



Segmentation and Annotation of Video Data: A Survey

Ba Tu Truong

Institute for Multi-Sensor Processing & Content Analysis (IMPCA)
Department of Computing
Curtin University of Technology
Western Australia

Svetha Venkatesh

Institute for Multi-Sensor Processing & Content Analysis (IMPCA)
Department of Computing
Curtin University of Technology
Western Australia

In this paper, we present a graph-based visualization concept called the Double-Ring Take-Transition-Diagram (DR-TTD) that can capture and express the internal structure of a film scene and its editing patterns. The DR-TTD representation exhibits essential properties such as fully automatic construction, compactness, clarity, temporal perseverance and explicitly links to semantics. It presents takes and their transitions via nodes and edges of a 'graph' consisting of two rings as its backbone. All steps in the DR-TTD construction, including node classification, connecting nodes via edges and unit linking, are motivated from the understanding of film grammar for shot arrangement. In addition, there are signatures for filmic and semantic elements made explicit in this representation and these include dialogue, moving between zones/dramatic progression, shot association, introduction and resolution, master shot, non-dialogue narration and film editing orchestration.

Department of Computing, Curtin University of Technology, Perth, Western Australia

Contents

1	Multimedia Content Management	1
1.1	Segmentation and Reconstruction of Structure	2
1.1.1	Shots	2
1.1.2	Logical Units	5
1.2	Content Annotation	13
1.2.1	Semantic Segmentation from Base Unit Classification	13
1.2.2	Clip/Program Level	16
1.2.3	Logical Unit/Sequence of Shots Level	19
1.2.4	Shot/Sequence of Frames Level	21
2	Summary	23

Institute for Multi-Sensor Processing & Content
Analysis (IMPCA)
Department of Computing
Curtin University of Technology
Western Australia

tel: +61 8 9266 7647
fax: +61 8 9266 2819
<http://impca.cs.curtin.edu.au/>

Corresponding author:
Ba Tu Truong
truongbt@cs.curtin.edu.au

As set out in the scope of this research, the domain of this work is Film and the mission is to investigate the automatic extraction of structural and expressive elements in Film. However, it is important and necessary to consider the problem within a broader context of Multimedia Content Management (MCM) because:

- All data classified as audio-visual documents including movies, sports, news, training videos, etc. shares common properties and relationships.
- The task of film understanding and film analysis share the same application setting (e.g., digital library, videos on demand) as with other video content management tasks such as video genre identification, sports event detection, video summary generation, etc.
- As a relatively new research area, comparing to computer science research in general, it is useful and interesting to have a complete look at the field from a global perspective. The overview reveals applicability as well as limitations of existing techniques when applied to the film domain.

This chapter is organized into two sections. Section 1 accounts for the majority of this chapter, in which a systematic review of state-of-the-art research in strongly correlated aspects of a MCM system is provided, and these aspects include temporal video segmentation and the reconstruction of structures, semantic labeling of extracted structural units and abstraction of video documents for content navigation and browsing. Section ?? describes the role of Film Grammar in automatic understanding and analysis of film content.

1 Multimedia Content Management

The ultimate purpose of an automatic MCM system is to facilitate the user access to video documents. There are two common mechanisms for achieving this purpose: automatic content annotation and video summarization.

The first mechanism is referred to as automatic video indexing and is defined as the process of automatically assigning content-based labels to video documents. A video indexing solution needs to deal with three issues:

1. Identifying the unit \mathbf{U} to be indexed. This issue is related to granularity. Other than clip/frame level of annotation, a temporal video segmentation process is required, which often divides a video document into shots or sequence of shots. This process is called segmentation and reconstruction of structures and shall be reviewed in Section 1.1.
2. Identifying perspective ρ . The perspective/attribute ρ of the unit under which a label, discrete or continuous, should be computed. This perspective will determine the domain range for computed labels. For example, if the unit is *shot*, perspective ρ is *size*, then the value domain can be {Close-Up, Medium Close-Up, ..., Long Shot, Very Long Shot}.
3. Computing v . How can we compute the perspective value v for a particular unit \mathbf{U} with respect to perspective ρ , $v = \mathbf{U}^\rho$. If the value is *nominal*, it is often considered a classification problem. Otherwise, v is often computed as a continuous function of constructive elements, e.g., tempo is computed from motion and shot length. This process requires the extraction of appropriate features from the video sequence and supervised learning mechanisms and shall be target in Section 1.2.

So far our discussion seems to indicate that identifying a structural unit \mathbf{U} and assigning a label v are two separate steps in a video indexing solution. In practice, both these steps can blend into one single integrated process with no such sequential ordering, which we shall term *semantic segmentation*. The difference between semantic segmentation and ‘pure’ segmentation is that in semantic segmentation, the perspective ρ itself determines the boundaries between structural units. For example, shot transition detection is a pure segmentation process, as extracted shots do not assume any semantic label. The same applies to scene extraction in a movie. On the other hand, the segmentation of a soccer program into *play* and *break* sequences is a semantic segmentation process, because the perspective $\rho = \text{GameState}$ decides how the game should be segmented. The same is applied to segmentation of commercial blocks from a TV broadcast stream.

Apart from serving as an intermediate step in video indexing, video segmentation and reconstruction of the structures play a significant role in facilitating video content browsing and navigation, allowing organization of video data according to their temporal structures and relations, e.g., table of content generation. One equally important mechanism for enhancing the user access to video documents is video abstraction which requires no content labeling and annotation. The purpose of video abstraction is to retain only portions of video content (as a set of images or a video skim) that are relevant to a certain perspective ρ or to maximize the relevance of the extracted abstract given a time constraint. This shall be detailed in Section ??

It should be noted that, in this review, we opt to overlook work in the area of video retrieval/query/search by examples. Video retrieval can be considered as a two-class classification problem (i.e., relevant and non-relevant), in which perspective the ρ needs to be abstracted from an example clip given by the user. Significant in its own right, however, research in this area is considered irrelevant to this thesis.

In addition, this review shall focus on the MCM problem targeted in each work as well as the techniques and features it employ. It does not focus on the strength, limitations and performance of a specific work, as meaningful characterization of these aspects is almost impossible since experimental setups are widely different.

1.1 Segmentation and Reconstruction of Structure

1.1.1 Shots

Definition 1 *Shot* - A shot $\mathbf{S} = \{\mathbf{F}_i, \mathbf{F}_{i+1}, \dots, \mathbf{F}_j\}$ is defined as a stream of $j - i + 1$ frames continuously recorded by a single camera. That is, $\mathbf{F}_i^\rho = \mathbf{F}_{i+1}^\rho = \dots = \mathbf{F}_j^\rho = v$, $\mathbf{F}_{i-1}^\rho \neq v$, where $\mathbf{F}_{j+1}^\rho \neq v$ and perspective ρ is camera operation ID recorded during film shooting.

Definition 2 *Shot Transition* - A transition \mathbf{T}_i between shot \mathbf{S}_i and \mathbf{S}_{i+1} is defined as a set of consecutive frames, possibly empty set, used to join these two shots to create a continuous video stream.

There are different kinds of transitions between shots. If $\mathbf{T}_i = \emptyset$, the transition is called a *cut*, otherwise the transition is *gradual*. The cut is an instantaneous change from one shot to another and can be seen as the shortest distance between two shots. Gradual transitions are mainly of three kinds $\{\textit{fade}, \textit{dissolve}, \textit{wipe}\}$. There are two types of fades: *fade-in* and *fade-out*. A fade-out occurs when the picture information gradually disappears, leaving a blank screen. A fade-in occurs when the picture gradually appears from a blank screen. A fade-out to or fade-in

from black is the most common; however, it is possible to fade-out to or fade-in from any other color. A dissolve occurs when one whole picture fades away while another whole picture appears. A wipe is an optical effect in which successive full-strength images from one shot is progressively replaced or pushed away or compressed by successive full-strength images from another shot. Images from the second shot may appear at the side of the screen, the corner, the middle or several places at once, gradually taking over the screen by following some geometric pattern.

Detecting shot transitions is an important task in video content analysis because the shot is a fundamental unit of manipulation in video production and hence video indexing, representation and retrieval. Therefore, a large body of work has been published over the years addressing the shot transition detection problem and many comprehensive reviews have been presented. Here we shall present only a brief overview, and readers are referred to [1, 2] and references therein for a more comprehensive treatment of the problem.

Successive frame differences The earliest and simplest technique for hard cut detection uses pixel-wise comparison of successive frames as done in [3]. Some well known hard scene cut detection techniques employ color histograms which are compared using different metrics [4, 5,] or with *twin-comparison* [6,]. The algorithms primarily differ in the selection of an appropriate color space to compute the histograms and in the functions used to determine the similarity between two histograms.

Production models Much research in gradual transition detection has been carried out by analyzing the production models of these effects [7, 8,]. Instead of exploiting intensity changes in individual pixels in video frames, [9, 10, 11, 12,] investigate the effect of the production models on frame luminance mean and variance. During an ideal dissolve or fade, the mean changes in a linear manner, while the variance has a parabolic or half parabolic shape. [9] detects dissolves by first recording all negative spikes in the second order derivative of frame variances and ensuring that the luminance means within a dissolve region do not change sign. [10] record all successive peaks and the valleys between them on the variance curve as indications of parabolic regions caused by dissolves. Conditions are further imposed to verify that the dissolves are wide enough and the valleys are deep enough. [11] detects fades by recording all negative spikes in the second derivative of the variance curve and ensuring that the first derivative of the mean curve is relatively constant next to a negative spike. [12] detects fades by fitting a regression line on the frame standard deviation curve.

Compressed domain techniques Some techniques utilize information available in MPEG compressed video streams to detect scene discontinuities such as DCT coefficients [13, 14,], bit rate [15, 16, 17,], motion vectors/macro block information [10, 18, 19, 20, 21, 22,]. These compressed-domain techniques often trade the detection accuracy for speed and efficiency by eliminating the need for full-decoding of a compressed video stream.

Threshold selection mechanisms Equally important to finding appropriate features and metrics to compare two consecutive frames is the problem of interpreting the frame difference values such that certain values can be selected to indicate shot changes. Proper selection of difference values is usually done by setting thresholds. [6] propose a statistical approach for

determining the threshold, based on the mean value μ and the standard deviation δ of frame-to-frame differences. A window based approach in which shot detection takes place at the central value of a temporal window [4, 23, 24,] can improve thresholding since it is more appropriate to treat a shot change as a local activity. One requirement with the window-approach is that the window size should be set so that it is unlikely that two shots occur within the window. Therefore, the center value in the window must be the largest frame-to-frame difference in the window. [4] select the threshold based on the second largest value within the window. [23] divide all points within the window into two clusters depending on the distance from each point to the largest and the smallest points. The threshold is then set based on the distance between the two clusters. If the distance is below a threshold, the threshold will be set just above the upper cluster; otherwise, it will be set as half of the distance between the clusters. [24] combine the sliding-window approach and general statistical models for the frame-to-frame difference curve to detect hard cuts.

Probabilistic solutions In moving toward threshold-free solutions to shot transition detection, probabilistic approaches have recently been adopted. [25] propose a statistical framework for the shot transition detection based on minimization of the average detection error probability criterion. The statistical functions are formulated from motion compensation features and the knowledge about shot length distribution (assumed to be a Poisson distribution). Different kinds of shot transitions can be detected under this framework. A similar approach is employed in [26] in which two distribution models of the shot length are experimented with: *Erlang* and *Weibull*. Instead of modeling shot length, [27] statistically solves the threshold selection problem when the measurement for shot transition and non shot transition are modeled as two *Gaussian* distributions or two *Gamma* distributions. [28] build a Hidden Markov Model (HMM) with states being {shot, fade, dissolve, cut1, cut2, pan, zoom} to segment a video sequence into shots together with camera motion regions. In addition to the image-based distance between two successive frames, the audio distance based on acoustic difference between intervals just before and after the frames is used as the discriminating feature. [29] present a shot transition detection method based on the representation of visual contents in a video using coupled Markov chains. Shot transitions are detected by computing of how well the observations attached to the next frame in the sequence fit a probability distribution obtained from the previous frames.

Temporal slice analysis Temporal slice analysis has been proposed as an alternative to the frame difference approach. A slice is an 1D image taken from a frame, while a spatio-temporal image is a collection of slices in the sequence at the same position. Vertical, horizontal, primary diagonal and subprimary diagonal slices are most frequently used. Table 1 shows the patterns of the temporal-slice image for different shot transition types. In general, a camera cut results in vertical boundary lines; a wipe results in slanted or curved boundary lines; while a dissolve connects two regions slowly and does not have a clear boundary [30,]. The task of detecting shot transitions is therefore equivalent to the task of segmenting the image into regions. [30] proposed a Markov-based image segmentation algorithm to locate the color texture discontinuities at region boundaries to detect wipes and cuts. Dissolves are detected by checking that the luminance means of slices are approximately constant and while their variances forms a concave upward parabolic shape. They also utilize MPEG motion vectors and DCT coefficients to eliminate false positives in detection hard cuts. [31] first form the derivative image of a temporal-slice image by taking the absolute difference of two adjacent pixels on the same vertical

line. Cuts are detected by looking for peaks in the sums of all columns of the derivative image. Wipes are detected by first recording all peaks in the vertical lines of the derivative image and then the connectivity of these peaks are checked. Rather than using specific horizontal, vertical or diagonal slices, [32] make use of every slice of the spatio-temporal image, which is generated from a histogram-based differencing scheme rather than from the pixel values themselves. Cuts, wipes and dissolves are detected in this work.



















Slices	Cut	Dissolve	Wipe			
			l-to-r	r-to-l	t-to-b	b-to-t
Horizontal						
Vertical						
Diagonal						

Table 1: Illusion of edit effects on temporal slice images (Adapted from [33,]).

1.1.2 Logical Units

Partitioning a video sequence into shots can be considered as a fundamental or a low-level video segmentation process. This process does not rely on the semantic content of the underlying video. Unlike shots, there are no physical markers to indicate scene boundaries; instead, high-level units are determined logically by semantic boundaries. In this section, we characterize the research dealing with the problem of finding such semantic boundaries which we shall assign the generic term, *logical units*. This term is adapted from the term *logical story unit* by omitting the word ‘story’ - allowing it to refer to any attributes consistently observed in each video segment/unit.

Definition 3 *Logical Unit* - A logical unit \mathbf{U} through a perspective ρ is a set of consecutive shots $\{\mathbf{S}_i, \mathbf{S}_{i+1}, \dots, \mathbf{S}_j\}$ that have the same value for ρ . That is, $\mathbf{S}_i^\rho = \mathbf{S}_{(i+1)}^\rho = \dots = \mathbf{S}_j^\rho = v$ while $\mathbf{S}_{i-1} \neq v$ and $\mathbf{S}_{j+1} \neq v$.

This definition also indicates that all logical unit segmentation techniques use shots as the building blocks and assume that shot indices are already available.

Specific names Logical units can assume different names depending on the perspective ρ as well as the convention used in a particular work and underlying video genre.

- *Scene*. The most appropriate application of the term scene is to narrative-driven fictional video content such as movies and TV sitcoms. In this domain, the common definition of the term scene is as follows:

Definition 4 *Scene* - A scene is a logical unit where the perspective ρ is time, locale and dramatic incident. In other words, a scene is a set of contiguous shots that are unified by time, locale and dramatic incident.

- *Topic*. Often associated with news, documentaries, and training and educational videos, the common perspective in this case is the topic under discussion in the video sequence. The definition of a topic can be expressed in terms of a logical unit as follows:

Definition 5 *Topic* - A topic is a logical unit where the perspective ρ is the subject of discussion.

- *Story unit*. Sometimes, the term story is used. In the narrative driven fictional video domain, it is equivalent to a scene. In other domains, it is similar to a topic.
- The literature also contains various other terms including logical story unit, episode, video paragraph and macro segments. These terms often refer to the scene or the topic. At a higher level than the scene, a sequence can also be seen as a logical unit. It is defined as a unit of film composed of a number of interrelated shots or scenes that together comprise an integral segment of the film narrative. For example, rehearsal, performance, and review of a concert together constitute a sequence.

Figure 1 shows an overview of different aspects of a logical unit extraction technique drawing upon our literature survey and each is discussed in detail below.

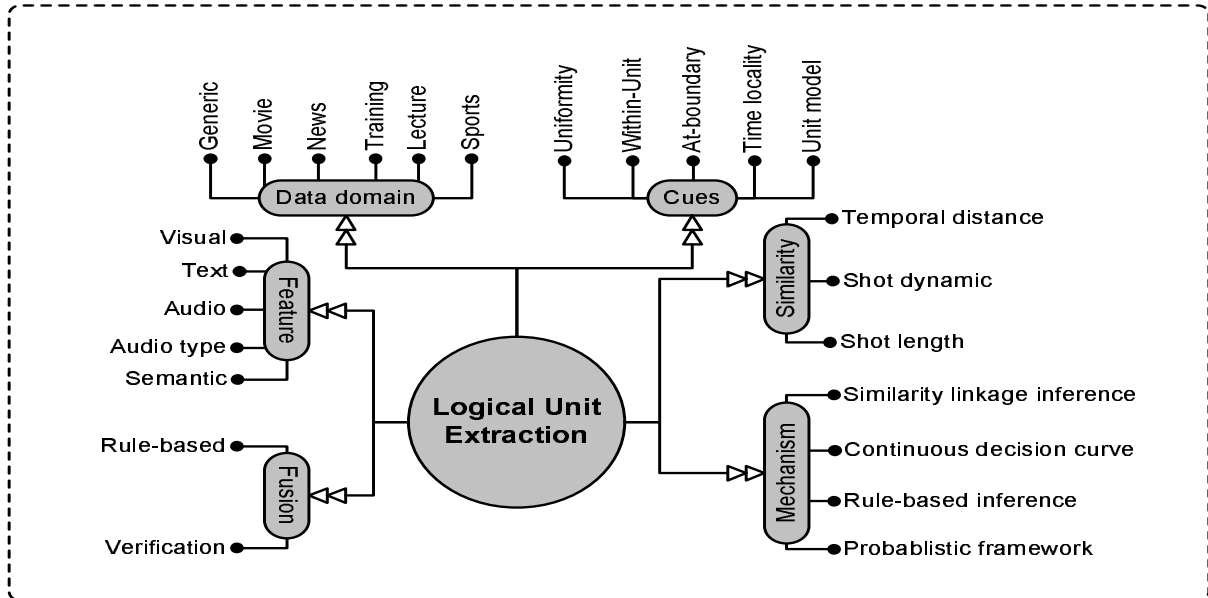


Figure 1: Attributes of logical unit segmentation techniques.

Data domain It is important to know what data domain a specific work has been developed and experimented on, since the proposed features and techniques are often fitted to the underlying data. For example, extracting topics in news often relies on the detection of anchorperson

shots. Although the greater the knowledge we have about the attribute of the data domain, the better the understanding we can develop about the portability of developed techniques (e.g., algorithms developed for ABC news may not work well on SBS news, techniques developed on drama film may not be as effective on action film), we will limit the domain knowledge to genre. Segmentation of a video sequence into logical units has been applied to *movies/sitcoms* (e.g., [34, 35]), *news* (e.g., [36, 37,]), *sports* (e.g., [38]) and *training videos* (e.g., [39, 40]). Some work has aimed to work across many genres, which we will classify as *generic*. Genre classification for individual work can be found in Tables 2 and 3. Furthermore, the literature shows that knowing the data domain of the work allows us to identify the perspective ρ being investigated in the work. For example, all work in the movie/sitcoms/generic domain deals with the scene extraction problem, while work in the news/documentaries/training domain tackles the topic segmentation problem.

Cues Once the target data domain is specified, the next step in developing a logical segmentation technique is researching the domain to discover important cues. We identify the following cues that have been exploited in the literature for logical unit segmentation tasks.

- *Unit uniformity.* Because each individual shot of a unit must contribute towards the perspective ρ or express the unit attribute, there must be similarity and commonality amongst them. The uniformity of a logical story unit manifests as the similarity of visual, audio and textual content amongst its shots. The visual similarity results from the fact that most, if not all, shots capture the same setting conforming to the 180-degree rule. Also, shots from the same take appear throughout the scene. With respect to audio, the similarity of audio content manifest as similar background noises that exist with a particular scene setting. Furthermore, similar acoustic characteristics should be observed in shots that contain the speech of the same person. For textual content, certain key-words, e.g. “Iraq”, can be spotted at different times during a news article/topic.
- *Rule-based knowledge about boundary shots.* The above cues are about the global property of a unit. However, the starting and ending shots of a logical unit are special shots and they may exhibit certain unique characteristics. Semantically, the first few shots of a logical unit often serve as its introduction, while the last shots resolve the unit. Therefore, it has been known that the filmmaker deliberately inserts certain devices/transitions into these shots to enhance the flow of the narration. News topics are often opened and closed with anchorperson shots. According to [41], if there is no music during 30 seconds and music appears in the soundtrack, then there is a scene boundary at the beginning of the music. [42] observe that a relatively long silence duration exists at the boundary of two news topics.
- *Rule-based knowledge about within-unit shots.* In contrast to the previous cues, cues of this kind tell us about specific attributes of the shots that eliminate them from being at the boundary. For example, [41] identify the rule about camera work that states if there are 3 or more simple shots with the same camera work (other than still shot) then they belong to the same sequence. Rule-based knowledge about within-unit shots helps in reducing the false positive boundaries.
- *Logical unit model.* There may exist the complete model for a logical unit. For example, it is a widely exploited assumption that a news story consists of an introduction shot followed

by the report shots. A more detailed news model is proposed by [43] which includes other elements such as shot transitions, their types, interview sequences, weather forecast, news introduction and news conclusion.

- *Time locality.* This attribute is related specifically to unit uniformity and almost all scene segmentation techniques exploit it to a certain degree. Basically, it indicates that within a temporal window, we should be able to find a shot similar to the current shot or the scene has changed. In other words, in computing shot association or unit uniformity it is not necessary and probably more robust not to include shots too far away.

Underlying mechanisms Given segmentation cues and extracted features, the next step is to develop an effective mechanism for extracting candidate boundaries. We identify the following three basic mechanisms for logical unit segmentation that can be used alone, combined or further refined in subsequent steps.

- *Similarity linkage inference.* This approach is sometimes referred to as the *overlapping links* method [34,]. Based on unit uniformity cues, this early approach proceeds by first identifying pairs of similar shots in the video sequence and linking them. The start of a new scene is often declared at the shot where no shot before it is linked to a shot after it, e.g., shot 9 in Figure 3. One special case of this approach uses cut edges of Scene Transition Graph formed from shot clusters as an indicator of shot scene boundaries [44,]. In this case, links are assigned to all pairs of shots belonging to the same cluster. Other work employing the overlapping link mechanism includes [41, 45, 46, 47], and the graph-based mechanism is also used by [48, 49, 50, 51].

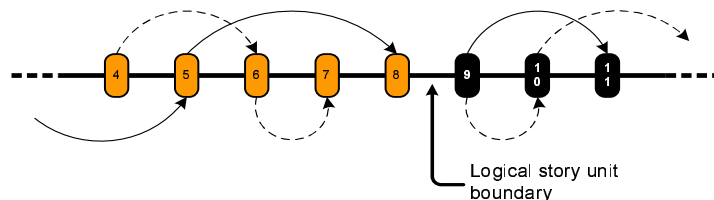


Figure 2: An example of similarity linkage.

- *Decision curve formulation.* Like the similarity linkage inference, this approach relies on the unit uniformity cues discussed previously. It evolves from the use of histogram difference for detecting shot transitions. Rather than relying on the discrete link inference of the previous mechanism, it first extracts a certain continuous attribute for each shot that links with whether or not a logical story unit boundary occurs at that shot. Often, this feature indicates the level of uniformity around a shot either in *visual terms* [52, 53, 35, 54, 55, 56, 57, 58, 59,], *audio terms* [60, 61, 62, 63, 64,] or *textual terms* [65,]. The boundaries between logical story units are detected by searching for local minima/maxima, thresholding, or a combination of these schemes. For visual features, this attribute is

computed by evaluating the association of shots preceding the current shot and those that follow. Note that, in terms of base units, shots may be replaced by a fixed size clips [63,] or sentences [65,] for audio features and texts respectively. Different methods arise by varying the choice of shot pairs, their association measure and the combination of resulting association values. For example, in [56], the shot pairs consist of the current shots and one of those shots within a temporal window; the similarity measure is a linear combination of color histogram difference and number of in-between shots; and the combination is the ratio of total difference in the left window over the total difference in the right window. Methods employing memory models [64, 35, 52,] can also be considered as belonging to the decision curve formulation approach and the uniformity attribute is termed *recall*. In [66], a formulation of tempo for individual shots is proposed based on an linear combination of motion and shot length characteristics. Story boundaries are identified by detecting edges of the tempo function, acknowledging that a story change often involves a large tempo change.

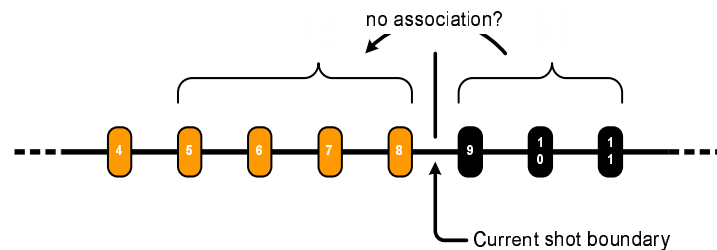


Figure 3: Formulating uniformity value across a shot boundary.

Various uniformity measures can also be used to complement each other. In this case, a function for combining these measures to create the final decision curve is required. To date, the linear function is used in the literature [64, 55, 67,]. [64] compute the overall uniformity in the audio signal by linearly combining three component uniformity features: periodicity, envelope and randomness. [55] formulate two features called “merging force” and “splitting force”. [67] compute two sets of attributes called “cohesions” and “disjunctions”. The former indicates the consistency of the content across a shot boundary. The latter indicates the inconsistency and is computed from editing information or special shots, which contain some semantic hints for the break points of the video sequence, e.g., black frames, silences and wipes.

- *Rule-based interpretation of shot semantics.* The above two approaches exploit the uniformity cues. This approach mainly exploits specific characteristics of boundary and within-unit shots. News topic segmentation based on the occurrence of anchorperson falls into this class. It also includes techniques that rely on the changes in audio types [68, 63,]. [41] propose an extensive set of rules regarding motion, audio types and shot transition types for locating scene boundaries in movies. Rules can also be incorporated into a Finite State Machine for the purpose of segmenting the news programs into a number of topics and

auxiliary segments [36,]. [69] uses sequence textual delimiters in production transcripts to locate scene boundaries. [68] identify shot boundaries when two shots exhibit different types of audio classes across the boundary or they exhibit the same class of audio signal except at the boundary across the two shots. No boundaries should be claimed when two adjacent shots exhibit the same audio signal for a certain period of time. [70] use rules regarding anchorperson shots and captions as the main mechanism for identifying news topics. For example, if there is only one video caption between two anchorperson shots, then shots between these two anchorperson shots belong to the same scene starting at the first anchorperson shot.

- *Probabilistic models.* Some work has focused on building a probabilistic framework for modeling the internal working of logical units, facilitating the detection of their boundaries. Based on an explicit news model, [43] propose a HMM approach that detects topic boundaries using audio and semantic features. [37] improve the Finite State Machine (FSM) approach in [36] by automatically constructing a HMM that has four hidden states (topic start, topic end, commercial and other) to model the news programs. Topic boundaries are declared when the executing HMM transits into the topic-start and topic-end states. Using features extracted from caption text, [38] employ a Bayesian network to classify an American football shot into live, replay, etc. The start of logical unit of the game is identified at the first live shot of any sequence of successive live shots.

It is worth noting that in order to exploit the time locality cue, most techniques based on the first two approaches employ a temporal window in computing the similarity/uniformity at a certain shot boundary to limit the influence of shots too far away from the scene boundary decision.

Features After identifying significant cues, one must identify computable features that best compute these cues. Work in the literature has exploited various features ranging from low-level to relatively high-level features for solving a particular logical story unit segmentation task. We identify the following five feature categories.

- *Visual.* Color histograms are the most frequently used features for computing unit uniformity cue [44, 61, 50, 35,]. Differences arise from the choice of color model and quantization method. Other visual representations include dominant color histograms [54, 55, 63,], spatial structure histogram [55,], and color auto-correlograms [51,]. [34] divide the image into blocks and subsequently use block-matching to compute the similarity between shots. [47] only extract visual features from pre-determined regions of the image. The Phase Correlation Function (PCF) is used in [61] to detect motion-based candidate boundaries.
- *Text.* Various textual sources have been exploited including closed caption, superimposed caption, speech and external transcripts. Textual data can be used either in measuring the uniformity via a similarity metric or in identifying boundary shots via the identification of key phrases and markers. News captions may contain special topic markers such as “>>>” [36, 71,]. The appearance of key-phrases in the introduction and conclusion of a news topic can also be exploited [36, 72, 73,]. [69] align production transcripts to the shot indices of a movie based on longest subsequent matching and then use the scene markers/keywords to locate scene boundaries. Word similarity can be used for topic segmentation in news,

documentaries and instructional videos. [65] compute the similarity between sentences extracted from close caption to detect news story boundaries. In their work, text analysis is performed to provide four keyword vectors: organizational, personal, locational and temporal. [40] use the frequency of superimposed text as one of the features for extracting main topics in training videos.

- *Audio*. This category refers to primitive audio features such as energy, envelope and etc. These low-level features are often used for constructing the decision curve or verification of candidate boundaries. [74] use volume, power and spectrum features to compute the uniformity. [61, 62] extract 12 primitive audio features for each clip to compute audio uniformity. Features used by [64, 35] include cepstral flux, multi-channel cochlear decomposition and cepstral vectors. In addition, envelope fit is used for some of these features.
- *Audio type*. The most common approach for integrating audio information into the logical unit segmentation process requires the audio stream to be segmented and/or classified into various types such as silence, music, speech, noise, environmental. The audio type change can be incorporated as an indicator of logical unit change or to verify candidate boundaries in a rule-based manner [68, 71, 63, 75, 76, 50, 58,]. [39, 40] use audio type together with visual elements such as the presence of faces to create more abstract features for topic boundary segmentation.
- *Semantics*. In realizing boundary and within-unit shot cues for logical unit segmentation task, the audio type is not the only semantic feature exploited. The anchorperson shot is the most common feature in story segmentation of news programs [57, 43, 70,]. [41] use shot transition devices and camerawork characteristics in their rule-based macro segmentation of movie. [58] use information about focal length (manually labeled) in the cinematic-based framework for the extraction of film scenes. [39] classify shots into Onscreen-Narration and Speech and use them to locate the topic changes in instructional videos from candidate segments. The presence of faces is used as one of features for identifying the main topic [40,]. [43] rely on the content class labeling (Begin, End, Newscaster, Report, Interview, and Weather Forecast) developed by [77] as the feature for topic segmentation of a broadcast news. [49] compute the similarity based on a combination of color histograms, dominant colors, camera motion types, number, size and position of faces.

Shot similarity attributes Logical unit segmentation techniques relying on the first two base mechanisms require a shot similarity measure. A good measure should discriminate the uniformity of shots within an unit against the dis-uniformity across the unit boundary. Various measures have been employed and we chose to characterize them according to the below three important aspects:

- *Temporal distance*. This kind of measure takes the temporal distance between two shots into consideration. It works on the basis that the association between two shots decreases if the temporal distance between them increases. [45] add to shot similarity measure a factor called *temporal attraction* that falls to zero according to the temporal distance between frames. They eliminate the use of temporal window (discrete in its nature) by a more continuous treatment of the time locality cue. [35] add a factor of $(1 - \Delta t/W)$ (Δt and W are the temporal distance between two shots and the window size respectively) to the

decision curve. [52] add a factor of $e^{-(\Delta t)/W}$ to account for the temporal distance between shots. [55, 54] introduce the factor $\frac{1}{1+\Delta t/C}$ into the similarity measure, where d is the temporal distance between two shots and C is a constant set to 20. [53] experiment with three different weighting functions: linear, parabolic and sinusoidal. Whilst the above authors treat temporal distance as the number of frames between two shots, it can also be measured in terms of the number of shots which separate them in [57, 56,].

- *Shot dynamics.* The activity level of two shots influences their similarity. [45] use average frame-to-frame difference to represent the activity of a shot. [46] compute the average motion over the temporal window and use it as the weighting factor in computing shot similarity.
- *Shot duration.* The similarity in the shot durations contributes toward the similarity between shots. For example, [35] make their decision curve proportional to the length of each of the shots on the grounds that if a shot is in a memory for a long period of time it will be recalled more easily. A similar approach is used in [52]. Other work that incorporates shot duration into the similarity measure includes [50].

Candidate boundary fusion and refinement This discussion is only applicable to methods that employ multiple cues separately. There are two basic ways for integrating these cues in order to produce the final list of logical unit boundaries.

- *Rule-based fusion of candidate boundaries.* Here, candidate boundary sets produced by separate cues using base mechanisms described earlier are fused together in a rule-based manner. For example, [35] combine audio scene breaks and visual scene breaks by merging the two sets and removing audio boundaries found within a window from a visual scene boundary. Similarly, [63] simply look for the coincidence of audio and visual breaks to identify the final boundary set, while more complex rules are used by [76] for fusing audio-visual candidate boundaries. [55] combine the candidate boundaries produced by the local minima of the merging force and local maxima of the splitting force by checking for their simultaneous occurrence and actual minimum/maximum values. [61] combine color breaks, motion breaks and audio breaks in a rule-based manner to produce the final scene boundary list. Fusing rules can also be learned in a probabilistic framework. [72] use the *maximum entropy model* (ME) to systematically fuse separate candidate story boundary sets produced by mid-level features of various types (visual, audio and semantic). This work is further extended in [73] by expanding the feature set and including a feature wrapper to bridge the raw multi-modal features and the ME model. Rather than fusing only candidate boundaries, [41] combine a list of candidate boundaries, non-boundary segments and “distinguish shots” to create the final sequence boundary list.
- *Verification of primary candidate boundaries based on secondary cues.* Here, primary cues are used to produce a set of candidate boundaries via one of the above mechanisms. Each of the candidate boundaries is then tested and classified as a logical unit boundary or non-boundary using secondary cues. A good candidate set, therefore, should have a very high recall rate on the reference boundaries. For example, [58] refine the rough scene boundaries located based on visual similarity by using cinematic rules (e.g., concentration rule and enlargement rule) regarding the shot distance and scene construction. [46], on the other

hand, merge short segments with larger ones based on the motion similarity between two shots across the boundary. Similarly, [59] integrate shot length and motion into a measure called *scene dynamic* (analogous to tempo feature proposed by [66]), merging two consecutive scenes with large scene dynamic values. [50] merge isolated scenes, i.e., those that contain one single shot, by examining the audio-types and audio uniformity of boundary shots. For example, two scenes are merged if the two boundary shots contain the music class or they contain similar speech signals.

Taxonomy Tables 2 and 3 show the classification of methods from literature according to the above framework. Note that a reference is repeated if two or more separate methods are proposed in the work.

1.2 Content Annotation

In this section, we first describe generic techniques for semantic segmentation via the classification of base units. The actual features and learning techniques for content annotation tasks are described in the next three sub-sections, and we overview the literature regarding semantic content labeling at different structural levels: (a) clip/program, (b) logical unit/sequence of shots, and (c) shot/sequence of frames.

1.2.1 Semantic Segmentation from Base Unit Classification

In the classification work, it is assumed that segmentation has been accomplished previously and/or the data comes from different sources, and the input to the classifier belongs to only one class. Content classification is often done on short-term base units such as shots or fixed-size clips. However, a real video may be constructed from long multiple-shot sequences of the same content class, forming an overall structure that alternates between these classes. For example, a complete soccer broadcast is composed of two halves, studio segments and commercial breaks. In order to enable fully automatic approach to content annotation, the boundaries between these sequences also need to be identified.

In an ideal case in which each base unit is correctly labeled by the classifier, the segmentation is straightforward, i.e., sequence boundaries are coincident with the changes in class labels of the base units. However, in reality, the likelihood value for the correct class can be temporarily reduced due to the mismatch between the features and its model, which often occurs in a short time interval. At these points, the likelihood for some incorrect class takes over. A number of techniques have been proposed for smoothing out this temporary noise in the class labeling, enabling semantic segmentation:

1. *Candidate sequence delimiters.* This approach exploits specific cues about the starting and ending of a sequence, and has been used to extract commercial blocks from a broadcast program, which are delimited by black frames and short periods of silence [86, 71, 87,].
2. *Minimum temporal duration.* A lower-bound threshold t is set on the number of base units for a particular class, v . Proceeding from left to right, for a sequence of successive units with potential labeled v , if the sequence duration is more than t , it will be assigned the label v , otherwise it will be assigned the label of the preceding sequence. This technique is employed in [88] to segment a video sequence into dialogue and non-dialogue scenes via the classification of individual shots.

Work	Data domain						Feature					Mechanism				Similarity			Fusion	
	Generic	News	Movies	Training	Sports	Instr.	Video	Text	Audio	Audio type	Semantic	Similarity linkage inference	Decision curve formulation	Rule-based	Probabilistic framework	Temporal distance	Shot dynamic	Shot duration	Verification	Rule-based fusion
[68]	x								x					x						
[55]	x						x						x			x				
[54]	x						x						x			x				x
[61]	x						x		x				x							x
[67]	x						x	x	x				x							
[56]	x						x						x			x				
[41]	x						x		x		x	x		x						x
[49]	x						x				x		x							
[44]	x						x					x								
[60]	x								x				x							
[74]	x								x				x							x
[43]		x							x				x							
[43]		x									x			x						
[43]		x							x		x				x					
[70]		x					x	x			x		x	x					x	
[63]		x					x			x			x	x						x
[78]		x						x			x			x						
[65]		x						x					x							
[71]		x						x	x					x						
[36]		x						x	x	x	x			x						
[37]		x					x	x	x						x					
[79]		x									x				x					
[72]		x					x	x	x		x			x						x
[73]		x					x	x	x		x			x						x
[80]		x					x		x		x			x						
[81]		x					x	x	x		x		x	x					x	
[42]		x								x	x			x						x

Table 2: Taxonomy of logical unit extraction techniques: Part I.

Work	Data domain						Feature					Mechanism				Similarity			Fusion	
	Generic	News	Movies	Training	Sports	Instr	Video	Text	Audio	Audio type	Semantic	Similarity linkage inference	Decision curve formulation	Rule-based	Probabilistic framework	Temporal distance	Shot dynamic	Shot duration	Verification	Rule-based fusion
[82]		x						x	x		x			x						x
[75]		x							x	x			x	x						x
[57]		x					x				x	x	x	x		x				
[69]		x						x						x						
[35]			x				x		x				x			x		x		x
[64]			x						x				x							
[48]			x				x					x								
[51]			x				x					x								
[46]			x				x					x					x			x
[59]			x				x						x							x
[83]			x				x		x			x								
[34]			x				x					x								
[84]			x				x						x							
[53]			x				x						x			x				
[66]			x				x						x							
[45]			x				x					x				x	x			
[76]			x					x		x			x	x						x
[50]			x				x		x	x		x						x		x
[47]			x				x					x								
[85]			x						x				x							x
[58]			x				x				x		x							x
[52]			x				x					x				x		x		
[40]				x			x			x	x	x	x	x						
[39]				x			x		x	x	x	x	x	x						x
[38]					x			x							x					
[62]		x							x				x							

Table 3: Taxonomy of logical unit extraction techniques: Part II.

3. *Optimal path between classes.* Proposed in [89] for segmenting a news program into commercial blocks and news reporting, this approach can be applied to any classification mechanism that produces a likelihood value for a unit with respect to a particular class, e.g., HMMs. First, the likelihood that a short segment, e.g., shot, belongs to a particular content class for every video segment is computed. The likelihood values are not used to determine the class for that segment, i.e., the class receiving the highest likelihood score, but as intermediate values. The optimum state transition path for the entire input video sequence is the one that has the highest accumulated likelihood. Note that this approach requires a penalty function for transitions between content classes. In a similar approach [90, 91, 92,], the likelihood values for each segment corresponding with respect to all content classes form the input to a ‘super’ HMM, whose states represent content classes and transition probabilities are computed from prior knowledge, e.g., counting the number of transitions in the data.
4. *Optimal path within a class and between classes.* [89] propose an extension to the above method, and in this approach a super HMM is built by concatenating HMM’s for different content classes. The optimal content class transition path is obtained by searching the path within a class and between classes using a dynamic programming technique. Note that this approach is only applicable to state-based temporal models such as the HMM.

1.2.2 Clip/Program Level

Content annotation at this level attempts to associate program-related semantics, e.g., genre and sub-genre to the entire video clip, which can be a single frame, single shot, a sequence of shots or a sequence of story units. Identifying program-related semantics is important for organizing a large video corpus as well as enabling the application of domain-specific analysis to the classified video data. For example, if an incoming video is identified as containing a soccer game, soccer-specific analysis can then be applied to the data to extract deeper soccer-related semantics.

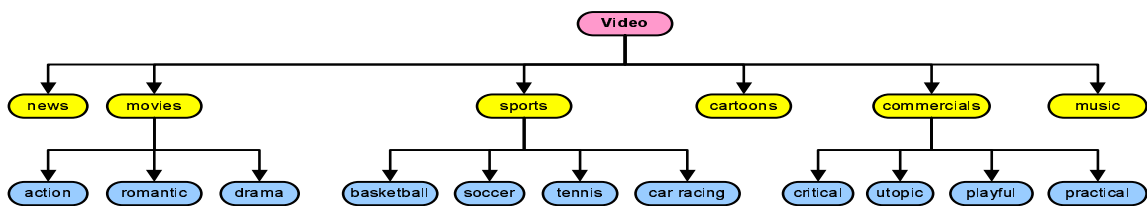


Figure 4: The video genre hierarchy.

Genre Genre classification is the highest level¹ of classification in which a video sequence is categorized exclusively into one of the pre-defined broad classes, namely {sports, movies, sitcoms, news, cartoons, etc.}. As pointed out in [94], genre classification of an arbitrary video is

¹[93] suggest that video sequence can be classified in a level higher than genre, namely *purpose*, {entertainment, information, communication}. However, purpose classification only requires the genre of the video to be identified.

sometimes a difficult problem and may come down to subjective views and semantic subtleties, however, researchers have often deliberately choose genres that are relatively well defined and commonly recognized in their experiment. Among the first studies in automatic genre classification is [95] in which the authors propose a three-step approach to video classification (1) syntactic property analysis, (2) film style attribute abstraction, and (3) mapping and recognition. In demonstrating this approach, they extract various low-to-mid level visual and audio features that include shot length, camera motion intensity, audio types and detected captions for classifying a video sequence into {news, commercial, cartoon, sports (tennis and car racing)}.

In our previous work, we classify a video clip into {cartoons, commercials, music, news, sports} relying purely on visual-based information including average shot length, the percentage of shot transition types, together with six color content features [96,]. A C4.5 decision tree classifier is employed. In a somewhat complementary work, [97] use only audio related features to classify roughly the same data set. They demonstrates the superior performance of wavelet-based audio features including centroid, bandwidth, sub-band energy, sub-band variance, zero crossing rate} to features extracted from time and frequency domains {loudness standard deviation and dynamic range, etc.} in discriminating between {news, commercial, cartoons, music shows, concerts and sports (motor racing)}. The results from three different classifiers, Decision Trees, Support Vector Machines (SVM) and k-Nearest neighbors are reported and compared.

In their early work, [98] extract background camera motion and foreground object motion to classify three video genres {sports, cartoons, news} via Gaussian Mixture Models (GMM). This work is extended in [99] and extracts features from shot-term spectral estimates of audio signals and motion dynamics of video signals and uses a GMM to classify a video clip into one of five classes {sports, cartoons, news, commercial, music}. The classification results for the two modalities - audio and visual - used separately, and for a linear fusion of the two modes at the GMM score level are compared. [100] extend this work by concatenating visual and audio features into a single feature vector, whose dimensionality is reduced via Principle Component Analysis (PCA). Experiments in [98, 99] confirm that the classification performance increases with the duration of the input video clips.

[101] propose a technique for classifying {sports, sitcoms, talk show, c-span} videos. Two features {shot length, motion (the average number of macro-blocks with motion vectors)} are modeled using GMMs. The number of clusters produced from an agglomerative clustering algorithm is used to estimate the parameters of GMMs. Each shot is assigned a symbol corresponding to the cluster ID it is most likely to belong to and the sequence of such symbols is used to build a Ergodic HMM classifier. [102] extend this work by including the magnitude of the motion vector as well as the color and texture of the shot frame in the feature set. As for classifiers, Hybrid HMM-SCFG (stochastic context-free grammars) models, instead of HMMs, are constructed in an attempt to capture the hierarchical structure of video sequences from {sports, sitcoms, comedy, c-span}.

Rather than performing classification on a set of predefined genres, [103] focus on characterizing a sports video based on the presence of action replays, the amount of scene text, and statistics (magnitude, direction, and amount) on camera and/or object motion to discriminate against other video kinds.

Sub-Genre A genre is often comprised of several sub-genres. For example, a sports video should show a specific kind of sports, e.g., football, basketball, volleyball, etc. Therefore, the next task in assigning program-related labels to a video sequence is to classify it into one of the

domain sub-genres. Also, it is to be noted that some studies have actually classified a video clip into a mixed set of genre and sub-genres, e.g. in [95], as described above, the label set is {newscast, cartoon, commercial, tennis, car racing}.

The most active domain for sub-genre classification is sports. It is made possible due to a number of factors including the fact that most sports are held in a fixed field and the field can be characterized with its color, texture and edge patterns [104,] as well as having specific audio attributes associated with the particular sports, e.g., ball hits in tennis. In [105], TV sports news articles are classified into one of 9 sub-genres based on the multiple subspace method. Discrete Cosin Transform (DCT) components for the first frame of a shot are used as the feature to construct multiple subspaces for the representation of each genre. A genre is assigned to an article, if the average project amount of all shots belong to the article to all subspaces of the genre is maximum. In [106], MPEG-1 motion vectors from P-frames are processed via PCA analysis for dimensionality reduction. The transformed feature vectors are then used to build a HMM for classifying sports sequences into one of three sub-genres {basketball, ice hockey, soccer}. [104] classify sports into one of the following kinds {floor, high diving, field hockey, long horse, javelin, judo, soccer, swimming, tennis, track} based on visual and edge features.

There have been a number of attempts at classifying movie sequences, typically previews/trailers, into different genres such as action, comedy, drama, etc. In [107], movie trailers are classified into two roughly genre-based categories {character (comedy, drama, romance) and action}. The ratio of shot duration over motion energy is used to construct the HMMs for these two classes. A relatively similar work is reported in [108], in which shot length and average shot activity are used to characterize the level of violence of a movie which is then used to classify movie trailers into action and romance/comedy. In a more detailed investigation, [109] propose a method to classify movie previews into genres {action/drama, action/comedy, comedy/drama, drama, comedy, horror} using low level computable features including {average shot length, color variance, motion content (computed from temporal slice analysis) and lighting key (product of means and standard deviation of gray level of a \mathcal{R} -frame)}. Classification is performed using the mean shift clustering method. Unlike above studies, which work on previews/trailers, [110] use the actual movie data in their study of horror film genre typing. [110] attempts to characterize horror film genre via the frequency of sound events that include {surprise or alarm, apprehension, surprise followed by sustained alarm, apprehension building up to a climax}. These events correspond to specific sound patterns that result from the manipulation of sound energy dynamics, the change in audio energy over time. Their experiments demonstrate the correlation between the occurrences of above events and the presence of horror themes within film.

[111] propose an approach to indexing commercials. Color characteristics, the distribution of lines, and editing effects are features used to calculate the likelihood score that a commercial belongs to one of four semiotic categories {practical, playful, utopic, critical}.

In [112], audio features including {volume distribution, pitch contour, frequency-related measures} are input into an Artificial Neural Network (ANN) to classify a video document into a mixture of genres and sub-genres and sub-constructs within a program of a specific genre {news reports, weather forecasts, commercials, basketball, and football games}. This work is extended in [113] in which ergodic HMMs models are used to improve the classification results. Using the same data set, [82] aim to compare the performance of multi-modality over single-modality (audio) in the genre classification task. Four multi-modal integration methods are experimented with and these include direct concatenation, product HMM, 2-stage HMM and ANN. They report significant improvements in classification results when a multi-modal feature set is used,

and the product HMM is identified as the best overall integration method.

Commercial detection and segmentation So far we have discussed genre identification and labeling as a pure classification task, however, it is not always the case. A long sequence of broadcast video may contain several video programs, each belonging to a different genre, or one program can be interrupted by the insertions of other video sequences of a separate genre, i.e., commercial breaks. Identifying the boundaries between these segments can be seen as a semantic segmentation problem. The most popular problem is the detection and filtering of commercials, which relies on the follow set of cues [87,].

- *High cut frequency.* The advertisers try to deliver as much information (both audio and visual) to the viewer as possible in a restricted time slot by quickly cutting between shots.
- *High visual dynamics.* Due to fast moving objects, complex camera operations and fast changing color patterns in a scene, a high level of visual dynamics is observed.
- *Duration constraints.* A single commercial usually lasts for 30 seconds to 1 minute. The entire commercial break is typically 2 to 4 minutes.
- *Black frames and silence delimiters.* Commercial breaks are usually delimited from the main program content by some black frames together with a short period of silence.
- *Others.* The absence of station logos, the raised volume of audio signals, the appearance of text and still images with different sizes and at different locations are some of the other multi-modal cues for identifying commercial breaks from the main program.

[86] use the frequency of “strong” hard cuts (joining two visually distinctive shots) and black frames to locate candidate commercial delimiters, which is then verified by action indicators, edge change ratio and motion vector length. In [71], information about black frames and cutting rate is combined through a set of heuristics rules for filtering out commercial breaks from a news sequence. In [87], the visual dynamics magnitude and cutting rate are used to locate coarse level candidate boundaries. Information about silence and black frames is then used to confirm if a candidate boundary is an actual commercial break delimiter. [73] detect commercial blocks from a news program via frame matching, which is based on image templates such as station logos and caption titles. The absence of such entities indicates that the frame belongs to a commercial segment. Morphological operators are further applied to the results to filter out noise due to the dynamic content within the commercial and the variance of production rules. [89] use the dynamic programming technique to detect commercial blocks from a news program, in which the HMM video genre classifier developed in [113] is used to compute the genre likelihood values for each short segment (see Section 1.2.1).

1.2.3 Logical Unit/Sequence of Shots Level

Break vs. play in sports In [91], visual and audio streams associated with *break* and *play* soccer sequences are modeled separately using HMMs. The *late fusion* method (through the combination of decisions regarding the audio and visual streams to form a single recognition) demonstrates better performance than the *early fusion* method (through concatenation of the feature vector). Features extracted include motion (how well the estimated global motion model

can actually model the displacement of points between two consecutive frames, average motion amplitude, ratio of the likelihood of no background motion and the likelihood of background motion) and audio (12 LPC Cepstral coefficients with the log energy, delta and acceleration coefficients). Semantic segmentation into *break* and *play* sequences is achieved from the classification of fixed-size segments via a HMM of two states. The input to this HMM is the likelihood of *break* and *play* and transition parameters are computed by counting the number of transitions in real games. The same problem is addressed in [114], in which visual features are used to detect grass and grass orientation. Each shot is classified as {close-up, zoom-in, global} based on grass characteristics. Play/break boundaries are assumed to align with view transitions. The main idea is to classify long, consecutive global view shots as plays, consecutive close-up view as breaks and resolving the fuzziness in short segments of zoom-in views by considering the majority labels in the neighborhood of current segments. This work is further extended in [90] wherein dominant color ratio and motion intensity features are used and multiple HMM models are built for each game state. Dynamic programming is used to combine the likelihood values produced by HMM classifiers for each fixed size segments to eventually segment the soccer game into break and play sequences. This is similar to the use of HMM in [91]. In addition, unsupervised Hierarchical Hidden Markov Model (HMMM) is employed in [115], achieving results comparable with supervised learning HMMs for segmenting a soccer game into break and play sequences.

Domain-specific sequences in sports [92] segment and classify tennis video sequence into four classes {first missed serves, rally, replay, game break }, allowing the complete reconstruction of a tennis video. Individual HMMs are built for each content class based on the observation of audio types {speech, applause, ball hits, noise, music}, shot duration and the visual similarity between the shot \mathcal{R} -frame and a reference shot that represents the global view. Long-term structure of a tennis game is modeled by connecting these HMMs, creating a high-level HMM with transition probabilities crafted from prior knowledge. [116] develop separate HMMs to detect important baseball events including {home run, catch, hit, infield play}. States for these models come from seven view classes {pitch view, catch overview, catch close-up, running overview, running close-up, audience view , touch-base close-up}. A shot-based observation given to the models comprises of the likelihood values that the shot belongs to one of these view classes

News topics In classification of news stories, [117] classify news articles into different subjects {politics, economy, science, etc} based on character Recognition of embedded captions. Similarly, [118] classify a news stories into one of 10 subjects {politics, military, sports, weather, transport, business, health, entertainment, science & technology, and daily}. In this work, SVMs are exploited to compute text confidence scores from low-level textual features and GMMs are employed to compute audio-visual confidence score from low-to-mid level audio-visual features (face, close-caption, black frame, shot duration and motion energy). The decision strategies use multi-modal confidence scores as input and perform topic classification in a text-biased manner.

Dialogue vs. action in movies Two HMM topologies have been applied in [119] for the diction of dialogue scenes in a movie: left-to-right model (establishing, dialogue, transitional) and ergodic 2-state model (dialogue and non-dialogue). Note that left-to-right HMM requires

the pre-segmentation of the video sequence so that each clip contains one establishing shot sequence and one dialogue sequence. Features input into HMM include audio types, faces and visual change to previous shot and the shot before the previous shot. [88] extract dialogue scenes from a movie sequence by first classifying all shots into dialogue and non-dialogue based on face detection, camera motion estimation and audio classification. Then, shot labels are used to extract dialogue sequences via a Finite State Machine (FSM) which consists of three states {dialogue, probably dialogue, non-dialogue}. The middle state ensures the transition from dialogue state to non-dialogue state happens only if at least three non-dialogue shots are detected after the sequence of dialogue shots. [120] aim to identify different kinds of *scenes* in the movie such as 2-speaker dialogue, multiple-speaker dialogue and hybrid scenes. Shots are first grouped into different clusters via a method called *window base sweep* and each cluster is classified into 3 classes: periodic, partly-periodic and random. Scene boundaries are determined by progressive scenes, whilst the scene classes are determined based on the inference about the cluster class of every shot in the scene and the audio characteristics. [121] detect violent scenes in a movie from the recognition of associated elements such as gunfire, explosion and blood. Activity descriptors, color and audio features are used in their work. [122] detect violent events and car chases in a movie based on an analysis of environmental sounds. These high-level events are recognized based on the dominance of low level sound elements such as engines, horns, explosions or gunfire.

Affective events in movies In [123], three emotional events {sadness, fear, joy} are extracted from a movie based on their association with color, motion and the cutting rate of a scene. For example, sadness is signified by dark, unsaturated colors, little camera motion and slow cutting rate. Candidate scenes with emotional events are hypothesized to be different to previous scenes and are extracted by measuring the similarity between the current scene to previous ones based on a FIFO short-term memory model. Fitness values for the above emotional events are then computed for a candidate scene to eventually determine if it contains a certain emotional event. This work is further extended in [124] when HMMs are used as the base classification mechanism. Two HMM topologies are used in this work. In the first one, each emotional event is explicitly modeled by one state in the HMM and a decision on whether a particular emotional event exists or not is based on the number of emotional states representing this event in the state trajectory produced by dynamic programming. In the second approach, each emotional event is modeled by a separate HMM. The best results are obtained with the second approach. In addition to studying the association between sound events and horror themes in the film as the whole, as discussed in Section 1.2.2, [110] investigate the occurrences of sound energy events {surprise or alarm, apprehension, surprise followed by sustained alarm, apprehension building up to a climax} at the scene level of film, drawing conclusions between the number of sound energy events present and the thematic content of scenes with respect to horror.

1.2.4 Shot/Sequence of Frames Level

Camera framing in sports Assigning labels such as size and subject to sport is important in automatic analysis of sports videos due to the association between certain shot classes and high-level semantics (e.g., close-ups for exciting action) as well as between shot class patterns and sports events (e.g., close-ups followed by replay for soccer goal events). It is also made possible from the fact that in sports broadcast, there are often a limited number of cameras at

fixed positions. Identified shot classes for a given sports is often a mixture of shot perspectives such as subject, size and action, e.g. player close-up or outfield overview. Motion features (local motion activity, persistent camera panning, motion activity ratio in court model) are used in [125] to classify a tennis shot into one of five categories {courtview-playing, medium-view-player-following, close-up head-tracking, site-bird view, audience}. Classification is performed via domain specific rules, e.g., court-view playing shots often contain small local motion activity with interlaced pan left and right. [126] propose a baseball scene classification method which automatically labels each shot with one of eight labels {pitching, running, base, audience, outfield overview, infield overview, others (commercial, interview, etc.)}. The maximum entropy philosophy is applied to fuse multimodal clues from which a semantic label is predicted, these clues relate to {audio types, textual keywords (from speech), color distribution, edge distribution, camera motion}. Temporal progression of each shot is captured by dividing it into three equal-length segments. [127] classify soccer shots into three size-based categories: {long shot, in-field medium shot, out-of-field, close-up shots}. Features used include the ratio of grass-colored pixels in a frame, and the size and number of soccer objects detected in a shot. In [128], the percentage of green pixels, motion and shot duration are used to classify a shot from American football coaching videos into one of three types {score board, end-zone, sideline}. In addition, this work demonstrates the superior performance of the probabilistic approach based on Bayesian inference over a deterministic approach.

Live play vs. replay Identifying replay shots in sports videos has been addressed in a number of studies due to the association between important game-related events and the occurrence of these shots. [38] classify each segment of a broadcast American football game delimited by speaker changes as {live, replay, others (report, studio, etc) and commercial} using textual information extracted from closed-captions. Rather than using caption texts, replay is identified in [129] based on features associated with a compressed MPEG stream including macroblock, motion and bit rate. [130, 127] detect slow-motion replay shots via a zero crossing measure that evaluates the amplitude of the fluctuations in the frame differences within a temporal window.

News In [73], anchorperson shots are detected by first detecting faces and checking for its consistent position within the frame. Histograms of regions of interest, extended from detected face regions are used to group all shots via an agglomerative clustering algorithm. The dominant cluster identifies all anchorperson shots in the sequence. [80] propose a model-free detection of anchorperson shots based on the graph-theoretical clustering method. Clusters formed with at least three member shots are considered comprising of anchorperson shots. Additional clues such as the minimum length of news story and spatial discrimination is used to eliminate false positive clusters. In addition, the information about graphics shots and superimposed captions have also been used. In [79], news shots are classified into {intro/highlight, anchor, 2 anchor, gathering, still image, live-reporting, speech, weather, sport, text-scene, and special} based on a set of multimodal features including color histograms, background scene change, speaker change, audio, motion activity, shot duration, face, shot type (estimated from face size, close-up, medium-distance), number of caption text lines and centralized videotext. A combination of Decision Trees and template matching is employed as the classifier. Similarly, [131] classify news shots into 5 categories {speech/report, anchor, walking, gathering, computer graphics} via characteristics of detected faces (size, position, and motion) and motion (e.g., motionless frames are used as an indicator of computer graphics shots). [132] segment (and classify) news program

into shots and edit segments of the following types {anchor, report, begin, end, weather forecast, interview, cut, frame translate, window change}. Their method proceeds by constructing separate HMMs for the above content classes from six motion-related features, and then a top-level HMM for the complete TV news program is built from these HMMs.

Movie With respects to the semantic classification of movie shots, [133] use a Bayesian network to recognize four semantic shot attributes: the presence/absence of a close-up, the presence/absence of a crowd in the scene, the type of set (nature vs. urban), and the high/low level of action. Three features used in the work are shot activity, texture energy and amount of skin tones in the shot. [134] attempt to classify shots into two locale-related classes {interior, exterior} based on luminance intensity variance caused by natural and artificial lights.

2 Summary

In this chapter, we have reviewed existing work in various areas of MCM, and in doing so positioned our objective of extracting expressive and structural elements in film in this broad context. In this overview, we have attempted to provide taxonomy of work addressing the various aspects of MCM that include temporal and segmentation of structure (shot and logical unit), content labeling (classification, semantic segmentation and event detection) and video abstraction (\mathcal{R} -frame extraction and video skim generation). In addition, we have briefly discussed the role of film grammar in formulating solutions for automatic understanding and analysis of film.

In the next chapter, we can move toward setting up the foundation which includes a discussion of the feature extraction and experimental data set, which is essential in all component tasks in this research.

References

- [1] G. Ahanger and T. Little, “A survey of technologies for parsing and indexing digital video,” *Journal of Visual Communication and Image Representation*, vol. 7, no. 1, pp. 28–43, 1996.
- [2] R. Lienhart, “Reliable transition detection in videos: A survey and practitioner’s guide,” *International Journal of Image and Graphics*, vol. 1, no. 3, pp. 52–56, 2001.
- [3] A. Nagasaka and Y. Tanaka, “Automatic video indexing and full-motion search for object appearance,” in *Proceedings of Second Working Conference on Visual Databases Systems*, 1992, pp. 113–127.
- [4] B.-L. Yeo and B. Liu, “Rapid scene analysis on compressed video,” *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 2, pp. 533–544, 1995.
- [5] N. V. Patel and I. K. Sethi, “Compressed video processing for cut detection,” in *IEE Proceedings: Vision, Image and Signal Processing*, vol. 134, Oct. 1996, pp. 315–322.
- [6] H. Zhang, A. Kankanhalli, and S. Smoliar, “Automatic partitioning of full-motion video,” *Multimedia System*, vol. 1, pp. 10–28, 1993.

-
- [7] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proceedings of the Second ACM International Conference on Multimedia (MULTIMEDIA '94)*, New York, Oct. 1994, pp. 357–364.
- [8] L. Gu, K. Tsui, and D. Keightley, "Dissolve detection in MPEG compressed video," in *Proceedings of IEEE International Conference on Intelligent Processing Systems*, vol. 2, 1997, pp. 1692–1696.
- [9] A. M. Alattar, "Detecting and compressing dissolve regions in video sequences with a DVI multimedia image compression algorithm," in *Proceedings of 1993 IEEE International Symposium on Circuit and Systems*, Chicago, Illinois, 1993, pp. 13–16.
- [10] J. Meng, Y. Juan, and S.-F. Chang, "Scene change detection in a MPEG compressed video sequence," in *IS&AT/SPIE Symposium Proceedings Vol 2419*, Feb. 1995.
- [11] A. M. Alattar, "Detecting fade regions in uncompressed video sequences," in *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP'97)*, Munich, Germany, April 1997, pp. 3025–3028.
- [12] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Image and Video Processing VII*, ser. Proceedings of SPIE, vol. 3656, 1999, pp. 290–301.
- [13] F. Arman, A. Hsu, and M.-Y. Chiu, "Image processing on compressed data for large video databases," in *ACM Multimedia (ACMMM'93)*, California, July 1993, pp. 267–272.
- [14] H. Zhang, C. Low, Y. Gong, and S. Molliar, "Video parsing using compressed data," in *IS&T/SPIE, Image and Video Processing II*, Feb. 1994, pp. 142–149.
- [15] E. Deardorff, T. Little, J. Marshall, D. Venkatesh, and R. Walzer, "Video scene decomposition with the motion picture parser," in *Digital Video Compression on Personal Computers: Algorithms and Technologies*, ser. Proceedings of SPIE, vol. 2187, Feb. 1994, pp. 44–45.
- [16] J. Feng, K.-T. Lo, and H. Mehrpour, "Scene change detection algorithm for MPEG video sequence," in *Proceedings of IEEE International Conference on Image Processing (ICIP'96)*, vol. 2, Lausanne, Switzerland, 1996, pp. 821–824.
- [17] W. Heng, K. Ngan, and M. Lee, "Validity of scene cut detection using bit rate information of VBR video," in *Symposium on Image, Speech, Signal Processing and Robotics*, vol. II, Hongkong, Sept. 1998, pp. 133–138.
- [18] T. C. Kuo, Y. Lin, A. L.P.Chen, S.-C. Chen, and C. Ni, "Efficient shot change detection on compressed video," in *Proceedings of International Workshop on Multimedia Database Management Systems*, 1996, pp. 101–108.
- [19] V. Kobla, D. Doermann, and A. Rosenfeld, "Compressed domain video segmentation," Center for Automation Research, College Park, MD 20742-3275, Tech. Rep. 839, 1996.
- [20] N. Gamaz, X. Huang, and S. Panchanathan, "Scene change detection in MPEG domain," in *IEEE Southwest Symposium on Image Analysis and Interpretation*, Tucson, Arizona, 1998, pp. 12–17.

-
- [21] M. Sugano, Y. Nakajima, H. Yanagihara, and A. Yoneyama, "A fast scene change detection on MPEG coding parameter domain," in *Proceedings of IEEE International Conference on Pattern Recognition*, 1998, pp. 889–892.
- [22] S.-C. Pei and Y.-Z. Chou, "Effective wipe detection in MPEG compressed video using macroblock type information," *IEEE Transactions on Multimedia*, vol. 4, no. 3, pp. 309–315, sep 2002.
- [23] M. Ardebilian, X. Tu, and L. Chen, "Improvement of shot detection methods based on dynamic threshold selection," in *Multimedia Storage and Archiving Systems*, ser. Proceedings of SPIE, vol. 3229, Dallas, USA, Nov. 1997, pp. 14–22.
- [24] A. Hanjalic, R. L. Legendijk, and J. Biemond, "A novel video parsing method with improved thresholding," in *Third Annual Conference of the Advanced School for Computing and Imaging (ASCI'97)*, Heijen, Neitherland, 1996.
- [25] A. Hanjalic and H. Zhang, "Optimal shot boundary detection based on robust statistical models," in *Proceedings of IEEE International Conference on Multimedia and Systems*, vol. 2, Florence, Italy, June 1999, pp. 710–714.
- [26] N. Vasconcelos and A. Lippman, "Statistical models of video structure for content analysis and characterization," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 3–14, jan 2000.
- [27] Y. Altunbasak, "A statistical approach to threshold selection in temporal video segmentation algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, vol. 6, Istanbul, Turkey, Jun 2000, pp. 2421–2424.
- [28] J. S. Boreczky and L. D. Wilcox, "A Hidden Markov Model for video segmentation using audio and image features," vol. 6, pp. 3741–3744.
- [29] J. M. Sanchez, X. Binefa, and J. Kender, "Coupled markov chains for video contents characterization," in *IEEE International Conference on Pattern Recognition (ICPR'2002)*, Quebec, Canada, aug 2002.
- [30] C. Ngo, T. Pong, and R. Chin, "Detecting gradual transitions through temporal slice analysis," in *IEEE Internaltional Conference on Computer Vision and Pattern Recognition*, vol. 1, Colorado, June 1999, pp. 750–755.
- [31] H. Kim, S.-F. Park, J. Lee, W. M. Kim, and S. M.-H. Song, "Processing of partial video data for detection of wipes," in *Storage and Retrieval for Image and Video Databases VII*, ser. Proceedings of SPIE, vol. 3656, 1999, pp. 280–289.
- [32] M. S. Drew, Z.-N. Li, and X. Zhong, "Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences," in *IEEE International Conference on Pattern Recognition (ICPR'2002)*, Quebec, Canada, aug 2002.
- [33] B. T. Truong, "Shot transition detection and genre identification for video indexing and retrieval," Honours Thesis, School of Computing, Curtin University of Technology, Western Australia, nov 1999.

- [34] A. Hanjalic, R. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video retrieval systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 580–588, June 1999.
- [35] H. Sundaram and S.-F. Chang, "Video scene segmentation using video and audio features," in *IEEE International Conference on Multimedia and Expo*, IEEE. New York: IEEE Press, August 2000.
- [36] A. Merlino, D. Morey, and M. T. Maybury, "Broadcast news navigation using story segmentation," in *ACM Multimedia 97*, 1997, pp. 381–391.
- [37] S. Boykin and A. Merlino, "Machine learning of event segmentation for news on demand," *Communications of the ACM*, vol. 43, no. 2, pp. 35–41, 2000.
- [38] N. Nitta, N. Babaguchi, and T. Kitahashi, "Story based representation for broadcasted sports video and automatic story segmentation," in *IEEE International Conference on Multimedia and Expo (ICME'02)*, Lausanne, Switzerland, August 26-29 2002, pp. 58–64.
- [39] D. Q. Phung, S. Venkatesh, and C. Dorai, "High level segmentation of instructional videos based on the content density function," in *ACM Multimedia (ACMMM'02)*, Juan Les Pin, France, Dec 1-6 2002, pp. 296–298.
- [40] —, "Hierarchical topic segmentation in instructional films based on cinematic expressive functions," in *ACM Multimedia (ACMMM'03)*, Berkeley, CA, Nov 2-8 2003, pp. 287–290.
- [41] P. Aigrain, P. Joly, and V. Longueville, "Medium knowledge-based macro-segmentation of video into sequences," in *Intelligent Multimedia Retrieval*, M. T. Maybury, Ed. AAAI Press/The MIT Press, 1998, ch. 8, pp. 159–174.
- [42] C. Wang, Y. Wang, H. yong Liu, and Y. xiang He, "Automatic story segmentation of news video based on audio-visual features and text information," in *2003 International Conference on Machine Learning and Cybernetics*, vol. 5, nov 2-5 2003, pp. 3008–3011.
- [43] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll, "New approaches to audio-visual segmentation of TV news for automatic topic retrieval," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Salt Lake City, Utah, may 2001.
- [44] M. Yeung, B.-L. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 7, no. 1, pp. 94–109, July 1998.
- [45] Y. Rui, T. S. Huang, and M. S., "Constructing table-of-content for videos," *ACM Multimedia System Journal: Special Issue in Multimedia Systems on Video Libraries*, vol. 7, no. 5, pp. 359–368, 1999.
- [46] Y.-M. Kwon, C.-J. Song, and I.-J. Kim, "A new approach for high level video structuring," in *IEEE International Conference on Multimedia and Expo (ICME'00)*, vol. 2, New York, August 2000, pp. 773–776.

- [47] J. Zhou and W. Tavanapong, "Shotweave: A shot clustering technique for story browsing for large video databases," in *Int'l Workshop on Multimedia Data Document Engineering (MDDE'02)*, Prague, March 2002.
- [48] R. Hammoud, L. Chen, and D. Fontaine, "An extensible spatial-temporal model for semantic video segmentation," in *1st International Forum on Multimedia and Image Processing*, Anchorage, Alaska, Mar. 1998.
- [49] P. Bouthemy, C. Garcia, G. Tziritas, E. Veneau, and D. Zuga, "Scene segmentation and image feature extraction for video indexing and retrieval," in *International Conference on Visual Information Systems, Visual' 99*, ser. Lecture Notes in Computer Science, vol. 1614. Springer Verlag, June 1999, pp. 245–252.
- [50] Y. Li, W. Ming, and C.-C. J. Kuo, "Semantic video content abstraction based on multiple cues," in *IEEE International Conference on Multimedia and Expo (ICME'01)*, Tokyo, August 2001, pp. 623–626.
- [51] R. Hammoud and D. G. Kouam, "A mixed classification approach of shots for constructing scene structure for movie films," in *Irish Machine Vision and Image Processing Conference*, Sept. 2000, pp. 223–230.
- [52] J. R. Kender and B.-L. Yeo, "Video scene segmentation via continuous video coherence," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98)*, Jun 23-25 1998, pp. 367–373.
- [53] E. Veneau, R. Ronfard, and P. Bouthemy, "From video shot clustering to sequence segmentation," in *ICPR'00*, vol. 4, Barcelona, sep 2000, pp. 254–257.
- [54] T. Lin, H. Zhang, and Q.-Y. Shi, "Video scene extraction by force competition," in *IEEE International Conference on Multimedia and Expo (ICME'01)*, Tokyo, Japan, Aug 22-25 2001, pp. 960–963.
- [55] T. Lin and H. Zhang, "Video content representation for shot retrieval and scene extraction," *International Journal of Image Graphics*, vol. 3, no. 1, pp. 507–526, 2001.
- [56] L. Zhao, W. Qi, S. Yang, and H. Zhang, "Video shot grouping using best-first model merging," in *Proc. 13th SPIE Symposium on Electronic Imaging - Storage and Retrieval for Image and Video Databases*, San Jose, Jan 2001, pp. 262–267.
- [57] N. O'Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy, and A. Smeaton, "News story segmentation in the Fischlar video indexing system," vol. 3, pp. 418–421.
- [58] J. Wang, T.-S. Chua, and L. Chen, "Cinematic-based model for scene boundary detection," in *Multimedia Modeling Conference (MMM'01)*, Amsterdam, Nov 2001, pp. 3–18.
- [59] Z. Rasheed and M. Shah, "Scene detection in hollywood movies and TV shows," in *CVPR'03*, Madison, Wisconsin, June 16-22 2003.
- [60] J. Nam and A. H. Tewfik, "Combined audio and visual streams analysis for video sequence segmentation," in *ICASSP-97*, vol. 4, April 1997, pp. 2665–2668.

-
- [61] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for content-based video segmentation," in *IEEE International Conference on Image Processing (ICIP'98)*, Chicago, Illinois, oct 1998.
- [62] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI Signal Processing Systems for Signal Processing and Video Technology*, vol. 20, no. 1/2, 1998.
- [63] H. Jiang, T. Lin, and H.-J. Zhang, "Video scene segmentation with the assistance of audio content analysis," in *IEEE International Conference on Multimedia and Expo (ICME'00)*, New York, August 2000.
- [64] H. Sundaram and S.-F. Chang, "Audio scene segmentation using multiple models, features and time scales," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00)*, Istanbul, Turkey, Jun 2000.
- [65] I. Ide, H. Mo, N. Katayama, and S. Satoh, "Topic-based inter-video structuring of a large-scale news video corpus," in *IEEE International Conference on Multimedia and Expo (ICME'03)*, Baltimore, July 6-9 2003.
- [66] B. Adams, C. Dorai, and S. Venkatesh, "Automatic extraction of expressive elements from motion pictures: Tempo," *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 472–481, December 2002.
- [67] X. Lu, Y.-F. Ma, H.-J. Zhang, and L. Wu, "An integrated correlation measure for semantic video segmentation," in *IEEE International Conference on Multimedia and Expo (ICME'02)*, Lausanne, Switzerland, August 26-29 2002.
- [68] C. Saraceno and R. Leonardi, "Identification of successive correlated camera shots using audio and video information," in *IEEE International Conference on Image Processing*, 1997.
- [69] R. Ronfard and T. T. Thuong, "A framework for aligning and indexing movies with their script," in *IEEE International Conference on Multimedia and Expo (ICME'03)*, vol. 1, Baltimore, July 6-9 2003, pp. 21–24.
- [70] X. Zhu, L. Wu, X. Xue, X. Lu, and J. Fan, "Automatic scene detection in news program by integrating visual feature and rules," in *The second IEEE Pacific-Rim conference on multimedia*, Beijing, Oct. 2001, pp. 837–842.
- [71] A. G. Hauptmann and Witbrock, "Story segmentation and detection of commercials in broadcast news video," in *Proceedings of Advances in Digital Libraries Conference*, Santa Barbara, CA, apr 1998.
- [72] W. Hsu and S.-F. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *IEEE International Conference on Multimedia and Expo (ICME'03)*, Baltimore, July 6-9 2003.
- [73] W. Hsu, S.-F. Chang, C.-W. Huang, L. Kennedy, C.-Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *Storage*

- and Retrieval Methods and Applications for Multimedia 2004*, ser. Proceedings of SPIE, vol. 5037, San Jose, CA, Jan 2004, pp. 244–258.
- [74] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang, “Scene change detection by audio visual clues,” in *IEEE International Conference on Multimedia and Expo (ICME’02)*, Lausanne, Switzerland, August 26-29 2002.
- [75] T. Kemp, T. Lin, and H. Zhang, “Strategies for automatic segmentation of audio,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’01)*, vol. 3, Salt Lake City, Utah, May 2001, pp. 1423–1426.
- [76] A. Yoshitaka and M. Miyake, “Scene detection by audio-visual features,” in *IEEE International Conference on Multimedia and Expo*, IEEE. Tokyo: IEEE Press, August 2001.
- [77] S. Eickeler and S. Muller, “Content-based video indexing of TV broadcast news using hidden markov models,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’99)*, Salt Lake City, Utah, 1999, pp. 2997–3000. [Online]. Available: citeseer.ist.psu.edu/iurgel01new.html
- [78] A. Hanjalic, R. Lagendijk, and J. Biemond, “Semi-automatic news analysis, indexing, and classification system based on topics pre-selection,” in *Storage and Retrieval of Image and Video Databases*, ser. Proceedings of SPIE, vol. 3656, San Jose, CA, January 1998, pp. 86–97.
- [79] L. Chaisorn, T.-S. Chua, and C.-H. Lee, “The segmentation of news video into story units,” in *IEEE International Conference on Multimedia and Expo (ICME’02)*, vol. 1, Lausanne, Switzerland, August 26-29 2002, pp. 73–76.
- [80] X. Gao and X. Tang, “Unsupervised and model-free news video segmentation,” in *IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL’01)*, Dec 14 2001, pp. 813–816.
- [81] S. Raaijmakers, J. den Hartog, and J. Baan, “Multimodal topic segmentation and classification of news video,” in *IEEE International Conference on Multimedia and Expo (ICME’02)*, vol. 2, Lausanne, Switzerland, August 26-29 2002, pp. 33–36.
- [82] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zahib, “Spatial color indexing and applications,” *International journal of Computer Vision*, vol. 35, no. 3, pp. 245–268, 1999.
- [83] R. Lienhart, S. Pleiffer, and W. Effelsberg, “Scene determination based on video and audio features,” in *Proceeding of IEEE Conference on Multimedia Computing and Systems*, Florence, Italy, 7-11, June 1999, pp. 685–690.
- [84] A. Hanjalic, R. Lagendijk, and J. Biemond, “Detection of global story units in full-length movies,” in *IEEE Workshop on Content-based Access of Image and Video Libraries*, Puerto Rico, Jun 1997.
- [85] Y. Cao, W. Tavanapong, K. Kim, and J.-H. Oh, “Audio-assisted scene segmentation for story browsing,” in *International Conference on Image and Video Retrieval (CIVR’03)*,

- ser. Lecture Notes in Computer Science, vol. 2728. Urbana, IL: Springer, 2003, pp. 446–455.
- [86] R. Lienhart, C. Kuhmunch, and W. Effeslberg, “On the detection and recognition of Television commercials,” in *Proceedings of IEEE Conference on Multimedia Computing and Systems*, 1997, pp. 509–516.
- [87] Y. Li and C.-C. J. Kuo, “Detecting commercial breaks in tv programs based on audio-visual information,” in *Conference on Internet Multimedia Management Systems, SPIE’s International Symposium on Voice, Video and Data Communications*, Boston, MA, Nov 5-8 2000.
- [88] M. D. Santo, G. Percannella, C. Sansone, and M. Vento, “Dialogue scenes detection in mpeg movies: A multi-expert approach,” in *Multimedia Databases and Image Communication, Second International Workshop (MDIC 2001)*, ser. Lecture Notes in Computer Science, vol. 2184. Springer, Sep 17-18 2001, pp. 192–201.
- [89] J. Huang, Z. Liu, and Y. Wang, “Joint video scene segmentation and classification based on hidden markov model,” in *IEEE International Conference on Multimedia and Expo (ICME’00)*, New York, August 2000, pp. 1551–1554.
- [90] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, “Structure analysis of soccer video with hidden markov models.”
- [91] M. Barnard, J.-M. Odobez, and S. Bengio, “Multi-modal audio-visual event recognition for football analysis,” in *IEEE Workshop on Neural Networks for Signal Processing (NNSP’03)*, Toulouse, 2003.
- [92] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, “HMM based structuring of tennis videos using visual and audio cues,” in *IEEE International Conference on Multimedia and Expo (ICME’03)*, Baltimore, July 6-9 2003.
- [93] C. Snoek and M. Worring, “A review on multimodal video indexing,” in *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, August 2002, pp. 21–24.
- [94] M. Roach, J. Mason, W. D. Evans, L.-Q. Xu, and F. Stentiford, “Recent trends in video analysis: A taxonomy of video classification problems,” in *Internet and Multimedia Systems and Applications*, 2002.
- [95] W. Effelsberg, S. Fischer, and R. Lienhart, “Automatic recognition of film genres,” in *The Third ACM International Multimedia Conference and Exhibition (MULTIMEDIA’95)*, New York, Nov. 1995, pp. 367–368.
- [96] B. T. Truong, S. Venkatesh, and C. Dorai, “Automatic genre identification for content-based video categorization,” in *ICPR 2000*, Barcelona, Sep 2000, pp. 230–233.
- [97] D. Q. Phung, C. Dorai, and S. Venkatesh, “Video genre categorization using audio wavelet coefficients,” in *The Fifth Asian Conference on Computer Vision*, Melbourne, Australia, 23-25 January 2002, pp. 69–74.

-
- [98] M. Roach, J. Mason, and M. Pawlewski, "Video genre classification using dynamics," in *Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2001, pp. 1557–1560.
- [99] M. Roach, L.-Q. Xu, and J. Mason, "Classification of non-edited broadcast video using holistic low-level features," in *Tyrrhenian Workshop on Digital Communications*, 2002.
- [100] L.-Q. Xu and Y. Li, "Video classification using spatial-temporal features and pca," in *IEEE International Conference on Multimedia and Expo (ICME'03)*, vol. 3, Baltimore, July 6-9 2003, pp. 485–488.
- [101] C. Taskiran, C. A. Bouman, and E. J. Delp, "Discovering video structure using the pseudo-semantic trace," in *SPIE Conference on Storage and Retrieval for Media Databases*, vol. 4315, San Jose, CA, jan 2001, pp. 571–578.
- [102] C. Taskiran, I. Pollak, C. A. Bouman, and E. J. Delp, "Stochastic models of video structure for program genre detection," in *SPIE Conference on Visual Content Processing and Representation (VLBV'03)*, vol. 4315, Madrid, Spain, sep 2003, pp. 571–578.
- [103] V. Kobla, D. DeMenthon, and D. Doermann, "Identification of sports videos using replay, text, and camera motion features," in *Storage and Retrieval for Media Databases*, ser. Proceedings of SPIE, vol. 3972, jan 2000, pp. 332–343.
- [104] M. Mukunoki, M. Bettini, J. Assfalg, and A. del Bimbo, "Classification of raw material sports videos for broadcasting using color and edge features," in *IEEE International Conference on Multimedia and Expo (ICME'01)*, Tokyo, Japan, Aug 22-25 2001, pp. 665–668.
- [105] Y. Ariki and Y. Sugiyama, "Classification and retrieval of TV Sports News by DCT features using multiple subspace method," pp. 1488–1491.
- [106] E. Sahouria and A. Zakhor, "Content analysis of video using principal components," *IEEE Transactions on Circuits and Systems for Video Technology (CVST)*, vol. 9, no. 8, pp. 1290–1298, 1999.
- [107] G. Iyengar and A. B. Lippman, "Models for automatic classification of video sequences," in *Storage and Retrieval for Image and Video Databases VI*, ser. Proceedings of SPIE, vol. 3312, San Jose, January 1998, pp. 216–227.
- [108] N. Vasconcelos and A. Lippman, "Towards semantically meaningful feature space for the characterization of video content," in *International Conference on Image Processing ICPR'97*, vol. 1, Santa Barbara, California, June 1997, pp. 25–28.
- [109] Z. Rasheed, Y. Sheikh, and M. Shah, "Semantic film preview classification using low-level computable features," in *3rd International Workshop on Multimedia Data and Document Engineering (MDDE'03)*, Berlin, Germany, Sep 2002.
- [110] S. Moncrieff, S. Venkatesh, and C. Dorai, "Horror film genre typing and scene labeling via audio analysis," in *IEEE International Conference on Multimedia and Expo (ICME'03)*, Baltimore, USA, 6-9 July 2003. [Online]. Available: http://impeca.cs.curtin.edu.au/pages/simon_pub.html

- [111] C. Colombo, A. D. Bimbo, and P. Pala, "Retrieval of commercials by semantic content: the semiotic perspective," *Multimedia Tools and Applications*, vol. 13, no. 1, pp. 93–118, 2001.
- [112] Z. Liu, J. Huang, Y. Wang, and T. Chen, "Audio feature extraction & analysis for scene classification," in *IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing*, New Jersey, June 1997.
- [113] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden markov model," in *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, Dec. 1998, pp. 27–32.
- [114] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and systems for segmentation and structure analysis in soccer video," in *IEEE International Conference on Multimedia and Expo (ICME'01)*, Tokyo, Japan, Aug 22-25 2001.
- [115] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Unsupervised discovery of multilevel statistical video structures using hierarchical hidden markov models," in *IEEE International Conference on Multimedia and Expo (ICME'03)*, Baltimore, July 6-9 2003.
- [116] P. Chang, M. Han, and Y. Gong, "Extract highlights from baseball game video with hidden markov models," in *IEEE International Conference on Image Processing (ICIP'02)*, Rochester, NY, sep 22-25 2002.
- [117] Y. Ariki and K. Matsuura, "Automatic classification of TV news articles based on telop character recognition," in *Proceedings of IEEE International Conference on Multimedia and Systems*, vol. 2, Florence, Italy, June 1999, pp. 148–152.
- [118] P. Wang, R. Cai, and S.-Q. Yang, "A hybrid approach to news video classification with multi-model features," in *Fourth International Conference on Information, Communications & Signal Processing, Fourth IEEE Pacific-Rim Conference On Multimedia (ICICS-PCM 2003)*, Singapore, Dec 15-18 2003.
- [119] A. A. Alatan, A. N. Akansu, and W. Wolf, "Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing," *Multimedia Tools and Applications*, vol. 14, no. 2, pp. 137–151, 2001.
- [120] Y. Li and C.-C. J. Kuo, "Movie event detection by using audiovisual information," in *IEEE Pacific-Rim Conference on Multimedia (PRCM'01)*, Beijing, October 2001, pp. 198–205.
- [121] J. Nam, M. Alghoniemy, and A. H. Tewfik, "Audio-visual content-based violent scene characterization," in *ICIP'98*, vol. 1, Chicago, Illinois, Oct. 1998, pp. 353–357.
- [122] S. Moncrieff, C. Dorai, and S. Venkatesh, "Analysis of environmental sounds as indexical signs in film," in *The Second IEEE Pacific-Rim Conference on Multimedia*, Beijing, China, October 2001. [Online]. Available: http://impca.cs.curtin.edu.au/pages/simon_pub.html
- [123] H.-B. Kang, "Video abstraction techniques for a digital library," pp. 120–132, 2002.
- [124] —, "Affective content detection using HMMs," in *ACM Multimedia (ACMMM'03)*, Berkeley, CA, Nov 2-8 2003, pp. 259 – 262.

- [125] X.-D. Yu, L.-Y. Duan, and Q. Tian, "Shot classification of sports video based on features in motion vector field," in *IEEE Pacific Rim Conference on Multimedia*, ser. Lecture Notes in Computer Science, Y.-C. Chen, L.-W. Chang, and C.-T. Hsu, Eds., vol. 2532. Springer, 2002, pp. 253–260.
- [126] W. Hua, M. Han, and Y. Gong, "Baseball scene classification using multimedia features," in *IEEE International Conference on Multimedia and Expo (ICME'02)*, Lausanne, Switzerland, August 26-29 2002, pp. 347–350.
- [127] A. Ekin and A. M. Tekalp, "A framework for tracking and analysis of soccer video," in *Visual Communication and Image Processing 2002*, ser. Proceedings of SPIE, vol. 4671, San Jose, CA, Jan 2002, pp. 763–774.
- [128] B. Li and I. Sezan, "Semantic sports video analysis: approaches and new applications," in *IEEE International Conference on Image Processing (ICIP'03)*, vol. 1, Barcelona, Spain, sep 14-17 2003, pp. 17–20.
- [129] V. Kobla, D. DeMenthon, and D. Doermann, "Detection of slow-motion replay sequences for identifying sports videos," in *IEEE 1999 International Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, Sept. 1999.
- [130] H. Pan, P. Beek, and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Salt Lake City, Utah, may 2001. [Online]. Available: citeseer.nj.nec.com/386396.html
- [131] I. Ide, K. Yamamoto, and H. Tanaka, "Automatic video indexing based on shot classification," in *International Conference on Advanced Multimedia Content Processing (AMCP'98)*, Osaka, Japan, nov 1998, pp. 87–102.
- [132] S. Eickeler, A. Kosmala, and G. Rigoll, "A new approach to content-based video indexing using Hidden Markov Model," in *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Louvain-la-Neuve, Belgium, June 1997, pp. 149–154.
- [133] N. Vasconcelos and A. Lippman, "A bayesian framework for semantic content characterization," in *IEEE International Conference on Pattern Recognition*, Santa Barbara, June 1998, pp. 566–571.
- [134] W. Mahdi, M. Ardebilian, and L. Chen, "Automatic scene segmentation based on exterior and interior shots classification for video browsing," in *SPIE Conference on Multimedia Storage and Archiving Systems IV*, Boston, sep 1999.