# Information Theory-Based Shot Cut/Fade Detection and Video Summarization

Zuzana Černeková, Ioannis Pitas, *Senior Member, IEEE* and Christophoros Nikou *Member, IEEE*

***Abstract*— New methods for detecting shot boundaries in video sequences and for extracting key frames using metrics based on information theory are proposed. The method for shot boundary detection relies on the mutual information (MI) and the joint entropy (JE) between the frames. It can detect cuts, fade-ins and fade-outs. The detection technique was tested on the TRECVID2003 video test set having different types of shots and containing significant object and camera motion inside the shots. It is demonstrated that the method detects both fades and abrupt cuts with high accuracy. The information theory measure provides us with better results because it exploits the inter-frame information in a more compact way than frame subtraction. It was also successfully compared to other methods published in literature. The method for key frame extraction uses mutual information as well. We show that it captures satisfactorily the visual content of the shot.**

***Index Terms*— shot boundary detection, entropy, mutual information, detection accuracy, video segmentation, video analysis, key frame extraction.**

## I. Introduction

INDEXING and retrieval of digital video is an active research area. Video segmentation is a fundamental step in analyzing video sequence content and in devising methods for efficient access, retrieval and browsing of large video databases. Shot boundary detection is an important task in managing video databases for indexing, browsing, search, summarization and other content-based operations. A video shot is defined as a sequence of frames captured by *one camera in a single continuous action in time and space* [1]. Usually it is a group of frames that have consistent visual characteristics (including color, texture and motion). Video shots are the basic structural building blocks of a video sequence. Their boundaries need to be determined possibly automatically to allow content-based video manipulation.

Early work on shot detection was mainly focused on abrupt cuts. A comparison of existing methods is presented in [2], [3]. In some early approaches a cut is detected when a certain difference measure between consecutive frames exceeds a threshold. The difference measure is computed either at a pixel level or at a block level. Noticing the weakness of pixel difference methods (high sensitivity to object and camera motions), many researchers suggested the use of other measures based on global information, such as intensity histograms or color histograms [4]–[7]. The standard color histogram-based algorithm and its variations are widely used for detecting abrupt cuts. Even these histograms do not explicitly model the image difference caused by large camera motion, and thus, strictly speaking, are incapable to differentiate between smooth camera motion/parameter changes and gradual scene transitions. While the use of more complex features, such as image edges or histograms or motion vectors [8] improves the situation, it will only relieve but do not solve this problem [9]. The possibility of detecting abrupt cuts by measuring information changes between adjacent images, quantized by mutual information in gray-scale space of the images is followed in [10]. In order to detect the video shot cut they are using affine image registration for compensation of camera panning and zooming. This makes the approach very computationally expensive.

Gradual transitions such as dissolves, fade-ins, fade-outs and wipes are examined in [11]–[17]. Such transitions are generally more difficult to detect, due to camera and/or object motions within a shot. A *fade* is a transition of gradual decrease (fade-out) or increase (fade-in) of visual intensity. Fades are widely used in TV footage and their appearance generally signals a shot or story change. Therefore, their detection is a very powerful tool for shot classification and story summarization. Existing techniques for fade detection proposed in the literature rely on twin thresholding [18] or standard deviation of pixel intensities [2]. Hampapur et al [19] suggested a shot detection scheme based on modeling video edits. They computed a chromatic images by dividing the intensity change of each pixel between two successive frames by the intensity of the same pixel in the second frame. During dissolves and fades, this chromatic image assumes a reasonably constant value. This technique is very sensitive to camera and object motion. Lienhart [2] proposed to locate first all monochromatic frames in the video as potential start/end points of fades. Monochromatic frames were identified as frames with standard deviation of pixel intensities close to zero. Fades were then detected by starting to search in both directions for a linear increase in the standard deviation of pixel intensity/color. An average true positive rate of 87% was reported at a false alarm rate of 30%. An alternative approach also based on the variance of pixel intensities was proposed by Alattar [20]. Fades were detected first by recording all negative spikes in the time series of the second order difference of the pixel intensity variance, and then by ensuring that the

Manuscript received February 12, 2004; revised November 3, 2004. This study has been supported by the Commission of the European Communities in the framework of the Methods for Unified Multimedia Information Retrieval (MOUMIR) project - RTN-1999-00177 MOUMIR. This paper was recommended by Associate Editor R. Lienhart.

Z. Černeková and I. Pitas are with Aristotle University of Thessaloniki, Department of Informatics, Artificial Intelligence and Information Analysis Laboratory, 54124 Thessaloniki, Greece (e-mail: zuzana@aiia.csd.auth.gr; pitas@aiia.csd.auth.gr).

C. Nikou is with University of Ioannina, Department of Computer Science, GR 45110 Ioannina, Greece (e-mail:cnikou@cs.uoi.gr).

first order difference of the mean of the video sequence was relatively constant next to the above mentioned negative spike. A combination of both approaches is described in Truong et al. [21]. These methods have a relatively high false detection rate. Moreover, even if the existing methods based on histograms [22] detect scene changes correctly, they cannot detect fades and other gradual video transitions.

Key frames provide a suitable video summarization and a framework for video indexing, browsing and retrieval. The use of key frames greatly reduces the amount of data required in video indexing and provides an organizational framework for dealing with the video content. A lot of research work has been done in key frame extraction [23]–[25]. The simplest proposed methods are choosing only one frame for each shot (usually the first one), regardless of the complexity of visual content. The more sophisticated approaches take into account visual content, motion analysis and shot activity [26]. These approaches either do not effectively capture the major visual content or are computationally expensive.

In this paper, we propose a new approach for shot boundary detection in the uncompressed image domain based on the mutual information and the JE between consecutive video frames. Mutual information is a measure of information transported from one frame to another one. It is used for detecting abrupt cuts, where the image intensity or color is abruptly changed. A large video content difference between two frames, showing weak inter-frame dependency leads to a low MI. The entropy measure provides us with better results, because it exploits the inter-frame information flow in a more compact way than a frame subtraction. In the case of a fade-out, where the visual intensity is usually decreasing to a black image, the decreasing inter-frame joint entropy is used for detection. In case of a fade-in, the increasing JE is used for detection. The application of these entropy-based techniques for shot cut detection was experimentally proven to be very efficient, since they produce false acceptance rates very close to zero.

The proposed method was also favorably compared to other recently proposed shot cut detection techniques. At first, we compared the JE metric for fade detection to the technique proposed by Lienhart [2] relying on the standard deviation of pixel intensities. Finally, we compared our algorithm for abrupt cut detection based on MI to two techniques relying on histograms. The first one [22] combines two shot boundary detection schemes based on color frame differences and color vector histogram differences between successive frames. The second technique [2] uses one of the most reliable variants of histogram-based detection algorithms.

We propose also a method for extracting key frames from each shot using already calculated MI values. The mutual information expresses the changes in the shot and thus, the selected key frames capture well the visual content of the shot.

The remainder of the paper is organized as follows: In Section 2, a brief description of the MI and the JE as well as a definition of abrupt cuts and fades are presented. The description of our approach and its computational complexity are addressed in Section 3. In Section 4, the method for key frame extraction is described. Experimental results are presented and commented in Section 5 and conclusions are drawn in Section 6.

## II. Background and definitions

### A. Mutual information

Let $X$ be a discrete random variable with a set of possible outcomes $A_X = \{a_1, a_2, ..a_N\}$ having probabilities $\{p_1, p_2, ..p_N\}$, with $p_X(x = a_i) = p_i, p_i \geq 0$ and $\sum_{x \epsilon A_X} p_X(x) = 1$. Entropy measures the information content or ”uncertainty” of $X$ and it is given by [27], [28]:

$$H(X) = -\sum_{x \in A_X} p_X(x) \log p_X(x). \tag{1}$$

The *joint entropy* of $X$, $Y$ is expressed as:

$$H(X,Y) = -\sum_{x,y \in A_X, A_Y} p_{XY}(x,y) \log p_{XY}(x,y) \tag{2}$$

where $p_{XY}(x,y)$ is the joint probability density function. For two random variables $X$ and $Y$, the *conditional entropy* of $Y$ given $X$ is written $H(Y|X)$ and is defined as:

$$\begin{aligned} H(Y|X) &= \sum_{x \in A_X} p_X(x) H(Y|X=x) = \\ &= -\sum_{x,y \in A_X, A_Y} p_{XY}(x,y) \log p_{XY}(x|y) \end{aligned} \tag{3}$$

where $p_{XY}(x|y)$ denotes conditional probability. The conditional entropy $H(Y|X)$ is the uncertainty in $Y$ given knowledge of $X$. It specifies the amount of information that is gained by measuring a variable and already knowing another one. It is very useful if we want to know if there is a functional relationship between two data sets. Conditional entropy has the following properties:

- $H(X|Y) \leq H(X)$
- $H(X|Y) = H(Y|X)$
- $H(X|Y) = 0 \Leftrightarrow X = f(Y)$

The *mutual information* between the random variables $X$ and $Y$ is given by:

$$I(X,Y) = -\sum_{x,y \in A_X, A_Y} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \tag{4}$$

and measures the amount of information conveyed by $X$ about $Y$. Some important properties of the MI are:

- $I(X,Y) \geq 0$.
- For both independent and zero entropy sources $X$ and $Y$: $I(X,Y) = 0$.
- $I(X,Y) = I(Y,X)$
- The relation between the MI and the JE of random variables $X$ and $Y$ is given by:

  $$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{5}$$

  where $H(X)$ and $H(Y)$ are the marginal entropies of $X$ and $Y$.
- The MI is a measure of the additional information known about one expression pattern when given another as is given by

  $$I(X,Y) = H(X) - H(X|Y). \tag{6}$$

According to (5), the MI not only provides us with a measure of correspondence between $X$ and $Y$ but also takes into account the information carried by each frame at their overlap. By these means, MI decreases when the amount of shared information between $H(X)$ and $H(Y)$ is small. We can also see from (6) that the MI will reduce if $X$ carries no information about $Y$.

### B. Video Cuts and Fades

A video shot cut (abrupt cut) is an instantaneous content transition from one shot to the next one. It is obtained by simply concatenating two different shots without the use of any other transition effect. The cut boundaries show an abrupt change in image intensity or color. Cuts between shots with little content or camera motion and constant illumination conditions can be easily detected by looking for sharp intensity changes. However, in the presence of fast object motion, camera motion or illumination changes, it is difficult to distinguish if intensity changes are due to shot content changes or a shot cut [18].

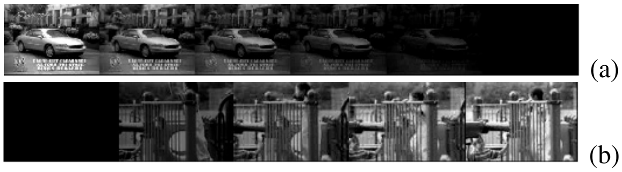

Fig. 1. *Consecutive frames from "news" video sequence showing: (a) a fade-out;(b) a fade-in.*

A fade-out is a video transition determining the progressive darkening of a shot until the last frame becomes black, as can be seen in Figure 1a. A fade-in allows the gradual transition from a black frame to the fully illuminated one, as shown in Figure 1b. Fades spread the boundary between two shots across a number of consecutive video frames. They have both start and end frames identifying the transition sequence. In both cases (fade-in, fade-out) fades can be mathematically modeled as luminance scaling operations. If $G(x, y, t)$ is a gray scale sequence and $l_s$ is the length of the transition sequence, an intensity scaling of $G(x, y, t)$ is modeled as [18]:

$$f(x, y, t) = G(x, y, t) \cdot (1 - \frac{t}{l_s}) \qquad t \in [t_0, t_0 + l_s] \tag{7}$$

Therefore, fade-out is modeled by:

$$f(x, y, t) = G_1(x, y, t) \cdot (\frac{l_1 - t}{l_1}) \tag{8}$$

and fade-in by:

$$f(x, y, t) = G_2(x, y, t) \cdot (\frac{t}{l_2}) \tag{9}$$

## III. Shot detection

In our approach, the MI and the JE between two successive frames are calculated separately for each of the RGB components. Let us consider that the video sequence gray levels vary from 0 to $N - 1$. At frame $\mathbf{f}_t$ three $N \times N$ matrices $\mathbf{C}^R_{t,t+1}$, $\mathbf{C}^G_{t,t+1}$ and $\mathbf{C}^B_{t,t+1}$ are created carrying information on the gray level transitions between frames $\mathbf{f}_t$ and $\mathbf{f}_{t+1}$. In the case of the $R$ component, the element $\mathbf{C}^R_{t,t+1}(i, j)$, with $0 \leq i \leq N - 1$ and $0 \leq j \leq N - 1$, corresponds to the probability that a pixel with gray level $i$ in frame $\mathbf{f}_t$ has gray level $j$ in frame $\mathbf{f}_{t+1}$. By the other words, $\mathbf{C}^R_{t,t+1}(i, j)$ is a number of pixels which change from gray level $i$ in frame $\mathbf{f}_t$ to gray level $j$ in frame $\mathbf{f}_{t+1}$, divided by the number of pixels in the video frame. Following equation (4), the mutual information $I^R_{t,t+1}$ of the transition from frame $\mathbf{f}_t$ to frame $\mathbf{f}_{t+1}$ for the $R$ component is expressed by:

$$I^R_{t,t+1} = - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \mathbf{C}^R_{t,t+1}(i, j) \log \frac{\mathbf{C}^R_{t,t+1}(i, j)}{\mathbf{C}^R_t(i)\mathbf{C}^R_{t+1}(j)}. \tag{10}$$

The total MI is defined as:

$$I_{t,t+1} \triangleq I^R_{t,t+1} + I^G_{t,t+1} + I^B_{t,t+1}. \tag{11}$$

By using the same considerations, the JE $H^R_{t,t+1}$ of the transition from frame $\mathbf{f}_t$ to frame $\mathbf{f}_{t+1}$, for the $R$ component, is given by:

$$H^R_{t,t+1} = - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \mathbf{C}^R_{t,t+1}(i, j) \log \mathbf{C}^R_{t,t+1}(i, j). \tag{12}$$

The total JE is defined as:

$$H_{t,t+1} \triangleq H^R_{t,t+1} + H^G_{t,t+1} + H^B_{t,t+1}. \tag{13}$$

### A. Abrupt shot cut detection

A small value of the MI $I_{t,t+1}$ indicates the existence of a cut between frames $f_t$ and $f_{t+1}$. Basically, in this context, abrupt cut detection is the outlier detection in an one-dimensional MI signal given by (11) [29]. In order to detect possible shot cuts, an adaptive thresholding approach was employed. Local MI mean values on an one-dimensional temporal window $W$ of size $N_W$ are obtained at each time instance $t_c$ without considering the current value $I_{t_c,t_c+1}$ at the current window center $t_c$ [29]:

$$\bar{I}_{t_c} = \frac{1}{N_W} \sum_{\substack{t \in W \\ t \neq t_c}} I_{t,t+1} \tag{14}$$

The quantity $\bar{I}_{t_c}/I_{t_c,t_c+1}$ is then compared to a threshold $\epsilon_c$ in order to detect the peaks, which correspond to the shot cuts. Threshold $\epsilon_c$ was chosen experimentally. An example of abrupt cut detection using MI is illustrated in Figure 2.

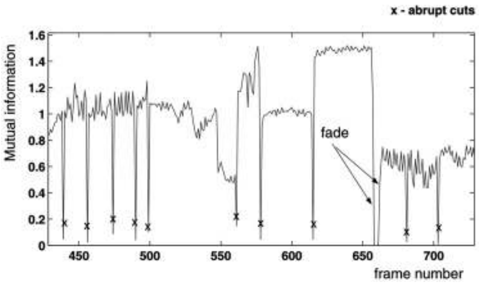

Fig. 2. *Time series of the MI from "ABC news" video sequence showing abrupt cuts and one fade.*

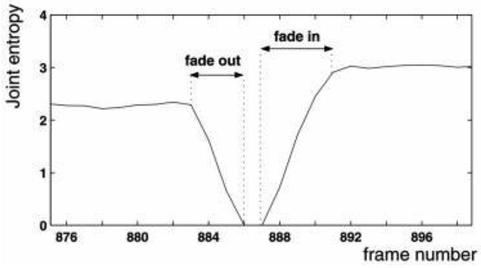

Fig. 3. *The joint entropy signal from "CNN news" video sequence showing a fade-out and fade-in to the next shot.*

### B. Fade detection

Since MI decreases when the transmitted information from one frame to another is small (in case of cuts and fades) the JE (13) is employed, to efficiently distinguish fades from cuts. The JE measures the amount of information carried by the union of these frames. Therefore, its value decreases only during fades, where a weak amount of inter-frame information is present. The pattern showing fade out and fade in is shown on Figure 3. Thus, only the values of $H_{t,t+1}$ below a certain threshold $\epsilon_f$ are examined. These values correspond to the black frames. The instance, where the JE is at a local minimum, is detected and is characterized as the end time instance $t_e$ of the fade-out. At this point the frame has become black, it does not carry any information. The next step consists in searching for the fade-out start point $t_s$ in the previous frames using the criterion:

$$\frac{H_{t_s,t_s+1} - H_{t_s-1,t_s}}{H_{t_s-1,t_s} - H_{t_s-2,t_s-1}} \geq T \tag{15}$$

where $T$ is a predefined threshold which guarantees that, at start point $t_s$, the JE starts decreasing. In order to handle a specific type of video sequences where the frame content remains exactly the same for two or three consecutive frames, due to the chosen video digitalization procedure (typically a reduced digitization frame rate), we check the possible increase of JE values up to the third frame. The same procedure also applies for fade-in detection (with $t_s$ being detected at first). Finally, since the fade has boundary spread across number of frames, the segment is considered as a fade only if $t_e - t_s \geq 2$, otherwise it is labeled as a cut. An example of JE signal showing a fade-out detection is presented in Figure 4.

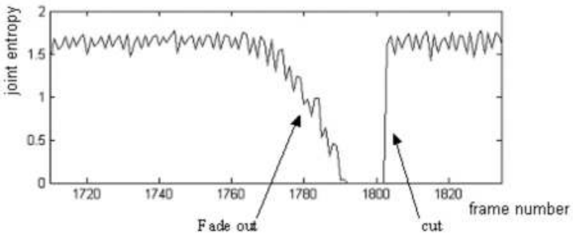

Fig. 4. *The joint entropy signal from "basketball" video sequence showing a fade-out and a transition from a black frame to the next shot.*

### C. Computational complexity

The computational complexity of these algorithms consist of calculating three histograms for each color component $R$, $G$, $B$ for two consecutive frames. If the frame size is $n$ pixels than we need $n$ additions to calculate one histogram. First we need to calculate 3 histograms, which consist of $3n$ additions. We also need another $3N^2$ multiplications, $(N-1)^2$ additions and $N^2$ logarithm calculations for computing (10), where $N$ is number of video sequence gray levels separately for $R$, $G$ and $B$ color channels. Thus, for calculation the MI between two frames $O(n+N^2)$ additions, $O(N^2)$ multiplications and $O(N^2)$ logarithm calculations are needed. The same order of computational complexity is needed for the calculation of JE, namely $3(n+(N-1)^2)+2$ additions, $3N^2$ multiplications and $3N^2$ logarithm calculations.

## IV. Key frame selection

After the temporal segmentation of a video sequence to shots, the key frames can be selected from each shot for video indexing. Our approach uses MI values, which provided us information about content changes between consecutive frames in the shot. Let us have a video shot having $N_L$ frames $s = \{f_1, f_2, ..., f_{N_L}\}$ obtained by our method for shot cut detection [30] described in the section III. Let the MI values in this shot be $I_s = \{I_{1,2}, I_{2,3}, ..., I_{N_L-1,N_L}\}$. In order to find if the content in the shot changes significantly, the standard deviation $\sigma_{I_s}$ of the MI within this shot is calculated. The value $\sigma_{I_s}$ is compared to predefined threshold $\epsilon$. If $\sigma_{I_s} < \epsilon$, we assume that the content did not change significantly during the shot. Thus, any frame can effectively represent the visual content. Such a shot can present for example, a anchor person in a news video sequence. In this case, the first or a middle frame is selected as key frame.

In the case of bigger content changes within the shot, the shot must be described by more than one key frame. This is done by a split-merge approach. The MI values in the shot are divided into clusters $\{c_i\}_{i=1}^{K}$, where $K$ is a number of clusters obtained after the split-merge algorithm. The threshold parameter $\delta$ controls the number of created frame clusters. Initially, all the MI values in the shot are assigned to the one cluster $c_1 = \{I_{1,2}, I_{2,3}, ..., I_{N_L-1,N_L}\}$. The standard deviation $\sigma_{c_1}$ of these values in the cluster is compared to the predefined threshold $\delta$. If it exceeds the threshold, $\sigma_{c_1} > \delta$, all the MI values are split in two clusters: $c_{11} = \{I_{1,2}, I_{2,3}, ..., I_{\frac{N_L}{2}-1,\frac{N_L}{2}}\}$ and $c_{12} = \{I_{\frac{N_L}{2},\frac{N_L}{2}+1}, I_{\frac{N_L}{2}+1,\frac{N_L}{2}+2}, ..., I_{N_L-1,N_L}\}$. The algorithm works recursively till the standard deviation of the MI values in the cluster is smaller than the given threshold $\delta$. Then, consecutive clusters are checked for possible merging. If the standard deviation of two consecutive clusters is smaller than threshold $\delta$, these clusters are merged. This way, all frames from the given shot are split to clusters, depending on the MI values. After this procedure, only those clusters having enough frames are considered as *key clusters* [26] and a representative frame is extracted from this cluster as the potential key frame. In this paper, key cluster should have more than $N_L/(K*2)$ frames, where $N$ is number of frames in the shot and $K$ is number of clusters. In [26] a key cluster should have at least $N_L/K$ frames. However, in the case when we have only 2 clusters, this method will discard the smaller one, which is not acceptable. For each key cluster the most representative frame is taken as a potential key frame. The most representative frame is the one which maximizes inter-frame mutual information in the cluster:

$$f_{key} = \arg\max_{\bar{f}} \left( \frac{1}{N_j} \sum_{\substack{i \\ \bar{f} \neq f_i}}^{N_j} I_{\bar{f},f_i} \right) \tag{16}$$

where $N_j$ is number of frames in the cluster, we call this frame median frame. An example of this procedure can be seen in Figure 5.

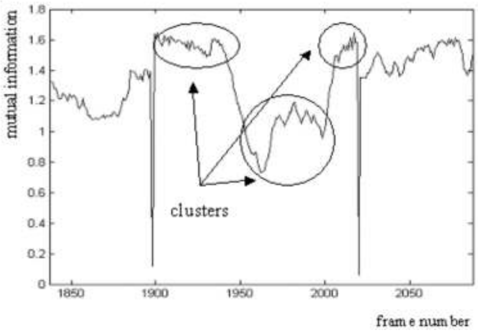

Fig. 5. *The MI signal from "star" video sequence presenting the clusters created by split-merge method. The selected potential key frames from each cluster is shown in Figure 6.*

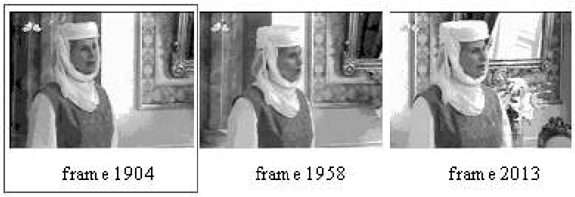

Fig. 6. *Potential key frames from "star" video sequence extracted from each cluster of the shot.*

After extracting potential key frames $\{k_i\}_{i=1}^{K}$ from the shot $s$, we try to reduce the number of key frames that represent the shot. To do this, these key frames are compared to each other by calculating their MI using (11). If the content of the key frames is similar enough (as indicated by a high MI value), the shot can be presented by less key frames. Therefore, if $I_{k_i,k_{i+1}} > \epsilon$, where $\epsilon$ is a predefined threshold, only the frame $k_{i+1}$ is considered to be a key frame and is compared to the next potential key frame $k_{i+2}$. Otherwise, both frames $k_i$ and $k_{i+1}$ are taken as key frames and $k_{i+1}$ is further compared to the others potential key frames $\{k_{i+2}, \ldots, k_K\}$. An example can be seen in Figure 6. After calculating the MI between these frames, only the first frame (frame number 1904) was selected as a key frame to represent content of the shot. This procedure of the reduction of key frame number enhances the robustness of our method to the choice of threshold $\delta$ used in the split-merge procedure. Let us note that we are not interested only in picking the fewest possible key frames but also in picking good key frames that are visually and possibly semantically important.

The reduction of the number of key frames can be done by using Median LVQ [31]. This way, we can handle the problem when the potential key frames are similar in the beginning and in the end of the shot. At first we assign all the potential key frames to one cluster $c_1$. Then the key frame that maximizes inter-frame MI is chosen as representative key frame $k_1$ of this cluster. We find the key frame $k$, which minimizes the MI between the representative key frame $k_1$ and other key frames. If this MI is below a given threshold, we split the cluster to the two. The key frame $k$, is chosen as a representative $k_2$ for new cluster $c_2$. Then we reassign the rest of key frames to these two clusters. If the MI of a key frame and the representative $k_2$ is higher than MI of this key frame and the representative $k_1$ of the cluster $c_1$ we assign this key frame to the cluster $c_2$. We repeat this procedure recursively, till we find all key frames.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed method was tested on several TV sequences (see Table I) containing many commercials, characterized by significant camera parameter changes like zoom-ins/outs, pans, abrupt camera movements as well as significant object and camera motion inside single shots. The video sequences contain film, sport, studio news, advertisements, political talks and TV series logos. These 4 video sequences "basketball", "news", "football" and "movie" totaled about 30 min, their frame size varying between 176 × 112 and 176 × 144 pixels. These videos have a specific digitization feature: they frame content changes every 3rd frame instead of every frame. For each video sequence, a human observer determined the precise locations and duration of the transitions to be used as ground truth.

TABLE I
THE VIDEO SET USED IN OUR EXPERIMENTS.

| video | frames | cuts | fade-ins | fade-outs |
|---|---|---|---|---|
| **basketball** | 3882 | 44 | 7 | 4 |
| **news** | 9446 | 40 | 6 | 6 |
| **football** | 5589 | 28 | 0 | 0 |
| **movie** | 19722 | 147 | 0 | 0 |
| **TREC video sequences** | | | | |
| **6 debate videos** | 125977 | 230 | 0 | 0 |
| **4 CNN news videos** | 209978 | 1287 | 57 | 57 |
| **4 ABC news videos** | 206144 | 1269 | 64 | 69 |

To enable future comparison with other boundary detection techniques, newscasts from the reference video test set TRECVID 2003 was added to the testing set, containing video sequences of more than 6 hours duration that has been digitized with a frame rate of 29.97fps at a resolution of 352 × 264 pixels. We used spatially downsampled frames for our experiments with resolution 176 × 132 pixels to speed up calculations. The ground truth provided by TRECVID was used for these video sequences.

In order to evaluate the performance of the shot cut detection method presented in Section III, the following measures were used, inspired by the receiver operating characteristics in

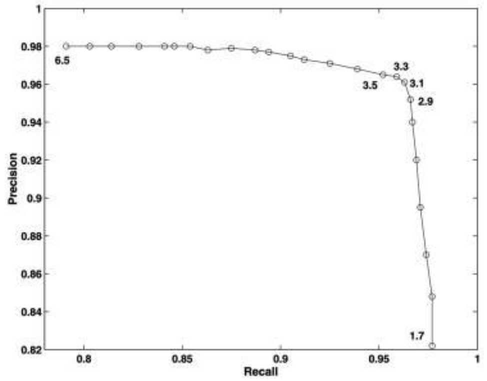

Fig. 7. *The recall-precision graph obtained for shot cut detection method by varying threshold $\epsilon_c$ in the range* $[1.7, 6.5]$.

statistical detection theory [3], [32]. Let $GT$ denote the ground truth, $Det$ the detected (correct and false) shots cuts using our methods. The following performance measures have been used:

- the *Recall* measure, also known as the true positive function or sensitivity, that corresponds to the ratio of correct experimental detections over the number of all true detections:

$$Recall = \frac{|Det \bigcap GT|}{|GT|}; \qquad (17)$$

where $|GT|$ denotes the cardinality of set $GT$

- the *Precision* measure defined as the ratio of correct experimental detections over the number of all experimental detections:

$$Precision = \frac{|Det \bigcap GT|}{|Det|}. \qquad (18)$$

For our experiments we used the size of temporal window $N_W = 3$. We have tested the method with several choices of threshold $\epsilon_c$. The recall-precision curve obtained by changing threshold $\epsilon_c$ is shown in Figure 7. The experimental tests, performed using a common prefixed threshold ($\epsilon_c = 3.1$) for all video sequences are summarized in Table II (MI method). The elapsed time for obtaining results (abrupt cuts and fades) for one video sequence having 51384 frames was 1517 seconds. Thus, the algorithm can operate in real time for the raw video sequence. The large majority of the cuts were correctly detected even in the case of the video sequences, which contain fast object and camera movements. A part of the video sequence showing a cut between two shots involving high content motion that was successfully detected by the proposed method is presented in Figure 8. A snapshot of the "football" sequence is shown in Figure 9, where a big object appears in front of the camera. This case is wrongly characterized by existing methods as a transition, whereas our method correctly does not detect a transition.

Sensitivity to camera flashes is shown on Figure 10. Figure 10a, 10b show the color histogram difference and MI respectively, calculated for the same part of video sequence containing a lot of camera flashes. Notice that the MI metric is significantly less sensitive to shot illumination changes (camera flashes) even in the RGB color space compared to histogram-based methods. This is clear from (6) and the properties of conditional entropy. In case of camera flash occurrence, the main information transported from one frame to the next one is preserved, which means that they differ only in luminance by a certain amount. Thus, from properties of conditional entropy $H(X|Y) \cong 0$ and accordingly $I(X,Y) \cong H(X)$. In the case of shot cuts the two frames are independent and there is no information transported between them, which means $I(X,Y) = 0$. In the case of histogram comparisons, the camera flashes sometimes produce more significant peaks than the peaks corresponding to cuts, which cause false detections.

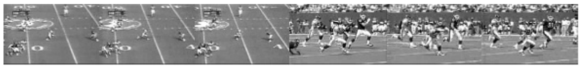

Fig. 8. *Consecutive frames from "football" video sequence showing an abrupt cut between two shots coupled with large video object motion.*

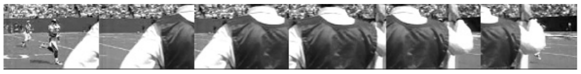

Fig. 9. *Consecutive frames from "football" video sequence showing a big object appearance in front of the camera during panning.*

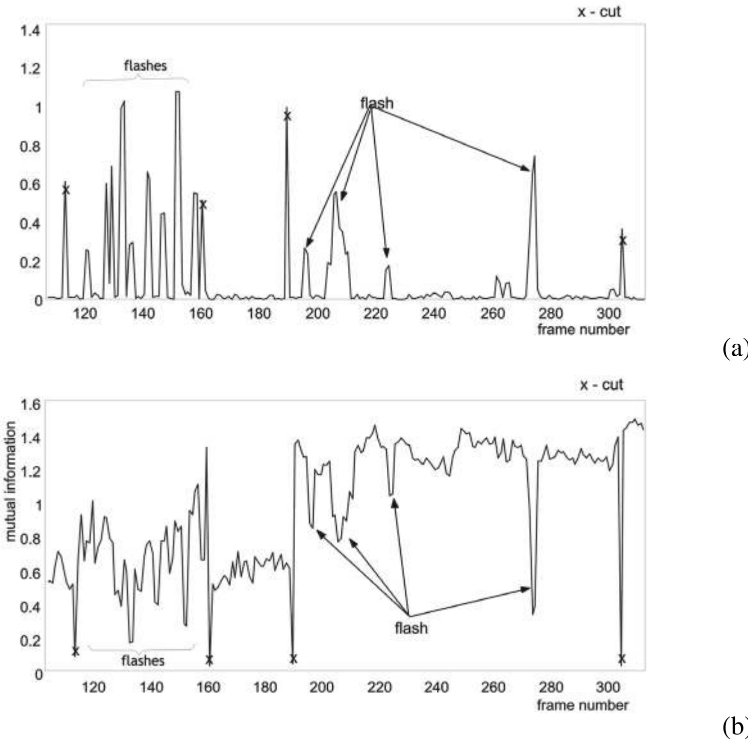

Fig. 10. *A part of video sequence containing many camera flashes. (a) color histogram comparison and (b) mutual information calculated for the same part of video sequence.*

The obtained results are better than the results for shot cut detections reported in the TRECVID2003 competition [33]. The best reported abrupt cut detection results for recall and precision are 93% and 95% respectively, whereas our method produces 97% recall and 95% precision. Most false detections

TABLE II

FIXED THRESHOLD SHOT CUT DETECTION RESULTS.

| | **MI method** | | **Combined histogram method** | | **Color histogram method** | |
|---|---|---|---|---|---|---|
| **video** | Recall | Precision | Recall | Precision | Recall | Precision |
| **basketball** | **1.00** | **1.00** | 0.91 | 0.97 | 0.52 | 0.85 |
| **news** | 0.96 | **1.00** | 0.96 | 0.98 | 0.80 | 0.89 |
| **football** | 0.93 | 1.00 | 0.96 | 1.00 | 0.68 | 0.68 |
| **movie** | **1.00** | **1.00** | 0.93 | 0.98 | 0.87 | 0.93 |
| **6 debate videos** | 1.00 | 0.99 | 1.00 | 1.00 | 0.87 | 1.00 |
| **4 CNN news** | **0.96** | **0.96** | 0.87 | 0.83 | 0.85 | 0.84 |
| **4 ABC news** | **0.97** | **0.94** | 0.85 | 0.81 | 0.87 | 0.73 |
| **TREC total** | **0.97** | **0.95** | 0.87 | 0.83 | 0.86 | 0.80 |

were caused by corrupted parts of video sequences, where a sequence of exactly the same frames is followed by a significantly different frame. This case is shown in Figure 11. In some cases, false detections appeared in the case of commercials where artistic camera edits were used. The missed shot cut detections were caused mainly by shot changes between two images with very similar spatial color distribution (Figure 12a) or if the shot change occurred only in a part of the video frame (Figure 12b). Some video sequences contained a special type of hard cut, which contained one transitional frame ("cut in two frames"). In this case, the MI shows a double peak, which was not always a strong one. Therefore, in some cases, its use caused misdetection. Such fake peaks could be caused, for example, by flash. This drawback could be improved by a pre-processing stage, where all double MI peaks would be checked for the possibility of corresponding to a hard cut, by comparing the previous and the successive frame. In case the MI value is still small, we can modify it to become a single peak.

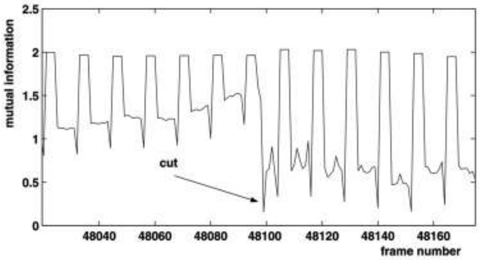

Fig. 11. *A mutual information of temporally subsampled video sequence.*

We compared our algorithm to the technique proposed in [22]. This approach combines two shot boundary detection schemes based on color frame differences and color vector histogram differences between successive frames. It is claimed to detect shot boundaries efficiently even under strong video edit effects and camera motion. This method operates in the HLS color space and ignores luminance information in order to overcome the possible drawback of histogram sensitivity to shot illumination changes. The results of this algorithm applied on the same video sequences are summarized in Table II (second and third columns). Several false shot cut detections were performed due to camera flashes. Although this approach has a high shot cut detection rate, it is generally lower compared to that of our proposed technique (Table II; MI method). Our technique is robust to the detection of shots with small length, occurring particularly during TV advertisements. Out of the falsely detected cuts, about $10\%$ of them come from other types of transitions which were detected by chance.

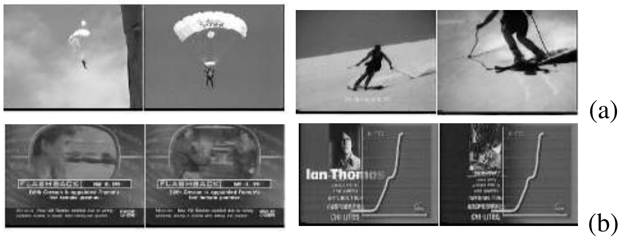

Fig. 12. *Consecutive frames from video sequences presenting abrupt cuts, which caused missed shot cut detection: a) shot changes between two images with very similar spatial color distribution and b) shot change occurs only in a part of the video frame.*

Another algorithm, that was used for comparison purposes is the so-called color histogram-based shot boundary detection algorithm [2]. It is one of the most reliable variants of histogram-based detection algorithms. Hard cuts and other short-lasting transitions are detected as single peaks in the time series of the differences between color histograms of contiguous frames or of frames lying at a certain distance $k$ apart. A hard cut is detected if only the $i$-th color histogram difference value exceeds a certain threshold $\phi$ within a local environment of a given radius of frame $f_i$. A global thresholding is used. In order to cope with the particular type of hard cut, which consists of one transitional frame, double peaks were modified into single peaks in a pre-processing stage. Table II (fifth and sixth columns) shows the results obtained by this method with threshold $\phi = 0.4$. One can see that, in general, the results are poorer than those of our proposed method using MI, they are even lower than the results obtained by the combined histogram method.

The same measures (17) and (18) have been used for the evaluation of fade detection performance. The spatial overlap precision has been used as well defined as:

$$Spatial\ Overlap\ Precision = \\ = \frac{1}{|GT \bigcap Det|} \sum_{i=1}^{|GT|} \frac{|Det_i \bigcap GT_i|}{|Det_i \bigcup GT_i|}, \quad (19)$$

where $GT_i$ is a subset of $GT$, which represents the duration of one fade. Overlap is considered to be a strong measure for

TABLE III

EVALUATION OF FADE DETECTION BY THE PROPOSED JOINT ENTROPY METHOD.

| video | fade-ins | | | fade-outs | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | Overlap | Recall | Precision | Overlap |
| **basketball** | **1.00** | 1.00 | **0.78** | **1.00** | 1.00 | **0.90** |
| **news** | **1.00** | 1.00 | **0.71** | **1.00** | 1.00 | 0.85 |
| **4 CNN news** | **1.00** | 0.84 | **0.83** | 1.00 | **0.84** | 0.86 |
| **4 ABC news** | 0.93 | **0.90** | **0.74** | 0.92 | **0.92** | 0.75 |

detection accuracy, since, for example, a shot of length $N_L$ shifted by one frame results in only $\frac{N_L-1}{N_L}$ overlap.

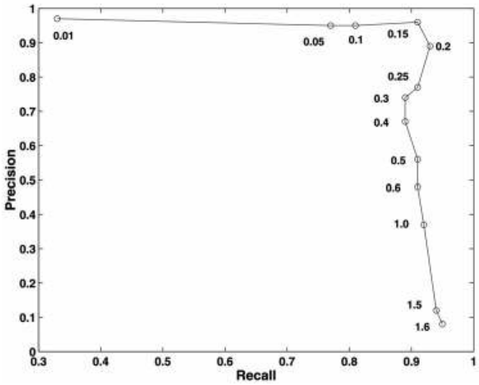

Fig. 13. *The recall-precision graph obtained for fade detection method by varying threshold* $\epsilon_f$ *in the range* $[0.01, 1.5]$ *and choosing* $T = 3$.

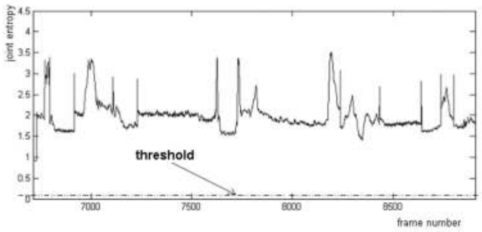

Fig. 14. *The joint entropy signal from "star" video sequence representing no fades.*

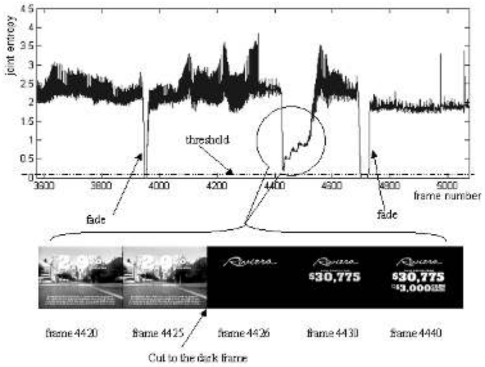

Fig. 15. *The joint entropy signal from "news" video sequence having 2 fades.*

The fade detection method was tested for several choices of threshold $\epsilon_f$. The recall-precision curve obtained by changing threshold $\epsilon_f$ is shown in Figure 13. The experimental tests, performed using a common prefixed thresholds ($\epsilon_f = 0.15$ and $T = 3$) for all video sequences are summarized in Table III. Using this setup, the fade boundaries were detected within a precision of ±2 frames. In most cases, the boundaries toward black frames were recognized without any error. The robustness of the JE measure in fade detection and, especially, in avoiding false fade detections is illustrated in Figures 14 and 15. The use of JE for detecting fades is robust in case big objects move in front of the camera, that can cause severe occlusion and a blank frame in the video sequence. In our experiments, the threshold was set to a very low value to avoid false detections (see Figure 15). Some fade detections were missed when there was noise in the black frame or when the fading was not complete and the end frame was just very dark gray instead of black. In the TRECVID competition, only the abrupt cuts and the gradual transitions are evaluated separately. Therefore, we cannot make comparison of our results to those of other TRECVID participants.

Our method for fade detection was compared to the approach proposed in [2] that is based on the standard deviation of pixel intensities (SD method) and claims to detect fades with high detection rate and to determine fade boundaries with high precision. In order to obtain results three different parameters have to be set: minimal required length of a fade, minimal required correlation and maximal allowed standard deviation of the pixel color values in the first/last monochrome frame of a fade. We used the default setup of the parameters in the algorithm, as proposed in [2], since it was referred that this choice attains the best performance. As can be seen in Table IV, several fades were not correctly detected by this method. You can notice a very low recall in case of "basketball" and "news" video sequences, which was caused by the video content change every 3rd frame. The above mentioned observations were also confirmed by the starting/ending frame location estimation provided by our JE approach and the SD technique and their error statistics are presented in Tables V and VI respectively. For all fades (Table I), the starting and ending frames were detected by both methods and the location errors were calculated. The JE proposed method provided superior performance than the SD method in median and mean error values and presented no errors in fade-out end point detection. Furthermore, the significantly smaller maximum errors of the JE technique with regard to those of the SD method illustrate the robustness of our algorithm.

After the video was segmented to video shots, we applied our method for key frame selection on the video sequences. In the case of shots without significant content changes, our method successfully chose only one frame, even if more

TABLE IV

EVALUATION OF FADE DETECTION BY THE SD METHOD [2].

| **video** | **fade-ins** | | | **fade-outs** | | |
|---|---|---|---|---|---|---|
| | Recall | Precision | Overlap | Recall | Precision | Overlap |
| **basketball** | 0.14 | 1.00 | 0.41 | 0.50 | 1.00 | 0.85 |
| **news** | 0.17 | 1.00 | 0.54 | 0.17 | 1.00 | 0.65 |
| **4 CNN news** | 0.96 | 0.84 | 0.70 | 1.00 | 0.71 | 0.87 |
| **4 ABC news** | 1.00 | 0.88 | 0.68 | 0.97 | 0.87 | 0.81 |

TABLE V

FADE LOCATION ERROR STATISTICS FOR THE PROPOSED JE METHOD (IN FRAME NUMBERS).

| **effects** | **fade-outs** | | **fade-ins** | |
|---|---|---|---|---|
| **frame** | $f_s$ | $f_e$ | $f_s$ | $f_e$ |
| **median bias** | 2 | **0** | **0** | 2 |
| **mean bias ± s. dev.** | 2.3 ± 1.7 | **0.2 ± 0.4** | **0.3 ± 0.6** | **2.6 ±2.6** |
| **max bias** | 9 | **1** | **2** | 10 |

TABLE VI

FADE LOCATION ERROR STATISTICS FOR THE SD METHOD (IN FRAME NUMBERS).

| **effects** | **fade-outs** | | **fade-ins** | |
|---|---|---|---|---|
| **frame** | $f_s$ | $f_e$ | $f_s$ | $f_e$ |
| **median bias** | 1 | 1 | 1.5 | 1 |
| **mean bias ± s. dev.** | $1.7 \pm 2.3$ | $1.3 \pm 0.9$ | $1.8 \pm 2.2$ | $2.9 \pm 3.0$ |
| **max bias** | 8 | 4 | 4 | 9 |

potential key frames were extracted after clustering. For shots with big content changes, usually due to camera or object motions, more key frames were selected, depending on visual complexity of the shot. An example of key frames extracted from one shot with more complicated content can be seen in Figure 16. The selected key frames of a certain video sequence are shown in Figure 17. Thus, our method captures the shot content very well.

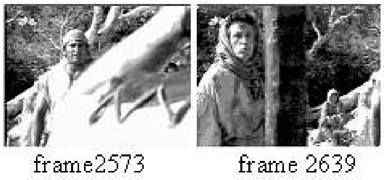

Fig. 16. *Examples of key frames selected by our method from "star" video sequence to represent visual content of one shot.*

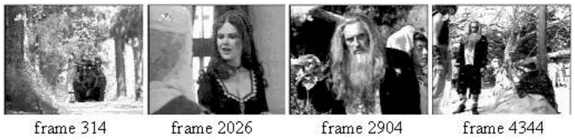

Fig. 17. *Examples of key frames extracted by our method from "star" video sequence.*

## VI. CONCLUSION

A novel technique for shot transitions detection is presented. We propose a new method for detecting abrupt cuts and fades using the MI and the JE measures respectively. The accuracy of our approach was experimentally shown to be very high. Experiments illustrated that fade detection using the JE can efficiently differentiate fades from cuts, pans, object or camera motion and other types of video scene transitions, while most of the methods reported in the current literature fail to characterize these kinds of transitions. The method was successfully compared to other methods reported in literature.

## ACKNOWLEDGMENT

The C-SPAN video used in this work is provided for research purposes by C-SPAN through the TREC Information-Retrieval Research Collection. C-SPAN video is copyrighted.

## REFERENCES

[1] X. U. Cabedo and S. K. Bhattacharjee, "Shot detection tools in digital video," in *Proceedings of Non-linear Model Based Image Analysis 1998, Springer Verlag, Glasgow*, July 1998, pp. 121–126.

[2] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," in *Proc. of SPIE Storage and Retrieval for Image and Video Databases VII, San Jose, CA, U.S.A.*, vol. 3656, January 1999, pp. 290–301.

[3] P. Browne, A. F. Smeaton, N. Murphy, N. O'Connor, S. Marlow, and C. Berrut, "Evaluation and combining digital video shot boundary detection algorithms," in *Proceedings of the Fourth Irish Machine Vision and Information Processing Conference, Queens University Belfast*, 2000.

[4] A. Dailianas, R. B. Allen, and P. England, "Comparison of automatic video segmentation algorithms," in *Proceedings, SPIE Photonics East'95: Integration Issues in Large Commercial Media Delivery Systems, Oct. 1995, Philadelphia*, vol. 2615, 1995, pp. 2–16.

[5] G. Ahanger and T. Little, "A survey of technologies for parsing and indexing digital video," *Journal of visual Communication and Image Representation*, vol. 7, no. 1, pp. 28–43, 1996.

[6] N. V. Patel and I. K. Sethi, "Video shot detection and characterization for video databases," *Pattern Recognition*, vol. 30, no. 4, pp. 583–592, April 1997.

[7] S. Tsekeridou and I. Pitas, "Content-based video parsing and indexing based on audio-visual interaction," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 522–535, 2001.

[8] C.-L. Huang and B.-Y. Liao, "A robust scene-change detection method for video segmentation," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11 no.12, pp. 1281–1288, 2001.

[9] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?" *IEEE Trans. Circuits and Systems for Video Technology*, vol. 12 no.2, pp. 90–105, 2002.

[10] T. Butz and J. Thiran, "Shot boundary detection with mutual information," in *Proc. 2001 IEEE Int. Conf. Image Processing, Greece*, vol. 3, October 2001, pp. 422–425.

[11] R. Lienhart, "Reliable dissolve detection," in *Proc. of SPIE Storage and Retrieval for Media Databases 2001*, vol. 4315, January 2001, pp. 219–230.

[12] M. S. Drew, Z.-N. Li, and X. Zhong, "Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences," in *Proc. 2000 IEEE Int. Conf. on Image Processing*, vol. 3, 2000, pp. 929–932.

[13] R. Lienhart and A. Zaccarin, "A system for reliable dissolve detection in video," in *Proceeding of IEEE Intl. Conf. on Image Processing 2001 (ICIP'01), Thessaloniki, Greece*, Oct. 2001, pp. 406-409.

[14] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying production effects," *ACM Journal of Multimedia Systems*, vol. 7, pp. 119–128, 1999.

[15] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, November 2000.

[16] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. Volume: 12, no. 2, pp. 90 – 105, Feb. 2002.

[17] W. J. Heng and K. N. Ngan, "Shot boundary refinement for long transition in digital video sequence," *Multimedia, IEEE Trans. on*, vol. Volume: 4, no. 4, pp. 434 – 445, Dec. 2002.

[18] A. D. Bimbo, *Visual Information Retrieval*. Morgan Kaufmann Publishers, Inc, San Francisco, California, 1999.

[19] A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proc. ACM Multimedia 94, San Francisco, CA*, October 1994, pp. 357–364.

[20] A. M. Alattar, "Detecting fade regions in uncompressed video sequences," in *Proc. 1997, IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1997, pp. 3025–3028.

[21] B. T. Truong, C. Dorai, and S. Venkatesh, "New enhancements to cut, fade, and dissolve detection processes in video segmentation," in *ACM Multimedia 2000*, November 2000, pp. 219–227.

[22] S. Tsekeridou, S. Krinidis, and I. Pitas, "Scene change detection based on audio-visual analysis and interaction," in *Proc. 2000 Multi-Image Search and Analysis Workshop*, March 2001, pp. 214-225.

[23] B. G. unsel and A. M. Tekalp, "Content-based video abstraction," in *Proc. 1998 IEEE Int. Conf. on Image Processing, Chicago IL, October*, 1998.

[24] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," in *Visual Database Systems*, vol. II, 1992.

[25] W. Wolf, "Key frame selection by motion analysis," in *Proc. 1996 IEEE Int. Conf. Acoust., Speech and Signal Proc.*, 1996, vol. 2, pp. 1228-1231.

[26] Y. Zhuang, Y. Rui, T. S. Huang, and S. Metrotra, "Adaptive key frame extraction using unsupervised clustering," in *In Proc. of IEEE Int. Conf. on Image Processing, Chicago IL, October*, 1998, pp. 886–890.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[28] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, Inc., 1991.

[29] I. Pitas and A. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications*. Kluwer Academic, 1990.

[30] Z.Cernekova, C.Nikou, and I.Pitas, "Shot detection in video sequences using entropy-based metrics," in *Proc. 2002 IEEE Int. Conf. Image Processing, Rochester N.Y., USA, 22-25 September*, 2002, vol. III, pp. 421-424.

[31] C.Kotropoulos and I.Pitas, "A variant of learning vector quantizer based on split-merge statistical tests," in *in Proc. of Lecture Notes in Computer Science:Computer Analysis of Images and Patterns, Springer Verlang, 1993*, 1993.

[32] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, pp. 283–298, 1978.

[33] "Trec video retrieval evaluation," 2003. [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

**Zuzana Černeková** received the Diploma of Master of Science in 1999 from Comenius University, Bratislava, Slovakia.

She has studied informatics with specialization on Mathematics methods of informatics and Computer Graphics. She took Doctor of Natural sciences (RNDr.) in 2000. She was a researcher and lecture assistant at the Department of Computer Graphics and Image Processing, Faculty of Mathematics and Physics, Comenius University. Her research interests lie in the areas of computer graphics, visualization, 3D animations, multimedia, video processing and pattern recognition. She is currently a pre-doc researcher and Ph.D. student at the Department of Informatics, Aristotle University of Thessaloniki, Greece.

Ms. Cernekova is a member of the SCCG organizing committee.

**Ioannis Pitas** (SM'94) received the Dipl. Elect. Eng. in 1980 and the Ph.D. degree in electrical engineering in 1985, both from the University of Thessaloniki, Thessaloniki, Greece.

Since 1994, he has been a Professor at the Department of Informatics, University of Thessaloniki, Greece. His current interests are in the areas of digital image processing, multimedia signal processing, multidimensional signal processing and computer vision. He has published over 450 papers, contributed in 17 books and authored, co-authored, edited, co-edited 7 books in his area of interest. He is the co-author of the books Nonlinear Digital Filters: Principles and Applications (Norwell, MA: Kluwer, 1990) and 3D Image Processing Algorithms (New York: Wiley, 2000), is the author of the books Digital Image Processing Algorithms (Englewood Cliffs, NJ: Prentice Hall, 1993), Digital Image Processing Algorithms and Applications (New York: Wiley, 2000), Digital Image Processing (in Greek, 1999), and is the editor of the book Parallel Algorithms and Architectures for Digital Image Processing, Computer Vision and Neural Networks (New York: Wiley, 1993) and co-editor of the book Nonlinear Model-Based Image/Video Processing and Analysis (New York: Wiley, 2000). He is/was principal investigator/researcher in more than 40 competitive R&D projects and in 11 educational projects, all mostly funded by the European Union.

Dr. Pitas is/was Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEICE, Circuits Systems and Signal Processing, and was co-editor of Multidimensional Systems and Signal Processing. He was Chair of the 1995 IEEE Workshop on Nonlinear Signal and Image Processing (NSIP95), Technical Chair of the 1998 European Signal Processing Conference, and General Chair of IEEE ICIP2001. He was co-chair of the 2003 International workshop on Rich media content production. He was technical co-chair of the 2003 Greek Informatics conference (EPY).

**Christophoros Nikou** (M'02) was born in Thessaloniki, Greece, in 1971. He received the Ph.D. degree in image processing and computer vision in 1999, the DEA degree in optoelectronics and image processing in 1995, both from Louis Pasteur University, Strasbourg, France, and the Diploma degree in electrical engineering from the Aristotle University of Thessaloniki in 1994.

During 2001, he was a Senior Researcher with the Department of Informatics, Aristotle University of Thessaloniki, where he conducted research in image processing in the framework of various European projects. From 2002 to 2004, he was with Compucon S.A., Thessaloniki, Greece, managing research projects in 3-D medical image processing and analysis. Since 2004 he is Lectuer at the Depatment of Computer Science, University of Ioannina, Greece. His research interests mainly include computer vision, pattern recognition, biomedical image processing, image registration and segmentation, deformable models, statistical image processing.

Dr. Nikou is a member of the Technical Chamber of Greece.