REIHE INFORMATIK
16/96

On the Detection and Recognition of Television Commercials

R. Lienhart, C. Kuhmünch and W. Effelsberg
Universität Mannheim
Praktische Informatik IV
L15,16
D-68131 Mannheim

On the Detection and Recognition of Television Commercials

Rainer Lienhart, Christoph Kuhmünch and Wolfgang Effelsberg

University of Mannheim, Praktische Informatik IV, 68131 Mannheim, Germany
{lienhart,kuhmuenc,effelsberg}@pi4.informatik.uni-mannheim.de

ABSTRACT

TV commercials are interesting in many respects: advertisers and psychologists are interested in their influence on human purchasing habits, while parents might be interest in shielding their children from their influence. In this paper, two methods for detecting and extracting commercials in digital videos are described. The first method is based on statistics of measurable features and enables the detection of commercial blocks within TV broadcasts. The second method performs detection and recognition of known commercials with high accuracy. Finally, we show how both approaches can be combined into a self-learning system. Our experimental results underline the practicality of the methods.

1 Introduction

Commercials play an important role in our lives - whether we like it or not. Popular institutions such as TV are mainly sponsored by advertisers or supported by advertising. For companies commercials are marketing instruments essential to drawing attention to their products and increasing their sales. These companies generally charge other companies to verify that their TV commercials are actually broadcasted as contracted. Presently, employees/humans must watch TV to carry out such verification. It would be desirable to transfer this task to a computer system. Such a computer system would watch TV and record precisely the spot time and date of broadcasting and the channel identifier. Perhaps companies would also like to observe automatically what their competitors are doing. Marketing companies may be interested in relating the measurable features of the different spots to their success in the market. These are all objectives of potential interest to producers and advertising agencies. On the consumer side, parents might want to shield their children from the commercials' influence by interrupting TV during commercial breaks.

With a commercial detection system applications such as commercial broadcast logging and commercial-free TV could be achieved. Some of the possible applications require only the detection of commercials as such whereas others also require recognizing a particular spot. In the paper we therefore describe two different approaches to commercial detection: one is feature-based and the other is recognition-based. The first approach only detects commercial blocks as such, while the latter also allows recognition of known commercials and can even distinguish slightly different versions of the same spot.

The paper is structured as follows: Section 2 presents the most important technical features of TV commercials and TV commercial blocks, using German television as an example. For all these features we derive detection indicators in Section 3 and combine them into a complete feature-based commercial detection system. Section 4 presents our second, recognition-based approach. It is capable of commercial detection as well as of commercial recognition. In Section 5 we combine both approaches into a reliable self-learning commercial detection system. Finally, Section 6 concludes our paper with an outlook on future work.

2 Technical Features of TV Commercials

In this section we lay down the features of commercials. Although the focus is on German television, most of the features are also valid for commercials in general or an equivalent can be found in other countries. These features ordinarily distinguish commercials from other film genres such as feature films, newscasts and sportscasts.

2.1 Structure of a Commercial Block

Generally, commercials are grouped into commercial blocks, which are simply a sequence of several consecutive commercials. A typical (German) commercial block contains the following elements (Figure 1):

- a commercial block introduction,
- a sequence of commercials (spots),
- a broadcasting station's advertisements and previews, and
- optionally a film introduction or short repetition of the cast.

A commercial block is always preceded by a transitional sequence leading from the broadcast into the commercial block itself. This limiting sequence of 3 to 5 seconds' length makes the difference in content clear to the viewer and is called "*commercial block introduction*" in German broadcasting. Broadcast stations are required by law to visually distinguish the broadcast clearly from the interrupting commercial block (we will address this legal point

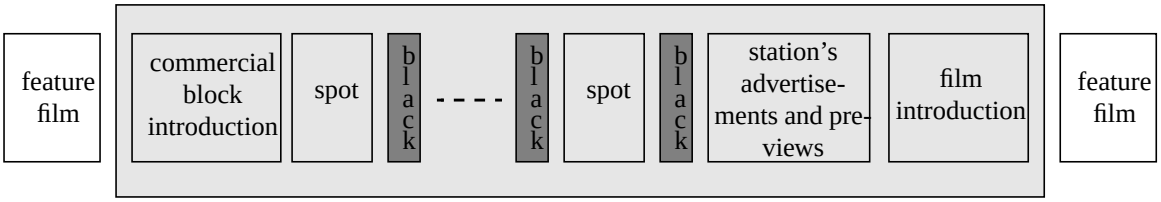

Figure 1: Structure of a German commercial block.

later). The introductions change frequently, e.g. in correspondence to the four seasons or special events such as the Olympic Games. On the other hand, the transitional sequence usually never changes during the transmission of a telecast. Once recognized it can be used for the detection of subsequent insertions of commercial blocks during the same telecast.

A *film introduction* has properties similar to those of a commercial block introduction. It is a short transition back to the program whose aim is to signal to the observer that the movie is continuing, e.g. by a film title. However, it is often omitted since there exists no pertinent legal regulation. As a substitution in the case of movies, some channels replay the last shots of the movie broadcasted right before the commercial break.

A *broadcasting station's advertisements and previews* announce upcoming or future telecasts on that channel. Typically, they last between 15 and 30 seconds.

The commercial *spots* themselves are video sequences lasting between 5 and 90 seconds. Several are broadcasted consecutively. Individual spots are separated from each other by dark monochrome frames. A particularity of German telecasts is that a stations's screen logo is turned off during commercials and turned on again afterwards.

2.2 List of Technical Features

The features of commercial blocks and individual spots can be divided into two groups: Those directly measurable and those measurable only indirectly. Directly measurable are low-level features which can easily be detected by the computer, while indirectly measurable features are of a higher level of abstraction and more difficult to compute. Moreover, some features are valid for all commercials blocks and/or spots, while others are only valid for a subset of them.

Directly Measurable Features

1. A directly measurable feature of commercial blocks and spots is their restricted temporal length. Generally, a spot lasts no longer than 30 s and a block no longer than 6:30 min. The maxima ever observed were 90 s for a spot and 8 min for a block.
2. Two consecutive commercials are separated by a short break of 5 to 12 dark monochrome frames [5][10].
3. The volume of the audio signal is turned up during commercials and turned down afterwards.

Indirectly Measurable Features

1. A human observer perceives commercials as full of motion, animated, and full of action. This sensation is supported by a high frequency of cuts and quick changes in color contents.
2. A commercial contains many still images. In particular, the last scene is often a "still" image presenting the product, the company and/or product name.

3. There exist special editing habits which are frequently used and can be recognized automatically.
4. Often text appears within commercials. The text shows the product or company name and other useful semantic information. It can be identified and evaluated [9][11].

In Section 3 we will show how these features can be computed and how relevant they are for detecting commercials.

Legal Regulations
In Germany the ratio of commercials to other televised material is regulated by law. The regulations differ for private and public TV stations with the regulations for public TV being more restrictive. The following table summarizes the restrictions for private TV stations:

Restriction	Value
Maximum share of commercial time	20% of the daily broadcasting time
Maximum share per hour	12 minutes
Minimum distance between two commercial blocks	20 minutes
Commercial block introduction	clearly visible distinction between broadcast and commercial

Table 1: Legal regulations about commercials on German private TV.

Additional regulations restrict commercials during movies which is of importance in our case because we concentrate on those in this paper: Movies may not be interrupted more than once in 45 minutes.

3 Feature-based Detection of Commercials

In this section we investigate how the technical features of commercials can be measured. We present our computational indicators and analyze experimentally their ability to identify commercial blocks and spots. In general, the features will not simultaneously hold true for each commercial block: Not every commercial spot will end with a freeze image, contain text or depict a moving action. Moreover, some feature films may also exhibit features which are typical of commercial spots. Therefore, our feature-based commercial detection system operates in two steps: First, potential commercial block locations within a video sequence are located; then they are analyzed in more detail.
Eight sample video sequences recorded from television are used to prove the characteristics of the features:

name	length	block start	block end	commercial start	commercial end	# of spots	avg. spot length
Aliens 2	22:53	07:57:20	14:54:05	08:00:23	14:21:17	17	00:22
Dancing with the Wolves	22:37	07:58:01	14:35:00	08:02:06	14:04:16	17	00:19
Scent of Women	18:33	05:58:23	12:20:06	06:08:01	12:16:04	16	00:21
The Firm	17:40	04:53:06	11:47:10	04:56:08	11:26:21	18	00:20
Black Rain	17:24	05:01:14	12:21:14	05:04:18	11:20:17	18	00:18
Superman	29:20	08:40:00	11:36:00	08:42:19	11:07:21	7	00:16
Sneakers	20:00	04:59:03	12:40:18	05:02:23	11:52:02	16	00:24
Star Trek 4	25:46	10:39:13	19:0:22	10:43:24	18:01:18	16	00:26

Table 2: The sample video set (times in *minutes:seconds* or *minutes:seconds:frame#*).

Whenever parameters (e.g. thresholds) have to be determined for the feature indicators they are derived from the first five sample videos and validated by the three remaining ones. Note that in all our examples the commercials were embedded in feature films; we have no experience yet with commercials embedded in other genres, such as sports or newscasts.

3.1 Monochrome Frames

In Section 2 we have pointed out that individual commercial spots within a commercial block are always separated by several dark monochrome frames. They can be identified easily by calculating the standard intensity deviation σ_I of the pixels of each frame. Here, intensity equals the gray-level of the pixels. For a perfect monochrome frame σ_I should assume zero. In practice, in the presence of noise, a frame is regarded as monochrome if σ_I drops below the small threshold $t_{MF\sigma}$. In order to detect dark monochrome frames only, the average intensity μ_I is also required to be below the small threshold $t_{MF\mu}$. In formulas:

$$MF(I) = \begin{cases} \text{dark monochrome frame} & (\sigma_I \le t_{MF_\sigma}) \wedge (\mu_I \le t_{MF_\mu}) \\ \text{other monochrome frame} & (\sigma_I \le t_{MF_\sigma}) \wedge (\mu_I > t_{MF_\mu}) \\ \text{polychrome frame} & else \end{cases}$$

$$\text{with} \quad \mu_I = \frac{1}{N} \cdot \sum_{n=1}^{N} I_n \quad \text{and} \quad \sigma_I = \sum_{n=1}^{N} (I_n - \mu_I)^2 \quad .$$

In the formulas the original image is represented as a list of intensity values I_n, one for each pixel, N pixels in total. In Figure 2 we depict the dark monochrome frame occurrences during a feature film interrupted by a commercial break. Note the large difference in the frequency of such frames during the commercial blocks and during the feature films. We also measured the distribution of the length of such dark-frame sequences. As you can see in Table 3, the length of commercial separation is usually between 0.12 s and 0.4 s. We conclude that any monochrome frame sequence shorter than 0.12 s or longer than 0.4 s is therefore not a commercial separator.

Name	**Monochrome frame sequence length distribution used for commercial separation**						
	<.12	**<.2**	**0.2**	**0.24**	**0.28**	**<=.4**	**>.4**
Aliens 2	0	.06	.82	.06	0	0	.06
Dancing with the Wolves	0	0	0	0	.63	.37	0
Scent of Women	.07	0.8	.13	0	0	0	0
The Firm	.06	.29	.41	.06	.06	.12	0
Black Rain	0	0	.12	.59	.06	0.17	0.06
Superman	0	0.17	.83	0	0	0	0
Sneakers	0	0	.27	.07	.47	.2	0
Star Trek 4	.07	0	.07	0	.2	.53	.07

Table 3: Monochrome frame sequence length distribution.

On German TV commercial blocks of at least 4 spots can reliably be detected by the following simple detection scheme: Find each sequence of at least three monochrome sequences of 0.12 to 0.4s which are not further apart than 60 seconds. In result, 99,98% of the candidate sequences in our test set were part of a commercial block, no block was missed, but 15.3% of the overall length of a commercial block was not detected, i.e. the commercial block introduction, the first and last spot and the station's advertisements and previews. Thus, monochrome frame sequences are a strong commercial block indicator. However, they generally miss a substantial part of a commercial block.

3.2 Scene Breaks

In this subsection we analyze the style and frequency of scene breaks used in commercials and feature films. Here, we will concentrate exclusively on hard cuts and fades.

Hard Cuts

While watching commercials you may notice the high editing frequency. Since most scene transitions are hard cuts, a high hard-cut frequency can be observed during commercials. Hard cuts are scene breaks which result from splicing two shots together without any transition. They are perceived as an instantaneous change from one

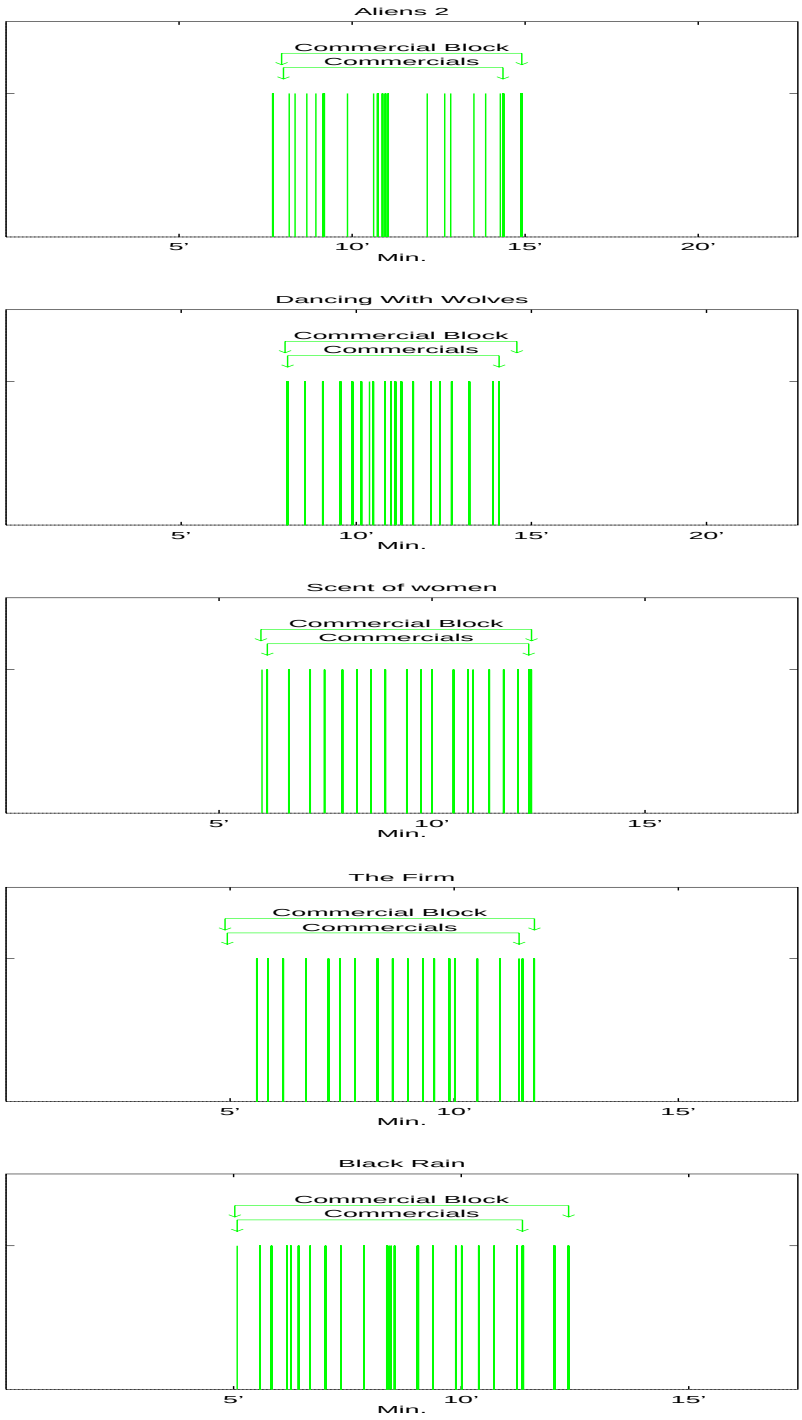

Figure 2: Monochrome frame distribution for the sample video set. To save space only the graphs of the first five sample videos are shown. The other three graphs look similar.

shot to another [1]. The difference in color histograms between consecutive frames has been proven to work successfully in detecting hard-cuts [3]. Thus, we compute a 64-bin color histogram over the entire frame considering only the two most significant bits of each color band, and normalize it by the number of pixels in the frame. Then, the color histogram difference between two successive frames is calculated. A shot boundary is declared if the difference exceeds threshold $t_{HardCut}$.

However, even within a set of hard cuts you can distinguish between stronger and weaker ones. Weak hard cuts are characterized by a low difference of histogram values close to $t_{HardCut}$, while the difference for strong hard

cuts is significantly above $t_{HardCut}$. Since commercial blocks consist of a set of non-related spots, we expect strong hard cuts between them. Moreover, they want to give the observer the impression of dynamics and action. Hence, even within a spot, the scene of action changes frequently, resulting again in strong hard cuts. We detect them by applying a second, higher threshold to the difference in histograms between consecutive frames: $t_{Strong-Hardcut}$. Figure 3 shows the averaged values for the video samples. Notice that the hard-cut frequency is not as significant as the strong hard-cut frequency for our purpose. Thus, only the frequency of strong hard cuts is further considered as a discriminator.
It is obviously difficult to determine the right values for the thresholds. In our studies so far we have "manually" set all threshold values based on the first five sample videos. In principle it would be correct to compute the optimal threshold values based on a statistical analysis of the features of the sample video set [16].

The average strong hard-cut frequency, in strong hard cuts per minute, is 20.9 for spots and only 3.7 for the rest of our video samples.

To detect potential commercial block locations, we select each connected subgraph of the strong hard-cut graphs in Figure 3 as a potential commercial block. The graph is regarded as disconnected at all locations where it drops below 5 strong hard cuts per minute. Each candidate sequence is rejected if it does not exceed 30 strong hard cuts per minute at least once. Applying this rule to our test video set, all commercial blocks are found. On the average, the detected ranges covered 93.43% of the commercial blocks and 0.09% of the non-commercial block sequences. Thus, strong hard cuts are a good pre-filter for commercial blocks.

Fades
Fades are scene breaks which gradually blind out from a scene into a monochrome frame or blind in from a monochrome frame into a scene [1] [6]. Either the first or last frame is monochrome and exhibits a standard intensity deviation σ_I close to zero. In contrast to that, the alternate end point shows the scene in full intensity and, thus assumes a large standard intensity deviation value. In between these two extremes σ_I is either monotone increasing or monotone decreasing. For nearly all fades the graph can be specified in more detail: during a fade the graph of σ_I plotted against the frame number is either linear or concave. This characteristic temporal behavior of the standard deviation of intensity enables fade patterns to be reliably detected. Thus, our indicator detects a fade if the following conditions hold for a sequence of consecutive σ_I values:

- it consists of linear segments with a minimum correlation of 0.8
- each linear segment has a minimum length of 10 frames
- the gradients of the segments are either decreasing and positive or increasing and negative
- either the last or the first σ_I value of the sequence is monochrome.

For our commercial blocks the fade rate was 0.5 fades/minute contrasting to 0.02 fades/minute for our feature film set.

Note that the subsequent features are only calculated for non-monochrome frames, and from scene break to scene break. The scene transition frames are no longer considered.

3.3 Action

Typically spots have a high level of "action". The human perception of action is influenced by many different aspects: The impression of action can be caused by fast-moving objects in the scene, e.g. people running around or fighting with each other. But the impression of action can also be caused by certain editing and camera control operations [1]. For instance, frequent hard cuts and many zooms also result in an impression of action. Moreover, a calm scene with pumping and changing colors is perceived as action, too. These are many different aspects of "action" which we want to measure (partially) by the following indicators:

- edge change ratio, and
- motion vector length.

Initially, we also investigated the motion energy (= the sum of the pixel differences between consecutive images) since it seemed to be optimally suited to registering all three different aspects, especially change of colors and camera operations. It turned out, however, that the quality of information generated is below that of the other action indicators. Thus, motion energy is no longer considered.

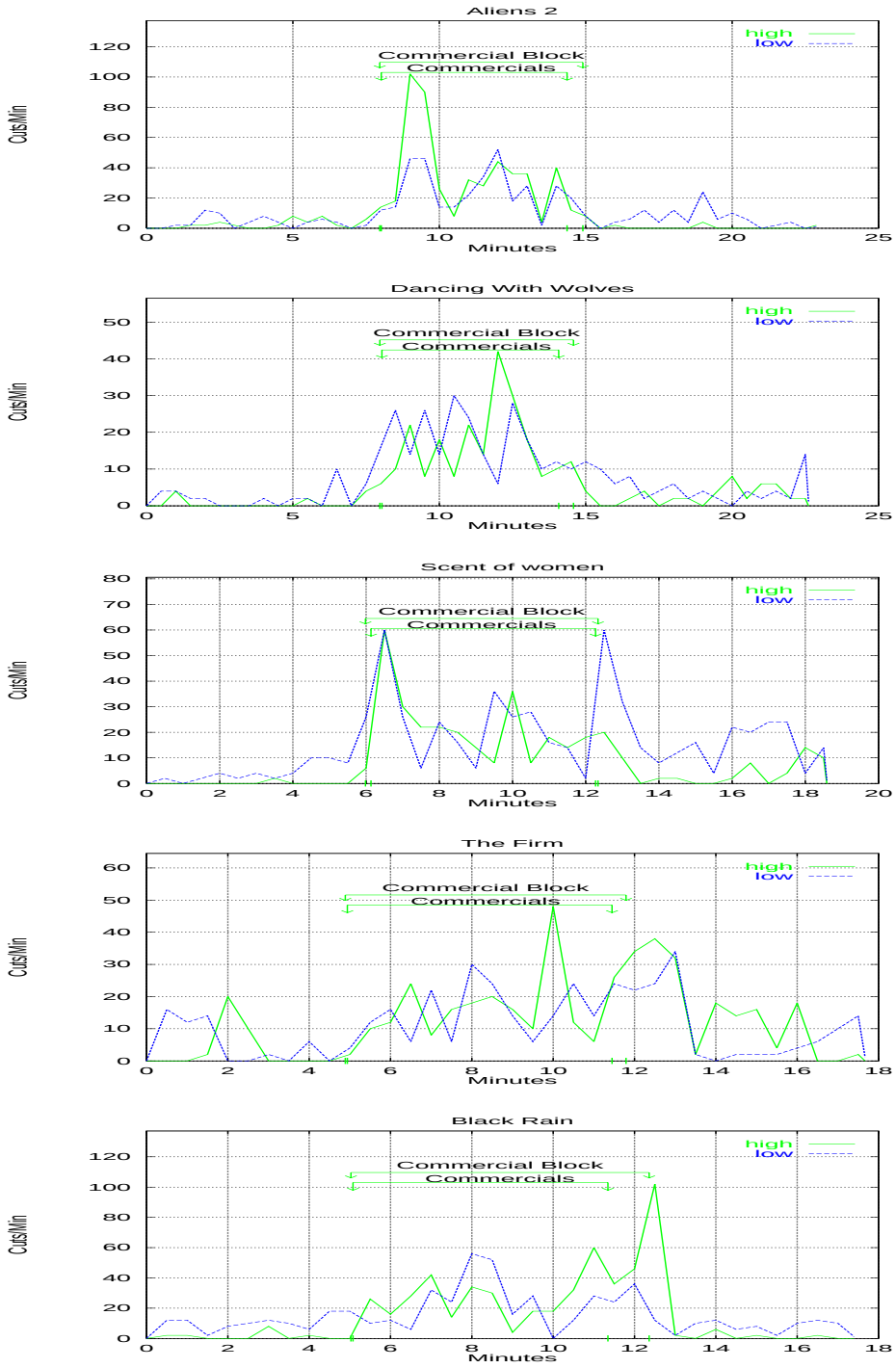

Figure 3: Hard cuts per minute for the sample video set. Notice the difference between “strong” and “soft” hard cuts. To save space only the graphs of the first five sample videos are shown. The other three graphs look similar.

Edge Change Ratio (ECR)

The edge change ratio (ECR) was proposed as a characteristic feature by Zabih, Miller, and Mai [18]. They used the well-known Canny edge detection algorithm [2], although, in principle any edge detection algorithm could be used.

Let σ_n be the number of edge pixels in frame n and X_n^{in} and X_{n-1}^{out} the number of the entering and exiting edge

pixels in frame n and n-1, respectively. Then the edge change ratio ecr_n between frame n-1 and n is defined as

$$ecr_n = max\left(\frac{X_n^{in}}{\sigma_n}, \frac{X_{n-1}^{out}}{\sigma_{n-1}}\right)$$

The advantage of the edge change ratio as a characteristic parameter is that it registers structural changes in the scene such as entering, exiting and moving objects as well as fast camera operations. However, it is somewhat independent of variations in color and intensity since it relies on sharp edges only. Consequently pumping images have no effect on the indicator. Thus, it registers two of the three parts of "action" listed at the beginning of the subsection. Notice that the edge change ratio is only calculated within each shot. It is not used here for detecting scene breaks.

As can be seen from the graphs in Figure 5 the edge change ratio for commercial blocks is dynamic; it is often much more static for feature films. Thus, a commercial block candidate can be detected by frequent changes above a threshold t_{ECR}. The indicator's extended finite state machine is depicted in Figure 4. When we applied this indicator to our set of test videos, all commercial blocks were detected. In average, the detected ranges covered 96.14% of the commercial blocks and 0.09% of the non-commercial block sequences.

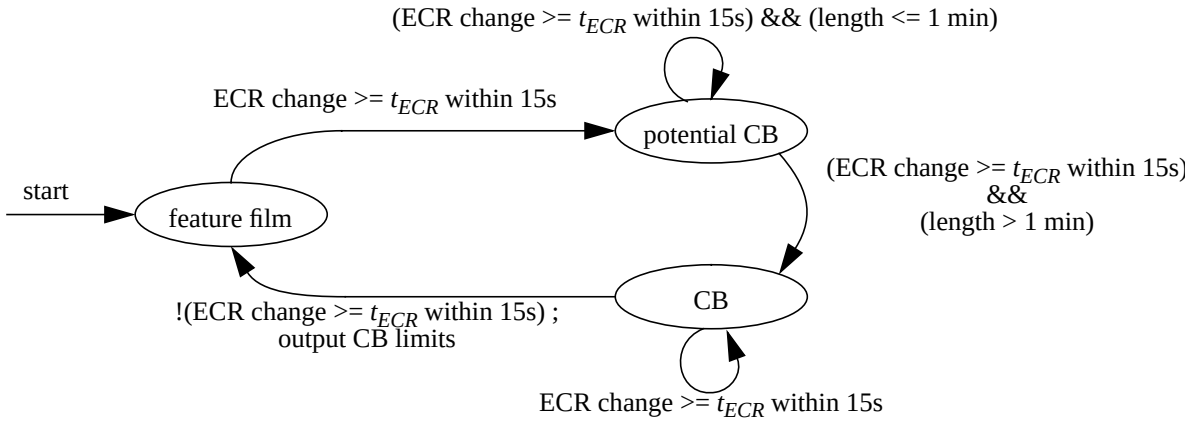

Figure 4: Extended finite state machine of the indicator for dynamic subgraph ranges in the action graphs (i.e. in the ECR and in the motion vector length graphs).

Motion Vector Length

An important feature of action is fast object movement. The motion vector length measures object movement by using an algorithm similar to a motion compensation algorithm used by MPEG encoders [4] [6] called "Exhaustive Search Method".

Each single frame of the video is divided into so-called macroblocks of 16x16 pixels. The best matching position for each macroblock of a frame is calculated by comparing the block with each possible position within an area of 20 pixels around the original location.

The result of the matching operation is a motion vector with the length of the distance between the position of a block in two consecutive frames. With (x_1,y_1) the position of the macroblock in the first frame and (x_2,y_2) its position in the consecutive frame, the length of the vector for a macroblock i is calculated as follows:

$$MB_i = \begin{cases} \sqrt{20^2 + 20^2} & \text{if position cannot be located} \\ \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} & \text{else} \end{cases}$$

A closer look at Figure 5 shows that a commercial block candidate can also be detected by our indicator for dynamic subgraph ranges. Applying this indicator to our test video set, all commercial blocks are found. On the average, the detected ranges covered 96.2% of the commercial blocks and 0.2% of the non-commercial block sequences.

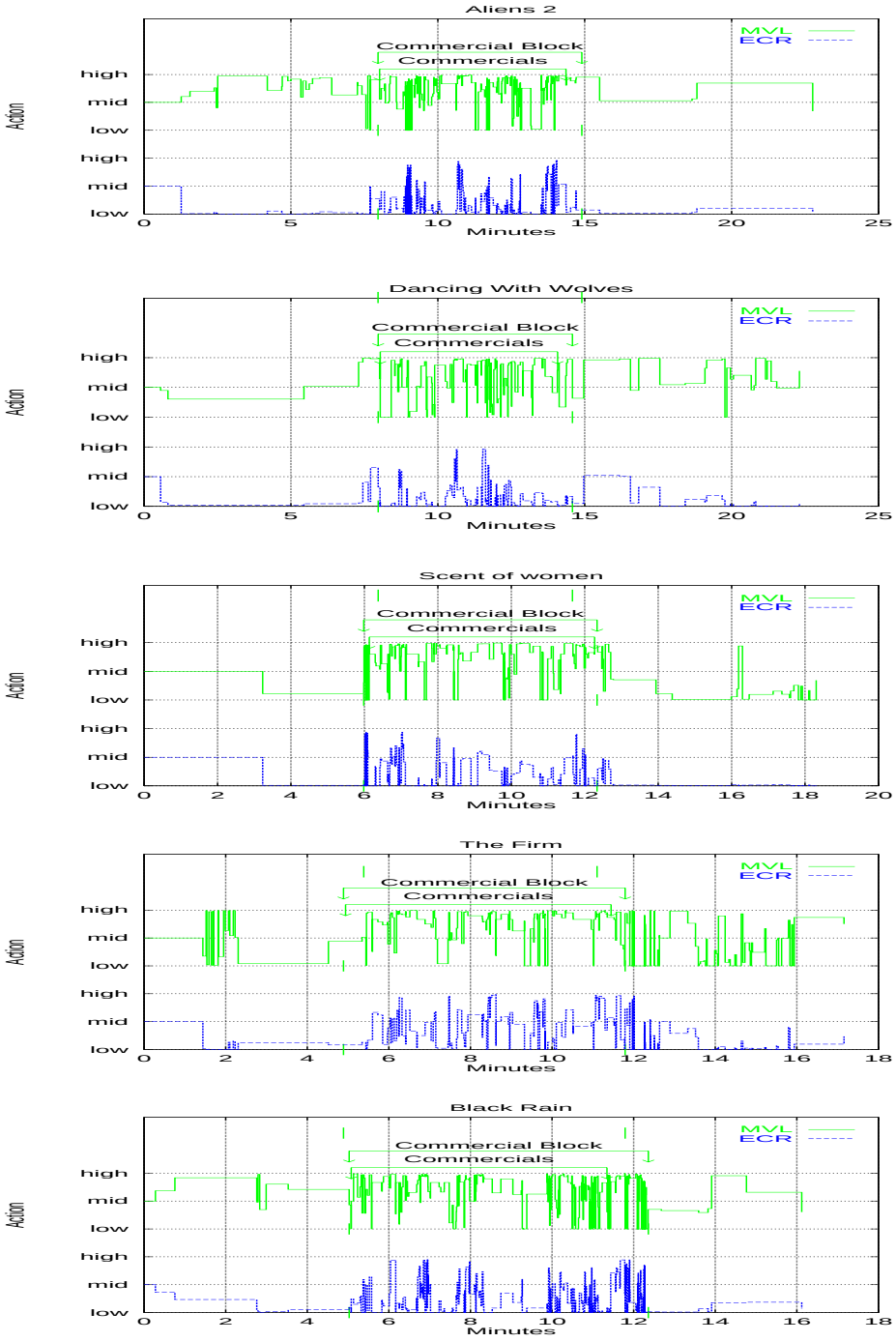

Figure 5: The different indicators for "action" and typical patterns during a commercial block. To save space only the graphs of the first five sample videos are shown. The other three graphs look similar.

3.4 A Feature-based Commercial Detection System

Having introduced the characteristic features, we now explain their composition into a system accomplishing accurate results at reduced computational costs. Our commercial detection system uses the monochrome frame sequence feature and the strong hard cut feature as fast pre-selectors; the accurate, but computationally expensive

action detector is utilized to determine the precise limits. We distinguish between the following cases:

- If both pre-selectors indicate a commercial block, i.e. the intersection of their detected candidate ranges is not empty, or only the strong hard cut feature, a commercial block is detected. The action criterion is now used to find its precise limits. If the action criterion at one limit is below the commercial block criterion, the search for the precise limits is performed towards the inner range, otherwise outwards from the range.
- If only the monochrome frame sequence detector indicates a commercial block the range is regarded as a false detection.

Applying this detection system to our test video set, all commercial blocks were selected. On the average, the ranges detected covered 96.14% of the commercial blocks and 0.09% of the non-commercial block sequences. However, computation time is reduced by a magnitude of one in comparison to that required to calculate all features for all frames.

If our objective is to save all commercial block frames while discarding as many feature film frames as possible (O1) the following supplementary rules must be added to the above outline of the feature-based commercial detection algorithm. The detected commercial block ranges are extended by 30 seconds at both ends. If it is our objective to save all feature film frames while discarding as many commercial block frames as possible (O2), the detected commercial block ranges are shortened by 10 seconds at both ends. In practice, these values result in excellent outcomes.

4 Recognition of Known Commercials

The feature-based commercial detection system presented so far allows a rough localization of the commercial blocks. However, to determine their precise limits, i.e down to a single shot, the system would have to be capable of grouping semantically related shots [16]. Additionally, some of the features used may easily be changed in the future, for example, the delimiting monochrome frames between spots: They can easily be omitted by the television stations.
Furthermore, in other film genres such as sports or in other countries such as the USA, programs are sometimes interrupted by a single commercial without any transition to or from it. Reliable detection of a single commercial is difficult since the feature-based approach expects it to have a minimum length. If the block is too short, features either do not change enough due to averaging, or the change is too short to distinguish it from accidental runaways. A second approach, able to cope with the stated situations, is described here. It is based on the fact that commercials often run on TV for an extended period of time, and it is thus possible to store and recognize features of *known* commercials.

Recognition-based detection of commercials depends on a database of an initial set of spots whose recognition in the current program is the aim. Individual spots are stored and compared on the basis of comprehensive fingerprints. Two questions will be investigated in the following: What is a suitable and comprehensive fingerprint for shots, and how should two fingerprints be compared?

4.1 Fingerprint

A commercial spot consists of a sequence of images. Accordingly, we construct a fingerprint of each spot by calculating important features per frame and then represent the spot's fingerprint as a sequence of these features. We call the representation of the value of a feature a *character*, the domain of possible values an *alphabet*, and the sequence of characters a *string*.

A feature used for a fingerprint should meet the following requirements:

- It should tolerate small differences between two fingerprints calculated from the same spots, but broadcasted and digitized at different times. The differences are caused by slight inaccuracies in rate and color, or by TV and digitizer artifacts.
- It should be easy/fast to calculate and rely on only few values, so that computation, storage and comparison of fingerprints remain inexpensive.
- It should show a strong discriminative power.

As an example we use the following simple feature as a fingerprint: the color coherence vector (CCV) [14]. In our opinion it fulfills the requirements: CCVs are fast to calculate, show strong discriminative power and tolerate slight color inaccuracies. However, rate inaccuracies (such as dropped frames) must be absorbed by the compari-

son algorithm.

Color Coherence Vectors
The color coherence vector (CCV) [14] is related to the color histogram. However, instead of counting only the number of pixels of a certain color, the color coherence vector also differentiates between pixels of the same color depending on the size of the color region they belong to. If the region (i.e. the connected 8-neighbor component of that color) is larger than t_{ccv}, a pixel is regarded as coherent, otherwise, as incoherent. Thus, in an image there are two values associated with each color j:

- α_j, the number of coherent pixels of color j and
- β_j, the number of incoherent pixels of color j.

A color coherence vector then is defined as the vector

$$\langle(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)\rangle \quad .$$

Before calculating the color coherence vector we scaled the input image to 240x160 pixels and smoothed the image by a Gaussian filter of sigma 1 as also done by Pass et. al. [14]. t_{ccv} was set to 25, and the color space used only the two most significant bits of each RGB color component.

4.2 Comparison

Let us now introduce our fingerprint matching algorithm. Given a query string *A* of length *P* and a longer subject string *B* of length *N*, the *approximate substring matching* finds the substring of *B* that aligns with *A* with minimal substitutions, deletions and insertions of characters [10] [13]. The minimal number of substitutions, deletions and insertions transforming A into B is called the minimal distance *D* between *A* and *B*. Two fingerprint sequences *A* and *B* are regarded as identical if the minimal distance *D* between query string *A* and subject string *B* does not exceed the threshold $t_{stringDist}$, and the difference in length does not exceed 90%, i.e. *P/N* is greater than or equal to 0.9. At first glance use of approximate substring matching rather than approximate string matching seems questionable since we want to identify identical spots; however, in our experiments we noticed that commercials are sometimes slightly shortened at the beginning and/or end, and the distance *D* should not be increased by this effect. The approximate matching procedure guarantees that sequences recorded with minor rate and color inaccuracies can still be found.
In addition, it cannot be expected that the same commercial spot recorded at different times and from different broadcasting stations have identical fingerprints. Long sequences are more likely to contain erroneous characters, and thus $t_{stringDist}$ is set in relation to the length of the search string *A*. There exist several fast approximate substring matching algorithms with worst-case time complexity O(DN) requiring only $O(P^2)$ to $O(D^2)$ space. We use the one proposed by Landau and Vishkin [8].

4.3 Recognition-based Commercial Detection

We use the fingerprint and comparison techniques as follows to identify individual spots precisely. A sliding window of length *L* seconds runs over the video, stepping forward from shot to shot (see Figure 6), each time calculating the CCV fingerprint of the window. At each position the window fingerprint is compared with the first *L* + *S* seconds of each spot fingerprint stored in the database. If two are similar, the window is temporally expanded to the whole length of the candidate fingerprint in the database and the two are compared (see Figure 7). If a commercial is recognized, the window jumps to the end of that commercial, otherwise it only shifts forward to the next shot.
Recognition-based detection, like feature-based detection, consists of two steps: Step one aims to reduce the computational cost by shortening the fingerprints to be compared at the expense of less discrimination power. Therefore, this step can only detect candidate spots. Step two determines whether the candidate is identical to a stored spot. The reason for setting the subject string to be of length *L* + *S* in the first step is to avoid an increase of the approximate distance by frames dropped at the start of the commercial, which might occur in practice. Therefore, *S* will always be chosen as low as possible and should be zero in the ideal case. For our test spots *S* = 2 frames was fine.

We do not require *L* to be less than the length of the shortest possible commercial since in that case the role of the

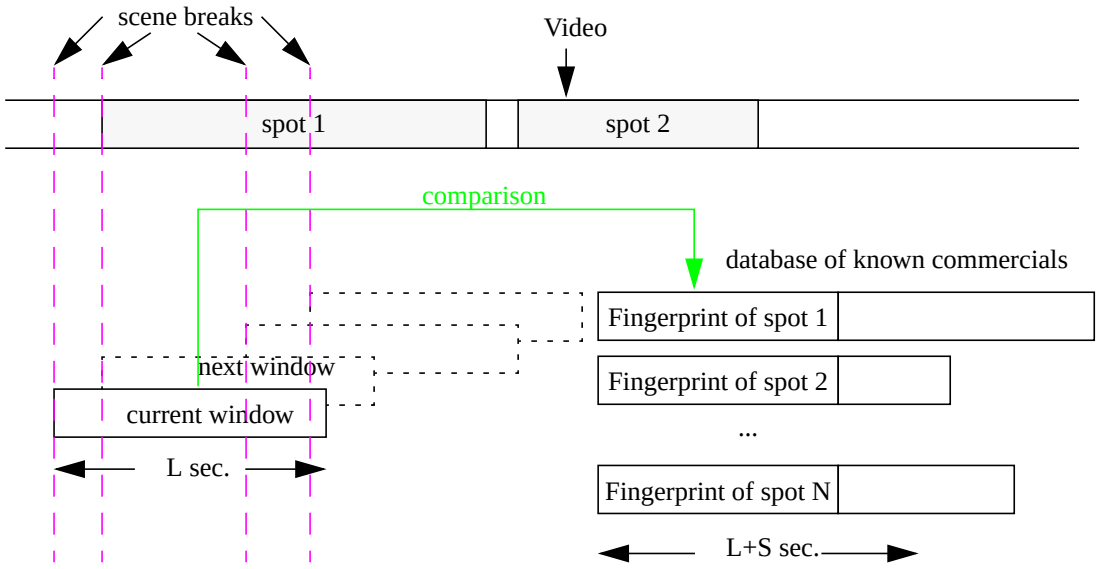

Figure 6: First step of the recognition-based commercial detection system.

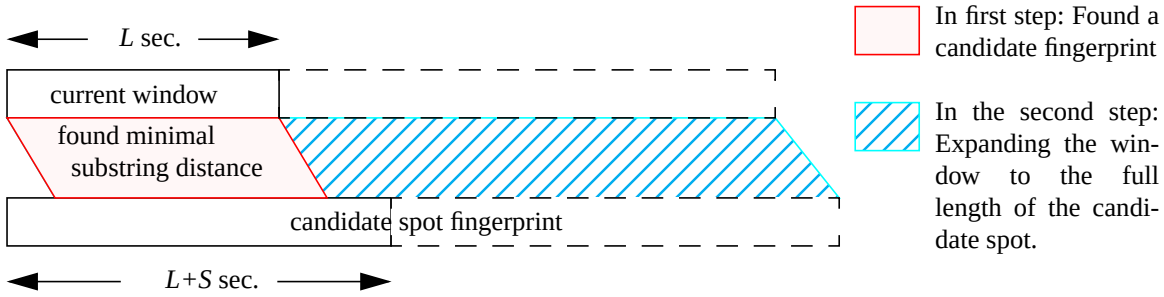

Figure 7: Expansion process in the second step of the recognition-based commercial detection system.

two fingerprints would be swapped. However, a computationally optimal value for *L* is difficult to determine. Two factors affect the computation time:

- Firstly, the cost of finding candidate commercial spot windows. Given an O(DN) comparison algorithm, M as the number of commercials in the database, D_{rel} as the maximum difference expressed in percent, the time needed for comparing the fingerprint of the window with one in the database is $O((L*D_{rel}) * (L+S))$ and the time for determining all candidate commercial spots for a window is proportional to $O((L*D_{rel}) * (L+S) * M)$. In the formula, (L+S) specifies the length of each fingerprint used from the database for comparison, and D*L the maximal difference allowed.
- Secondly, the test whether a candidate is a known commercial or not. It is obvious that the complexity is fixed for any candidate, and the total complexity thus depends on the number of candidates determined per window position by the first step. This number increases reversely proportional to the length *L* of the fingerprint: The lower L is, the less discriminate is the window fingerprint. But it is difficult to specify this change in probability by formulas since the consecutive values are highly correlated. Thus we determined heuristically good values for L and S by run-time analysis with the video test set.

Experimental Results

We digitized 200 commercial spots from several German TV channels. This set contained a number of new spots as well as the spots from our sample video set, but recorded from different TV channels and/or at different times. We applied our spot recognition approach to each sample in our video set. All commercials were recognized, none was missed and none was falsely detected. The localization was also very high. On the average the difference between the precise and detected locations was only 5 frames.

The processing time for our recognition system was only 90% of the duration of the video sample in real-time, once the CCV values for the video had been calculated. Therefore, by using a fast assembler implementation of the CCV computation and string comparison, the whole process could be performed in real-time.

5 Combining the two Approaches into a Self-learning Commercial Detection System

In this chapter we describe a commercial detection system that makes use of both formerly described approaches. It is our objective to build a system of the same precision in localization and recognition of spots as the second approach while reducing the computational cost and keeping the system automatically up-to-date, i.e the database should "learn" new spots autonomously. For the following we assume that our system is always on-line; or in other words: the system is constantly running, checking all TV channels for commercial blocks.

The two approaches are combined hierarchically: In the first step we use the feature-based approach to reduce the number of candidate commercial areas by means of the "monochrome frame sequence" and "strong hard cut" criteria. Since the feature-based approach is used as a pre-filter, the objective is to miss as few spots as possible. Therefore, we follow the objective O1 as described in Section 3.4. In the second step the recognition-based approach is used to identify the individual spots and determine their exact borders. That way the computationally expensive approximate string comparison must only be applied to a small subset of the video; the scene breaks and thus the hard cuts have to be computed anyway for the recognition-based approach and detection of the monochrome frame sequences is very inexpensive.

Furthermore we try to find unknown spots automatically, i.e. let the system learn new spots autonomously. We assume that our database contains almost every broadcasted spot. So if a new commercial is broadcasted for the first time we can assume that it is usually surrounded by known spots. If so, the commercial recognition algorithm will find the end of the known spot sent before and the beginning of the known spot sent after the new one, defining the exact position of the unknown spot. After removing the dark monochrome frames at the head and tail the new spot can be inserted into the database. Problems will arise in the following cases:

- The unknown spot is either the first or last one in a commercial block. But it is not very likely that a new spot will always be the first or the last in the commercial blocks. It can be assumed that sooner or later the spot will be surrounded by two well-known spots in one of the next commercial blocks and then be inserted into the database.
- More than one unknown spot is surrounded by known spots because in that case we cannot distinguish between the different spots. The system would assume that the two spots are one - although quite long - and would insert the whole piece into the database. We deal with this case by searching for the appearance of monochrome frame sequences of characteristic length to break the sequence up into the individual spots.

To overcome the problem of erroneously inserted clips we suggest the following: First we can let the system require confirmation by the user each time a new spot is detected. Second we can insert the clips only provisionally. The clip will only then be inserted permanently when it is found in other commercial blocks, too. If the clip cannot be detected in other block after a certain time it will be removed automatically.

Finally, the precise borders of the commercial blocks must be determined. Thus, the first 2 minutes of each commercial block are searched for repeatedly appearing sequences of 3 to 5 seconds. If such a sequence can be found in several commercial blocks of the same channel the sequence is regarded as a commercial block introduction and added to a database labeled "commercial block introduction". Consequently, the commercial block candidates are searched not only for known and new commercials but also for known and new commercial block introductions. This procedure allows precise determination of the beginning of the commercial blocks. Unfortunately, this does not work for the end of a commercial block due to the lack of legal regulations. Thus the end of a commercial block must be determined roughly via the features as described in Section 3.
We have not yet done any experimental studies with the integrated algorithm, but are planning to do so.

6 Conclusions

This paper describes two methods for detecting and extracting commercials in digital videos. The first approach is based on the heuristics of measurable features. It uses features fundamental to TV advertising such as high action

rate and short shot length. These features cannot easily be changed by the advertising industry. Only the short dark monochrome sequences used for as commercial separators have - strictly speaking - nothing to do with a commercial spot in itself and can therefore easily be replaced by an other separator. This feature must therefore be adjusted to local habits. For instance, on our sample tapes from the US the commercials have been separated by one to three dark monochrome frames, often surrounded by a fast fade.
The performance of the feature-based commercials detection system was quite high: 96.14% of all commercial block has been selected, while misclassifying only 0.09% feature film frames. The system can easily be adjusted to the local commercial features in different countries.

The second approach relies on a database of known commercial spots. Due to its design - it recognizes commercials known in advance - it attains high precision. Moreover, the method is also capable of recognizing individual spots. No adjustments for different countries are needed. The performance is very high: all spots were recognized with no false hits at all.
Both approaches have been combined into a reliable self-learning TV commercials detection and recognition system.

So far we have only tested our feature-based detection approach on a limited number of samples. In the upcoming months we will use the system with the parameters derived from the initial set to analyze new video material, in particular genres other than feature films.

In near future we will also extend our work into the audio domain and explore the different application domains in which our commercial detection and recognition system could be used.

Acknowledgements

We would like to thank Stephan Fischer for sharing with us his experience with genre recognition.

References

[1] David Bordwell, Kristin Thompson. Film Art: An Introduction. *McGraw-Hill*, Inc., 4th ed., 1993.
[2] John Canny, “A Computational Approach to Edge Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, No. 6, pp. 679-697, Nov. 1986.
[3] John S. Boreczky and Lawrence A. Rowe. Comparison of video shot boundary detection techniques. In Storage and Retrieval for Still Image and Video Databases IV, Proc. SPIE 2664, pp. 170-179, 1996.
[4] Eric Clan, Arturo Rodriguez, Rakeshkumar Gandhi, and Sethuraman Panchanathan. *Experiments on block-matching techniques for video coding*. Multimedia Systems, 2(5):228-241, December 1994.
[5] Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. “Automatic Recognition of Film Genres”. *Proc. ACM Multimedia 95*, San Francisco, CA, pp. 295-304, Nov. 1995.
[6] D.L. Gall, “MPEG: A Video Compression Standard for Multimedia Applications”, Communications of the ACM, 34, 4, April 1991.
[7] Arun Hampapur, Ramesh Jain, and Terry Weymouth. Production model based digital video segmentation. *Journal of Multimedia Tools and Applications*, Vol 1, No. 1, pp. 1-38, March 1995.
[8] G. M. Landau and U. Vishkin. Introducing efficient parallelism into approximate string matching and a new serial algorithm. *Symp. on Theory of Computing*, pp. 220-230, 1986.
[9] Rainer Lienhart. Automatic Text Recognition for Video Indexing. *Proc. ACM Multimedia 96*, Boston, MA, pp. 11-20, Nov. 1996,
[10] Rainer Lienhart, Silvia Pfeiffer, and Wolfgang Effelsberg. The MoCA Workbench: Support for Creativity in movie content analysis. *Proc. of the IEEE Conference on Multimedia Computing & Systems*, Hiroshima, Japan, pp. 314-321, June 1996.
[11] Rainer Lienhart and Frank Stuber. Automatic Text Recognition in Digital Videos. In *Image and Video Processing IV 1996*, Proc. SPIE 2666-20, pp. 180-188, Jan. 1996.
[12] Eugene W. Meyers. A Sublinear Algorithm for Approximate Keyword Matching. *Algorithmica* 12, 4-5, pp. 345-374, 1994.
[13] T. Ottmann and P. Widmayer. Algorithms and Data Structures. *BI-Verlag*, Mannheim, 1993. (in German)
[14] Greg Pass, Ramin Zabih, Justin Miller. Comparing Images Using Color Coherence Vectors. Proc. *ACM Multimedia 96*, Boston, MA, pp. 65-73, Nov. 1996.

[15] A. Murat Tekalp, “Digital Video Processing”, Prentice Hall Signal Processing Series, 1995.
[16] Charles W. Therrien. Decision, Estimation, and Classification: An Introduction to Pattern Recognition and Related Topics. John Wiley & Sons, Inc. 1989.
[17] Minerva Yeung, Boon-Lock Yeo, and Bede Liu. Extracting Story Units form Long Programs for Video Browsing and Navigation. *Proc. of the IEEE Conference on Multimedia Computing & Systems*, Hiroshima, Japan, pp. 296-305, June 1996.
[18] Ramin Zabih, Justin Miller, and Kevin Mai. A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. *Proc. ACM Multimedia 95,* San Francisco, CA, pp. 189-200, Nov. 1995.