video sequence. These techniques are classified into two wide categories: *specific methods* and *generic methods*.

Specific methods exploit the prior knowledge of the domain in order to construct a structured model of the analyzed video. They can be applied only to very specific types of programs like news, sports programs, series, advertisements, etc. On the other hand, generic methods try to find a universal approach for the structuring of videos, independent of their type and based only on their content features.

The rest of the chapter is composed of two main parts presented in the next two sections. They discuss in more detail each of the two types of approaches, the different techniques that were proposed in the literature for their implementation, their strengths and weaknesses.

## 1.2 Specific approaches

Authors who propose specific methods consider that a universal solution for high-level video analysis is very difficult, if not impossible, to achieve [13]. Low level features generally used for indexing video content, are not sufficient to provide a semantically meaningful information. In order to achieve sufficient performance, the approaches have to be specifically adapted to the application. Bertini et al. [3] consider that, in order to ease the automatical extraction of high level features, the specific knowledge of the different types of programs must be considered. Thus, many researches have focused on very specific types of programs like sports programs, movies, advertisements and news broadcasts. The methods they propose are specific methods. Specific methods make use of prior knowledge of the type of the

analyzed TV program in order to extract relevant data and construct its structure model. They are supervised as they generally require the prior creation and manual annotation of a training set used to learn the structure. A class of TV programs that are often analyzed by specific methods is sports programs. These have a very well defined structure. The rules of the game provide prior knowledge that can be used to provide constraints on the appearance of events or the succession of those events. These constraints are very helpful to improve the accuracy of the event detection and categorization. Another class of programs that is most appropriate for this kind of systems are news programs, as they have also a very clear structure. They are produced using almost the same production rules: they consist generally in a succession of reports and anchorperson shots. Specific methods analyze the temporal (the set, reports, advertising, forecast, etc.) and spatial structures (images with the anchorperson, logos, etc.). Most of the work relies on finding the anchorperson shots and then deduce the sequences of shots representing the reports.

We focus on these two categories of TV programs (sport and news) for the specific approaches as they are the most representative for their class and discuss them in more detail in the next subsections.

### 1.2.1 Approaches for sports programs structuring

Sports videos have been actively studied since the 1980s. Due to the fact that sports events attract a large audience and have important commercial applications, they represent an important domain of semantics acquisition. Sports videos are generally broadcasted for several hours. People who miss the live broadcast are often interested in the strongest moments only. Motivated by this need of the users, broadcasters should automatically select and annotate the relevant parts of the video in order to recover the structure and create a table of content that would allow a direct access to the desired parts. Sports videos are characterized by a defined structure, specific domain rules and knowledge of the visual environment. All these make possible the extraction of the video structure, allowing non-linear browsing. We distinguish two levels in the sports video analysis: segment and event. In the first case, the objective is to segment the video into narrative segments like *play*

and *break* through a low level analysis of the video. The second one, assumes a higher level analysis of the video and its objective is to identify the interesting moments(highlight events) of the sports video. Both of them are predetermined by the type of sport. For example, an event in the case of soccer might be the detection of the goal while for tennis this will be the match points. Therefore, in order to accomplish these objectives the use of prior knowledge is needed. This prior knowledge may be related to the type of the analyzed sport(i.e. game surface, number of players, game rules) but also to the production rules for the video program (i.e. slowmotion replay, camera location and coverage, camera motion, superimposed text [45]).

**Low level analysis of sports videos**

As already said, the first level of semantic segmentation in sports videos is to identify the *play* and *break* segments. For the class of field ball games, a game is in play when the ball is in the field and the game is going on. Out of play is the complement set, when the action has been stopped: ball outside the field, score, audience, coach, play stopped by the referee, etc. Generally any sports game can be segmented into these two mutually exclusive states of the game. The interest in obtaining such a segmentation is that it allows play-by-play browsing and editing. It also facilitates the further high level analysis and detection of more semantically meaningful units as events or highlights. It is claimed that highlights are mainly contained in a play segment [41]. The occurrence of audio visual features in play-break segments show remarkable pattern for different events. For example, a goal event usually happens during a play segment and it is immediately followed by a break. This break is used by the producers to emphasize the event and to show one or more replays for a better visual experience. As a result, the goal event can be described by a pattern which can be defined using the play and break structure. Another benefit of using play-break segmentation is that it reduces significantly the video data (no more than 60% of the video corresponds to a play), as stated in [46]. A sports video is actually composed of continuous successions of play and break sequences. Two major factors influence the sports video syntax: the producer and the game itself. A sports video benefits of typical production rules. Most of the broadcasted sports videos, use a variety of camera view shots and additional editing

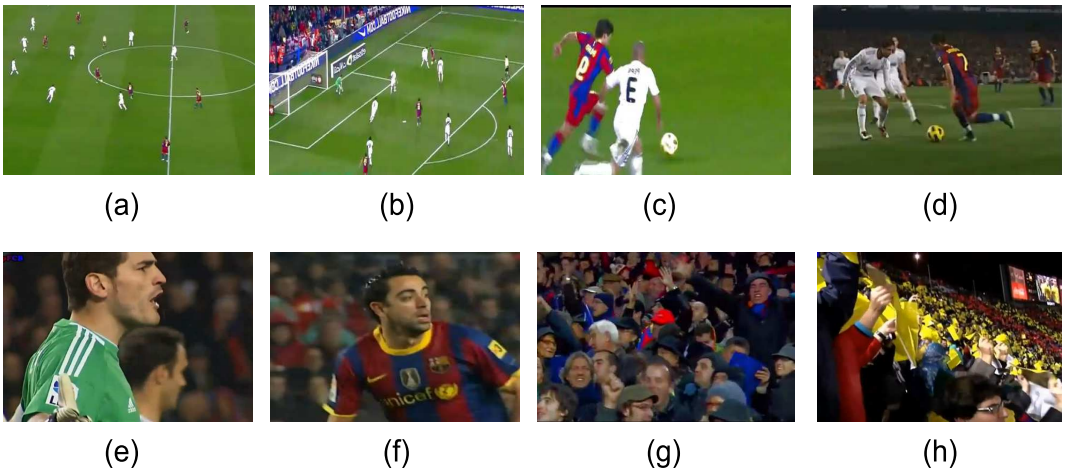


Figure 1.1: View types in soccer: (a,b) Long view, (c,d) in-field medium view, (e,f) close-up view, and (g,h) out of field view.

effects such as replay and on screen captions to describe the sports video content. The different outputs of camera views have successfully been used for play-break segmentation. An example is in [11] where shots are classified in three classes : long shot, in-field shot and close-up or out of field shot (see Figure 1.1). A long shot displays the global view of the field hence it serves for localization of the events in the field. An in-field medium shot is a zoomed-in view of a specific part of the field and it contains usually a whole human body. A single isolated medium length shot between two long shots corresponds to a play while a group of nearby medium shots corresponds to a break in the game. The replays are also considered medium shots. A close-up shot shows the view of a person and in general indicates a break in the game. Also, out of field shots that show the audience, the coach, etc., indicate a break in the game. The classification is made automatically, based on the ratio of grass color pixels (G) for the close-up and out of field shots which have a small value of G. Because the limit between medium or long shots is questionable for the frames with high value of G, a cinematographic algorithm (Golden Section Composition) is used in order to classify them as long view or in field medium shots. An analysis of frame to frame change dynamics is performed in order to detect the replays. The same approach is used in [41] where play-break sequences are segmented using the output from camera view classification and replay detection. The camera view classification is performed on each frame using the playing field (or dominant) color ratio which measures the amount of grass pixels in the frame. As replays are often recognized as play shots because of the global view, even though

they should be included as breaks (as they are played during breaks), replay detection is applied to locate additional breaks and to obtain more accurate results. The algorithm is tested on soccer, basketball and Australian football videos.

In [9], Duan*et al.*, use a mid level representation, between low level audiovisual processing and high level semantic analysis. First, low level visual and audio features are extracted and a motion vector model, a color tracking model and a shot pace model(that describes the production rules effect on the shot length) are developed. These models are then used in a supervised learning algorithm in order to classify shots into predefined classes, e.g. Player Close-up,Field View, Audience. SVM is also used to classify game specific sounds e.g. applause, whistling. The semantic classes are specific to the nature of the analyzed sport and are further used to construct a temporal model for representing the production rules of the sport. The sports video shot sequences are partitioned into two logical segments, namely, *in play segments* (IPS) and *out of play segments* (OPS), that occur in successive turns. Based on the temporal model and the semantic shot classes, the in play and out of play segments are determined by checking the shot transition pattern. The approach has been tested on five field ball type sports. In Xie *et al.* [46] a stochastic method based on HMMs is proposed for play-break sequence detection. As in previous approaches, dominant color ratio is used, this time along with motion intensity in order to identify the view type of each frame of a soccer video. Motion intensity gives an estimate of the gross motion in the whole frame including object and camera motion. A wide shot with high motion intensity often results from a play, while static wide shot usually occurs during the break in a game. Six HMM models are created for the respective game sequences. Using a sliding window, the likehood for each of the pre-trained (play and break) models is retained. A dynamic programming algorithm is used for the temporal segmentation depending on the transition probabilities between the play-break and the likehood of each segment to belong to one of the two classes (play/break). The experimental evaluations showed improvement of the performance of the HMM based algorithm compared to the discrete rule based one studied in a previous work of the author [49]. In the last cited one, heuristic rules were used in detecting the view type of the sequences in order to obtain play-break segmentation. The rules took in consideration the time duration of each view and their relative position in time.

HMMs can also be modeled in a multistream approach [48] to classify shots of soccer or volleyball videos. A multistream is obtained by combining multiple single stream HMMs and introducing the weight for each stream. For soccer videos, the first stream is the motion feature vector and the second is the color feature vector of each frame. For volleyball videos the processing is similar excepting the features used.

**High level analysis of sports videos**

Highlight events in sports videos are composed of the most interesting moments in the video that usually capture the users attention. They represent the exciting scenes in a game, i.e. the goal in a soccer match, the tennis match points or the pitches in baseball. A lot of research has focused on the detection of interesting events in sports videos, most of it employing rule based annotation and machine learning or statistics based annotation. The modalities used to acquire the semantics are either individual or multiple modalities. The main sources of information used for the analysis are the video track and the audio one (because crowd cheers, announcer's excited speech, ball hits are highly correlated with the exciting segments). Also, editing effects such as slow motion replay, or close caption information that accompanies the sports program and that provides information of the game status, are used for detecting important events. Within each of these, the features used for the program segmentation may be general (i.e. shot boundary detection) or domain specific (i.e. the center line on a soccer field is vertical or almost vertical in frame [52]). As the highlights are specific to each type of sport, most of the approaches make use of sport specific features. The features used for semantic analysis of sports videos can also be considered as cinematic or object based features. The cinematic ones are referred to those that result from the production rules , such as camera views and replays. Sport video production is characterized by the use of a limited number of cameras on fixed positions. The different type of views are very closely related to the events in the video and for each type of sport a pattern can be identified. As an example, during a rally in a tennis video, a global view of the camera, filming the entire court is selected. Right after the rally, a close-up of the player that just scored a point is captured [20]. In the soccer videos a goal is indicated by a close-up on the player who just carried out the important scene, followed by a view of the audience and a slow motion

replay. Moreover, starting from close-up shots, the player's identity can be automatically annotated using face recognition techniques and based on textual cues automatically read from the player's jersey or from text caption [4].

For the case of object based features, they are used for high level analysis of sports videos. Objects are described by their color, texture, shape and motion information. Using even more extensive prior knowledge, they are localized in different frames and their motion trajectories are identified and used for the detection of important events. As example in diving videos, player body shape segmentation and transition is used to represent the action [26].

As a result, high level analysis in sports videos can be done based on the occurrences of specific audio and visual information that can be automatically extracted from the sports video.

**Rule based approaches** Many of the existing works on event detection use rule based approaches to extract a predictable pattern and construct, manually or automatically, a set of rules by analyzing the combination of specific features. In order to analyze different sports, Duan *et al.* [9] used a mid level representation, between low level audiovisual processing and high level semantic analysis. As described in the previous subsection, predefined semantic shot classes are used to facilitate high level analysis. Based on production rules, a general temporal model is developed and coarse event (in play)/non event (out of play) segments are identified. Richer events are classified in a hierarchy of *regular events*, which can be found in the transitions between in play and out of play segments, and *tactic events* that usually occur inside the in play segments. To identify the regular events, the semantic shot classes and audio keywords are exploited in order to derive heuristic rules according to game specific rules. For example, a soccer goal occurs if there are many *close-up* shots, *persistent excited commentator speech* and *excited audience*, and long duration within the OPS segment. For tennis, events like serve, return, score, etc. are detected by analyzing audio keywords like *hitting ball* and *applause* to compute the ball hitting times and the intervals between two ball hits, and the applause sound at the end of court view shots. Because of the use of audio keywords, the precision in detecting some events (i.e. goal for soccer) can be low due to the confusion from the loud environmental audience sound. Tactic events are strongly dependent

on the game specific knowledge and are detected using object features. To identify *take the net* and *rally* in tennis games, the player tracking is employed. Player tracking and its movement over time is also used by Kim *et al.* [23] to detect the locations where the play evolution in soccer games will proceed, e.g. where interesting events will occur. A ground level motion of players at each step is extracted from individual players movement using multiple views and a dense motion field is generated. Using this field the locations where the motion converges are detected. This could represent an important application for automated live broadcasting. Player localization is viewed as a K partite graph problem by Hamid *et al.* [18]. Nodes in each partite of the graph are blobs of the detected players in different cameras. The edges of the graph are a function of pair wise similarity between blobs observed in camera pairs and their corresponding ground plane distances. The correspondence between a player's blob observed in different cameras is treated as a K-length cycle in the graph. Yu *et al.* [52] use players motion intensity, along with other object features and low level features to rapidly detect the boundary of interesting events indicated by the gamelog. The instant semantics acquisition and fast detection of event boundaries is important for the two interactive broadcast services for live soccer video they propose. The first one is a *live event alert* service that informs mobile viewers of important events of a live soccer game, by providing a video clip of the event with a time lag between 30s and 1.5 min. The second one is *the on-th-fly language selection* and allows users to choose their preferred contents and preferred language. To improve the accuracy of event boundaries detection, object features are used. The center line of a soccer field is detected based on the knowledge that this line is vertical or almost vertical in every frame. The goal mouth is also detected based on the prior knowledge : the two posts are almost vertical, the goal posts and goal bar are bold line segments and compared with the center line and side lines and the goal posts are shorter line segments. Gueziec [17] tracks also specific objects, as the ball during pitches in baseball games, to show as replays if the strike and ball decisions were correct. The implemented system was launched in 2001 during *ESPN's Sunday Night Baseball* and it tracked pitches with an extremely low failure rate. The real time tracking involves the use of extensive prior knowledge about the game and system setup (i.e. camera locations and coverage). Considering that object based features are computationally costly, Ekin *et al.* [11] try to use more cinematic features to detect certain events in soccer videos and employ the object

based ones only when needed, to increase the accuracy. Therefore, after classifying shots and segmenting the video into play/break segments(see previous subsection), events like goal, red-yellow cards and penaltys are detected. For goal detection, a pattern is computed using only cinematic features, resulted from common rules used by producers after goal events. The red-yellow cards are indirectly identified by locating the referee by its distinguishable colored uniform. The penalty box is detected based on the *three-parallel-line rule* that defines the penalty box area in the soccer field. All these approaches use an extensive prior knowledge of the game and production rules. Most of them are based on manual observation and heuristic knowledge. In order to reduce the amount of knowledge used for sports video analysis and make the framework more flexible for different sports, Tjondronegoro *et al.* proposed in [41] a knowledge-discounted event detection approach in sports videos. The highlights contained in each play/break (P/B) sequence are classified using a set of statistical features (e.g. sequence duration, break ratio, play ratio, replay duration, near goal ratio, excitement ratio, close-up view ratio) calculated from the P/B sequence. During a training phase, each of the predefined events were characterized using the set of statistics and the heuristic rules were constructed. To classify which highlight is contained in a P/B segment, a score is used for each statistical feature. The value of each calculated feature is compared to the trained one, and if the value falls within the trained statistics of the highlight, the corresponding score is incremented. The highest score will indicate the most likely highlight contained in the P/B segment.

**Machine learning based approaches** Other approaches that try to use a modest quantity of prior knowledge, only needed for selecting the features that match the most with the event, are approaches based on machine learning. For example Hidden Markov Models, that have been proven to be effective for sequential pattern analysis, can be used to model tennis units like *missed first serve*, *rally*, *break* and *replay*, based on the temporal relationship between shots and with respect to editing and game rules [21]. Visual features are used to characterize the type of view of each shot. HMMs are also utilized to recognize the action and index highlights in diving and jump game videos [26]. The highlight in diving videos is described by the entire diving action. As the action occurs many times in the video and the distance between two action clips is regular, these are detected by motion segmentation

and a hierarchical agglomerative clustering. Then the action is recognized based on body shape segmentation and shape transitions modeled by continuous HMMs. The Baum-Welch algorithm is used to train the HMMs and the Viterbi algorithm is used to calculate each model's output probability. The model with the maximum probability is the recognition result. The HMMs can also be modeled in a multilayered approach to detect events like *offence at left/right court*, *fast break at left/right court*, etc., in basketball videos [48]. Instead of detecting and tracking objects, the statistical learning approach is used to build semantic models based on a set of motion features. HMMs are also used to model audio sequences. In [45] the plopping sound is modeled with HMMs in order to identify the highlights in diving sports videos. As it has been proven that browsing highlight with contextual cues together is preferred, slow motion replay and game statistics information in superimposed captions are extracted. It might happen that announcer's exciting speech and audience applause to exceed the plopping sound, making it hard to take a correct decision. Special sound effects like announcer's excited speech (detected using learning machines) and ball-hits (detected using directional templates) have also been employed to detect highlights in baseball programs [37]. A probabilistic framework is used to combine the two sources of information. If an audio segment has a high probability to be an excited speech segment and occurs right after a frame that has a high probability to contain a baseball hit, than it is very likely for the segment to be an exciting (highlight) segment. Audio and video markers together are used in [47] to detect highlights in sports videos like soccer, baseball and golf. A Gaussian Mixture Model has been generated for each predefined audio class (applause, cheering, commentator's excited speech, music) and used to identify the audio markers. For the visual ones, an object detection algorithm such as Viola and Jones has been used to detect the pattern in which the catcher squats waiting for the pitcher to pitch the ball in baseball games, the players bending to hit the golf ball in golf sports and appearance of the goal post in soccer games. Other approaches for event/highlight detection in sports videos propose the use of neural networks to classify shots into predefined classes and utilize them for further processing [1], or AND-OR graphs to learn a storyline model from weakly labeled data using linguistic cues annotations and visual data [16].

### 1.2.2 Approaches for news programs structuring

A second class of programs very appropriate for the specific systems are the news programs. Compared to the sports programs, the work related to news programs is much more extensive and dates for a long time. This might be partially due to their regular structure which makes the analysis much easier. Indeed, most news videos exhibit a similar and well defined structure (Figure 1.2). They usually start with a general view of the set during which the

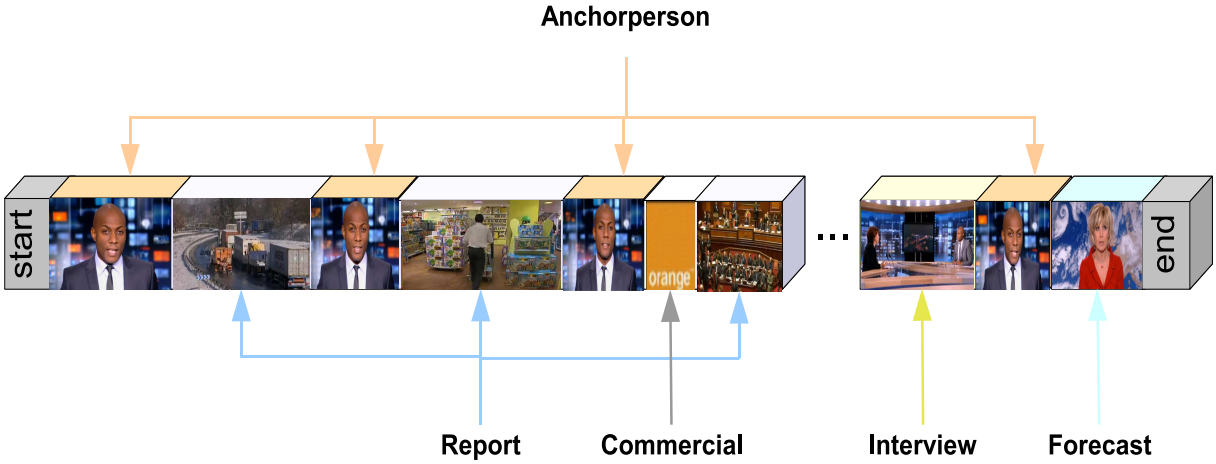


Figure 1.2: The structure of a TV news program.

anchorperson announces the main titles that will be developed later (the highlights). The rest of the program is organized as a succession of TV reports (stories) and segments where the anchorperson announces the next topic. Each story generally begins with an anchorperson segment that provides the general description of the event and continues with a more detailed report and sometimes interview segments. At the end of a story an anchor segment may reappear to give a summary or conclusion. Most news broadcasts end with reports on sport, weather and finance. Commercials may also appear during the broadcast.

**Shots classification** The fundamental step in recovering the structure of a news program is thus to classify the shots into different classes (anchorperson, report, weather forecast, interview etc.). Then, based on this classes video segmentation into story units can be performed. Detecting anchorperson shots plays an important role in most of the news video

segmentation approaches. The anchorperson shots exhibit a lot of characteristics that make their detection easier. Early approaches used properties like the fact that the same anchorperson appears during the same news program and the background remains unchanged during the entire shot. They also considered that, anchorperson shots are usually filmed with a static camera so the anchorperson will always be situated in the same place of the image. Furthermore, each TV channel has representative semantic objects, like logos or captions with the anchorperson name, that are displayed only during news program. All of these, made possible the detection of anchorperson shots by **template matching techniques**, based on weighted similarity measures and model images. Zhang *et al.* [56] proposed a spatial structure model for the anchorperson frame, based on a composition of distinctive regions: the shape of the anchorperson, the caption of reporters names and the logo of the broadcast channel that appears in the top right part of the image. Recognizing an anchorperson shot involves testing every frame over a frame model which in turn means testing each frame against the set of region models. However, constructing a set of models images is difficult due to variations from one TV station to another and the matching algorithm is time consuming. Gunsel *et al.* [15] proposed also semantic object detection and tracking (i.e. faces and logo) using a region matching scheme, where the region of interest is defined by the boundary of the object. The detection of anchorperson shots lies here on a face detection technique, highly used in this purpose. The possible regions where anchorpersons might be situated are extracted using skin detection (color classification) and histograms intersection. These regions are then compared to templates stored in the application database. The prior knowledge of the spatial configuration of the objects included in these shots is used to limit the region of interest during skin detection. The system has some limitations, as it is difficult to track occluded or transparent semantic objects. Also, the interview shots can be confused with anchor shots if the interviewed person has the same spatial position as an anchorperson. In [24], once faces have been detected, procedures including face position analysis, size filter and dress comparison are used to reduce the possibilities of erroneous identification. In [2] a robust face detector approach is implemented by means of color segmentation, skin color matching and shape processing. Heuristic rules are then used to detect anchorperson, report/interview and outdoor shots (e.g. one or two face close-ups with a static background are classified as single or double anchor). Uncontrolled illumination conditions can interfere

with skin-color model, making the classification rates smaller for report/interview shots. In order to avoid using a model for the anchorperson shot recognition, another property of the anchorperson shots is exploited by researchers, for detecting these shots: this is their recurrent appearance during an entire news broadcast. Based on the fact that shots of the anchorperson are repeated at intervals of variable length and that their content is very similar, Bertini *et al.* [3] propose a **statistical approach**. They compute for each shot the *shot lifetime*, as "the shortest temporal interval that includes all the occurrences of a shot with similar visual content, within a video". The shots that have the lifetime greater than a threshold are classified as anchorperson shots. Based on the assumption that both the camera and anchorperson are almost motionless in anchor shots, the statistical approach is refined by the use of motion quantity in each shot. A subclassifications of anchorperson shots (like *weather forecast*) is obtained considering the speech content of anchorperson shots. High level information about the shots being broadcasted is also extracted from close captions text. The frequency of occurrence of anchorperson shots and their similarity is also used in [35] where a clustering method is used to group similar key frames in two groups. The smallest group will represent the anchorperson group. The neighboring shots that have key frames with the same label are merged together. A different approach based on frame statistics, models frames with HMMs [10]. Feature vectors deriving from color histogram and motion variations across the frames are computed and used to train the HMMs. A HMM is built for each of the predefined six content classes (i.e. newscaster, begin, end, interview, weather forecast, report) and four editing effect classes. Frames are then classified by evaluating the optimal path resulted from the Viterbi algorithm. The system has real time capabilities as it works three times faster than real time. Other approaches use SVM to identify anchorperson shots. In [28] the first anchorperson shot in the video is detected using key frames frequency and similarity. For each shot, a region of three other neighboring shots on both sides is taken into account for features computation. An SVM is then trained to identify other anchor shots, using features like distance from anchorperson template, semantic text similarity, shot length distance, average visual dissimilarity and minimum visual dissimilarity between the shots of left and right region. Chaisorn *et al.* [5, 7, 6] defines thirteen categories (e.g. anchor, 2anchor, interview, sport, finance, weather, commercial) to cover all essential types of shots in typical news videos. First, commercials are eliminated

using a heuristic approach based on black frames, still frames,cut rate, and/or audio silence. Weather and Finance are then identified using a histogram matching algorithm. Finally, a Decision Tree is employed to perform the classification of the rest of the shots using a learning based approach and temporal (audio type, motion activity, shot duration) and high level features (human faces and video text) extracted from video, audio and text. The same principle is used in [12]. As anchorperson is one of the most important shot categories, they add a similarity measurement module that uses the background similarity detection to reduce the errors that might appear when identifying the anchorperson shots (anchor shots can easily be confused with speech/interview as they all have similar face features). Thresholds need to be defined for each tested broadcast channel. Gao *et al.* [13] classify video shots into anchorperson shots and news footage shots by a graph-theoretical cluster (GTC) analysis. Since anchorperson key frames with identical model have the same background and anchorperson, they thus have similar color histogram and spatial content. Based on this similarity the anchorperson key frames of each model will be grouped together into subtrees of the graph and distinguished from the individually distributed news footage shots.

**TV news story segmentation** Once the shots are classified, a more challenging task is to segment the broadcast into coherent news stories. This means finding the boundaries of every story that succeeds in the video stream. Using the prior knowledge on the structure of a news broadcast and the classified shots, Zhang *et al.* [55] and Gao *et al.* [13] use a temporal structure model to identify the different stories in the broadcast. Ko *et al.* [24] extracts boundaries of anchorperson and non-anchorperson shots and use them next to metadata from the video structure to classify the news segments into six predefined categories (e.g. opening animation, headlines, preview, anchorperson greeting, news stories, weather forecast, closing scene). As TV channels have different programming and scheduling formats for the news videos, before performing the categorization, metadata from every channel need to be determined by separately analyzing each structure. In [15] consecutive anchorperson-news footage shots are combined into *news units* according to a set of predefined rules. For example, *the anchorperson shot(s) followed by a commercial are combined with the last news unit, the news footage shot(s) following a commercial are considered a part of the next news unit*, etc. In [30] Dynamic Programming is used to detect *relevant video segments*

*(important situations)* by associating image data and language data. For each of the two clues, several categories are introduced. Inter-modal coincidence between the two clues indicates important situations. A considerable number of association failures are due to detection errors and time lag between close caption and actual speech. Other approaches use Finite State Automatons [22, 27] or HMMs to model shot sequences and identify story boundaries. Chaisorn *et al.* [7, 6] represent each shot by a feature vector based on the shot category, scene/location change and speaker change and models it with ergodic HMMs. A similar approach is used in [12]. In addition, a pre-segmentation module based on heuristic rules is added to join together intro/highlights shots or weather shots that can be seen as a single story logical unit. Because in the output of the analysis using HMMs in [7, 6] there were embedded pattern rules, in [5] a global rule induction technique is used instead of HMMs. Pattern rules are learned, in a form similar to if-then-else syntax, from training data. The results are slightly lower compared to the use of HMMs but this method has reduced computational cost and complexity. Text is also highly used for story boundaries detection in news videos. An example is in [34] where a fully automatic television news summarization and extraction system (ANSES) is built. Shots are merged into story units based on the similarity between text keywords , extracted from the teletext subtitles that come along with the broadcast. The assumption made is that story boundaries always coincide with shot boundaries. Keywords are extracted from the subtitles and tagged with their part of speech (noun, verb, etc.). Each type of word has a score. A similarity measure is computed and each text segment is compared to each of its five neighbors on either side. When the similarity exceeds a certain threshold, the segments and their corresponding video shots are merged. In [28] closed captions text along with video stream is used to segment TV news into stories. For text, Latent Dirichlet Allocation (LDA) model is used to estimate the coherence of a segment and thus provide its boundaries. The assumption made is that *each document is represented by a specific topic distribution and each topic has an underlying word distribution.* In this manner, a coherent segment (containing only one story) will have only a few active topics, while a non coherent segment (that contains more than one story) will have a comparatively higher number of active topics. Using the same vocabulary as for the training corpus and a fixed number of topics, the likehood of a segment is computed and used as a score for performing the text segmentation task. For the

video based segmentation, anchorperson shots are identified (as described in the previous subsection) and story boundaries are identified based on the assumption that a story always begin with an anchorperson shot. The two detected boundaries from both approaches are fused using the "or" operator in order to improve the accuracy of story segmentation. The main drawback of the approach is that words that did not appear during training are dropped and whenever an anchorperson shot has been missed, the story boundary will also be missed.

## 1.3 Generic approaches

Generic methods are independent of the type of the video and try to structure it in an unsupervised way, without using any prior knowledge. Due to the fact that they don't rely on a specific model they are applicable to a large category of videos. The restraints of a certain domain are no longer available. As already mentioned in the introduction, the first step into building a structured description of a video is to segment it into elementary shots. A shot is the smallest continuous unit of a video document. But shots are too small and too numerous to assure an efficient and relevant structure of a video. Therefore, in order to segment the video into semantically richer entities, shots need to be grouped into higher level segments, namely *scenes*. A scene is generally composed of a small number of shots all related to the same subject, ongoing event or theme [45, 54]. However, the definition of a scene is very ambiguous and depends on the subjective human understanding of its meaning. In the literature they are also named "video paragraphs" [19], "video segments" [54, 43], "story units" [38, 51, 50] or "chapters" [39]. Existing scene segmentation techniques can be classified in two categories: the ones using only visual features and others using multimodal features. Both of them compute the scene segmentation either by clustering the shots into scenes based on their similarities (using the similarities inside a scene), or by emphasizing the differences between the scenes. Even though the goal is the same, the differences appear in the choice of the parameters and their thresholds. The challenge lies thus in finding the appropriate set of features that would lead to a correct identification of the scenes in a video. An objective evaluation of these methods assumes the existence of a ground truth at scene level. But this ground truth is human generated so it is hard to make a reliable comparison

of the performance of different approaches based on subjective judgments.

### 1.3.1 Visual Similarity-Based Scene Segmentation

A first group of scene segmentation approaches is based on the visual similarity between the shots of the video document. In this manner, Yeung et al. [50] use two metrics of similarity based on color and luminance information, to match video shots and cluster them into scenes. From these metrics, a dissimilarity index is derived and based on it, a proximity matrix for the video shots is built. The algorithm first groups the pairs of shots that are the most similar and then proceeds to group the other shots by their proximity values (dissimilarity indices). The main drawback of this approach is that it may happen that two shots belonging to different scenes are found visually similar and thus grouped into the same cluster (e.g. several scenes can take place in the same room, or several shots show the same person but were taken far apart in time). Therefore, the visual similarity of shots alone is not sufficient to differentiate the context of a shot and to produce a good structure of the video. In order to overcome this difficulty, time constrained clustering was proposed. The general idea is to group successive shots into meaningful clusters along the temporal dimension so that all the shots in a scene are relevant to a common event [45]. Different approaches have been proposed in the literature, most of them are based on the principle of Scene Transition Graph (STG). Yeung et al. [51] uses a STG in order to segment videos into story units. The nodes are clusters of visually similar shots and the edges indicate the temporal flow of the story. A fixed temporal threshold is used to delimitate distant shots. In this way only shots that fall within the time window can be clustered together. An edge exists between two nodes only if there is a shot represented by the first node that immediately precedes a shot represented by the second node. The *cut edges* are used in order to partition the graph into disjoint subgraphs. Each subgraph represents a story unit. The variation of clustering parameters (e.g. time window parameter) is also discussed. Rasheed et al. [36] exploits the same idea of a STG. They construct a weighted undirected graph called shot similarity graph (SSG). Each node represents a shot. Edges between the shots are weighted by their color and motion similarity and also with the temporal distance. Scene boundaries are detected by splitting this graph into subgraphs, so as to maximize the intra-subgraph similarities and minimize

the inter-subgraph similarities. The choice of the values for the visual dissimilarity threshold and temporal parameter can generate over/under segmentation depending on the length and type of video. This problem is discussed in [57] where a similar approach is implemented but moreover, the temporal parameter is estimated as depending on the number of shots in the video. When the shot number is large, the temporal parameter should increase so to avoid the over-segmentation and vice versa.

In [54] another approach for scene segmentation is proposed, based on Markov Chain Monte Carlo (MCMC) algorithm. Based on the visual similarity between all pairs of shots in the video, each shot is assumed to have a likehood of being declared a scene boundary. The scenes boundaries are first initialized randomly and then automatically updated based on two types of updates: diffusion (shifting of boundaries) and jumps (merging or splitting two adjacent shots). Finally, the shots with the highest likehood in their neighborhoods are declared as scene boundary locations. The method does not require the use of any fixed threshold (resolving the problem of over/undersegmentation).

As montage and cinematic rules are widely used by producers to put shots together into coherent stories, Tvanapong *et al.* [40] introduces a more strict definition of the scene based on continuity editing techniques in film literature. First, visual features are extracted from two predetermined regions of the keyframes. These regions were carefully chosen to capture the essential area of frames according to the scene definition. A "background" region will be used to identify shots in the same setting and a "two corners" region to detect events like the traveling event. For each keyframe a feature vector is computed. Guided by the strict definition and predefined editing rules (background criterion, upper corner criterion, lower corner criterion), the extracted features are compared in order to cluster together shots of the same scene.

### 1.3.2 Multimodal Similarity-Based Scene Segmentation

A second category of methods use the combination of features extracted from video, audio and/or text, in order to segment videos into scenes. Thereby, in [33] a scheme for identifying

scenes by clustering shots according to detected dialogs, similar settings and similar audio was developed. Shots are recovered from the video and values for each semantic feature are calculated: background and cuts are identified for audio, frontal face detection is used for recovering the shot/reverse shot pattern for dialog determination, color and orientation for settings determination. Shots are merged into scenes based on computed distances with respect to each feature. This results in different types of scenes depending on the underlying feature. The scenes of different types are also combined to construct better setting scenes, by merging the clusters that overlap into clusters of maximum size. Difficulties were encountered in audio sequence detection based on speech. The experiments showed also that the merging of all the features does not significantly improve the performance.

In [8] a different approach, based on video and audio attention is proposed to automatically detect scenes in videos. Attentive video features are extracted and used to segment the video into shots. Scene changes are detected based on the Euclidean distance among attentive audio features of two neighboring shots. The shots whose audio distance is lower than a threshold are merged into a scene.

In [38] two multi-modal automatic scene segmentation techniques are proposed (audio and video), both building upon the scene transition graph. First, a visual STG is created as in [51] and the cut edges are identified as the set of scene boundaries. Second, audio analysis is employed and a similar audio based STG is constructed, in parallel to the video STG. The technique proposed for combining the results of the two graphs involves the creation of multiple video and audio graphs using different construction parameters each time. A measure of confidence is computed for each boundary between shots, which was identified as scene boundary over the total number of generated video STGs and separately on the total number of audio STGs. These confidence values are linearly combined and all shot boundaries for which the resulted confidence value exceeds a threshold will form the set of scene boundaries. The approach was tested over only three documentary films and research still needs to be done for the optimization of the weights controlling the combination of the audio and visual graphs. In [14] audio and visual features are extracted for every visual shot. Using these features and a Support Vector Machine (SVM) each shot boundary is classified as scene change/non-scene change. The drawback of this method it that it requires

the availability of sufficient training data.

Based on the fact that scenes in videos are constructed by producers based on some cinematic rules, Wang *et al.* [44] use the concept of continuity to model these rules and extract the scene boundaries. After segmenting the video into shots, the framework successively (from lower level to higher level features) applies the concept of visual, position, camera focal distance, motion, audio and semantic continuity to group into scenes the shots that exhibit some form of continuity. An example of relation between cinematic rules and continuity might be that "visual continuity exists between shots with similar background" but "similar background models the scenes that happen at the same time and in the same location". For the case of audio continuity, "the same scene should possess the similar environment sound and dialog by the same speakers, especially for dialog scenes". The framework is tested using the first three levels of continuity to extract the scenes defined using most common cinematic rules.

Yamamoto *et al.* [39] propose another approach for semantic segmentation of TV programs based on the detection and classification of corner subtitles. These are considered as indicating the structure of a program based on the fact that they stay on the screen and switch at semantic scenes changes. Corner subtitles with similar features are grouped as relative subtitles, based on their color, location on the screen and segment location on the time axis. Chapter points are then detected according to the distribution of the relative subtitles. Three patterns of the program are manually defined according to broadcasted TV programs.

## 1.4 Conclusion

As stated in the introduction, TV program structuring is very important for a better understanding and an efficient organization of the video content. It facilitates tasks like video indexing, video classification and summarization and provides fast and nonlinear access to relevant parts of the programs (nonlinear video browsing). Its domain of applicability is very vast: in sports videos for the detection of moments of interest (e.g. goal in a soccer

match, pitches in baseball) and also for kinematic analysis for sports professionals training purposes [26]; in news broadcast programs to identify the different stories and/or build personalized TV news programs [29]; in entertainment TV shows to recover their structure and identify the main parts allowing the browsing inside such programs; in films to provide the chapters of the film; in home videos to permit the organization of the videos related to certain events. All the methods presented in the previous sections try to solve one ore more of these requirements. Both, specific and generic approaches have strengths and weaknesses. For the case of specific methods, although they give promising results, they can be applied to only very specific types of programs. They are characterized by a lack of generality due to the use of rules, templates or learning algorithms based on a previous analysis of the specific videos. Also when dealing with videos more complex than news or sports videos the computation cost could increase and an important knowledge in the domain will be needed.

On the other hand, generic methods try to structure a video without using any prior knowledge in the domain. But because the definition of a scene is very ambiguous and depends on the human understanding of its meaning, it is difficult to find an objective one, that would cover all users interests. This is why it is hard to compare the performance of the existing approaches and develop a better one. An objective evaluation would assume the existence of a ground truth (GT), whereas this GT is human generated so it depends on human judgment.

Even so, the advances in the field offer, as best as possible, solutions to the problems in the video content domain. A great amount of work continues to be done, with the goal to propose methods that are completely unsupervised and require as minimum as possible the human intervention. Services like Catch-up TV, that make use of these approaches, become more and more interesting for the TV providers and their clients, so their practical applicability is very important for the close future.

## References

[1] Jrgen Assfalg, Marco Bertini, Carlo Colombo, and Alberto Del Bimbo. Semantic annotation of sports videos. *IEEE MultiMedia*, 9(2):52 – 60, April 2002.

[2] Yannis S. Avrithis, Nicolas Tsapatsoulis, and Stefanos D. Kollias. Broadcast news parsing using visual cues: a robust face detection approach. In *IEEE International Conference on Multimedia and Expo*, number 3, pages 1469 – 1472, New York , USA, August 2000.

[3] M. Bertini, A. Del Bimbo, and P. Pala. Content-based indexing and retrieval of tv news. *Pattern Recognition Letters*, 22(5):503–516, April 2001.

[4] Marco Bertini, Alberto Del Bimbo, and Walter Nunziati. Automatic detection of player's identity in soccer videos using faces and text cues. In *14th annual ACM international conference on Multimedia*, pages 663–666, Santa Barbara, CA, USA, October 2006.

[5] Lekha Chaisorn and Tat-Seng Chua. Story boundary detection in news video using global rule induction technique. In *IEEE International Conference on Multimedia and Expo*, pages 2101–2104, Toronto, Ontario, Canada, July 2006.

[6] Lekha Chaisorn, Tat-Seng Chua, and Chin-Hui Lee;. The segmentation of news video into story units. In *IEEE International Conference on Multimedia and Expo*, number 1, pages 73 – 76, Lusanne, Switzerland, August 2002.

[7] Lekha Chaisorn, Tat-Seng Chua, and Chin-Hui Lee. A multi-modal approach to story segmentation for news video. *World Wide Web*, 6(2):187–208, 2003.

[8] Angelo Chianese, Vincenzo Moscato, Antonio Penta, and Antonio Picariello. Scene detection using visual and audio attention. In *ACM Int. Conf. on Ambi-Sys workshop on Ambient media delivery and interactive television*, Quebec, Canada, February 2008.

[9] Ling-Yu Duan, Min Xu, Tat-Seng Chua, Qi Tian, and Chang-Sheng Xu. A mid-level representation framework for semantic sports video analysis. In *ACM international conference on Multimedia*, pages 33 – 44, Berkeley, CA, USA, November 2003.

[10] S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll. Content based indexing of images and video using face detection and recognition methods. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1505–1508, Salt Lake City, Utah, USA, May 2001.

[11] Ahmet Ekin, A. Murat Tekalp, and Rajiv Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7):796–807, July 2003.

[12] Yong Fang, Xiaofei Zhai, and Jingwang Fan. News video story segmentation. In *12th International Multi-Media Modelling Conference*, pages 397–400, Beijing, China, January 2006.

[13] Xinbo Gao and Xiaoou Tang. Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(9):765 – 776, September 2002.

[14] Naveen Goela, Kevin Wilson, Feng Niu, and Ajay Divakaran. An svm framework for genre-independent scene change detection. In *IEEE International Conference on Multimedia and Expo*, pages 532–535, Beijing, China, July 2007.

[15] B. Gunsel, A. Ferman, and A. Tekalp. Temporal video segmentation using unsupervised clustering and semantic object tracking. *Journal of Electronic Imaging*, 7(3):592–604, 1998.

[16] Abhinav. Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. Understanding videos constructing plots learning a visually grounded storyline model from annotated. In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, Florida, USA, June 2009.

[17] Andr Guziec. Tracking pitches for broadcast television. *Computer*, 35(3):38 – 43, March 2002.

[18] Raffay Hamid, Ram Krishan Kumary, Matthias Grundmannz, Kihwan Kimz, Irfan

Essaz, and Jessica Hodgins. Player localization using multiple static cameras for sports visualization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 731 – 738, San Francisco, CA, June 2010.

[19] A. G. Hauptmann and M. A. Smith. Text, speech and vision for video segmentation : The infomedia project. In *AAAI Fall Symposium, Computational Models for Integrating Language and Vision*, Cambridge, Massachusetts, USA, November 1995.

[20] E. Kijak. *Structuration multimodale des vidos de sports par modles stochastiques*. PhD thesis, Universit de Rennes 1, 2003.

[21] E. Kijak, L.Oisel, and P.Gros. Hierarchical structure analysis of sport videos using hmms. In *International Conference on Image Processing*, volume 3, pages 1025–1028, September 2003.

[22] Jae-Gon Kim, Hyun Sung Chang, Young tae Kim, Kyeongok Kang, Munchurl Kim, Jinwoong Kim, and Hyung-Myung Kim. Multimodal approach for summarizing and indexing news video. *ETRI Journal*, 24(1):1–11, 2002.

[23] Kihwan Kim, Matthias Grundmann, Ariel Shamir, Iain Matthews, and Jessica Hodgin-sand Irfan Essa. Motion fields to predict play evolution in dynamic sport scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 840 – 847, San Francisco, CA, June 2010.

[24] Chien-Chuan Ko and Wen-Ming Xie. News video segmentation and categorization techniques for content-demand browsing. In *Congress on Image and Signal Processing*, pages 530 – 534, Sanya, Hainan, China, May 2008.

[25] Francis C. Li, Anoop Gupta, Elizabeth Sanocki, Li wei He, and Yong Rui. Browsing digital video. In *ACM Computer-Human Interaction*, pages 169 – 176, Hague, The Netherlands, 2000.

[26] Haojie Li, Jinhui Tang, Si Wu, Yongdong Zhang, and Shouxun Lin. Automatic detection and analysis of player action in moving background sports video sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(3):351 – 364, March 2010.

[27] Andrew Merlino, Daryl Morey, and Mark Maybury. Broadcast news navigation using story segmentation. In *ACM international conference on Multimedia*, pages 381–391, Seattle, WA, USA, November 1997.

[28] Hemant Misra, Frank Hopfgartner, Anuj Goyal, P. Punitha, and Joemon M. Jose. Tv news story segmentation based on semantic coherence and content similarity. In *16th International Multimedia Modeling Conference*, pages 347–357, Chongqing, China, January 2010.

[29] Bernard Mrialdo, Kyung-Tak Lee, Dario Luparello, and Jeremie Roudaire. Automatic construction of personalized tv news programs. In *7th ACM International Multimedia Conference*, pages 323–331, Orlando, Florida, USA, October 30-November 5 1999.

[30] Yuichi Nakamura and Takeo Kanade. Semantic analysis for video contents extraction - spotting by association in news nideo. In *ACM international conference on Multimedia*, pages 393–401, Seattle, WA, USA, November 1997.

[31] Paul Over, Alan F. Smeaton, , and George Awad. The trecvid 2007 bbc rushes summarization evaluation pilot. In *International workshop on TRECVID video summarization*, Augsburg, Germany, September 2007.

[32] Paul Over, Alan F. Smeaton, , and George Awad. The trecvid 2008 bbc rushes summarization evaluation. In *2nd ACM TRECVid Video Summarization Workshop*, pages 1–20, Vancouver, BC, Canada, October 2008.

[33] Silvia Pfeiffer, Rainer Lienhart, and Wolfgang Efflsberg. Scene determination based on video and audio features. *Multimedia Tools and Applications*, 15(1):59–81, September 2001.

[34] Marcus J. Pickering, Lawrence Wong, and Stefan M. Rger. Anses: summarisation of news video. In *Proceedings of the 2nd international conference on Image and video*

*retrieval*, pages 425–434, Urbana, IL, USA, July 2003.

[35] J-P. Poli. *Structuration automatique de flux televisuels*. PhD thesis, Universite Paul-Cezanne Aix-Marseille III, 2007.

[36] Zeeshan Rasheed and Mubarak Shah. Detection and representation of scenes in videos. *IEEE transactions on multimedia ISSN 1520-9210*, 7(6):1097–1105, December 2005.

[37] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *ACM international conference on Multimedia*, pages 105 – 115, New York, USA, October 2000.

[38] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, and Isabel Trancoso. Multi-modal scene segmentation using scene transition graphs. In *17th ACM International Conference on Multimedia*, pages 665–668, Beijing, China, October 2009.

[39] Koji Yamamoto Shunsuke Takayama and Hisashi Aoki. Semantic segmentation of tv programs using corner-subtitles. In *IEEE 13th International Symposium on Consumer Electronics*, pages 205 – 208, Kyoto, Japan, May 2009.

[40] Wallapak Tavanapong and Junyu Zhou. Shot clustering techniques for story browsing. *IEEE Transactions on Multimedia*, 6(4):517 – 527, August 2004.

[41] Dian W. Tjondronegoro and Yi-Ping Phoebe Chen. Knowledge-discounted event detection in sports video. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 40(5):1009–1024, September 2010.

[42] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1):1–37, February 2006.

[43] E. Veneau, R. Ronfard, and P. Bouthemy. From video shot clustering to sequence segmentation. In *15th International Conference on Pattern Recognition*, volume 4, pages 254–257, Barcelona, Spain, September 2000.

[44] Jihua Wang and Tat-Seng Chua. A cinematic-based framework for scene boundary detection in video. *The Visual Computer*, 19(5):329–341, 2003.

[45] Jinqiao Wang, Lingyu Duan, Qingshan Liu, Hanqing Lu, and Jesse S. Jin. A multimodal scheme for program segmentation and representation in broadcast video streams. *IEEE Transactions on Multimedia*, 10(3):393 – 408, April 2008.

[46] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recognition Letters*, 25(7):767 – 775, May 2004.

[47] Ziyou Xiong, Xiang Sean Zhou, Qi Tian, Yong Rui, and Thomas S. Huangm. Semantic retrieval of video - review of research on video retrieval in meetings, movies and broadcast news, and sports. *IEEE Signal Processing Magazine*, 23(2):18 – 27, March 2006.

[48] Gu Xu, Yu-Fei Ma, Hong-Jiang Zhang, and Shi-Qiang Yang. An hmm-based framework for video semantic analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(11):1422 – 1433, November 2005.

[49] Peng Xu, Lexing Xie, Shih-Fu Chang, A. Divakaran, A. Vetro, and Huifang Sun;. Algorithms and system for segmentation and structure analysis in soccer video. In *IEEE International Conference on Multimedia and Expo*, pages 721 – 724, Tokyo, Japan, August 2001.

[50] M. M. Yeung and B. Liu. Efficient matching and clustering of video shots. In *International Conference on Image Processing*, volume 1, pages 338–341, Washington D.C., USA, October 1995.

[51] M.M. Yeung and Boon-Lock Yeo. Time-constrained clustering for segmentation of video into storyunits. In *13th International Conference on Pattern Recognition*, volume 3, pages 375–380, Vienna, Austria, August 1996.

[52] Xinguo Yu, Liyuan Li, and Hon Wai Leong. Interactive broadcast services for live

soccer video based on instant semantics acquisition. *Visual Communication and Image Representation*, 20(2):117–130, February 2009.

[53] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(2):168 – 186, February 2007.

[54] Yun Zhai and Mubarak Shah. Video scene segmentation using markov chain monte carlo. *IEEE Transactions on Multimedia*, 8(4):686 – 697, August 2006.

[55] Dong-Qing Zhang, Ching-Yung Lin, Shi-Fu Chang, and John R. Smith. Semantic video clustering across sources using bipartite spectral clustering. In *IEEE International Conference on multimedia and expo ( ICME )*, pages 117– 120, June 2004.

[56] HongJiang Zhang, Yihong Gong, S.W. Smoliar, and Shuang Yeo Tan. Automatic parsing of news video. In *International Conference on Multimedia Computing and Systems*, pages 45 – 54, Boston, Massachusetts, May 1994.

[57] Yanjun Zhao, Tao Wang, Peng Wang, Wei Hu, Yangzhou Du, Yimin Zhang, and Guangyou Xu. Scene segmentation and categorization using ncuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 343 348, Minneapolis, Minnesota, USA, June 2007.