

Project number	: AC018
Project title	: SMASH
Deliverable Type	: Public

CEC Deliverable Number	: a0018/tud/it/ds/p/005/b1
Internal Project Number	: SMS-TUD-648-2
Contractual Deliverable Date	: 30 November 1996
Actual Date of Deliverable	: 25 November 1996
Title of Deliverable	: Report on Technical Possibilities and Methods for Security of SMASH and for Fast Visual Search on Compressed/Encrypted Data
Contributing Workpackages	: WP 320, WP 330
Nature of Deliverable	: report
Author(s)	: Alan Hanjalic (TU-Delft) Gerrit C. Langelaar (TU-Delft) Reginald L. Lagendijk (TU-Delft) Marco Ceccarelli (Philips Research Eindhoven) Mario Soletic (Philips Research Monza)

Abstract:

The main objective of this deliverable is to present and evaluate the state-of-the art concerning security and visual search on compressed data and new accomplishments in these areas which are of use for SMASH.

Keywords list:

copy protection, visual search on compressed video, multimedia database for consumer system applications.

<i>For further information please contact:</i>	<i>Dr. Eric Persoon, project leader SMASH Philips Research Eindhoven e-mail: persoon@natlab.research.philips.com</i>
--	---

CONTENTS

1	ROLE OF VISUAL SEARCH AND SECURITY IN THE COMBO	5
1.1	CONCEPT OF SMASH COMBO SYSTEM	5
1.2	IMPORTANCE OF SEARCHING AND SECURITY: STATE-OF-THE-ART CONTRIBUTIONS	6
1.3	SOLUTION CONCEPTS	7
2	STATE-OF-THE-ART IN VISUAL SEARCH OF VIDEO	8
2.1	INTRODUCTION	8
2.2	VIDEO-PARSING	9
2.2.1	FEATURES	11
2.2.2	METRICS	11
2.2.3	THRESHOLDING	12
2.3	VIDEO-CONTENT REPRESENTATION	13
2.3.1	INTRODUCTION	13
2.3.2	KEY-FRAME BASED VIDEO-CONTENT REPRESENTATION	13
2.4	CLUSTERING	14
2.4.1	INTRODUCTION	14
2.4.2	METHODS OF CLUSTERING	16
2.5	CONCLUSIONS	16
3	STATE-OF-THE-ART IN SECURITY	17
3.1	INTRODUCTION	17
3.2	COPY PROTECTION FOR DIGITAL AUDIO	17
3.3	CRYPTOGRAPHIC PROTECTION OF THE DIGITALLY BROADCASTED MATERIAL	19
3.4	COPY-PROTECTION OF ANALOGUE AND DIGITAL VIDEO	22
3.5	INTERFACE OF DIGITAL EQUIPMENT TO PC	23
3.6	PROTECTION OF PRIVATE DATA	24
3.7	CONCLUSIONS	24
4	REQUIREMENTS AND SOLUTION CONCEPT	26
4.1	VISUAL-SEARCH	26
4.1.1	GENERAL REQUIREMENTS ON VISUAL-SEARCH ENGINE	26
4.1.2	VIDEO-PARSING	27
4.1.3	KEY-FRAME EXTRACTION	28
4.1.4	CLUSTERING	30
4.2	SECURITY	31
4.2.1	REQUIREMENTS FOR THE SMASH PROTECTION SCHEME	31
4.2.2	THE COPY PROTECTION SCHEME FOR THE SMASH SYSTEM	31
4.2.3	PROTECTION OF PRIVATE DATA	34
4.2.4	EVALUATION OF THE PROTECTION SYSTEM	34

4.3 SMASH MMDB: WHEN RETRIEVAL IS MORE IMPORTANT THAN STORAGE	34
5 CONCLUSIONS	36
6 REFERENCES	37

1 Role of visual search and security in the COMBO

R.L. Legendijk

This chapter outlines the main features of the SMASH storage system (COMBO) and introduces the role of visual search and security. It also outlines the structure of the deliverable.

1.1 Concept of SMASH Combo System

The overall objective of the SMASH project is to develop and to show the technical feasibility of an integrated storage unit for multimedia applications in a domestic environment. Integral part of this storage unit are provisions for searching and protection of the stored information, and in particular for the visual information. This deliverable describes the technical possibilities and methods for these functions in the SMASH project.

Figure 1.1 shows the project's conceptual view on the place that the SMASH storage unit has in the consumer's home. The storage system is externally connected to the network through either a set top unit (DVB input) or through a PC (internet connection). In the domestic environment the SMASH storage unit serves as a remote server for multimedia applications. For this reason the unit itself does not have a user interface, but is accessed through either the PC application with Internet/Java interface or through the set top unit (STU) application running for instance an MHEG engine.

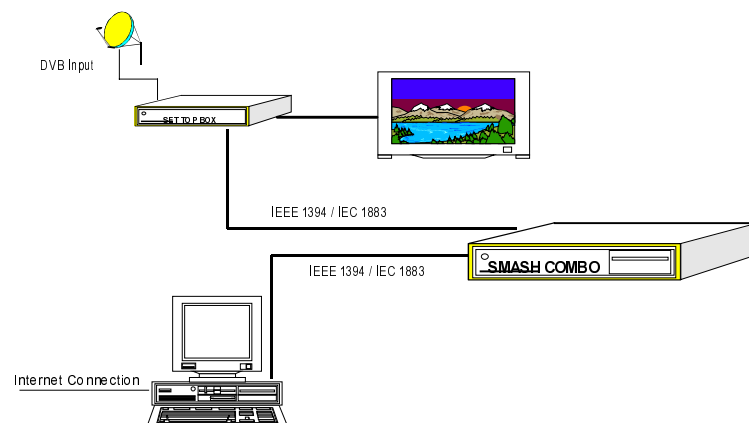


Figure 1.1: SMASH storage unit in the domestic environment

The architecture of the storage unit is built around two tightly coupled but technologically different storage media, namely a linear tape drive (LTD) and a hard disk drive (HDD). The linear tape drive is a cheap (in terms of Ecu/Mbyte) high capacity storage medium (at least 50 Gbyte) with relatively slow access time, while the hard disk drive provides a relatively expensive smaller capacity storage (e.g. 2 Gbyte) with fast random access. The two components of the SMASH storage unit are managed by the operating system of the unit

itself, while PC/STU applications can access the data residing on LTD or HDD through the Application Programming Interface (API).

The multimedia data used in applications will be dynamically stored in a distributed way, depending on the character of the multimedia data and on the application. The main distinction in storage will be between (audio-visual) stream data (to be stored on the LTD) and random access data files (to be stored on the HDD). To the user, however, the storage unit will be transparent, i.e. the user is not aware of how the data is distributed between the LTD and HDD. The chosen architecture of the storage unit combines the advantages of hard disk technology (fast and random access) with those of linear tape technology (huge storage capacity). The MultiMedia DataBase (MMDB), to be stored on the HDD, plays a crucial role in storing various kinds of service information, in relating information stored on the LTD and the HDD, and in providing efficient user access to the stored information.

The development of a storage system as outlined above requires the solving of many technical and conceptual challenges. In the SMASH project many of the architecture and application problems and choices have been discussed. Of course, these issues cannot and have not been discussed in isolation, since there are many interrelated issues in architecture, applications, and the systems special functionalities. Taking into account the developments in other branches of the project, this deliverable concentrates on specific software/hardware functionalities for managing the stored data, among which are searching of stored video streams, copy protection and encryption of private data, and the multimedia database (MMDB) for organizing and efficient retrieval of stored information.

Of particular interest in this deliverable is the DVB-VCR, which is one of the SMASH project's application studies: a recorder and player for satellite and cable DVB services, especially digital MPEG video with enclosed service information, which will use the storage unit for additional functionalities (break button, recording and playback of multiple stream) and increased user friendliness (browsing and selection of services using the program guides and visual contents of the services, management of tape collections).

1.2 Importance of Searching and Security: State-of-the-art contributions

With the development of mass storage devices such as the SMASH storage system, also comes the need for (i) efficient management and (ii) protection of the locally stored multimedia information. The importance of these two functions and the place where and how they are discussed in this deliverable are addressed in this section.

There is a growing world-wide interest in the development of browsing methodologies and tools tailored to the storage system and applications that accommodate for easily locating specific pieces of information in huge volumes of information. Existing navigation systems are all based on prior annotation of the information or search processes on textual information. It is widely recognized that there is a need for intelligent management and search methods in particular for the visual information in multimedia documents and in digital video and image libraries. The problems at hand are firstly that textual annotation can usually not be made beforehand because of the subjectivity and complexity of this task, and secondly that

visual information will be transmitted and stored in a compressed format complicating even the most simple image analysis operations.

In one of the applications selected by the SMASH project, digital video broadcasting (DVB) services will be recorded. To accommodate for an efficient navigation of the stored information, some sort of contents-based summary of the stored video streams needs to be made as the streams are received and stored. In Section 2 of this deliverable, the state-of-the-art in making such a summary of video information is described.

In the recent past we have experienced that the success of digital video broadcasting was not only determined by the existence of technical solutions (e.g. MPEG-2), but also by the ways in which service providers would be able to make money. The latter has led to a complex system of enciphering MPEG transport streams, deciphering, and billing. On similar grounds one can state that the success of multimedia storage systems not only depends on the technological advances, but also the existence of adequate copy protection methods. Service providers will not be willing to offer services in digital form without copy protection mechanism that limits duplication of the digital multimedia data.

Copy protection systems exist in which the protection signals are stored separately from the data. Since these protection signals can easily be removed without affecting the data quality, this system is insufficiently strong. More recent approaches embed the protection signals into the data, which is called labelling or watermarking. In Section 3, the state-of-the-art in copy protection of analogue and digital audio and video information will be presented.

1.3 Solution concepts

On the basis of state-of-the-art in searching, protection, and combo architecture, and taking into account the application requirements, the SMASH project has developed requirements and a solution concept for visual search, copy protection, and storage of the (summarized) information in the MultiMedia DataBase (MMDB). In Section 4 these requirements and the solution concepts will be outlined.

2 State-of-the-art in visual search of video

A. Hanjalic

2.1 Introduction

We are witnessing an immense growth in the development of digital libraries. Collected in such libraries are large volumes of digitized multimedia data, which are reachable via electronic networks by any user world-wide. Libraries containing a variety of well-organised digital data bring many advantages, such as preservation of quality of stored information and almost unlimited possibilities to manipulate and browse through data using comfortable user interfaces. However, with steadily increasing amount of data stored in such systems, the problem of efficiently navigating through these data and quickly and reliably finding and retrieving any part of it, becomes increasingly important.

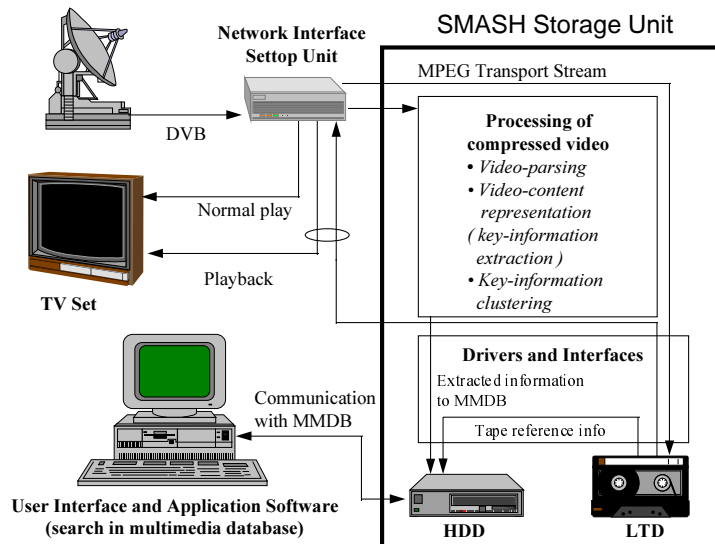


Figure 2.1: Scheme with video-processing steps needed for visual search of video and links to other components of the SMASH system as well as to the user's home environment

The most efficient way to interact with huge amounts of data, stored in modern digital storage systems, is to interact not with the data itself, but with the *abstract* of that data, which contains a compact representation of the whole stored content. The need for an abstract is especially strong in systems with *distributed storage*, like SMASH, where the user-interaction with the storage part (tape) containing the actual data is not efficient enough. Much easier user-interaction would be possible on the other storage component (disk) of the SMASH system, which then does not contain the actual data, but only the abstract. This data-representation should enable the user to get an impression about the actual stored content, i.e. to enable *content-based browsing/query* through data without the need to interact with the data itself. By making an abstract, special problems appear when dealing with stored *visual*

data, primarily due to their high semantical complexity. This chapter is confined to the visual search done on *stored video-data*.

Figure 2.1 shows the idea of making the video-abstract directly on the incoming (compressed) video-stream. There are three major steps in this process:

- *Video-parsing*, i.e. splitting the sequence in segments with uniform temporal content (video-shots), each being an object of representation through appropriate key-information and providing at its boundaries semantically logical entrance points for the retrieval process.
- *Video-content representation*, i.e. representing the content of each video-shot through a compact characteristic information (key-information).
- *Clustering*, i.e. organizing the total extracted key-information aiming at more efficient user-interaction.

In the area of *visual search of video* considerable amount of research results can be found in literature, providing steadily more efficient and reliable ways of performing the steps mentioned above. The importance of this research comes from the fact that the efficiency of the actual browsing/query process on stored video data is highly dependent on the way how video-processing in Figure 2.1 is done. In the following, the state-of-the art for each of these steps will be discussed in more details.

2.2 Video-parsing

The first video-processing item given in Figure 2.1 divides a given video stream in certain elementary segments, called *video-shots*. A video-shot can be defined as an unbroken sequence of frames, e.g. a zoom of a person talking [18]. There are several reasons why this process of *video-parsing* must be performed. Firstly, breakpoints between consecutive shots are semantically logical entrance points for video-retrieval. Secondly, shot-boundaries determine sequence parts each having one and the same content, i.e. they are *elementary content units* which are objects of representation, as it will be described in Section 2.3. Just like in the case of text, where elementary index units are words and phrases, also in the video some elementary entities are needed to be indexed [25]. Since one separate frame of the sequence has practically no meaning, the next longer elemental unit is the defined specific sequence of frames - video-shot. Defining these elementary units is done by detecting transitions between them.

There are three major classes of shot-transitions to be detected by video-parsing: sharp transitions (camera breaks), gradual transitions (dissolves, fade-out followed by fade-in) and special effects (fade-in, fade-out, wipe) combined with sharp transitions (e.g. fade-out followed by a camera break).

For each of these transitions, different techniques must be applied for their detection. The first step in the detection process is generally performed by measuring *content-changes* between each two consecutive frames of the sequence. The result is the curve of frame-to-frame differences (in the following text referred as FFD). This curve has, in case of sharp transitions, peaks at each place the difference is measured between frames from different shots, i.e. on the actual sharp shot-break. The logic behind this effect is that content changes between consecutive frames within a shot are much smaller than content changes between frames belonging to two different shots. If the transition is not sharp, its detection becomes more problematic. The dissolve is, for instance, spread over several frames (e.g. 15) resulting in FFD values between frames which are generally bigger than those within the shot but which cannot be defined as peaks. Better in this case is to talk about *plateau's*, i.e. gradual increase of FFD values followed by their gradual decrease.

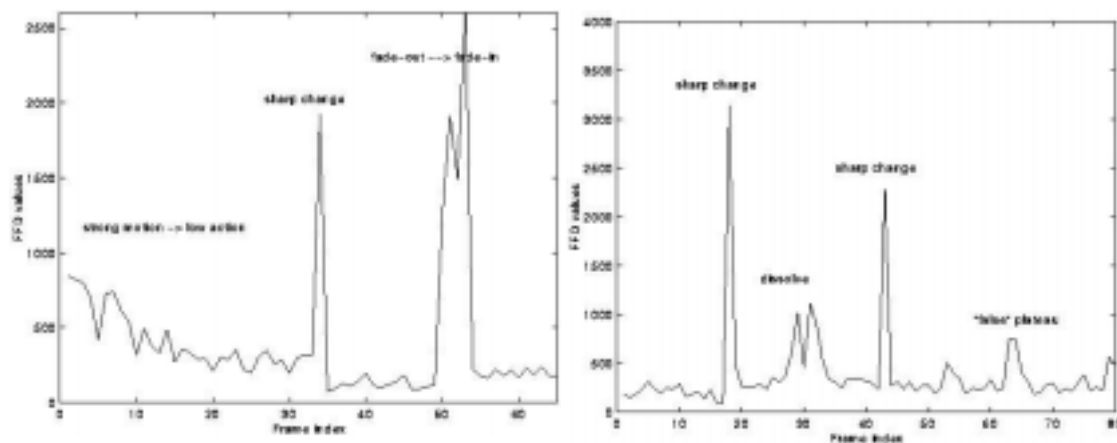


Figure 2.2: Some possible cases of shot-transitions and behaviour of FFD values along different sequences. The sequence left is a part of one music video-spot, and the one on the left the start of a soccer game. Both sequences are in MPEG-1 format, whereby the analysis is done on DC images of I frames and approximated DC-images of P-frames.

Other cases including special effects result again in other ways of FFD behaviour, which are first to be modeled in order to be detected. A further complicated factor by detecting shot-transitions are cases of fast camera movements causing large difference values. This causes a very similar behaviour of FFD-values like in the case of dissolve and could therefore be mixed up with an actual (gradual) shot-change. Also the case of a strong motion or fade followed by a sharp break or vice-versa causes considerable problems by the detection. Figure 2.2 shows the behaviour of FFD-values for two different sequences and with some characteristic transitions.

Two major issues are to be considered:

- The first issue is the way of measuring changes between each two consecutive frames. These changes should be based exclusively on the comparison of contents in these frames, i.e. only real *content-changes* should be taken into account. For instance, small movements of some objects by static backgrounds, camera movements, focal length

changes, or changes in luminance should not be registered as relevant changes for proclaiming shot breaks [30]. Frame-to-frame variations within a single shot will be dealt with by the process of video-content representation. The issue of video-parsing is related primarily to finding an appropriate feature describing the frame content and a metric to evaluate changes of that feature.

- The second issue is the actual detection of sharp peaks and plateau's out of all measured FFD-values. This is done by setting an appropriate threshold to evaluate FFD values. How difficult the threshold setting is, depends partially on the way how the FFD curve is obtained. The more distinguishable peaks and plateau's, compared to FFD values within each shot, the weaker is the thresholding problem, since there is a certain safety distance between the highest FFD value of a shot and the lowest peak-value of a shot-change. However, the thresholding always remains a problem, since the height of the peaks and appearance of plateau's cannot be estimated in general. It is therefore not realistic to expect from any video-parsing algorithm the reliability of 100% - independent of the way how FFD values are obtained.

2.2.1 Features

In order to ignore small changes between subsequent frames of a sequence, which are not relevant for shot-transitions, large spatial averaging on frame data before the actual video-parsing process is proposed very frequently (e.g.[30]). For the same reason, the feature used to represent a frame should be characteristic for the entire frame. Among such *global* frame-features a big majority of authors prefers *histograms*, which can be computed for each of colour components separately or for the frame-colour in combination (e.g. 3D-RGB histograms). Histograms appear to be a widely used tool for describing a general frame content also due to easiness of their computation. In spite of many different metrics used in literature for comparing histograms and obtaining FFD values, good performance of this feature is often reported [22], [25], [30], [31]. The quality of performance is measured by sharpness of peaks, by sharp transitions and by distinguishability of plateau's of gradual transitions. However, due to the fact that such plateau's can appear also by strong object or camera motion, it is necessary to distinguish between these effects to avoid false detections in high-action scenes. Therefore an additional feature is needed which is specific for action effects: *the motion information*. Motion can also be used for video-parsing, e.g. for detecting shot-transitions occurring on P or B-frames in MPEG-compressed streams. When working with MPEG streams this information is easily available in form of motion vectors. Usage of this feature is demonstrated in [21], [31], [32].

2.2.2 Metrics

In several recent publications [33], [34], [30], [31], [25] different metrics for comparing frame features (histograms) have been evaluated.

In [22] and [23] a good performance is reported by using the absolute difference of histograms $h(i)$ of two consecutive frames (k , $k-1$) and merging characteristics of histograms of all three separate colour components. While in [22] this metric is applied to components of the RGB-space, in [23] the YUV space was used. The metric can be given as

$$d_{YUV}(k, k-1) = \sum_i \sum_{j=Y,U,V} |h_k^j(i) - h_{k-1}^j(i)|$$

Although many authors including [22], [25], [30] claim that it is enough to use the histogram of only one component, the experience shows that the merging effect leads to much higher detection reliability. Some experiments about this have been made in [22].

In their frequently referenced comparative study, Nagasaka and Tanaka concluded in [33] that the so-called χ^2 - comparison of colour-histograms, defined by

$$\chi^2 = \sum_i \frac{|h_k^{Col.Comp}(i) - h_{k-1}^{Col.Comp}(i)|^2}{h_k^{Col.Comp}(i)}$$

performs best, whereby [25] and [31] claim that its performance is not necessarily better than the performance of normal absolute differences.

As explained in the last section, motion information can also be used as a feature for video-parsing. For detecting shot-transitions occurring on P or B frames of an MPEG-stream, [32] proposes to measure ratios of numbers of macroblocks without and with motion compensation (P-frames) and numbers of backward and forward motion vectors (B-frames). By detecting dissolves, motion vectors are used to model different camera movements ([25] and [31]).

2.2.3 Thresholding

While for finding an appropriate feature and metric for comparison of consecutive frames a number of acceptable solutions already exist, the problem of interpreting difference values, i.e. actual selection of certain values to be associated with shot-changes, still remains the major obstacle in practice. The proper selection of difference values is done by setting *thresholds*. Furth et al. give in [25] a statistical approach for determining the threshold, based on measuring mean-value μ and standard deviation σ of frame-to-frame differences for the whole sequence. The global threshold T is then estimated as

$$T = \mu + \alpha\sigma \quad .$$

Experimental results in [25] suggest that α should have values between 5 and 6. Statistical parameters needed for the formula are obtained after analysing the complete sequence and claiming that the distribution of all FFD values along the sequence (not considering FFD values on shot-transitions and in segments with camera/object motion) is Gaussian. Problem with this and with other approaches with global threshold is the case where a distinguishable break-peak can be observed in one stationary part of the sequence, but whose height is similar to FFD values along a high-action-shot in an other sequence part. The concept of global thresholds can also not be used in systems where video-parsing process is to be done “on-the-fly”.

The described statistical approach can be adapted for “on-the-fly” processing in a sense that all statistical parameter are reset after each detected shot-change. This provides at the same

time a local threshold selection. However, every missed detection or “false alarm” influences threshold-values for coming shots in a negative way, causing a burst of detection-mistakes.

The approach given in [22] partially avoids this problem, because the investigation of frame-to-frame differences is done within a sliding window - not dependent on missed detection or false alarms before the starting point of the window. In this case the threshold is determined locally for a (non-)causal sliding window, and is therefore time variant.

2.3 Video-content representation

2.3.1 Introduction

The process of representing video-content starts with partitioning of a given sequence in elementary content units, as described in previous section. Characteristic information of all shots in a sequence form the abstract of a sequence, and will be used in the browsing or query process.

The browsing or query method to be applied on the abstract is directly dependent on the way how the shot-content is represented [31]. This representation can be done by manual annotation of the video-content using *key-words*, by extracting *characteristic features* (e.g. shape, colour, size, texture, temporal changes along each video-shot, etc.). We consider one such representation method, which is of large practical importance and based on *key-frames*.

2.3.2 Key-frame based video-content representation

Representation through characteristic frames (*key-frames*) has been addressed very frequently in literature (e.g. [21], [22], [20]). For each detected video-shot, a number of key-frames is chosen. All extracted key-frames out of the sequence are organized semantically (clustered) before being presented to the user. Video-browsing based on key-frames is done by going through cluster-tree, each step giving a set of key-frames representing corresponding video-parts. Due to the “action-based” video-sampling, the content-development of each video part can be recognized very easily only by looking at given samples. Therefore, this concept of video-content representation appears to be very suitable for video-browsing purposes. However, also when considering query processes where the search for video-parts containing some specific objects, persons or features is performed, the key-frame based concept can be very efficient since features collected from key-frames can be used.

A simple method to select key-frames is to take the first frame of each shot [36]. More reliable content-representation requires, however, non-uniform sampling of the frames in the video-shot. Current approaches [21], [20] mainly work by setting thresholds within each shot, measuring frame-to-frame differences and comparing them to that threshold. Most of the time, the first frame of the shot is chosen as one key-frame. For the rest of the shot, each time the difference value exceeds the threshold, the coming frame is proclaimed to be a key-frame.

However, the described procedure is typically a sequential process leading generally to unpredictable results. In particular, the final amount of key-frames for the whole sequence cannot be estimated for any given threshold. We can end up with a huge number of key-frames or simply with too few key-frames - not enough for browsing. This problem is also related to the available space for storing extracted key-frames (abstract). Secondly, it is rather difficult to relate any particular parameter value by threshold setting to the key-frame collection resulting from that setting. Furthermore, it is based on “subjective” thresholds which is not acceptable in fully automated and widely used systems.

Aiming at more objective and controlled key-frame based video-representation, suitable for applications in SMASH, we have developed a new key-frame allocation method ([23], [24]), which has following major properties:

- Regulation (limitation) of the maximal number of key-frames for the entire sequence and its content-based distribution along the sequence
- Optimal key-frame allocation, independent of any parameters (thresholds).

Details of this method are given in Section 4.

2.4 Clustering

2.4.1 Introduction

Key-frames are chosen to give the user a compact overview of all the stored video-material. The word “compact” is used here primarily to show the relation between sizes of the abstract and the original video-data. However, when analysing the absolute size of the abstract even for only one movie of a normal length (e.g. 2h) the usage of attribute “compact” seems to be not very suitable. By a frame-rate of 25 frame/sec and an approximate shot-length of 90 frames, 2000 video-shots appear as objects of representation through key-frames. After considering this fact, two possible solutions can be approached: either to find ways how to cope with a large number of key-frames or to try to represent the whole movie through a couple of key-frames which are positioned on “right” places. The problem is that also in the second case the total amount of key-frames will not be small, if an acceptable content-representation is desired, but also concerning the fact that the COMBO capacity is far larger than to store only one movie.

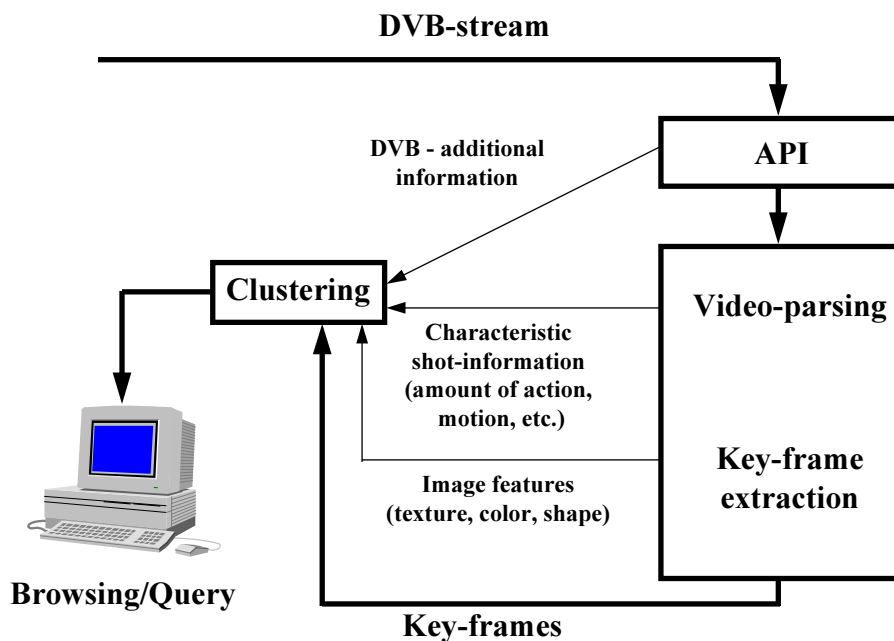


Figure 2.3: DVB application in the SMASH with the clustering block. All information needed for clustering video-shots can be seen

Defining efficient methods for organising the abstract data into meaningful clusters is a very important and research intensive area, either in case of stored image or video data. In case of video, there are two basic objectives of clustering:

- The first one is to ease the *browsing* through stored video. Browsing means mainly searching after a certain scene or part in the movie by going through the abstract. The clustering process is directed towards grouping of similar video-shots by using corresponding key-frames (their features) as well as specific temporal shot-characteristics. The attribute “similar” is related to certain grouping criteria, which can be application dependent (e.g. “all action scenes in one cluster”).
- The second objective is to enable *queries* on stored video-data. In this case, a search after a certain specific detail is possible (specific face appearing somewhere in the movie, all shots where some specific features - colour, texture, shape - can be found). The clustering process is directed stronger by feature-structure of key-frames representing each shot than by “global” shot-features. In comparison to the previous, the feature analysis of key-frames should be done much more carefully.

The Figure 2.3 shows the place of clustering in the visual search of video. This figure also shows that clustering procedures may use nonvisual information, for instance supporting DVB information.

2.4.2 Methods of clustering

All clustering methods can be divided into two major classes: *hierarchical methods* (leading to a tree-structure of data) and *partitioned methods* (leading to several parallel groups of data). Depending on the way the tree-structure is obtained, hierarchical methods can be further divided in *agglomerative methods* and *divisive methods* (going successively from coarse to fine splitting or vice versa).

Each clustering method is characterised by used features and metrics for comparing these features. The objective of each clustering method is to achieve maximal similarity of all objects belonging to one and the same cluster, where the similarity corresponds to the defined metrics and applied feature(s). Detailed study about colour-feature extraction can be found in [17]. In [19] a new image matching method in the subband domain is reported.

In [26] the creation of hierarchical scene transition graphs is proposed - a method of clustering all shots in a sequence for browsing purposes. Clustering is done by using colour, simple shape information and correlation of corresponding DC-images, as well as temporal variations in each shot to measure their similarity. Also a method for evaluating these mentioned similarity features, i.e. method of building proximity matrices for video-shots is presented. In [35], an overview of clustering methods for video-browsing is presented.

2.5 Conclusions

As discussed in all previous sections, there is a number of problems to be solved in order to obtain the video-abstract, whereby all mentioned requirements (“on-the-fly” processing, minimized parameter dependency, acceptable performance for general user, etc.) must be taken into account. Current research directions in the WP 320 deal with these problems, and some results - like those in section 2.3 - can be reported. Future plans include further improvement of the key-frame extraction algorithm through more sophisticated modelling of the actual content-development along the sequence, improvement of video-parsing related to its current dependency on subjective, sequence-specific parameters. However, notwithstanding the technical hurdles that remain to be taken for a fully automated key-frame extraction procedure, clustering of key-frames and the associated browsing/query process is still in its infancy but will determine the successful development of elegant user-interfaces to a great extent. Therefore, main emphasis within future plans for the WP 320 will be put on research in the clustering area.

3 State-of-the-art in security

G.C. Langelaar

3.1 Introduction

No universal copy protection system for all digital equipment exists yet. However, a copy protection mechanism has been defined for audio recorders. This mechanism is briefly discussed in this section.

Many different protection systems for digitally broadcasted material (analogue and digital video) are currently in use since every service provider uses his own system. A general overview of these systems is given in the next sections.

The role of the personal computer in the multimedia world becomes a major concern. To explain this, the interface of digital equipment to the personal computer is described.

Users must have the possibility to protect their own data against others (e.g. parental control). However, in some countries it is forbidden to use very strong cryptographic algorithms in consumer electronics. This problem is also addressed in this section. Finally conclusions are drawn.

3.2 Copy Protection for Digital Audio

On October 28, 1992, President Bush signed the Audio Home Recording Act into law [1]. The Act, an historic compromise between the consumer electronics and music industries, became effective immediately. The Act confirms consumers' right to use and retailers' right to sell all analogue and digital audio recording formats. As part of this compromise, digital audio recording devices must include a system that prohibits serial copying, and manufacturers or importers must pay a modest royalty on new digital audio recording devices and media.

In the U.S. digital audio recording or interface devices must contain one of the following:

- The Serial Copy Management System (SCMS), which permits first-generation digital-to-digital copies of pre-recorded music and other audio works, but prohibits multi-generation or "serial" copies of those copies (SCMS is implemented in DAT, MD and DCC recorders already on the market)
- A system with the same functional characteristics as SCMS, and which acts compatibly on the same copyright and generation status information as used by SCMS
- Any other system certified by the U.S. Secretary of Commerce as prohibiting unauthorised serial copying

Devices or services to circumvent the SCMS or any other serial copy control system may not be distributed. The Act applies only to "digital audio recording devices," defined as devices that are designed or marketed primarily for making digital audio recordings for private use (whether or not incorporated in some other device). The following devices are not generally subject to SCMS or royalty requirements:

- professional model products
- dictation machines, answering machines, and other audio recording equipment designed and marketed primarily for non-musical recording
- analogue audio recording devices or media
- personal computers
- VCRs and camcorders used primarily for video recording

All digital audio recorders like the DAT, DCC and mini-disc recorders, are equipped with the SCMS (Serial Copy Management Systems) to prevent consumers from making illegal copies of copyright protected material [2]. Using this system, a consumer can make digital copies of any digital source. However, such a copy can not be duplicated further using storage devices equipped with this protection method. The copy-prohibit-bits occur frequently in the data stream at fixed intervals.

S/PDIF is a serial one-line connection in one direction for the transport of digital stereo audio with the belonging subcode and error detection. To facilitate clock recovery from the data stream biphase-mark encoding is used. Each bit to be transmitted is represented by a symbol comprising two consecutive binary states. The first state of a symbol is always different from the second state of the previous symbol. The second state of the symbol is identical to the first if the bit to be transmitted is logical “0”, however it is different if the bit is logical “1” (see Figure 3.1).

Preambles are specific patterns providing synchronisation and identification of the subframes and blocks. These patterns violate the biphase mark code rules to avoid the possibility of data imitating the preambles. Three preambles are used (see Figure 3.1) to indicate the start of a sub-frame.

- X - indicate start of channel A
- Y - indicate start of channel B
- Z - indicate start of channel A and of a block of 192 frames

Each subframe contains two bits, which are part of the subcode data. The first bit is used for the user data block and is not used in most cases. The other bit (bit 30) is the same in each subframe (channel A and B) and is also responsible for the subcode block. From each frame (2 subframes) this bit is extracted to build a subcode block of 192 bits. Bit 2 (count: 0, 1, 2, ..) in this block is called the copy prohibit bit. So, by changing bit 30 and the last parity bit in frame 2, the copy protection can easily be removed [3]. The fixed position of the copy-prohibit-bit is therefore also the weakness of this protection.

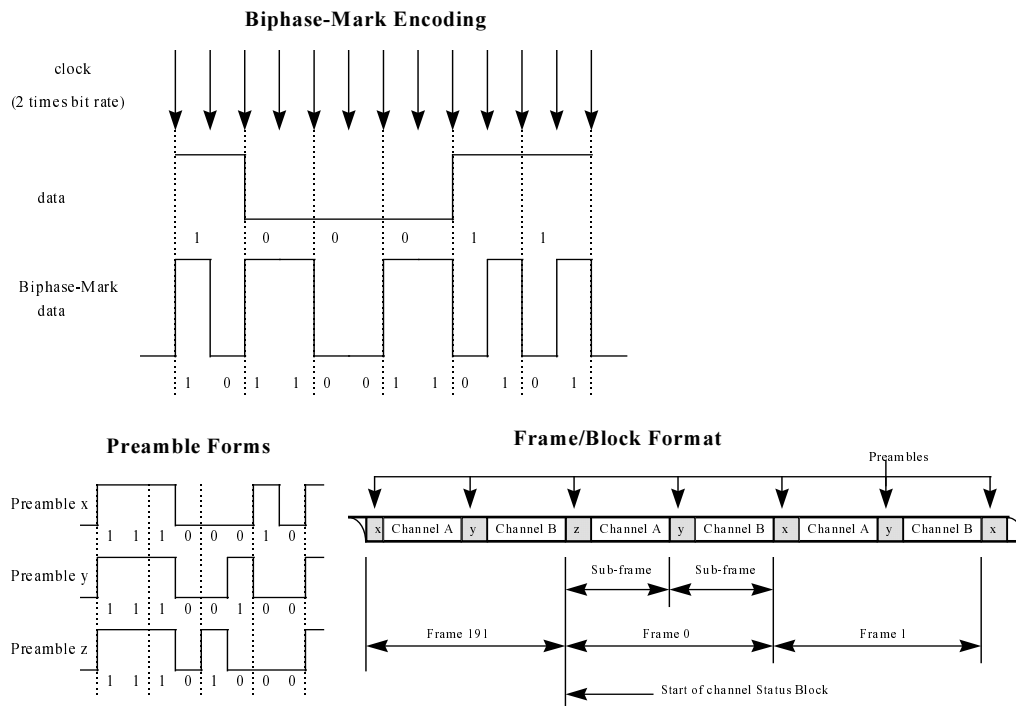


Figure 3.1: Biphasic format, preambles and frame/block format

3.3 Cryptographic Protection of the Digitally Broadcasted Material

Pay TV techniques rely on two independent mechanisms. On the first hand scrambling / encryption of the picture and of the sound, on the second hand management of commercial entitlements which have to be transmitted as secured messages to the descrambler box or Settop Box (control access). Encryption can easily be applied on a digital bitstream. In this case, all bits are encrypted by using for example a blockcipher like DES. Scrambling is used for analogue broadcasting. Using the latter method the signal format is changed, the synchronisation signals are suppressed and separately transmitted in an encrypted form. Sometimes, the audio signal is converted to a digital signal and encrypted. This digital encrypted audio signal can be embedded in the video signal.

The data is scrambled or encrypted using a control word (CW) or key. The control word or key will change after a short period. To send the new keys to the descrambler (STB) ECM's (Entitlement Control Messages) and EMM's (Entitlement Management Messages) are used. Those messages have a digital signature field which ensures the integrity of the message (e.g. a HASH-code). This prevents users to modify the context of the message.

An ECM is transmitted together with the scrambled signal. An ECM consists of three fields. The first field contains the access parameters. These parameters define the conditions under which access to the program is allowed. This field makes for example parental rating (additional PIN code is requested by the descrambler box) and geographical black out (a film may not be available in all European countries) possible. The second field contains the control word in encrypted form and the last field contains a data integrity check.

An EMM consists usually of four fields. Each EMM starts with an address field to select an individual descrambler box. There are two addressing modes, one for an individual descrambler box and one for a group of boxes. The second field contains the entitlement for the user. The third field contains the service keys in encrypted form and the last field contains a data integrity check. EMM's can also be used to send a command to the descrambler box (see VideoCipher and Videocrypt). Transmission of EMM's is generally the result from an explicit request from the user to the service provider. These messages are individual in general. Their content shall be interpreted by one descrambler box or by a limited number of descrambler boxes which are concerned by this particular entitlement.

EMM's do not have to be transmitted in a synchronous way with the program to which they apply. They have to be transmitted in advance in order to give access to the authorised consumer. Any network can be used to transmit them to the receiver: modem, mail or broadcast. Over air addressing means that the messages (EMM's) are broadcasted.

To be sure that an EMM is received by the user to renew a subscription for instance, there is no other way than to repeat the message sufficiently. EMM's are therefore organised in cycle for broadcasting. The length of the cycle is the major parameter determining the maximum time to wait to get an entitlement for a user, which has switched of his descrambler box for a long time.

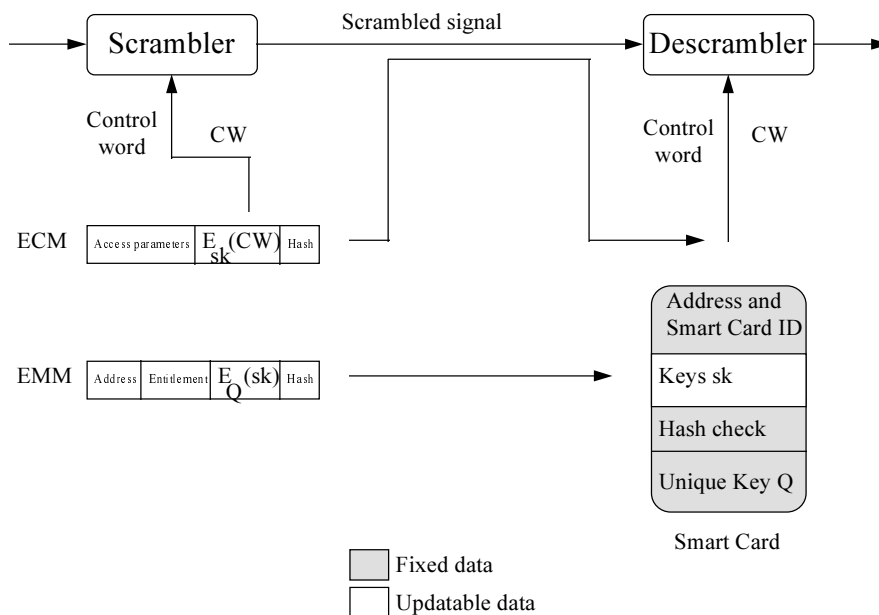


Figure 3.2: Key Management in a Pay TV system

The complete key management system is represented in Figure 3.2. The audio and video are scrambled using a cycling control word or key CW. Every fixed period (e.g. 10 seconds) an ECM is transmitted together with this scrambled signal. These ECM's contain the control words encrypted with the service keys SK, which must be present in the descrambler box. The service keys are less frequently updated by EMM's, for example once a month. The service keys are encrypted with one or more individual unique keys, which are safely stored inside the smart card or descrambler box.

The main problem is that every service provider uses his own scrambling / encryption algorithm and key management system. This means that many different systems are in use today.

In September 1994, the major European television producers, broadcasters, and manufacturers agreed on a new standard for the digital broadcast of video sequences called the Digital Video Broadcast (DVB) [4 - 8]. This new technology will gradually replace the current analog PAL and SECAM broadcast norms.

The first generation of DVB consumer receivers is expected to be a set top box called an Integrated Receiver Decoder (IRD). I.e. a small box which contains only a receiver and the above MPEG decoder. These IRDs will have the usual RF and SCART interfaces to the antenna, cable and TV/VCR. In addition IRDs are expected to have also data transmission interfaces for personal computers and other multimedia systems. One original point of the DVB system will be that the control access module (CA) will be a separated box which will be connected to the IRD using a PCM/CIA interface. A chip card slot will be optionally provided on the module.

There are two proposals for the conditional access module due to the different views of the participants. The established broadcasters, who already offer video services, would like make sure that their investment in their current descramblers is not lost. They will accept standardisation only up to a certain point. On the other hand, the newcomers, consisting mainly of network operators and the smaller broadcasters, would like to cooperate, since they know that not many people would buy a decoder box to watch only one or two channels. So, complete standardisation of the module would really be the solution for them. The equipment manufacturers also wish to standardise to come to cheap mass production.

In Figure 3.3 the proposed scheme for the DAVIC Conditional Access system is represented. This proposed model works with a Set Top Box implemented by a standard terminal and a detachable CA module (PCMCIA card). More information about Pay-TV systems can be found in [9].

Digital storage devices will enter the market soon like D-VHS, DVC, etc. [10] and of course our SMASH device in a later stage. To record digital signals the descrambler box (STB) must be equipped with a digital output. Actually, service providers are reluctant to accept digital interfaces and storage devices, but they may accept solutions in which the data is recorded in encrypted form to enable the basic time-shift function of the analogue VCR (Figure 3.4).

In this case they can still control the data, because the data must still pass the STB for descrambling and is nowhere in the system available in clear MPEG-2 format (only in analogue decompressed form). However, this solution has drawbacks since EMM's can change the service keys in the smart card. If a STB receives a new EMM, all recorded old data is lost. Because the keys needed to descramble this data are replaced.

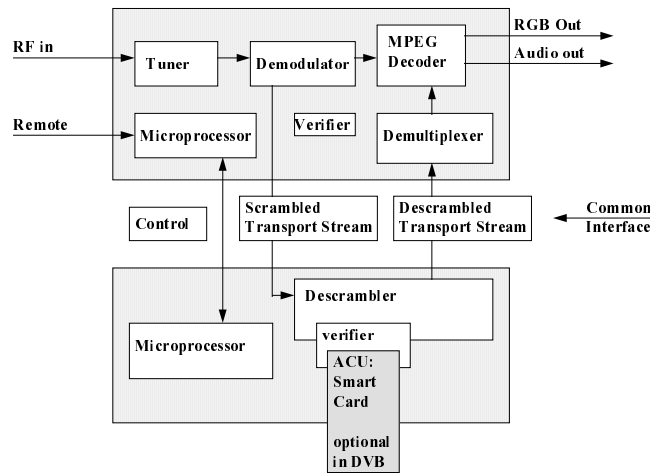


Figure 3.3: DVB proposal for decoder and CA system

3.4 Copy-Protection of Analogue and Digital Video

Macrovision is a videotape copy protection for analogue VHS video cassette recorders [11]. It is used on pre-recorded videotapes and in the newer STB's (see Figure 3.4). It seems to be more common in North America than in Europe. It is also used in the new Set Top Boxes to protect the outgoing signals against copying. When dubbing a protected tape, or copying a protected analogue signal, the picture that has gone through the recording VCR will get dark and then normal again periodically. This effect is caused by some new inserted false synchronisation pulses in the non-visible portion of the picture.

Nowadays, digital VCRs appear on the market. Representatives of the consumer electronics and motion picture industries have agreed to seek legislation concerning digital video recorders that would protect both intellectual property and consumers' rights in the digital age [12]. A recommendation is submitted to the US Congress. This recommendation would:

- assure that the ability of consumers to make home video recordings of anything shown over broadcast or basic cable television remains unimpeded.
- allow consumers to make at least one copy of subscription or other "pay-cable" programming, but digital copies of the copy may be prevented.
- allow copyright proprietors to prohibit copying from pay-per-view, video-on-demand programming and pre-recorded media

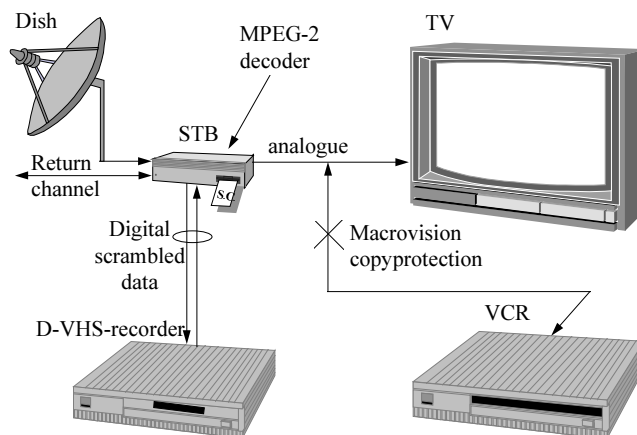


Figure 3.4: *D-VHS recorder connected to the STB*

These provisions would apply only to digital recording devices and recordings made from digital sources with conventional analogue VCRs.

The first digital consumer recorder is the Sony DHR-1000 [13]. This device can be used to edit digital home videos without restrictions. A Serial Copy Management System is implemented to prevent illegal copying of pre-recorded tapes. It is even possible for the broadcaster to indicate whether a movie can be recorded or not. So, the broadcaster can switch off the recorders. The recorder is equipped with an analogue tuner and stores the video in a digital intra frame coded format.

For pre-recorded material on CD-I and laserdisc there is still no copy protection mechanism. This is not necessary for the moment, since the interfaces to other devices are analogue. However, an interface for the CD-I to the PC exists and will be discussed in the next section.

The introduction of the Digital Video Disc (DVD) has been delayed due to a lack of a reliable copy protection system [12,14]. This high capacity disc (up to 18 Gigabyte) must replace the CD-I and laserdisc. The motion picture industry did not want that the DVD can be copied, but finally agreed with the manufacturers to allow the consumer to make one copy for own use only. The manufacturers must now implement a copy protection system in the DVD-discs, DVD-player / recorder and all other digital audio and video devices (maybe even in PCs).

3.5 Interface of Digital Equipment to PC

Currently, there are no copy protection mechanisms implemented in the PC. Software packages are not protected or have their own protection mechanisms. PCs can access data in another way than digital audio and video devices. Some software packages allow the user to extract digitally perfect copies of samples from audio CDs using a CD-ROM player. It also allows to extract MPEG streams from CD-I Digital Video CDs, and VideoCDs, and XA frames from CD-XA CDs. On every CD there is a bit which defines if copying a particular

track is permitted or prohibited. The software packages usually do not care about this bit and copy the data anyway. The data can be written back to for instance a CD-recordable without any problems.

Interfaces to the PC and legal software packages exist for the DCC and minidisc recorder. Music tracks can be copied to the harddisk and edited. So the music can also be copied from the harddisk to other tapes or discs, because the Serial Copy Management System is circumvented.

The same problem holds for the DAT-recorder. This device can also be used as tapestreamer for the PC. With some DAT-recorders it is also possible to copy music tracks to the harddisk and back [15].

3.6 Protection of Private Data

Users want to have the possibility to protect their own data against others (e.g. parental control). To protect the user data against theft a relative strong encryption algorithm should be used (e.g. DES). However, in some countries it is forbidden to use very strong cryptographic algorithms in consumer electronics. The law-enforcement agencies wish to have access to the communications of suspected criminals, which is threatened by secure cryptography. Industry and individual citizens, however, want to secure their private data, and look to cryptography to provide it. In the U.S. the Capstone project aims to develop a technology that attempts to balance these needs [16].

3.7 Conclusions

In the U.S. digital audio recording or interface devices must contain:

- The Serial Copy Management System (SCMS), which permits first-generation digital-to-digital copies of pre-recorded music and other audio works, but prohibits multi-generation or "serial" copies of those copies.

A recommendation for all digital recording devices is submitted to the US Congress. This recommendation would:

- assure that the ability of consumers to make home video recordings of anything shown over broadcast or basic cable television remains unimpeded.
- allow consumers to make at least one copy of subscription or other "pay-cable" programming, but digital copies of the copy may be prevented.
- allow copyright proprietors to prohibit copying from pay-per-view, video-on-demand programming and pre-recorded media

Service Providers are still reluctant to accept digital interfaces and storage devices, but they may accept solutions in which the data is recorded in encrypted form to enable the basic time-shift function. The service provider has still control over the broadcast data in this case. This solution only works with a Set Top Box and an analogue output to a TV-set.

Protection mechanisms exist for digital audio recorders, however a PC can easily circumvent the existing copy protection system. For video a more or less similar protection system is

implemented in the first digital VCR. But the motion picture industry is still very concerned about the PC. This is also the reason for the delay of the introduction of the DVD on the market.

Only a standardised copy protection system for all digital devices including the PC can be safe.

In some countries it is forbidden to use very strong cryptographic algorithms in consumer electronics. The law-enforcement agencies wish to have access to the communications of suspected criminals, which is threatened by secure cryptography. Industry and individual citizens, however, want to secure their private data, and look to cryptography to provide it. In the U.S. the Capstone project aims to develop a technology that attempts to balance these needs.

4 Requirements and Solution Concept

A. Hanjalic (4.1), G.C. Langelaar (4.2), M. Ceccarelli (4.3)

On the basis of the state-of-the-art in visual search, copy protection, and combo architecture and performance, this section will list requirements and conceptual solutions for various aspects of the combo system. Successively we discuss the requirements and solution concept of visual search, copy protection and multimedia database.

4.1 Visual-search

4.1.1 General requirements on visual-search engine

In large public digital libraries different preparatory annotation activities are carried out on the stored data so that its representation is optimal for subsequent user browsing processes. These activities are usually time-consuming and may include manual insertion of key-words, and the selection of audio and video clips.

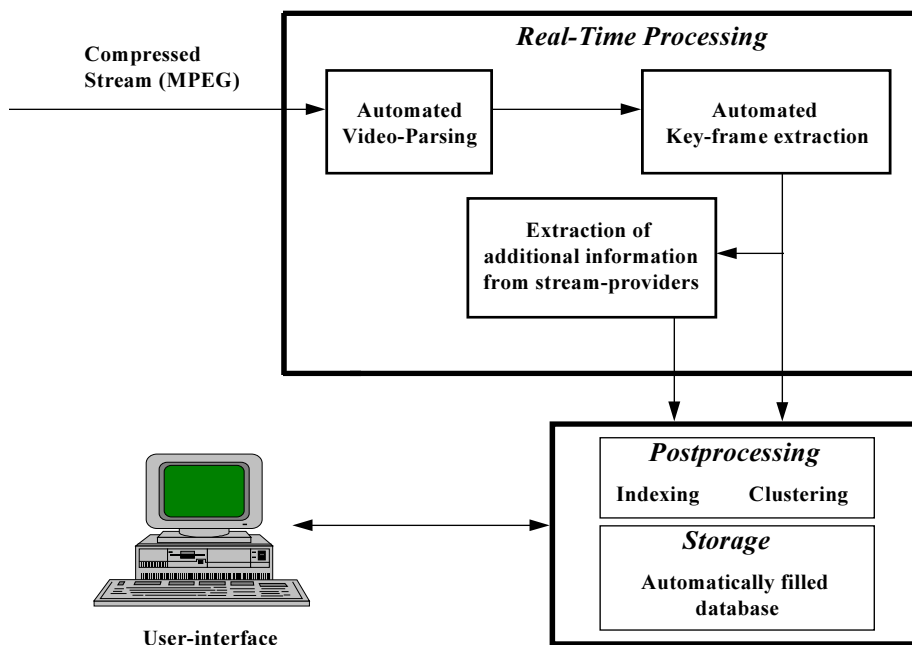


Figure 4.1: Automated system enabling search through stored video-data

For consumer storage systems, however, such annotation and selection activities should be carried out fully automatically, be it at a somewhat lower reliability and flexibility. This *automated* concept is chosen also to be used within the SMASH project.

In order to obtain a system suitable for consumer system applications, two requirements must be taken into account: (1) minimizing of user-involvement by system setting and (2) at the same time guaranteed good performance for any video-sequence incoming into the system. Also, the system must be able to perform the processing directly on the incoming stream, which is moreover to be done in real-time (or super-real-time - depending on the speed of downloading the stream on the storage unit). This “on-the-fly” processing aspect contributes strongly to the user-friendliness of the entire storage system.

In further text, the term *automated* will be used for both “on-the-fly” processing requirements and generalization of the performance through minimizing (and generalizing) the parameter dependency.

Further requirements on the visual search engine are particularly related to specific SMASH applications and architecture. An important application in SMASH is the recording of DVB services containing the MPEG-2 compressed video format. Since no content-based video-processing is possible directly on the compressed stream, its partial decoding up to a certain level is necessary. For the real-time processing, dedicated hardware is required. Important interfaces needed for the system realization are the bus for sending MPEG-2 transport streams to the COMBO, access to multimedia database and the COMBO-PC interface for developing browsing and postprocessing tools.

To fully automate the process of extracting the key information, two issues need to be studied, namely automated video parsing and automated key frame extraction. This can be observed in Figure 4.1. The following two sections contain proposed solutions for these two parts of the visual search engine.

4.1.2 Video-parsing

For consumer application systems, a video-parsing algorithm is needed, capable of “on-the-fly” video-analysis and with high detection reliability. Due to these requirements, the concept of global thresholding - as often proposed in the literature - cannot be applied. For SMASH purposes, a modified approach from [22] with *adaptive threshold* will be used where the investigation of frame-to-frame differences is done within a *sliding window* containing only a certain number of recently computed FFD values. The results are not dependent on missed detection or false alarms which happened before the starting point of the window. By this approach not the last computed but the middle FFD-value of the window is compared with the second largest window element and a sharp shot break is proclaimed if the ratio between these two values is larger than a given thresholding parameter. FFD values along the sequence are measured by comparing histograms of all three components of the YUV colour space. Histograms are formed with 32 bins for Y values and with 64 bins for both chrominance components. The used metric can be given as

$$d_{YUV}(k, k-1) = \sum_i \sum_{j=Y,U,V} |h_k^j(i) - h_{k-1}^j(i)|$$

employing histograms $h(i)$ of two consecutive DC-frames $(k, k-1)$, applied to Y, U and V components.

The original algorithm, as described in [22], detects sharp changes between consecutive shots, but reacts also by gradual transitions and special effects. The main disadvantage of this method is the need for specification of the thresholding parameter for distinguishing the maximal and next largest window element, as described before. The performance of the value proposed by authors in [22] can be guaranteed only for sequences used in tests. Our current efforts are directed towards minimizing this parameter dependency and obtain good performance quality for general sequences. First improved results have been obtained for detecting sharp shot-changes, and this is planned to be used in the first implementation phase of the project. Improved results for gradual changes and special effects are expected soon and will be implemented into the system in a later project phase.

4.1.3 Key-frame extraction

For the usage in the project, we propose our novel automatic key-frame extraction procedure [23], [24].

Figure 4.1 shows the SMASH reference framework as far as the key frame extraction and storage concept is concerned. Among other components of the automated storage unit, two essential components of our automated approach can be seen, namely *key frame assignment to each shot* and *key frame distribution along each shot*. In the first component, “on-the-fly” assignment of the number of key-frame per shot is carried out depending on the content of the shot and on the past content development. This key frame assignment is done such that the sum of all assigned key-frames along the sequence is close to a given maximal number of allowable key-frames N for the entire sequence. As a measure C_i of content, we have found *the sum of frame-to-frame differences along the shot i* sufficiently representative:

$$C_i = \sum_{n=2}^L d(n, n-1) \quad (1)$$

Here L is the number of frames within the shot and $d(n, n-1)$ the measured content-difference between frames n and $n-1$, here based on comparison of frame-histograms h using Y , U and V colour component:

$$d(n, n-1) = d_{YUV}(n, n-1) = \sum_i \sum_{j=Y, U, V} |h_n^j(i) - h_{n-1}^j(i)| \quad (2)$$

Such histogram-based measuring of content-changes along the sequence, although shown to be a powerful tool for video-parsing (Section 4.1.1), is not the final choice when performing the key-frame extraction. Currently, the work is being done within the WP 320 for defining much more sophisticated measures for content-development of a video-sequence (also called *action-measures*).

The “on-the-fly” assignment algorithm, using the defined content-measure C_i , takes on the following form:

$$K_i = \frac{C_i}{\sum_{u=1}^S C_u} N = \frac{C_i}{\frac{\sum_{u=1}^S C_u}{T}} \frac{N}{T} = C_i \frac{N}{T} \frac{T}{\sum_{u=1}^S C_u} \approx C_i \frac{N}{T} \frac{\sum_{u=1}^i T_u}{\sum_{u=1}^i C_u} \quad (3)$$

K_i represents the assigned number of key-frames to the shot i , T is the total sequence length, and T_u is the length of the shot u . C_u is the content of the shot u defined with (1) and S is the number of shots in the entire sequence.

The assignment step in (3) is followed by a threshold independent and objective procedure for content-based distribution of the assigned number of key-frames along each video-shot [23], [24]. Key-frame distribution along the shot results from minimising the following criterion function:

$$g(k_1, \dots, k_{K_i}, t_1, \dots, t_{K_i-1}) = \sum_{j=1}^{K_i} \int_{t_{j-1}}^{t_j} |C_i(x) - C_i(k_j)| dx \quad (4)$$

Here k_j ($j=1, \dots, K_i$) are the temporal positions of the key frames, while t_{j-1} and t_j are the *breakpoints* between the shot segments that are represented by key frame k_j . Note that t_0 and t_{K_i} are the (known) temporal begin and endpoints of i -th shot. The non-decreasing function $C_i(x)$ represents the content-development of the shot i , where each function value is obtained by accumulating frame-to-frame differences (2) along the shot up to the frame x . With (4) we indicate that we wish to approximate the actual content-development by using the curve $C_i(k_j)$ composed out of rectangles, each one defined by k_j and t_j and each corresponding to one key-frame. The minimisation process gives optimal positions of key-frames along the shot, such that their (variable) density best simulates the actual content-development. Figure 4.2a illustrates the distribution approach by 3 assigned key-frames to a shot with a given content-development. Figure 4.2b shows the result of the key-frame distribution over a single shot after applying the described approach.

Results of the combined key-frame assignment and key-frame distribution show that automated key-frame extraction is indeed feasible. Nonetheless, like in any other automated process, errors occur. The dominant type of errors that we have to deal with are failures of the video parsing process, i.e. detection of false shot changes and the missing of shot changes. Experimental results indicate that due to the use of objective criteria in the key-frame assignment and distribution, the system is fairly robust to errors in detecting shot changes.

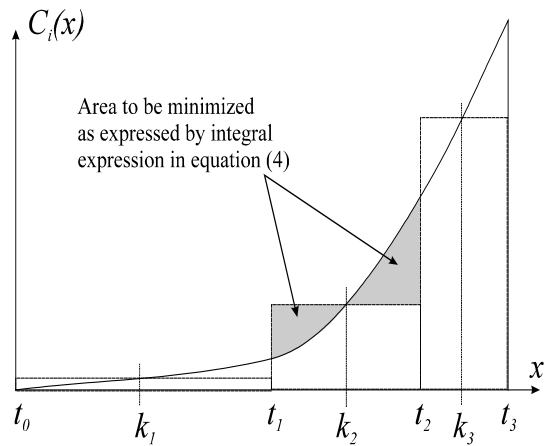


Figure 4.2a: Illustration of the proposed approach for key-frame allocation within the video-shot by assigned 3 key-frames. An approximation of the accumulation curve can be done through 3 flexible rectangles.

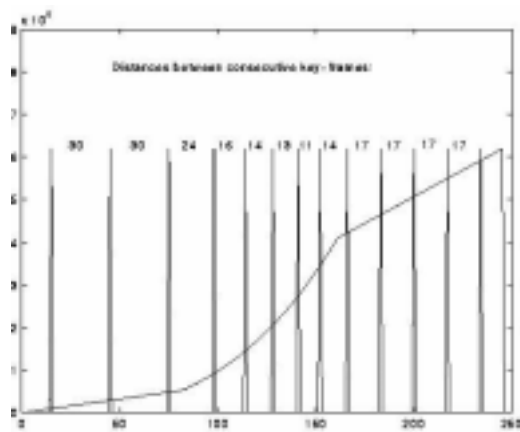


Figure 4.2b: Application of the approach to a fictive video-shot. Obtained variable key-frame densities picture the actual content-development

4.1.4 Clustering

A clustering procedure, suitable for the SMASH system, has not been proposed yet. Intensive research in this area is planned for the next phase of the project.

4.2 Security

The protection mechanism for the SMASH system must take care that service providers can keep control over their stored data and that copyrights are not violated. On the other hand must the system give consumers the possibility to protect their own data against others (e.g. parental control).

4.2.1 Requirements for the SMASH Protection Scheme

The protection scheme for the SMASH recorder should meet the following requirements:

- The system should support scrambled recording for broadcast material.
- Editing of home videos should be possible without limits.
- The system should be compatible with the S/PDIF protocol for digital audio (SCMS).
- It should be possible to make a copy of all digital audio / video services, only if the service provider allows the consumer to do so, however, it should not be possible to make copies of this master copy.
- Service providers should be able to prepare data in such a way that all recorders refuse to store the data.
- The PC is not equipped with any copy protection mechanism yet, so copying to a PC is always possible, but copying of already copied data from a PC to a storage device should be not possible.
- The system should be compatible with the new copy protection systems of other digital VCRs and recorders. These other recorders should be equipped with the same system.
- No strong encryption techniques should be used, to avoid problems with different laws in different countries.

4.2.2 The Copy Protection Scheme for the SMASH System

In Figure 4.3 the interface of the SMASH recorder to other devices is shown. For the interconnection a digital P1394 bus is used. Currently only the DVC camcorder and the Sony DHR-1000 VCR support the new communications protocol IEEE 1394. But in the future the Set Top Box, D-VHS, SMASH system and PC will also be equipped with this interface. Probably the rest of the devices will follow later. The PC has a special own interface to other digital storage devices. Most devices are connected to the internal PCI-bus or SCSI-bus. All devices access the data as real time bitstreams. The PC is the only device that also can access the data as files.

Nowadays, only interfaces exist between digital audio devices, camcorders and digital VCR's and between PC's and peripherals.

If the SCMS protocol [2] would be used for copy protection, all data must be divided into data packets for transmission over the databus. Together with each packet a separate subcode packet must be transmitted, containing the copyright status of this packet (among others the copy prohibit bit).

Every recording device connected to the bus checks the copyright fields in order to know if a stream may be recorded or not and refuses to record a stream or file in which a copy prohibit bit occurs. It is necessary that every packet has its own subcode packet, because the storage device can start recording on a random position in the stream.

This system has two disadvantages. If the copyright status is located in a subcode packet with a fixed position in the stream, a hacker can easily trace the subcode packets and toggle the copy prohibit bit to change the status of the stream. Another problem is the PC. If PC software can use a video stream, it accesses the data as file. But if the stream is accessed a file, it is also possible to copy it to the harddisk, without using the SCMS protocol (and without checking the subcode packets, containing the copyright status). So, if a PC stores a stream the stream is always accepted by the recorder, because no copyright information was found.

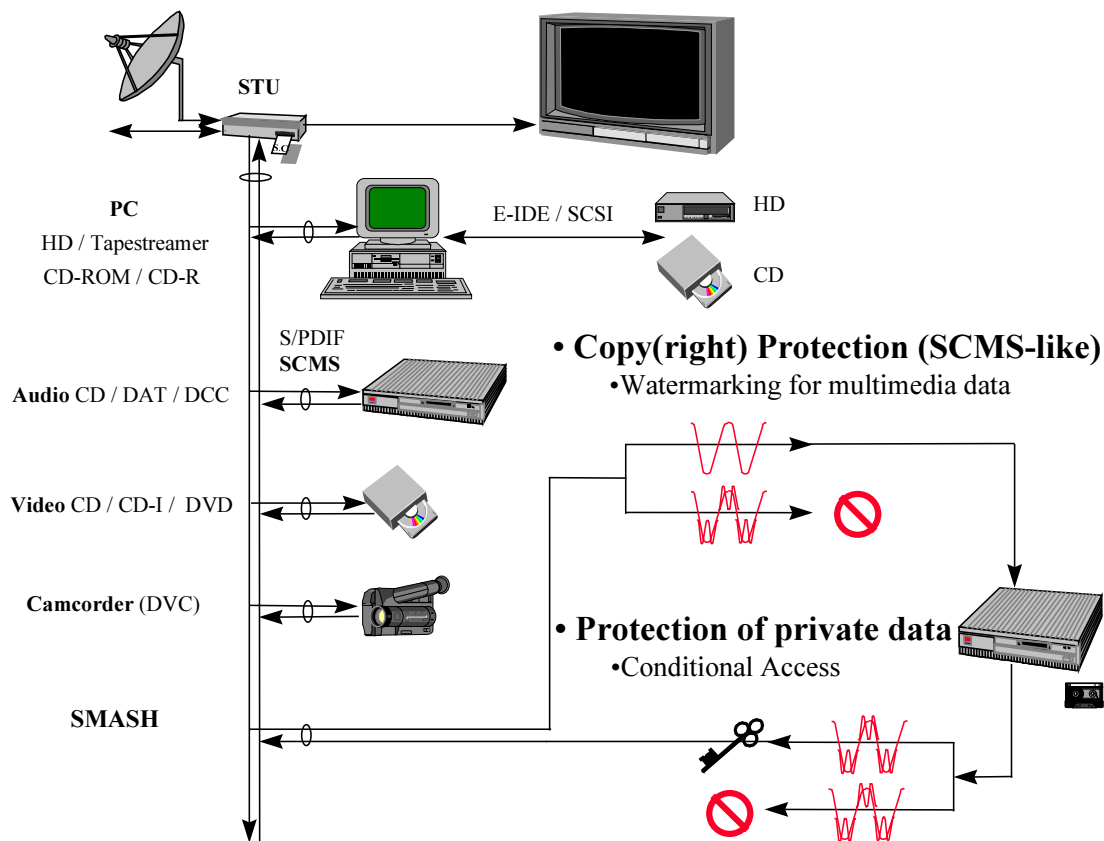


Figure 4.3 Interconnection and protection concept

Therefore the copyright status must be stored in the streams itself, a label or watermark must be added to the data. In this case the copy prohibit bit is inseparable of the data. This approach has many advantages:

- there is no new bus protocol needed, copyright status is stored in the data, not in the protocol, so every protocol can be used
- the copy prohibit bit can survive file format changes
- the copy prohibit bit stays in the data if it is copied to a device which has no protection mechanisms in it, and will be active again if the data is passed to recorders with the protection mechanism implemented.

- if a hacker tries to remove the copyright status, he has to tamper with the data itself, this means that in most cases the quality of the data is affected if the label is removed. If this new “lower-quality-data” is stored again, new copyright information is added. Removing this information, means again quality degradation.

The reliability of the copy protection system is of course dependent on a standardisation of the system. Every storage device must be able to extract the copyright information of the data and must refuse to record data in which a copy prohibit bit occurs.

The complete implementation of the copy protection mechanism is represented in Figure 4.4, which must be implemented in every storage device. The copy protection system is only active if data is stored.

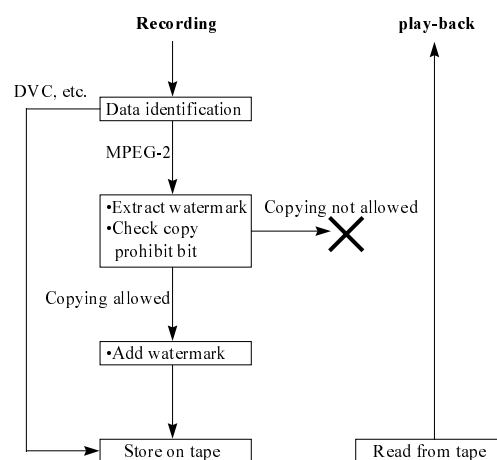


Figure 4.4: Copy protection scheme

- In the first block data-identification is performed. This block defines for what kind of data the protection holds. In this scheme this is only MPEG-2, but it can easily be extended to MPEG-1, JPEG, and other audio formats. Digital camcorders use a different format DVC. These streams do not have to be protected. If motion pictures appear on the market in this format, it is possible to distinguish between pre-recorded video and own recorded video [13]. For pre-recorded video in DVC format, the same path as for MPEG-2 can be used. For instance, if the data is scrambled according to the DVB standard, the format is not recognised as MPEG-2 or other protected formats and will be stored without any modifications.
- In the second block the watermark is extracted and the copyright bits in this watermark are checked. If the copy prohibit bit is set the recorder stops recording.
- If no copy prohibit bit is found, the copyright status, including a copy prohibit bit, is added to the audio and / or video data using watermark techniques.

To be compatible with the S/PDIF protocol the following scenario can be used. If data is offered according to the S/PDIF protocol, the copy prohibit bit is tested and the data is refused if this bit is set. If the data contained no copy prohibit bit, the data is labelled and stored. If

labelled data is played-back according to the S/PDIF protocol the copy prohibit bit is set. A DAT or DCC recorder will not copy the data further in this case.

If service providers wants to prepare data in such a way that all recorders refuse to store the data, they simply add a label to their data, in which the copy prohibit bit is set.

4.2.3 Protection of Private Data

If an encryption module is implemented in the system to protect the private data of the consumer and the recorder allows to play back encrypted data, the copy protection mechanism can be circumvented, because the watermark can not be detected in an encrypted stream. To avoid this, the recorder must not allow to play back such streams and must also not give PCs access to these files.

Since strong encryption algorithms are forbidden for consumer applications in some countries, it is better to protect the consumer data by a conditional access system. In this case there are also no problems with the labelling. Users can protect their data for instance with a password or PIN-code.

4.2.4 Evaluation of the Protection System

A simple representation of the complete protection system can be found in Figure 4.3. The system has the following properties:

- All recorded data, except scrambled data, can always be played back on every SMASH system.
- It is possible to make a copy of all digital audio / video services, only if the service provider allows to do so, however it is not possible to make copies of this master copy.
- Service providers can prepare data in such a way that the recorder refuses to store the data.
- The system supports scrambled recording for Pay-TV services and is compatible with the S/PDIF protocol.
- The PC is not equipped with any copyprotection mechanism yet, so copying to a PC is always possible, but copying of data from a digital storage device to a PC and back from a PC to a storage device is not possible.
This means that a PC can use all recorded data, but only as “read-only”-data. It is possible to delete data, but it is impossible to copy labelled data to the harddisk and than back to the SMASH device.
- Editing of home videos is possible without limits.
- Consumers can protect their private with a conditional access system.
- The strength of the system relies on the strength of the watermarking algorithm.

4.3 SMASH MMDB: when retrieval is more important than storage

The strongest point of the *SMASH combo* is given by the combination of its three main characteristics, that make it a candidate for potential market success: large capacity, fast

access and low cost. Once large capacity has been provided via a low cost tape system, the main challenge is to make the system react quickly to the user's commands, in order to transparently appear to him as a fast access device (as an 'expensive' array of disks, for example). This requires a careful and clever organisation of the stored information, in order to efficiently manage the resources the application requests.

Fast access and large amount of data are also the keywords of *Database Management Systems* (DBMS). A DBMS is a tool for manipulating a database, which is made available through special software for archiving, retrieving and modifying objects.

When data grow beyond a certain degree in quantity and complexity, a system for centralised management of data and applications is needed. This is certainly necessary in case of a storage system for multimedia applications as the SMASH Combo is, where plenty of data in different formats may enter the system. We can also say that, ultimately, *storage* would have no sense if no *retrieval* system was provided.

Database Management Systems provide not only data integrity, consistency, optimisation of both storage and processing resources, provide also the user with interesting features like access speed, flexibility of retrieval, total availability of the stored information. These capabilities have been reserved so far to companies with centralised computing systems, where large amount of information about items or people could be organised in alphanumeric tables. As the computer world has moved from large mainframes with centralised resources to networked, distributed and personal solutions, a new fast growing market is attracting the world of database to shift from large scale enterprise solutions to lightweight systems for small scale or distributed applications.

In the development of a Multimedia Database Management System for the SMASH Combo, the challenge is represented by the necessity of developing a fast, small footprint, embedded DBMS, capable of managing large and complex objects for applications with critical requirements.

Furthermore, the system must not need an administrator and must automatically take care of all the issues involved with management of data. It must store not only the objects, but also their attributes and their relationships. Extraction and annotations of metadata will be performed in real-time during storing operation, and queries will be embedded in the applications.

The DBMS will act as an interface between the physical repository and the client application. An interfacing mechanism for client-server communication also has to be defined.

5 Conclusions

R. L. Lagendijk

Besides application and technological choices and developments in the SMASH project, the functionalities of visual search and copy protection play an important role. To realize these functionalities, the project has built upon state-of-the-art techniques, and has defined three key topics that have been worked out in more detail, namely:

- Video parsing and key frame extraction,
- Storage of key frames in a multimedia data base,
- Watermarking.

The multimedia database will play a central role, as it will be filled upon recording time with many different types of information about the recorded data, while it will be consulted by applications (running on a settop box or PC) at the time the user wishes to access his data. In the project's continuation, work will concentrate on further development of watermarking techniques and clustering procedures of key frames, and the implementation of the MMDB and key frame extraction for use in real-time. Several of the results from this work will be brought into application (user and field) trial, either in full functionality or in scaled-down functionality, depending on the precise objectives of these trials.

6 References

- [1] The Audio Home Recording Act of 1992, WWW-pages of the Home Recording Rights Coalition, <http://www.access.digex.net/~hrcc/ahrasum.html>
- [2] Digital audio interface, International Standard IEC 958
- [3] Copybit-inverter, digitaal kopiëren zonder belemmeringen, W.Foede, *Elektuur* 1/96
- [4] Workpackage 3: ACCOPI Evaluation of Existing Systems, ACCOPI RACE project M1005, 19 April 1995
- [5] Implementation guidelines for the use of MPEG2 and content input to servers, DAVIC second call for proposals, CCETT
- [6] Common Interface Specification for Conditional Access and other Digital Video Broadcasting Decoder Applications, DVB, 16 February 1995
- [7] Access Control : Common Scrambling system and Common Interface for Conditional Access, Final Technical Report of the Conditional Access Specialist Group, DAVIC second call for proposals, CCETT, 17 November 1994
- [8] Standardisation in the DVB of conditional access systems for pay TV, D van Schooneveld, Philips Research Laboratories, Eindhoven, The Netherlands, *Tijdschrift van het Nederlands Elektronica- en Radiogenootschap* deel 60 - nr.3, 1995
- [9] Overview of protection methods in existing TV and storage devices, SMASH technical report SMS-TUD-609-1, Feb. 1996
- [10] Brücke zwischen analoger und digitaler Welt, D-VHS, Rainer Bucken, *Fernseh- und kino-technik* 49. Jahrgang Nr. 5/1995
- [11] Macrovision FAQ v1.0c, Antti Paarlahti, 1995, <http://www.paranoia.com/~filipg>
- [12] PC industry could delay DVD, Jean-Luc Renaud, *Advanced Television Markets*, Issue 47, May 1996
- [13] Pre-Test Digitale Videorecorder Sony DHR-1000, *Audio Video Totaal*, June 1996
- [14] Digital Video Disc, Industrie wil lokale wereldstandaards, Ruud van der Schaft, *Hifi video test* no. 6, June 1996
- [15] DAT-heads, Frequently Asked Questions, release 3.1, 1992, [http://www .ultranet.com/~jgm/dat-faq.txt](http://www.ultranet.com/~jgm/dat-faq.txt)
- [16] A proposed Federal Information Processing Standard for an Escrowed Encryption Standard (EES), National Institute of Standards and Technology (NIST), *Federal Register*, 58(145), July 1993
- [17] Smith, J.R., Chang, S.-F.: "Single Color Extraction and Image Query", *Proceedings of ICIP '95*, Washington DC, 1995.
- [18] R.W. Picard, "Light-years from Lena: Video and Image Libraries of the Future", *Proc. of ICIP 1995*, vol 1, pp. 310-313, Washington DC, USA, 1995
- [19] Wang, H., Chang, S.-F.: "Adaptive Image Matching in the Subband Domain", *Proceedings of VCIP '96*, San Jose, CA, 1996.
- [20] M.M. Yeung and B. Liu, "Efficient Matching and Clustering of Video Shots", *Proc. of ICIP 1995*, vol. 1, pp. 338-241, Washington DC, USA, 1995.
- [21] H. Zhang, C.Y. Low and S.W. Smoliar, "Video Parsing and Browsing using Compressed Data", *Multimedia Tools and Applications*, vol. 1, pp. 89-111, Kluwer Academic Publishers, 1995.
- [22] B. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Video", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.5, No.6, December 1995.
- [23] A. Hanjalic, R.L. Lagendijk, and J. Biemond, "A New Key-Frame Allocation Method for Representing Stored Video-Streams", *First International Workshop on Image Databases and Multi Media Search*, Amsterdam, The Netherlands, 1996
- [24] R.L. Lagendijk, A. Hanjalic, M. Ceccarelli, M. Soletic, and E. Persoon, "Visual Search in a SMASH System", *ICIP 1996*, Lausanne, Switzerland, 1996.
- [25] Furth, B., Smoliar, S.W., Zhang, H.: "Video and Image Processing in Multimedia Systems", Kluwer Academic Publishers, 1995
- [26] Yeung, M.M., Yeo, B., Wolf, W., Liu, B.: "Video Browsing using Clustering and Scene Transitions on Compressed Sequences"
- [27] W.Bender, D.Gruhl, N. Morimoto : "Techniques for Data Hiding", *Proceedings of the SPIE*, 2420:40, San Jose CA, USA, February 1995

- [28] Towards Robust and Hidden Image Copyright Labeling, E. Koch, J. Zhao, Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June, 1995
- [29] Copy Protection for Multimedia Data based on Labeling Techniques, G.C.Langelaar, J.C.A. van der Lubbe, J.Biamond, Proceedings 17th Symposium on Information Theory in the Benelux, Enschede, the Netherlands, May 1996
- [30] Sethi, I.K., Patel, N.: "A Statistical Approach to Scene Change Detection", Proceedings of SPIE Storage & Retrieval for Image and Video Databases, 1995
- [31] G. Ahanger and T.D.C. Little, "A Survey of Technologies for Parsing and Indexing Digital Video", *J. Visual Comm. and Image Representation*, vol. 7, no. 1, pp. 28-43, 1996.
- [32] J. Meng, Y. Juan, S.-F. Cheng, "Scene Change Detection in a MPEG compressed Video Sequence", *Proc. of the IS&T/SPIE Symp.*, vol. 2419, San Jose, CA, USA, 1995.
- [33] A. Nagasaka and Y. Tanaka, "Automatic Video Indexing and Full-Video Search for Object Appearances", *Visual Database Systems II*, Elsevier, pp. 113-127, 1992.
- [34] J.S. Boreczky and L.A. Rowe, "Comparison of Video Shot Boundary Detection Techniques", *Proc. of SPIE Conf. on Storage and Retrieval for Still Images and Video Databases IV*, 1996, pp. 170-179.
- [35] Zhong, D., Zhang, H., Chang, S.-F.: "Clustering Methods for Video Browsing and Annotation", *Proc. of the IS&T/SPIE Symp.*, vol. 2670, San Jose, CA, USA, 1995.
- [36] Arman, F., Hsu, A., Chiu, M.-Y.: "Image Processing on Compressed Data for Large Video Databases", *Proc. ACM Multimedia '93*, Anaheim, CA, 1993