

# A New Method for Key Frame based Video Content Representation\*

*Alan Hanjalic, Reginald L. Lagendijk, Jan Biemond*

Department of Electrical Engineering, Information Theory Group  
Delft University of Technology  
P.O. Box 5031, 2600 GA Delft, The Netherlands  
Phone: +31-15-2783084, Fax: +31-15-2781843  
*{alan,inald,biemond}@it.et.tudelft.nl*

## Abstract

Compact video representation is an important step when developing tools for search through large video data bases. It has been shown in many publications that the usage of representative video frames (key frames) for this purpose is indeed an appropriate way of preserving the entire temporal information flow of the sequence in a considerably smaller amount of data - if the key frame set is obtained properly. In this paper we describe a novel method for key frame based video representation. The main advantage of this approach is that the resulting set of key frames, as opposed to recent methods from literature, is not dependent on subjective thresholds or any other manually given parameters. It gives a key frame set based on "objective" model for the video information flow. Another advantage is that this key frame set contains not more than the maximal number of frames, which is set beforehand for the entire sequence.

## 1 Introduction

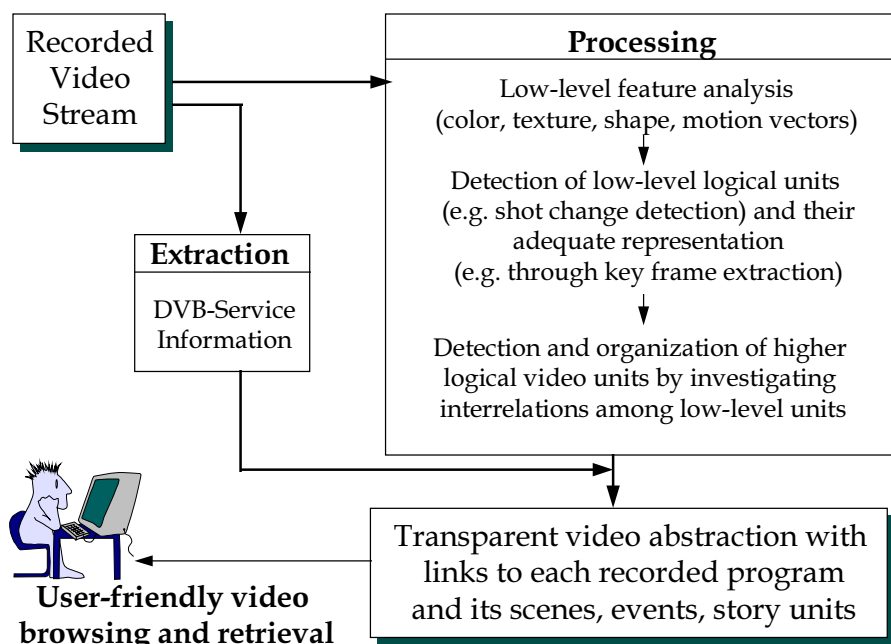
We are witnessing an immense growth in the development of digital libraries. Collected in such libraries are large volumes of digitized multimedia data, which are reachable via electronic networks by any user world-wide. Libraries containing a variety of well-organized digital data bring many advantages, such as preservation of quality of stored information and almost unlimited possibilities to manipulate and browse through data using comfortable user interfaces.

The development in the field of digital libraries depends on achievements in several areas. The US digital library initiative addresses several of these challenges [6]. Achievements in digital storage media technology cause a continuously increasing ratio between *storage capacity* and corresponding *costs*. With this increase of capacity the problem arises of locating desired parts of stored data in a quick and reliable way. The development of search and access methods is therefore as crucial as the storage technology itself for efficient usage of digital libraries. This has been recognized also by the International Organization for Standardisation, which started the new project "Multimedia Content Description Interface" (in short MPEG-7). This project should standardize the description of various types of multimedia information. The description should be related to the actual content and allow fast and efficient retrieval of any part of the stored information [12].

---

\* This work was supported in part by the EU ACTS program under the contract AC018: *SMASH* (Storage for Multimedia Applications Systems in the Home)

In the European project *SMASH* [<http://www-it.et.tudelft.nl/pda/smash>], new storage systems for multimedia are under investigation, as well as several search strategies. The goal of the project is to develop a large capacity home storage device, capable to store tens of Gbytes of various multimedia data on a combined tape-disk mass storage unit and being equipped with a tool for efficient and reliable tracing and retrieval of any desired part of the stored information. The main application focus of the SMASH system is on storage of digital *video* services: DVB (Digital Video Broadcasting) is the European standard for transmission of digital services [8], now being adopted also in other countries and organisations world-wide. Due to this focus, the development of search strategies within the project is oriented primarily to video sequences.



*Figure 1: One possible temporal scheme of all activities related to a user-friendly organisation of the recorded video, followed by the browsing and retrieval procedure on a transparent video abstraction. Reduced, key frame based video representation eases the third processing step in the scheme to a great extent*

In order to allow the user to efficiently browse for, select and retrieve a desired video part without having to deal directly with GBytes of stored (compressed) data, several activities have to be carried out as a preparation for such user interaction. Figure 1 illustrates this on the example of DVB-like information. The main goal of these procedures is to provide the user a compact and easy understandable overview of the complete stored video information. The user can browse through such

abstraction, easily build up an impression of the entire stored video and make a selection leading to a retrieval of the corresponding video segment.

In this paper we concentrate on one of the most important steps in making a video browsing tool: *compact representation of the video temporal information flow*. This representation is done by first segmenting the entire sequence in elementary content units called *video shots* (unbroken series of frames, e.g. a zoom of a person talking [11]), and then reducing each shot to a number of characteristic frames. The proposed novel method for extracting key frames has been developed to suite specific requirements of a fully automated video analysis system, being a part of the SMASH browsing tool. These requirements are

- high degree of process automation (parameter independence)
- sequence independence
- representation objectivity
- controllability of the total key frame number

In this paper we assume that a reliable video parsing algorithm (e.g. [4]) is used to detect changes between consecutive video shots.

## **2 Key frame based video content representation**

Video representation through characteristic frames (*key frames*) has been addressed very frequently in literature (e.g. [1, 3, 4, 5, 9, 10, 13, 14]) as an elegant and efficient way of preserving the whole temporal information flow of the sequence in a considerably smaller amount of data. The underlying assumption is that if these frames are extracted using an appropriate video sampling method, the visual content of each segment of the sequence can be easily recognized by looking at given samples. Such compact video representation appears thus to be suitable for the purpose of video browsing. Also when considering query processes where the search for video parts containing some specific objects, persons or features is performed, the concept based on key frames can be useful since features collected from key frames can be used. If detection of higher logical (“semantic”) units of a video is intended, which is to be done by investigating interrelations among different video shots, key frames can be useful as shot-representatives for comparison purposes.

### **2.1 Existing approaches**

A simple method to select key frames is to take the first frame of each shot [9]. More reliable content representation requires non-uniform sampling of the video shot. In [10], Pentland et al. have found the frames at the beginning and the end of a shot, in the middle of no-motion segments or in the middle of segments where the camera is tracking a foreground object, to be good key frames to represent the content of a shot. Some other approaches [1, 3, 14], based on measuring the differences between the last selected frame and the remaining frames and extracting a subsequent key frame if the measured difference exceeds the given threshold, are typically sequential processes leading generally to unpredictable results. In particular, the final number

of key frames for the entire 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 or other intended procedures. This makes it difficult to predict the capacity needed for storing extracted key frames (in spite of possible “key frame pruning”, as proposed in [14] for further reduction of already obtained key frame sets). Secondly, it is rather difficult, especially in [1] and [3] to relate any particular parameter value by threshold setting to the key frame collection resulting from that setting. Furthermore, the dependency of the approach on subjective and usually data dependent thresholds, limits its applicability in fully automated systems and leads to non-reproducible results.

### **3 New key frame extraction approach**

Aiming at a key frame based video representation, which fulfils the requirements indicated in the introduction to this paper, we have developed a new two-step key frame extraction method.

In the first step, the assignment of a number of key frames per shot is carried out depending on total “content” of a shot and also of the entire sequence. The term “content” is explained in the following section. 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. The number  $N$  can be adjusted depending on the type of the program to be stored.

The assignment step is followed by a threshold independent and objective procedure for optimal distribution of the assigned number of key frames along each video shot.

In the following sections we present results of our investigations concerning both of these steps. Important fact to be noticed is that the distribution of key frames along the shot in the second step is performed by a *numerical algorithm* delivering the best possible shot representation, given the number  $N$  and with respect to the used measure for the information flow dynamics along a shot [7].

#### **3.1 Measures for the information flow dynamics along a video shot**

There the possibility to simulate variations in the temporal information flow of the sequence by choosing an appropriate analytical function. This function measures relevant changes between each two consecutive frames of a sequence and indicates with its values the magnitude of such changes.

This leads to the problem of finding appropriate visual features and metrics to perform the described frame-to-frame comparison. For minimizing the influence of non-relevant temporal variations, “global” frame visual features should be used, such as color and intensity histograms. In our approach we adapted the method proposed in [4] and defined an analytical function for describing the relevant *frame-to-frame difference* (further referred to as *FFD*) between frames  $k$  and  $k-1$  as:

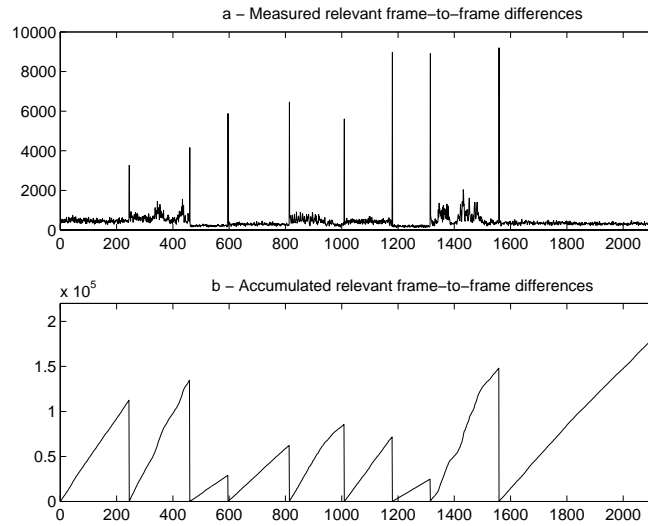
$$FFD(k) = d_{YUV}(k, k-1) = \sum_i \sum_{j=Y,U,V} |h_k^j(i) - h_{k-1}^j(i)| \quad (1)$$

This formula proved to be relatively robust with respect to the elimination of non-relevant temporal fluctuations, such as small object or camera movements, focal length changes, etc.

If the  $FFD(k)$  values in the shot  $i$  are accumulated from the beginning of the shot up to the shot frame  $k$ , i.e.

$$C_i(k) = \sum_{n=2}^k FFD(n) \quad (2)$$

than  $C_i(k)$  indicates the *total magnitude of temporal flow fluctuations up to the final summation point*.



**Figure 2a:** Measured relevant frame-to-frame differences for first nine shots of the sequence "Nature".

**Figure 2b:** Curves of accumulated frame-to-frame differences, obtained for each shot using (2).

$C_i(k)$  has a close-to-linear behaviour in shot parts with uniform temporal information flow and changes in steepness wherever changes in the flow dynamics occur. The function  $C_i(k)$  we take as the model for the *information flow dynamics* in the shot  $i$ . We will refer to it in the further text as the "content development" function of the shot  $i$ . Figure 2b illustrates the behaviour of  $C_i(k)$  for the  $FFD(k)$  curve from Figure 2b.

If the summation process stretches along the entire shot, we obtain the *total magnitude of temporal flow fluctuations in the shot*, that we can also refer to as the total “content” of the shot:

$$C_i = \sum_{k=2}^L FFD(k) \quad (3)$$

In this formula,  $k$  is the frame index and  $L$  is the number of frames in the shot.

### 3.2 Key frame allocation for each video shot

By spreading given maximal number of key frames  $N$  along the entire video sequence, each shot of the sequence gets assigned a fraction of given  $N$  key frames according to its relative share of “content” to the total “content” of the sequence. We therefore assign  $K_i$  key frames to shot  $i$  as:

$$K_i = \frac{C_i}{\sum_{j=1}^S C_j} N \quad (4)$$

$C_j$  is the “content” of the shot  $j$  and  $S$  is the number of shots in the entire sequence.

The resulting number of key frames delivers, after being normalised by the shot length, the *key frame density*. This density corresponds now to the relative amount of temporal variations of a shot, compared to all other shots in the sequence.

Shot Index	1	2	3	4	5	6	7	8	9
<b>Key frames and densities using (4)</b>	<b>14</b> 1:18	<b>16</b> 1:13	<b>4</b> 1:34	<b>8</b> 1:27	<b>11</b> 1:17	<b>9</b> 1:19	<b>3</b> 1:45	<b>18</b> 1:13	<b>22</b> 1:24
<b>Key frames and densities using (5)</b>	<b>12</b> 1:20	<b>12</b> 1:18	<b>3</b> 1:45	<b>8</b> 1:27	<b>10</b> 1:19	<b>9</b> 1:19	<b>3</b> 1:45	<b>17</b> 1:14	<b>21</b> 1:11

**Table 1:** Number of assigned key frames and resulting key frame rates per shot

Table 1 illustrates the key frame assignment for the sequence used to compute  $FFD(k)$  curve in Figure 2a. The given maximal number of key frames was  $N=100$ .

Equation (4) assumes that the entire sequence is available prior to the assignment process, so that the total “content” of the sequence (denominator in (4)) is known. If, however, the assignment procedure is to be done sequentially or on-the-fly, we propose the following approximation:

$$K_i = \frac{C_i}{\sum_{j=1}^S C_j} N = C_i \frac{N}{T} \frac{T}{\sum_{j=1}^S C_j} \approx C_i \frac{N}{T} \frac{\sum_{j=1}^i T_j}{\sum_{j=1}^i C_j} \quad (5)$$

Here,  $T$  is the total sequence length, and  $T_j$  is the length of the shot  $j$ . The intention by such approximation is to obtain similar assignment results as by (4), however only using the information available at the moment where  $K_i$  is computed. Since the information about the total “content” of the entire sequence (denominator in (4)) is not known, we can only summarize until the shot  $i$ . This action alone could change assignment results considerably if applied in (4), and we try to compensate it by taking into consideration also the time parameters, e.g. shot and sequence lengths. We assume that the ratio between the total sequence length and the total sequence “content”, can be well approximated by the ratio between the sequence length and “content”, both taken only up to the current shot  $i$ . Assignment results in Table 1 show only minor differences, compared to results where (4) was used.

### 3.3 Key frame distribution along a shot

After assigning a certain number of key frames to each video shot, the next step is to find locations for these key frames within a shot so that they approximately capture the entire temporal information flow of a shot.

We use the “content development” curve from (2) as reference, since it represents the process of building up the entire shot “content” by accumulating relevant temporal variations along all shot frames. By its varying steepness, the curve indicates exactly all locations “where something interesting happens”: steeper parts correspond to strong and flatter parts to more stationary temporal variations.

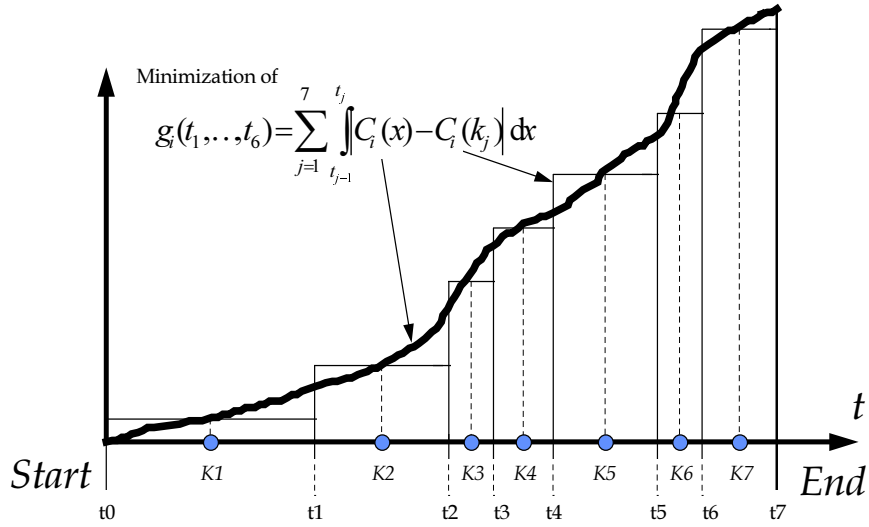
It is now our intention to perform the same “content building” operation, but using only a limited number of shot frames, e.g. assigned number of key frames. Basic idea can be seen in Figure 3, with 7 assigned key frames. The actual “content development” is approximated by the curve  $C_i(k_j)$  composed of rectangles, each one defined by  $k_j$  and  $t_j$  and each corresponding to one key frame. 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 approximation

process leads automatically to a key frame density which corresponds to the behaviour of the actual “content development” curve, e.g. higher density in steeper segments. In this way, the optimal representation of a (variable) temporal information flow along a shot can be achieved. It can be said, that each key frame represents all shot frames within its rectangle.

Technically, the key frame distribution along a shot results from minimizing the following criterion function:

$$g(k_1, \dots, k_{K_i}, t_1, \dots, t_{K_i-1}) = \sum_{j=1}^{K_i} \int_{t_{j-1}}^{t_j} |C_i(x) - C_i(k_j)| dx \quad (6)$$

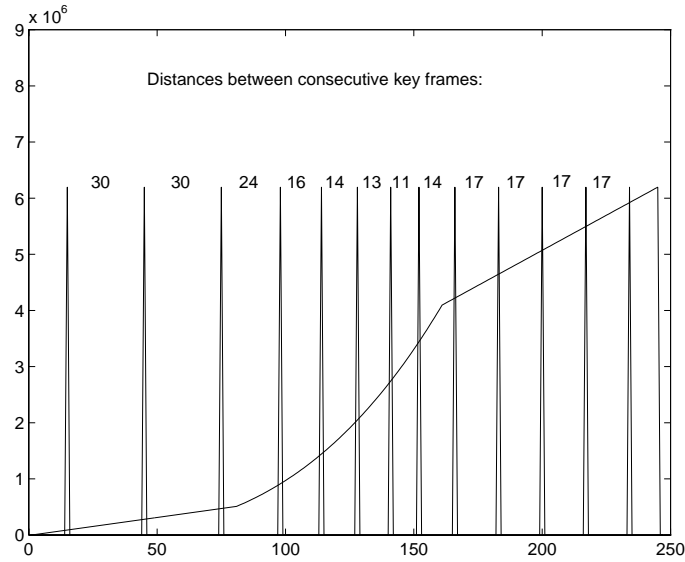
The numerical procedure for obtaining the optimal key frame distribution along a video shot by minimizing the criterion function (6) is explained in more details in [7].



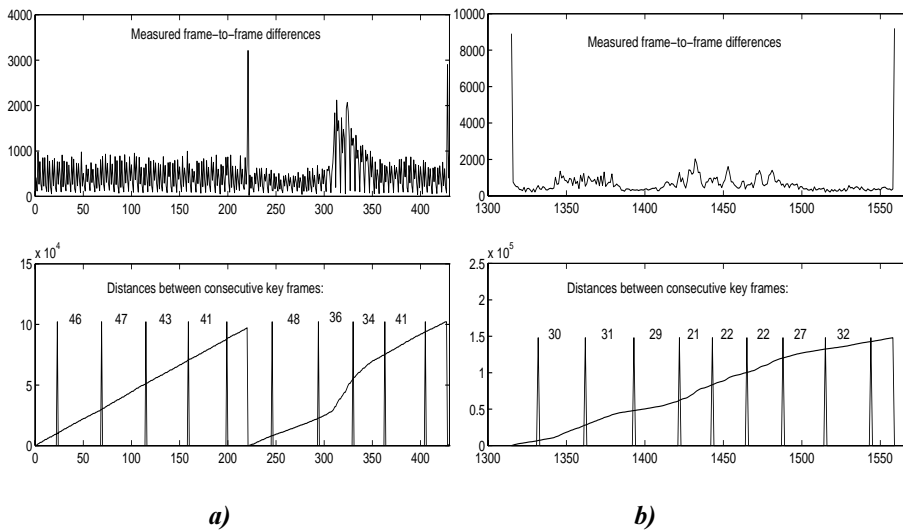
**Figure 3:** Illustration of the key frame distribution within a video shot by assigned 7 key frames. An approximation of the accumulation can be obtained using 7 flexible rectangles.

### 3.3.1 Key frame distribution experiments

Let us first check the performance of the distribution procedure on a fictive shot with the “content development” represented by the curve in Figure 4. The curve shows a variable temporal information flow along the shot. Three characteristic parts can be recognized, two with a stationary and the middle one with a non-stationary “content development”. 13 key frames were assigned to this shot and the result of their distribution along the shot can also be seen in Figure 4.



**Figure 4:** Application of the approach to a fictive video shot. Obtained variable key frame densities picture the actual content development



**Figure 5a:** Two shots of the movie *Four Weddings and a Funeral*. Frame-to-frame differences, accumulated differences and assigned key frames can be seen.

**Figure 5b:** Same analysis for one shot of the *Nature* movie

Figures 5a and 5b present results of the key frame allocation procedure on example of two real video sequences (*Four Weddings and a Funeral* -courtesy of Polygram, *Nature*).

In each case, the variability of the key frame density along a shot depends on the variability in the temporal information flow, e.g. behaviour of the “content development” curve. However, a certain level of homogeneity in spreading of key frames along the shot remains, providing the representation of the entire shot material.

#### **4 Conclusions**

In this paper a new two-step approach for key frame extraction is presented taking into account given maximal number of key frames for the whole sequence and spreading it corresponding to the measured temporal information flow along the sequence. The global key frame allocation method assigns a number of key frames to a shot depending on the total amount of changes in the information flow along the shot. These key frames are then optimally distributed over the shot in a threshold independent way.

In comparison to key frame extraction procedures proposed in the literature, this approach has following important properties:

- It allows the regulation of the maximal number of key frames per video sequence by setting the maximum of key frames  $N$ .
- Spreading of  $N$  given key frames along the sequence and their positioning within each shot is not based on any parameter (threshold) setting but on the analysis of the actual temporal information flow in the given sequence.

The quality of obtained key frame representation of the video sequence depends strongly on the robustness of frame-to-frame difference metric against non-relevant fluctuations in the analysed information flow.

#### **References:**

- [1] Zhang H., Low C.Y., Smoliar, S.W.: “Video Parsing and Browsing Using Compressed Data”, *Multimedia Tools and Applications*, 1, pp 89-111, Kluwer Academic Publishers, 1995.
- [2] Yeung M.M., Yeo B., Wolf W., Liu B.: “Video Browsing using Clustering and Scene Transitions on Compressed Sequences”, *IS&T/SPIE Multimedia Computing and Networking*, February 1995
- [3] Yeung M.M., Liu B.: “Efficient Matching and Clustering of Video Shots”, *Proceedings of ICIP*, Vol.1, pp 338-341, Washington, D.C., USA, 1995.
- [4] Yeo B., Liu B.: “Rapid Scene Analysis on Compressed Video”, *IEEE Transactions on Circuits and Systems for Video technology*, Vol.5, No.6, December 1995
- [5] Furth B., Smoliar S.W., Zhang H.: “Video and Image Processing in Multimedia Systems”, Kluwer Academic Publishers, 1995
- [6] COMPUTER - IEEE Computer Magazine, Vol. 29, Issue 5, May 1996

- [7] Legendijk R.L., Hanjalic A., Ceccarelli M., Soletic M., Persoon E.: "Visual Search in a SMASH System", Proc. ICIP '96, Lausanne, CH, 1996
- [8] ETS 300 421, "Digital broadcasting systems for television, sound and data services; framing structure, channel coding and modulation for 11/12 Ghz satellite services", EBU/ETSI JTC, December 1994.
- [9] Arman F., Hsu A., Chiu M.-Y.: "Image Processing on Compressed Data for Large Video Databases", Proc. ACM Multimedia '93, Anaheim, CA, 1993
- [10] Pentland et al.: "Video and Image Semantics: Advanced Tools for Telecommunications", IEEE MultiMedia, Summer 1994
- [11] Picard R.W.: "Light-years from Lena: Video and Image Libraries of the Future", *Proc. of the IEEE Int. Conf. on Image Processing 1995*, vol. 1, pp. 310-313, Washington DC, USA, 1995.
- [12] ISO/IEC JTC1/SC29/WG11: "MPEG-7: Context and Objectives (v.3)", Bristol, April 1997
- [13] Gerek O.N., Altunbasak Y.: "Key Frame Selection from MPEG Video Data", Proc. of SPIE Vol. 3024, San Jose 1997
- [14] Xiong W., Ma R., Lee J.C.-M.: "A Novel technique for Automatic Key Frame Computing", Proc. of SPIE Vol. 3022, San Jose 1997