Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 947–957

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Video segmentation based on 2D image analysis

Silvio Jamil Ferzoli Guimarães [a,b,*], Michel Couprie [b], Arnaldo de Albuquerque Araújo [a], Neucimar Jerônimo Leite [c]

[a] NPDI/DCC/UFMG, Caixa Postal 702, 30161-970 Belo Horizonte, MG, Brazil
[b] A²SI/ESIEE—Cité Descartes, BP 99, 93162, Noisy le Grand, France
[c] IC/UNICAMP, Caixa Postal 6176, 13083-970 Campinas, SP, Brazil

Abstract

The video segmentation problem consists in the identification of the boundary between consecutive shots. The common approach to solve this problem is based on dissimilarity measures between frames. In this work, the video segmentation problem is transformed into a problem of pattern detection, where each video event is transformed into a different pattern on a 2D image, called visual rhythm, obtained by a specific transformation. In our analysis we use topological and morphological tools to detect cuts. Also, we use discrete line analysis and max tree analysis to detect fade transitions and flashes, respectively. We present a comparative analysis of our method for cut detection with respect to some other methods, which shows the better results of our method.

Keywords: Video segmentation; Visual rhythm; Mathematical morphology; Image segmentation

1. Introduction

The video hierarchical model is usually divided into four levels according to the temporal resolution. The lowest level is the frame. Several frames are grouped into a shot that represents a continuous camera recording. Some shots with a story-telling coherence are grouped into scenes and different scenes constitute a digital video. Amongst the problems related to analysis and indexing of video, the video segmentation can be considered as an essential and first step. The video segmentation problem consists in the identification of the boundary between consecutive shots, called transition. The simplest transition between two consecutive shots is the sharp transition (cut) that is simply a concatenation of these shots. The common approach to cope with the cut detection is based on the use of a dissimilarity measure. Wang et al. (2000) and Del Bimbo (1999) review some of the most popular methods for cut detection, such as pixel-wise comparison, histogram comparison, etc. Unfortunately, the cut detection is complicated by the presence of effects, like gradual transitions, flashes and fast camera and object motions. The simplest gradual transition, the fade, consists in the gradual darkening (lightening) of a shot. Some works for fade detection can be

* Corresponding author.
E-mail addresses: sjamil@dcc.ufmg.br, guimaras@esiee.fr (S.J.F. Guimarães), coupriem@esiee.fr (M. Couprie), arnaldo@dcc.ufmg.br (A. de Albuquerque Araújo), neucimar@ic.unicamp.br (N. Jerônimo Leite).

0167-8655/03/$ - see front matter
PII: S0167-8655(02)00218-0

found in (Zabih et al., 1995; Fernando et al., 1999; Lienhart, 1999). Zabih et al. (1995) proposed an algorithm based on edge detection which is very costly due to the computation of edges for each frame of the sequence. Fernando et al. (1999) and Lienhart (1999) used a statistical approach considering features of the luminance signal. This approach presents high precision on long fades.

Another approach to the video segmentation problem is to transform the video V into a 2D image R, and to apply image processing methods on R to extract the different patterns related to each transition. Informally, each frame is transformed into a vertical line of R, as illustrated in Fig. 1(a). This approach can be found in (Tonomura et al., 1993; Chung et al., 1999; Ngo et al., 1999). Tonomura et al. (1993) defined the X-ray and Y-ray as the result of a video transformation obtained by a linear image transformation in each axis, and an edge analysis was performed to detect cuts. They also cited another video transformation based on the intensity histogram, but it was not well defined and not exploited. Chung et al. (1999) defined the visual rhythm and Ngo et al. (1999) defined the spatio-temporal slice, both are related to the same video transformation and a sub-sampling of each frame, like the principal diagonal sub-sampling (illustrated in Fig. 1(b)). Chung et al. applied statistical measures to detect some patterns, but the number of false detections is very high. Ngo et al. applied Markov models for shot transition detection, but it fails when there is low contrast between textures of consecutive shots.

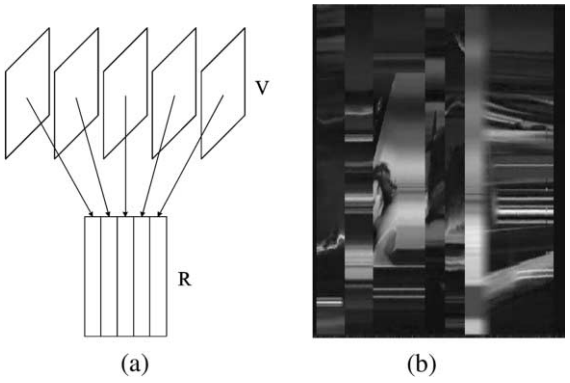

Fig. 1. Video transformation: (a) simplification of the video content by transformation of each frame into a column on R; (b) a real example of the principal diagonal sub-sampling.

We propose, in this work, different specific methods for video segmentation based on analysis of a 2D image, taking advantage of the fact that each video event is represented by a specific pattern in this image. This work is an extension of Guimarães et al. (2001) which used morphological and topological tools to detect cuts by analysis of the visual rhythm by sub-sampling. Here, we introduce the notion of visual rhythm by histogram, and we use it to detect different kinds of transition, mainly fades. On these two variants of visual rhythm, namely visual rhythm by sub-sampling and visual rhythm by histogram, we apply morphological, topological and discrete geometry tools to segment the video without the need of defining a dissimilarity measure between frames. In the general way, the simplicity of implementation, the low processing cost and the high quality of results can be considered as the main contributions of our work. Also, we verified that our methods are more robust than other implemented methods with respect to the tuning of threshold values. The fact that two different video events may be represented by the same visual rhythm pattern can be considered as the main drawback of our method. Fortunately, this problem is not frequent in real cases.

This paper is organized as follows. In Section 2 we define the video transformations, the visual rhythm by sub-sampling and by histogram. In Section 3 we present a methodology for cut detection. In Section 4 we propose a new method for fade detection based on analysis of the visual rhythm by histogram. In Section 5 we present two methods for flash detection. In Section 6 we report on a comparative analysis for cut detection involving our method and some other methods, using four different quality measures. According to these measures, we can verify that our method presents generally the best results. Some conclusions and a summary of future works are given in Section 7.

2. Video transformation

Let $\mathbb{D} \subset \mathbb{Z}^2$, $\mathbb{D} = \{0,\ldots,H-1\} \times \{0,\ldots,W-1\}$, where H and W are the height and the width of

each frame, respectively. A video V, in domain $2D+t$, can be seen as a sequence of frames f_t and can be described by $V=(f_t)_{t\in[0,T-1]}$ where T is the number of frames contained in the video.

2.1. Visual rhythm by sub-sampling

Informally, the visual rhythm by sub-sampling (or simply visual rhythm) is a simplification of the video content represented by a 2D image. This simplification can be obtained by a systematic sampling of points of the video, such as extraction of the diagonal points of each frame.

Definition 1 (*Visual rhythm* Chung et al., 1999 or spatio-temporal slice Ngo et al., 1999). Let $V=(f_t)_{t\in[0,T-1]}$ be an arbitrary video, in domain $2D+t$. The visual rhythm, in domain $1D+t$, is a simplification of the video in which each frame f_t is transformed into a vertical line of the visual rhythm image A, defined by $A(t,z)=f_t(r_x\times z+a, r_y\times z+b)$, where $z\in\{0,\ldots,H_A-1\}$ and $t\in\{0,\ldots,T-1\}$, H_A and T are the height and the width of the visual rhythm, respectively, r_x and r_y are ratios of pixel sampling, a and b are shifts on each frame.

Thus, according to these parameters, different pixel samplings could be considered, for example, if $r_x=r_y=1$ and $a=b=0$ and $H=W$, then we obtain all pixels of the principal diagonal. The choice of the pixel sampling constitutes a problem in the sense that different samplings produce different visual rhythms in which video events (cuts, fades, flashes, etc.) will appear as different patterns. Chung et al. (1999) presented different pixel sampling possibilities with their correspondent visual rhythms. They said that the best results are found when the sampling is based on a diagonal because it contains both horizontal and vertical features. In Fig. 1(b) and 2(a), we show examples of visual rhythm obtained by the principal diagonal sub-sampling.

2.2. Visual rhythm by histogram

To take advantage of the properties of an image histogram, such as global information, invariance to rotation and translation, we define here a new video transformation, called visual rhythm by histogram (VRH), where the video is transformed into a 2D image containing histogram frame information.

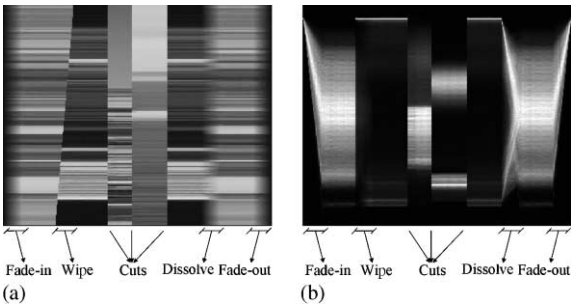

Fig. 2. Visual rhythm by sub-sampling (a) and by histogram (b) computed from the same video.

Definition 2 (*Visual rhythm by histogram* (*VRH*)). Let $V=(f_t)_{t\in[0,T-1]}$ be an arbitrary video, in domain $2D+t$ and $(H_{f_t})_{t\in[0,T-1]}$ the sequence of histograms computed from all frames of V. The visual rhythm by histogram B is a 2D representation of all frame histograms where each vertical line represents a frame histogram, B is defined by $B(t,z)=H_{f_t}(z)$, where $t\in[0,T-1]$ and $z\in[0,L-1]$, T is the number of frames and L the number of histogram bins.

The main problem of this representation is associated with the transformation of all histogram values into grayscale values. The simplest way for histogram representation is obtained by normalization of each histogram independently. Another possibility is the truncation of the values greater than $G-1$, where G is the number of grayscales. Due to the loss of information in the representation by truncation, we chose the histogram normalization for representing the visual rhythm by histogram. Furthermore, this normalization produces a filtering effect on the weakest histogram values, which mainly occurs when most pixels are grouped in only few bins. In fact, this kind of filtering is desirable for our application. In Fig. 2(b), we illustrate an example of visual rhythm by histogram where each value of the histogram is in the range $[0,255]$.

3. Cut detection

All cuts appear as "vertical lines" on the visual rhythm, as illustrated in Fig. 2. To facilitate the description of our method, the visual rhythm by sub-sampling or by histogram will be denoted by R, and each step will be separately described.

Filtering. In this step, we reduce noise on R using mathematical morphology filters (see Serra, 1988; Soille, 1999). The filtered visual rhythm is denoted by R_F. Here, we apply an opening (closing) by reconstruction to eliminate the small light (black) components. These morphological filters have the interesting property to preserve the sharp contours of the image.

Horizontal gradient. The aim of this step is to detect the locations where horizontal grayscale discontinuities occur in the filtered visual rhythm. These locations, when vertically aligned, can represent a cut. So, we calculate the norm of the horizontal gradient ∇_h of the filtered image by $|\nabla_h R_F(t,z)| = |R_F(t,z) - R_F(t-1,z)|$.

Thinning. This transformation is used here to simplify the peak detections (see Bertrand et al., 1997). Intuitively, a horizontal transition between two consecutive regions corresponds to a "peak" in the horizontal gradient of each line. In the case of a cut, the maximum of this peak is generally reduced to only one pixel but for gradual video transitions, for example, the maximum of a peak may consist of several neighboring pixels. In such cases, a simple maximum detection would result in multiple responses for a single transition. This is why we introduce the thinning step, with the aim of reducing every peak to a one-pixel-thin maximum.

Detection of the maximum points. After the thinning operation, we have a new image I_T where each horizontal peak is represented by a point, called maximum point. A point x in 1D image g is maximum if its two neighbors have values strictly smaller than $g(x)$. So, we must find all maximum points of the image I_T to identify the center points of the transitions. The image of maximum points is denoted by M.

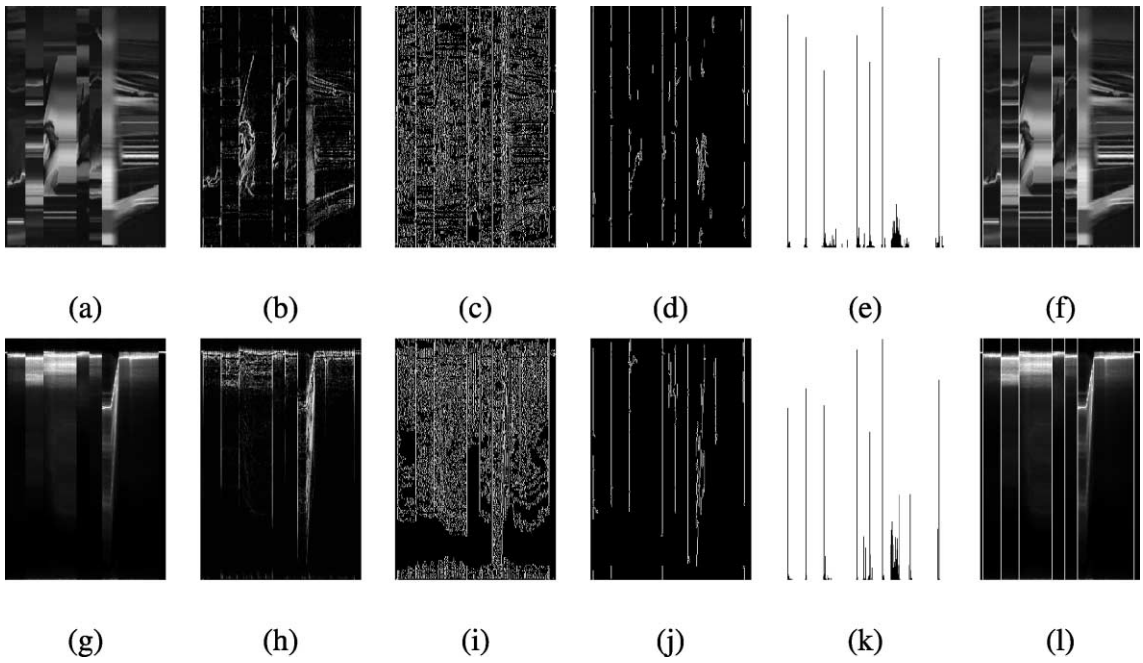

Fig. 3. Cut detection from a visual rhythm by sub-sampling (a)–(f) and by histogram (g)–(l): visual rhythms (a) and (g); thinning (b) and (h); maximum points (c) and (i); maxima filtering (d) and (j); normalized number of maximum points in the range $[0, 255]$ (e) and (k); detected cuts (white bars) superimposed on the visual rhythms (f) and (l).

Maxima image filtering. As illustrated in Figs. 3(c) and (i), the locations of the cuts appear as "vertical lines" in image M, embedded in irrelevant components (noise) which can be reduced by a morphological filtering. This filtering is an opening by reconstruction with a vertical structuring element of size $\lambda = 7$, defined empirically. The filtered maxima image is denoted by M_F.

Detection of cuts. From the filtered maxima image M_F, we create a 1D image P where each point t has a value $P(t)$ representing the number of maximum points of the vertical line t on M_F. Finally, when the value of the point $P(t)$ is greater than or equal to a threshold T, then a cut is detected at time t.

In Fig. 3, we illustrate the results of the main steps of our method when applied to a visual rhythm by sub-sampling (or histogram) obtained from the same real video.

4. Fade detection

The fade process is characterized by a progressive darkening of a shot until the last frame becomes completely black, or inversely (Del Bimbo (1999)). A more general definition is given by Lienhart (1999) where the black frame is replaced by a monochrome frame. A natural way for detecting the fades is to study the behavior of luminosity changing (Lienhart, 1999; Fernando et al., 1999) by modeling the process of fade creation with mathematical equations. Unfortunately, this process fails due to noise, when the extremal frames are not completely monochrome, and mainly, when we have short fades. If we study the behavior of image histograms in a fade, respecting the time coherence, inclined edges can be found on the visual rhythm by histogram (VRH), due to the shrinking (or expansion) of the histogram width during the fade, that corresponds to the number of non-zero bins. So, we propose a methodology for fade detection based on the VRH analysis which consists in the detection of inclined edges.

Fig. 4 illustrates an example of the fade detection by VRH analysis. In the first step, a thresholding is applied to the VRH (Fig. 4(b)). A gradient operator allows to extract the contours of the segmented image (Fig. 4(c)). The gradient operator used here is the external morphological gradient. Finally, an algorithm for line approximation described in (Dunham, 1986) is used. Afterwards, an analysis of size and inclination is realized (Fig. 4(d)), in this example, the permitted inclination is between $-44°$ and $44°$ and the minimal size of the fade edges is 100 pixels. These values were empirically defined.

5. Flash detection

The flash presence is very common in digital videos, mainly in television journal videos. When a flash occurs, an increase of the luminosity in a few frames is produced, as illustrated in Fig. 5. When we calculate a dissimilarity measure, we can see that this measure is very high in the frames affected by a flash. In fact, the presence of flashes often perturbs the cut detection. In this work, we propose two methods for flash detection without taking into account dissimilarity measures. The first one is a variant of the proposed method for cut detection and the second one considers a filtering of the max tree calculated from statistical measures of each frame (or frame sub-samplings).

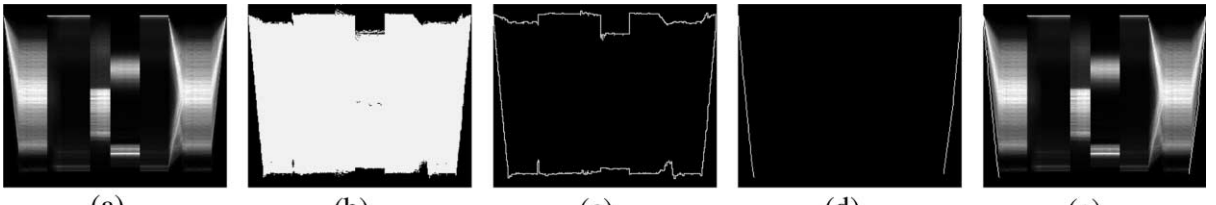

Fig. 4. Fade detection process: (a) visual rhythm by histogram, (b) thresholding, (c) gradient, (d) line filtering and (e) result superimposed on the visual rhythm by histogram (a).

Fig. 5. Flash video detection: (a) some frames of a sequence with the flash presence; (b) visual rhythm by sub-sampling; (c) detected flash.

5.1. Filtering by top-hat

On the visual rhythm, we can observe that the flashes are transformed into thin light vertical lines, as showed in Fig. 5(a). So, these lines can be extracted through a white top-hat by reconstruction operation. The white top-hat by reconstruction is a mathematical morphology operator defined as the difference between the original image g and the opening by reconstruction of g (Serra, 1988; Soille, 1999). Informally, this operator detects light regions according to shape and size specifications of the structuring element. Our first method for flash detection can be described as follows: (1) calculate the visual rhythm by sub-sampling using the principal diagonal pixel sampling; (2) apply the white top-hat by reconstruction with a square structuring element of size $\lambda = 5$. This size is associated with the potential duration of a flash; (3) apply a 1D thinning to each line of the above image; (4) find the maximum points; (5) apply an opening by reconstruction with vertical structuring element of size $\lambda = 7$, defined empirically; (6) calculate the number of maximum points of each column of the above image. As the method for cut detection, this method detects the center of the regions of interest, in this case, regions with peak luminosity. Thus, we can have false detections in regions of high luminosity changing that is not due to a flash. Usually, this method produces good results when the flash appears in the middle of the shot.

5.2. Max tree filtering

Usually, the frames affected by a flash are visually similar to their neighbors but with a higher luminosity. The analysis of flash presence can be given by the computation of some statistical measures like mean or median, because the frames affected by a flash present higher mean and median values with respect to their neighbors. From the computation of these statistical measures for all frames of the video, we can create a 1D image from the visual rhythm, or preferably from the original video data. From this 1D image, we need to find the "peaks" whose "height" is greater than a certain value h, and a "basis area" less than or equal to a value S that corresponds to the duration of the flash. In this work, we consider that the maximum flash duration is 5 frames, i.e., $S = 5$. The parameter h influences the sensitivity of the method and has a role similar to the threshold in Section 3. The notions of peak, height and basis area can be precisely defined thanks to a data structure called *max-tree* (Salembier et al., 1998) (refer to this paper for more details).

6. Experimental analysis

In this section, we show the experimental results for cut, fade and flash detection. Nowadays, our video database contains 450 videos, but we use only 32 videos for cut detection experiments, 46 videos for fade detection experiments and 10 videos for flash detection experiments. The choice of the videos was associated with the presence of the different events, such as cut, dissolve, wipe, flash, zoom-in, zoom-out, pan, tilt, object motion, camera motion, computer effects. In Table 1, we show some features of the chosen videos. To compare the different methods, we define some quality measures as follows.

Table 1
Features of the videos which were selected for the experiments

	Videos	Cuts	Fades	Flashes	Frames	Frames/event (mean)
Corpus 1 (Cut)	32	778	15	14	29 933	0
Corpus 2 (Fade)	46	–	59	–	41 881	14.2
Corpus 3 (Flash)	10	–	–	23	8392	3

6.1. Quality measures

The first step to realize a comparison analysis is to manually identify the video events that will be considered as reference for classifying the type of detection (e.g., correct, false or miss). We denote by *#Event* the number of events (cuts, fades, flashes, etc.), by *#Correct* the number of events correctly detected, by *#False* the number of detected frames that do not represent an event and by *#Miss* the number of the events that were not detected, defined by $\#Miss = \#Event - \#Correct$. From these numbers we can define two basic quality measures.

Definition 3 (*Recall and error rates*). The recall and error rates represent the ratios of correct and false detection, respectively, and are given by

$$\alpha = \frac{\#Correct}{\#Event} \quad \text{(recall)}, \quad \beta = \frac{\#False}{\#Event} \quad \text{(error)}.$$

Let τ be the threshold used for event detection normalized in the range $[0,1]$. If we consider that for each threshold τ we obtain different values for α and β, we can represent these relations as functions $\alpha(\tau)$ and $\beta(\tau)$, respectively. A new measure can be created to relate ranges in which α and β are adequate, according to the ratios of miss and of false detection that are permitted.

Definition 4 (*Robustness*). Let P_m and P_f be the ratio of miss and false detections that are permitted. The robustness μ measures the width of the interval where the recall and error rates have values smaller than $(1 - P_m)$ and P_f, respectively. This measure is in the range $[0,1]$ and is given by $\mu(P_m, P_f) = \alpha^{-1}(1 - P_m) - \beta^{-1}(P_f)$, where α^{-1} and β^{-1} are the inverses of the functions $\alpha(\tau)$ and $\beta(\tau)$, respectively.

In Fig. 6, we illustrate the robustness measure obtained from functions $\alpha(\tau)$ and $\beta(\tau)$. To follow, we define two other measures, E_m and R_f, that are associated with the absence of miss and false detection, respectively.

Definition 5 ("*Missless error*"). The missless error E_m is associated with the ratio of false detection when we have results without miss (a small ratio of miss P_m can be permitted, like 0.03). The missless error is given by

$$E_m(P_m) = \beta(\max\{\tau = \alpha^{-1}(q) | 1 - q \leqslant P_m\}).$$

Definition 6 ("*Falseless recall*"). The falseless recall R_f is associated with the ratio of correct detection when we have results without false detection (a small number of false detection P_f can be permitted, like 0.01). The falseless recall is given by

$$R_f(P_f) = \alpha(\min\{\tau = \beta^{-1}(p) | p \leqslant P_f\}).$$

When we use methods for event detection, we expect that the recall is highest with a smallest error

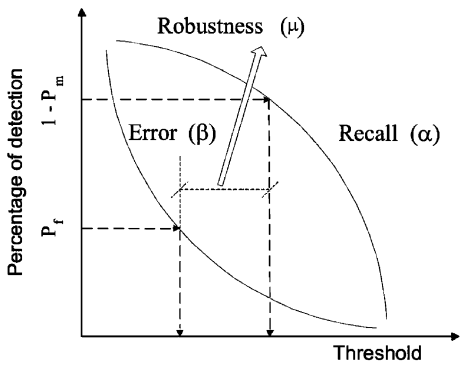

Fig. 6. Robustness (μ) measure.

rate. To find a compromise between these two requirements, we must define a "reward function" combining $\alpha(\tau)$ and $\beta(\tau)$. Since high values of α and low values of β have to be rewarded, the function $\alpha(\tau) \times (1 - \beta(\tau))$ is a natural choice.

Definition 7 (*Gamma*). The gamma measure γ represents the maximal value of the reward function defined above for all possible values of τ:

$$\gamma = \max\{\alpha(\tau) \times (1 - \beta(\tau)) | \tau \in [0, 1]\}.$$

The quality of the results is related to the values of the measures above defined. The highest values of robustness, falseless recall and gamma measure represent the best results of a method. The lowest values of missless error represent the best results of a method. In the next sections, we describe the experiments for cut, fade and flash detection.

6.2. Experiments for cut detection

In these experiments, we implemented three methods described in literature: a variant of pixel-wise comparison, a histogram intersection and a statistical technique based on visual rhythm. We chose these methods due to their simplicity and their effectiveness according to Demarty (2000), Del Bimbo et al. (2000) and Chung et al. (1999), respectively. We also implemented the proposed method with some variants. To follow, we describe all experiments applied to the corpus 1 followed by a global analysis of their results.

Experiment 1. This experiment uses the difference between pixels according to Demarty (2000) as the dissimilarity measure. A 1D signal is created from the dissimilarity values calculated in each frame of the video. Then a mathematical morphology operator, called inf top-hat, is applied to this signal, and finally, a threshold is used to detect the cuts.

Experiment 2. This experiment uses the histogram intersection according to Del Bimbo et al. (2000) as the dissimilarity measure. If the dissimilarity value is greater than a threshold, then a cut is detected. With the aim of improving the results, a subdivision of each frame is realized.

Experiment 3. This experiment uses the visual rhythm for cut detection based on a statistical method as described in (Chung et al., 1999). Here, the parameters are different from those used in the other mentioned methods, particularly the threshold. While in this method the threshold is locally adaptive and related to a parameter that varies from 1 to 10, in the other methods the threshold is fixed and global.

Experiment 4. In this experiment, we compute a 1D image associated with the mean of the difference between pixels in consecutive frames. We apply the following algorithm on this image: (i) apply a white top-hat by reconstruction with a flat structuring element of size 3; (ii) apply a thinning; and (iii) apply a thresholding.

Experiment 5. In this variant of our method introduced in Section 3, instead of applying the summation of the number of maximum points to each vertical line, we use the filtered maxima M_F image as a mask to select the grayscale values associated with all maximum points. Afterwards, we find the mean of these grayscale values in each vertical line. Then, a thresholding is applied to these results.

Experiments 6 *and* 7. These experiments are related to the method defined in Section 3 concerning the analysis of the visual rhythm by subsampling and by histogram, respectively.

In Fig. 7, we show graphically the experimental results for each experiment previously described. The graphics relate the threshold (rate with respect to the maximum value obtained from each experiment) to recall and error rates. From the curves shown in these graphics, it is possible to find the robustness, missless error rate, falseless recall rate and gamma measures, that are outlined in Table 2(a). From these experiments, we can verify that the proposed method based on visual rhythm analysis generally produces the best results, mainly according to the robustness and the missless error rate. The good value of the robustness means that the proposed method is not very sensitive to small variations around an "optimal value" of the main parameter. Another interesting aspect of our method concerns the missless error rate since, in general, we want results without miss and with a

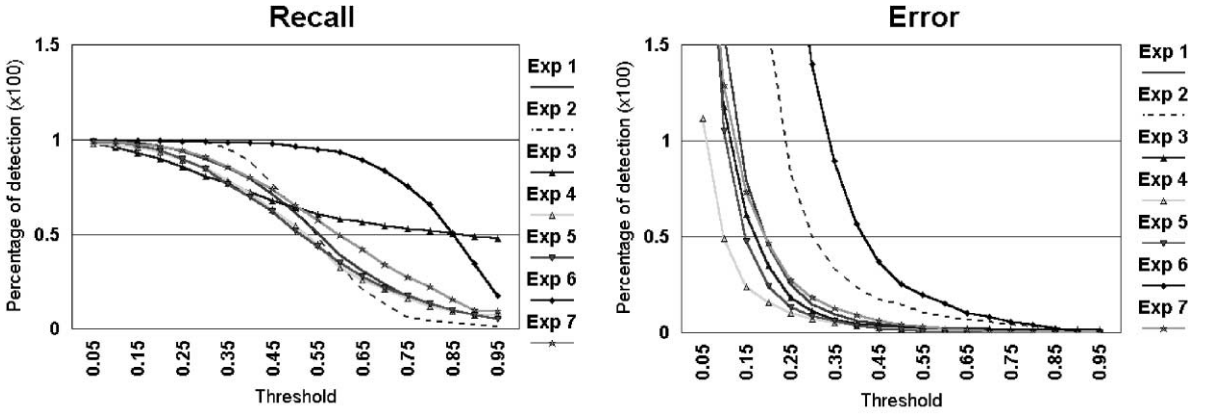

Fig. 7. Experimental results.

Table 2
Quality measures

	μ	E_m	R_f	γ
(a)				
Experiment 1	0.01	0.80	0.10	0.77
Experiment 2	0.00	0.51	0.00	0.68
Experiment 3	0.00	1.20	**0.51**	0.72
Experiment 4	0.06	0.49	0.21	**0.80**
Experiment 5	0.01	0.48	0.44	0.78
Experiment 6	**0.11**	**0.37**	0.35	**0.80**
Experiment 7	**0.11**	0.46	0.42	0.75
(b)				
Top-hat	0.05	**0.61**	0.26	0.56
Max tree	**0.11**	0.67	**0.43**	**0.69**

(c)	Fades	Recall	Error
All fades	59	0.95 (56)	0.30 (18)
Long fades	20	0.95 (19)	0.05 (1)

Cut detection (a) $\mu(0.10, 0.30)$, $E_m(0.03)$, $R_f(0.01)$ and γ; flash detection (b) $\mu(0.40, 0.30)$, $E_m(0.05)$, $R_f(0.01)$ and γ; fade detection (c). The best values are highlighted.

smallest possible ratio of false detections, so that we can eliminate them posteriorly. Indeed, a post-processing is essential to increase the quality of the results because many false detections are due to the presence of effects like flash, pan, zoom. Also, we can observe that the processing time for experiments based on visual rhythm by sub-sampling is significantly lower than for the experiments applied directly to the video. We notice that the results of the proposed method based on analysis of the visual rhythm by sub-sampling are better with respect to analysis of the visual rhythm by histogram.

6.3. Experiments for fade detection

In these experiments, we apply the method described in Section 4 to the corpus 2. An important aspect of this method is its robustness with respect to the fade size and to the quality of the fade frames, that is, our method can detect fades even when the extremal frames are not completely monochrome. On the other hand, fades with a small luminosity difference between extremal frames are not detected. In these experiments, we distinguished the long fades (duration $\geqslant 15$ frames) which are, in general, the easiest ones to detect. The quality measures are outlined in Table 2(b).

6.4. Experiments for flash detection

In these experiments, we apply the methods described in Section 5.1 and in Section 5.2 to the corpus 3. In Fig. 8, we illustrate some experimental results. The quality measures of the filtering of the max tree and the top-hat filtering are outlined in Table 2(c).

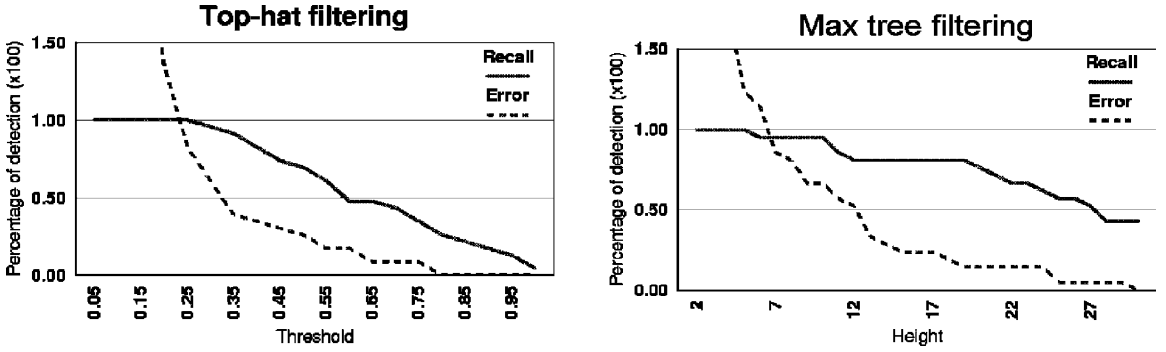

Fig. 8. Experimental results for flash detection.

7. Discussions and conclusions

In this work, we transform the video segmentation problem into a 2D image segmentation problem. Methods for cut, fade and flash detection are proposed. The main contribution of our work is the application of operators of mathematical morphology and digital topology to solve a problem of video segmentation. The exploitation of the visual rhythm by histogram, mainly for fade detection, also represents an original contribution of this work. We verified that we can identify short fades with a small ratio of false detections. For flash detection, the extraction of peaks by max tree analysis allows a detection of flashes in all positions of the shot.

The effectiveness of the results is associated with the choice of good parameters. Two types of parameters can be distinguished: fixed and variable. The fixed parameters, like the size of structuring elements, can be pre-determined for all applications. The use of a variable parameter, in our case, the threshold value, is interesting and sometimes necessary, because it plays an important role in the segmentation process where the user can adequate it according to the nature of data and the type of application. Also, the tuning of this parameter allows to find a compromise between over-segmentation and under-segmentation.

To realize a comparative analysis between different methods for event detection, we defined four quality measures: robustness, missless error, falseless recall and the gamma measure. According to these quality measures, we verified that the proposed method for cut detection has the best values of robustness, missless error and gamma measure, when compared experimentally to the other methods. We also computed the quality measures for flash detection, and according to these measures, the method based on max tree analysis generally presents the best results.

From this work, we observed that the visual rhythm by sub-sampling and by histogram present an adequate simplification of the video content, which can constitute the basis for future developments such as: (i) identify some other video events, like pan and zoom from the detection of their correspondent patterns; (ii) modify the proposed method for cut detection to detect gradual video transitions, using the multi-scale morphological gradient (Soille, 1999) to compute the horizontal gradient. We can also remark that considering the video sequence as 3D images, we could apply an extension of our method directly to the video data. We have to verify if the computation effort is rewarded by a better segmentation quality.

Acknowledgements

The authors are grateful to FAPEMIG, CAPES/COFECUB, CNPq and the SIAM DCC/PRONEX Project for the financial support of this work.

References

Bertrand, G., Everat, J.-C., Couprie, M., 1997. Image segmentation through operators based upon topology. J. Electron. Imaging 6, 395–405.

Chung, M.G., Lee, J., Kim, H., Song, S.M.-H., Kim, W.M., 1999. Automatic video segmentation based on spatio-temporal features. Korea Telecom J. 4 (1), 4–14.

Del Bimbo, A., 1999. Visual Information Retrieval. Morgan Kaufmann, Los Altos, CA.

Del Bimbo, A., Pala, P., Tanganelli, L., 2000. Retrieval of commercials based on dynamics of color flows. J. Visual Lang. Comput. 11, 273–285.

Demarty, C.-H., 2000. Segmentation et Structuration d'un Document Vidéo pour la Caractérisation et l'Indexation de son Contenu Sémantiqué. Ph.D. Thesis, École Nationale Supérieure des Mines de Paris.

Dunham, J.G., 1986. Optimum uniform piecewise linear approximation of planar curves. IEEE Trans. PAMI 8 (1), 67–75.

Fernando, W.A.C., Canagarajah, C.N., Bull, D.R., 1999. Fade and dissolve detection in uncompressed and compressed video sequences. In: Proc. of the IEEE ICIP, pp. 299–303.

Guimarães, S.J.F., Couprie, M., Leite, N.J., and Araújo, A.A., 2001. A method for cut detection based on visual rhythm. In: Proc. of the XIV Brazilian Symposium on Computer Graphics and Image Processing—SIBGRAPI, Brazil. IEEE Computer Society Press, pp. 297–304, ISBN 0769513301.

Lienhart, R., 1999. Comparison of automatic shot boundary detection algorithms. In: SPIE Image and Video Processing VII, pp. 25–30.

Ngo, C.W., Pong, T.C., Chin, R.T., 1999. Detection of gradual transitions through temporal slice analysis. In: Proc. of the IEEE CVPR, pp. 36–41.

Salembier, P., Oliveras, A., Garrido, L., 1998. Antiextensive connected operators for image and sequence processing. IEEE Trans. Image Process. 7 (4), 555–570.

Serra, J., 1988. In: Image Analysis and Mathematical Morphology: Theoretical Advances, Vol. 2. Academic Press, New York.

Soille, P., 1999. Morphological Image Analysis. Springer, Berlin.

Tonomura, Y., Akutsu, A., Otsuji, K., Sadakata, T., 1993. Videomap and videospaceicon: tools for anatomizing video content. In: ACM Interchi, pp. 131–136.

Wang, Y., Liu, Z., Huang, J.-C., 2000. Multimedia content analysis. IEEE Signal Process. Mag., 12–36.

Zabih, R., Miller, J., Mai, K., 1995. Feature-based algorithms for detecting and classifying scene breaks. In: ACM ICMCS, USA, pp. 12–13.