# A method for cut detection based on visual rhythm

SILVIO JAMIL FERZOLI GUIMARÃES[†§], MICHEL COUPRIE[§],
NEUCIMAR JERÔNIMO LEITE[‡], ARNALDO DE ALBUQUERQUE ARAÚJO[†]

[†]NPDI/DCC - Universidade Federal de Minas Gerais, Caixa Postal 702, 30161-970, Belo Horizonte, MG, Brasil
{sjamil,arnaldo}@dcc.ufmg.br
[§]A$^2$SI/ESIEE - Cité Descartes, BP 99, 93162, Noisy le Grand, France
{guimaras,coupriem}@esiee.fr
[‡]Instituto de Computação - UNICAMP Caixa Postal 6176, 13083-970, Campinas, SP, Brasil
neucimar@dcc.unicamp.br

**Abstract.** The visual rhythm is a simplification of the video content represented by a 2D image. In this work, the video segmentation problem is transformed into a problem of pattern detection, where each video effect is transformed into a different pattern on the visual rhythm. To detect sharp video transitions (cuts) we use topological and morphological tools instead of using a dissimilarity measure. Thus, we propose a method to detect sharp video transitions between two consecutive shots. We present a comparative analysis of our method with respect to some other methods. We also propose a variant of this method to detect the position of flashes in a video.

## 1 Introduction

The video segmentation problem can be considered as a problem of dissimilarity between images (or frames). Usually, the common approach to cope with this problem is based on the use of a dissimilarity measure which allows to identify the boundary between consecutive shots. The simplest transitions between two consecutive shots are sharp and gradual transitions [1]. A sharp transition (cut) is simply a concatenation of two consecutive shots. When there is a gradual transition between two shots, new frames are created from these shots [1]. In literature, we can find different types of dissimilarity measures used for video segmentation, such as, pixel-wise comparison, histogram-wise comparison, etc. If two frames belong to the same shot, then their dissimilarity measure should be small, and if two frames belong to different shots, this measure should be high, but in the presence of different effects, like zoom, pan, tilt, flash, this measure can be affected. So, the choice of a good measure is essential for the quality of the segmentation results.

Another approach to the video segmentation problem is to transform the video into a 2D image, and to apply methods of image processing to extract the different patterns related to each transition. This approach can be found in [2, 3], where the transformed image is called visual rhythm [2] or spatio-temporal slice [3]. Informally, the visual rhythm is a simplification of the video content represented by a 2D image. This simplification can be obtained from a systematic sampling of points of the video, such as, extraction of the diagonal points of each frame. In Fig. 1, we illustrate an example of point sampling from a video. So, the video segmentation problem is transformed into an image segmentation problem. In this work, we propose a method for cut detection based on analysis of visual rhythm. We also propose a variant of this method for flash detection. According to the comparative analysis involving our method and some other methods, we can verify that the proposed method for cut detection presents the best results.

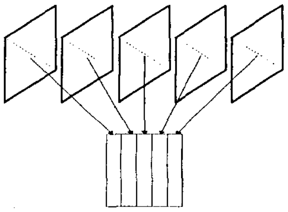


Figure 1: Visual rhythm.

This paper is organized as follows. In Sec. 2, we describe the basic method used in this work, the visual rhythm. In Sec. 3, we describe some related works. In Sec. 4, we propose a method for cut detection based on the analysis of visual rhythm. In Sec. 5, we show a variant of our method for flash detection. In Sec. 6, we have done a comparative analysis involving our method and some other methods using quality measures. Finally, some conclusions and a summary of future works are given in Sec. 7.

## 2 Visual rhythm

Let $D \subset Z^2$, $D = \{0, ..., M-1\} \times \{0, ..., N-1\}$, where $M$ and $N$ are the width and the height of each frame, respectively.

**Definition 2.1 (Frame)** A frame $f_t$ is a function from $D$ to $Z$ where for each spatial position $(x, y)$ in $D$, $f_t(x, y)$ represents the grayscale value of the pixel $(x, y)$.

1530-1834/01 $10.00 © 2001 IEEE

**Definition 2.2 (Video)** A video V, in domain $2D + t$, can be seen as a sequence of frames $f_t$ and can be described by

$$V = (f_t)_{t \in [0, T-1]} \tag{1}$$

where $T$ is the number of frames contained in the video.

When we work directly on the video, we have to cope with two main problems: the processing time and the choice of a dissimilarity measure. Looking for reducing the processing time and using tools for 2D image segmentation instead of a dissimilarity measure (as we will see in Sec. 4), we transform the video into a two-dimensional image, called visual rhythm [2, 3].

**Definition 2.3 (Visual rhythm (Spatio-temporal slice))** Let $V = (f_t)_{t \in [0, T-1]}$ be an arbitrary video, in domain $2D + t$. The visual rhythm $\vartheta$, in domain $1D + t$, is a simplification of the video where each frame $f_t$ is transformed into a vertical line on the visual rhythm that is defined by

$$\vartheta(t, z) = f_t(r_x * z + a, r_y * z + b) \tag{2}$$

where $z \in \{0, ..., M_\vartheta - 1\}$ and $t \in \{0, ..., N_\vartheta - 1\}$, $M_\vartheta$ and $N_\vartheta$ 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 $M = N$ then we obtain all pixels of the principal diagonal. If $r_x = -1$ and $r_y = 1$ and $a = M$ and $b = 0$ and $M = N$ then we obtain all pixels of the secondary diagonal. If $r_x = 1$ and $r_y = 0$ and $a = 0$ and $b = N/2$ then we obtain all pixels of a central horizontal line. If $r_x = 0$ and $r_y = 1$ and $a = M/2$ and $b = 0$ and then we obtain all pixels of a central vertical line.

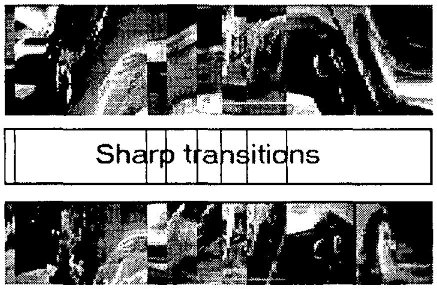


Figure 2: Visual rhythm obtained from a real video using different pixel samplings: principal diagonal (top) and central vertical line (bottom). The temporal positions of sharp video transitions are indicated in the middle image.

The choice of the pixel sampling is a problem because different samplings produce different visual rhythms with different patterns. [2] presents some pixel sampling with their correspondent visual rhythm, and it is said that the best results are found when the sampling based on a diagonal is used because it contains horizontal and vertical features. In Fig. 2, we illustrate two visual rhythms obtained from a same video but with different pixel samplings, in these cases we use the principal diagonal (Fig. 2-top) and central vertical axis (Fig. 2-bottom) sampling. We can observe that there are "vertical lines" in Fig. 2-bottom that do not correspond to sharp video transition but all sharp video transitions correspond to "vertical lines" on the visual rhythm.

## 3 Related works

In literature, we find different approaches for cut detection, amongst them, those that are applied directly to the video and those that are applied to a simplification of the video, called visual rhythm. In [4], we can find some methods for cut detection.

### 3.1 Methods applied to the video

These methods represent the most common approach for cut detection and are associated with dissimilarity measures. In general, dissimilarity measures (calculated between each pair of consecutive frames) are compared to a threshold to detect a transition, but the choice of a good threshold represents a problem, because the result of the video segmentation is highly dependent on the threshold value.

In [5], it is proposed a methodology for cut detection considering the mean of the pixel difference between two consecutive frames as the dissimilarity measure. Afterwards, a morphological filter is applied to the one-dimensional signal (signal computed by the dissimilarity measure). And finally, a thresholding with value 20% of the maximum value of this signal is applied. Another approach is to consider the histogram intersection [4] as dissimilarity measure. In theory, histograms of frames into the same shot would be similar, that is, their dissimilarity measure would be small.

### 3.2 Visual rhythm-based

On the visual rhythm $\vartheta$ obtained from the principal diagonal sampling, the cuts correspond to horizontal intensity discontinuities that are vertically aligned. These discontinuities may be easily observed on Fig. 2. In [2] is defined a statistical approach based on visual rhythm for cut detection. This approach considers the local mean and variance of the horizontal gradient. An adaptive thresholding is applied to detect a sharp video transition. In [3], we can find another method from visual rhythm based on concepts of Markov model for image segmentation.

## 4 A method for cut detection

Usually, the shot detection is the first step to automatically segment a video and it is associated with the detection of sharp and gradual transitions between two different shots [1]. In this work we consider only the sharp transition that is simply a concatenation of two consecutive shots.

With the aim of realizing a video segmentation without defining a dissimilarity measure, we can use a simplification of the video content, the visual rhythm, where the video segmentation problem, in domain $2D + t$, is transformed into a problem of pattern detection, in domain $1D + t$. So, we can apply methods of 2D image processing to identify different patterns on the visual rhythm because each video effect corresponds a pattern in this image, for example, each sharp video transition is transformed into "vertical lines" on the visual rhythm. Unfortunately, this correspondence is not one-to-one relation, i. e., a sharp video transition corresponds to a vertical line, but a vertical line is not necessarily a sharp video transition. This problem can be resolved by considering visual rhythms obtained from different pixel samplings. Afterwards, a simple intersection operation between these results may be used to correctly identify the sharp video transitions.

Fortunately, in general, we can use only a visual rhythm obtained from principal diagonal sampling because this problem rarely occurs in practice. Furthermore, this visual rhythm represents the best simplification of the video content, according to [2]. To follow, we will define a method for cut detection based on visual rhythm.

### 4.1 Steps of our method

Let V be an arbitrary video as defined in Sec. 2. To facilitate the description of our method, we will describe each step separately.

**Step 0. Visual rhythm creation** In this work, we use a principal diagonal pixel sampling, as described in Sec. 2, to create the visual rhythm $\vartheta$ from the video V.

**Step 1. Visual rhythm filtering** In this step, we eliminate the noise of the visual rhythm using mathematical morphology filters. The filtered image is denoted by $\vartheta_F$. We apply an opening (closing) by reconstruction to eliminate the small light (black) components. The readers are encouraged to read [6, 7] for more details about mathematical morphology. We choose this filtering method because it preserves the sharp contours of the image.

**Step 2. Horizontal gradient calculation** The aim of this step is to detect the horizontal boundary between two consecutive regions. This boundary (sharp contour) when vertically aligned can represent sharp video transition. So, we calculate the norm of the horizontal gradient $\nabla_h$ of the filtered image by

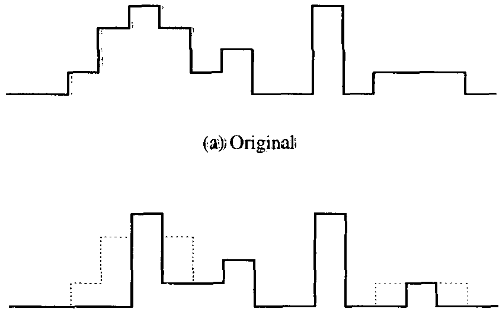
$$|\nabla_h \vartheta_F(t, z)| = |\vartheta_F(t, z) - \vartheta_F(t-1, z)| \qquad (3)$$

Other derivative operators could be considered here, we will discuss this point in Sec. 7.

**Step 3. Thinning operation** 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 in case of a gradual video transition 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 and thus to simplify peak detection.

Let us consider a point $x$ in a 1D image (or signal) $g$. We say that a point $x$ is destructible for $g$, if one neighbor of $x$ has a value greater or equal to $g(x)$ and the other neighbor has a value strictly smaller than $g(x)$. The thinning procedure consists in repeating the following steps until stability: i) select a destructible point $x$; ii) lower the value of $x$ downto the value of its lowest neighbor.

Selection of destructible points must be done in increasing order of value, so that each point is modified at most once. Points having the same value are scheduled with a fifo policy which guarantees that, in case of large flat maxima, the thinned signal is "well centered" with respect to the original one. This procedure is in fact a particular case, in 1D domain, of a topological operator introduced in [8]. Topological operators have as aim to simplify the image maintaining the topology. [8] presents operators for image segmentation based upon topology which generalizes to 2D grayscale images the notions of binary digital topology [9].


(a) Original

(b) Result

Figure 3: Example of thinning from 1D image. Dotted line, in (b), represents the original image.

This operator is applied to all horizontal lines of the gradient of the filtered visual rhythm, producing a new image $\mathcal{I}_\mathcal{T}$. In Fig. 3, we illustrate the thinning of a 1D image.

**Step 4. Detection of the maxima points** After the thinning operation, we have a new image $\mathcal{I}_\mathcal{T}$ with the horizontal peaks being 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 maxima points on the image $\mathcal{I}_\mathcal{T}$ to identify the center points of the transitions. This operation produces a new binary image $\mathcal{M}$ that is defined by

$$\mathcal{M}(t, z) = \begin{cases} 1, & \text{if } \mathcal{I}_\mathcal{T}(t, z) > \max(\mathcal{I}_\mathcal{T}(t-1, z), \mathcal{I}_\mathcal{T}(t+1, z)) \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

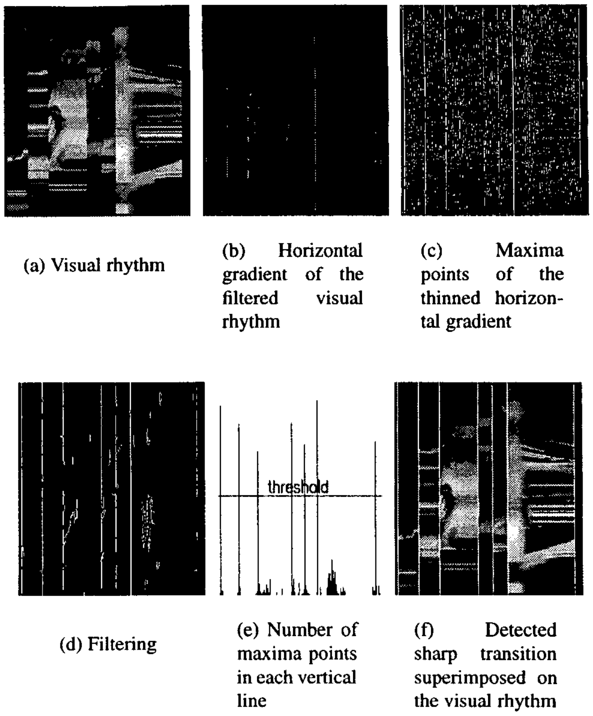
**Step 5. Maxima point filtering** If we observe on the image $\mathcal{M}$, the location of the sharp video transition is represented by vertical lines on $\mathcal{M}$. Unfortunately, irrelevant components (noise) are also present in this image $\mathcal{M}$, and considering that only the relevant vertical components are desired, we can use a morphological filter to eliminate the noise. This filter is an opening by reconstruction with a vertical structuring element of size $\lambda = 7$, defined empirically. The filtered maxima image is denoted by $\mathcal{M}_F$.

**Step 6. Calculation of the number of maxima points.** From the filtered maxima image $\mathcal{M}_F$, we create a 1D image $\mathcal{N}$ where each point $t$ has a value $\mathcal{N}(t)$ which represents the number of maxima points of the vertical line $t$ on $\mathcal{M}_F$. Thus, this 1D image $\mathcal{N}$ is given by

$$\mathcal{N}(t) = \sum_{z=0}^{M_\vartheta - 1} \mathcal{M}_F(t,z) \tag{5}$$

**Step 7. Detection of the sharp transition** Finally, we can detect the sharp video transition from the one-dimensional image $\mathcal{N}$ if we compare the values of each point to a threshold, i. e., when the value of the point $\mathcal{N}(t)$ is greater or equal to a threshold $T$, then a sharp video transition is detected.

In Fig. 4, we illustrate the results of some steps of our algorithm when we apply it to a visual rhythm obtained from a real video.



(a) Visual rhythm

(b) Horizontal gradient of the filtered visual rhythm

(c) Maxima points of the thinned horizontal gradient

(d) Filtering

(e) Number of maxima points in each vertical line

(f) Detected sharp transition superimposed on the visual rhythm

Figure 4: Sharp video detection. The threshold is equal to 50% of the maximum value.

## 5 Flash detection

The flash presence is very common in digital videos mainly in television journal videos. When a camera flash occurs, an increase of the luminosity in a few frames is produced, as illustrated in Fig. 5, and when we calculate a dissimilarity measure, like pixel-wise measure, we can see that in the frames affected by a flash, the dissimilarity measure is very high. In fact, a flash is confused with a sharp video transition. In the literature, we can find some methods for flash detection, like shot-reverse-shot [4]. In these cases, it is necessary to define a dissimilarity measure. In this work, we propose two methods for flash detection from the visual rhythm without defining a dissimilarity measure. The first is a variant of the proposed method for cut detection and the second considers a filtering of the component tree calculated from statistical measures computed by each frame (or frame sub-sampling).

### 5.1 Filtering by top-hat

On the visual rhythm, we can observe that the video flashes are transformed into thin light vertical lines, as showed in Fig. 6a. So, we can easily extract these lines from a white top-hat by reconstruction. The white top-hat by reconstruction is a mathematical morphology operator and represents the difference between the original image $g$ and the opening by reconstruction of $g$ [6, 7]. Informally, this operator detects light regions according to the shape and the size specifications of the structuring element. The method for flash detection can be described as follows.

1. Calculate the visual rhythm from the principal diagonal pixel sampling;
2. Apply the white top-hat by reconstruction with square structuring element of size $\lambda = 5$. This size is associated with the potential duration of a flash;
3. Apply a 1D thinning in each horizontal line;
4. Find the maxima points;
5. Apply an opening by reconstruction with vertical structuring element of size $\lambda = 7$, defined empirically;
6. Calculate the number of maxima points in each vertical line;
7. Apply a detection by thresholding.

We can observe that this method is very similar to the proposed method for cut detection. The difference here is the substitution of the morphological filter and horizontal gradient by the white top-hat by reconstruction. As the method for cut detection, this methodology detects the center of the regions of interest, in this case, regions with peak luminosity. Thus, we can have false detection in regions of high luminosity changing that do not represent a flash. Usually, this method produces good results when the flash appears in the middle of the shot.

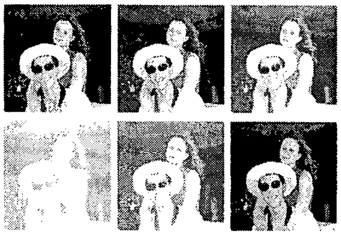


Figure 5: Some frames of a sequence with the flash presence.

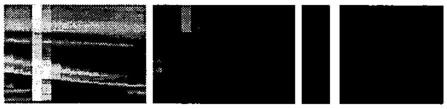


Figure 6: Flash video detection. Visual rhythm (left), white top by reconstruction (middle) and flash detected (right).

### 5.2 Component 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 realized by computation of some statistical measures like mean and median, where the frames affected by a flash present higher mean and median values with respect to their neighbors. From computation of these statistical measures for all frames of the video, we can create a 1D image for facilitating the flash detection. From this 1D image, we need to find the "peaks" with "height" greater than a value $H$, and with a "basis area" less or equal to a value $A$ that corresponds to the duration of the flash. In this work, we consider that the maximum flash duration is 5 frames, so $A = 5$. The parameter $H$ influences the sensitivity of the method and has a role similar to the threshold in Sec. 5.1. The notion of peak, height and basis area can precisely defined thanks to a data structure called *max-tree* [10] or *component tree* [11] (refer to these papers for more details on definitions and implementation).

## 6 Experimental results

In this section, we show the experimental results for cut detection and flash detection. Nowadays, our video database contains 150 videos, but we use only 32 videos for cut detection experiments and 10 videos for flash detection experiment. The choice of the sequences was associated with the presence of the different characteristics, such as, cut, dissolve, wipe, flashes, 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 quality measures in the next section.

### 6.1 Quality measures

We denote by $\#Cut$ the number of sharp (cut) transition, by $\#Correct$ the number of cuts correctly detected, by $\#False$ the number of detected frames that do not represent a cut and by $\#Miss$ the number of the cuts that are not detected defined by $\#Miss = \#Cut - \#Correct$. From these numbers we can define two basic quality measures.

| Experiment | Videos | Cuts | Dissolves | Flashes | Frames |
|---|---|---|---|---|---|
| Cut | 32 | 778 | 46 | 14 | 29933 |
| Flash | 10 | - | - | 23 | 8392 |

Table 1: Chosen video features for the experiments

**Definition 6.1 (Recall and error rates)** The recall and error rates represent the percentages of a correct and false detection, respectively, and are given by

$$\alpha = \frac{\#Correct}{\#Cut} \quad \text{(recall)} \tag{6}$$

$$\beta = \frac{\#False}{\#Cut} \quad \text{(error)} \tag{7}$$

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

**Definition 6.2 (Robustness)** Let $\alpha(\tau)$ and $\beta(\tau)$ be the functions that relate the threshold to recall and error rates, respectively. Let $m$ and $p$ be the percentage of miss and false detection that are permitted. The robustness $\mu$ is a measure related to the interval where the recall and error rates have the values smaller than $(1 - m)$ and $p$, respectively. This measure is in the range $[0, 1]$ and is given by

$$\mu(m,p) = \alpha^{-1}(1-m) - \beta^{-1}(p) \tag{8}$$

where $\alpha^{-1}$ and $\beta^{-1}$ are the inverses of the functions $\alpha(\tau)$ and $\beta(\tau)$, respectively. In Fig. 7, we illustrate the robustness measure obtained from functions $\alpha(\tau)$ and $\beta(\tau)$.

Next, we define two other measures, $E_m$ and $R_f$, that are associated with the absence of miss and false detection, respectively.

**Definition 6.3 ("Missless" error)** The missless error $E_m$ is associated with the percentage of false detection when we have results without miss (a small percentage of miss $P_m$ can be permitted, like 3%). The missless error is given by

$$E_m(P_m) = \beta(\max\{\tau = \alpha^{-1}(q) | 1 - q \leq P_m\}) \tag{9}$$

**Definition 6.4 ("Falseless" recall)** The falseless recall $R_f$ is associated with the percentage of correct detection when we have results without false detection (a small number of false detection $P_f$ can be permitted, like 1%). The falseless recall is given by

$$R_f(P_f) = \alpha(\min\{\tau = \beta^{-1}(p) | p \leq P_f\}) \tag{10}$$

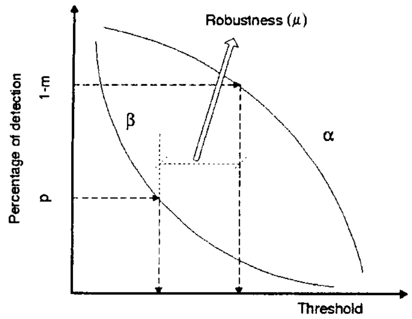


Figure 7: Robustness ($\mu$) measure.

When we use methods for cut detection, we expect that the recall is highest with a smallest error 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 $\alpha$ and low values of $\beta$ have to be rewarded, the function $\alpha(\tau) \times (1 - \beta(\tau))$ is a natural choice.

**Definition 6.5 (Gamma measure)** The gamma measure $\gamma$ represents the maximal value of the reward function defined above for all possible values of $\tau$:

$$\gamma = \max\{\alpha(\tau) \times (1 - \beta(\tau)) | \tau \in [0, 1]\} \qquad (11)$$

The quality of the results is associated with 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 detection 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, histogram intersection and a statistical technique based on visual rhythm. We chose these methods due to their simplicity and to present good results according to [5], [12] and [2], respectively. We also implemented the proposed method with some variants.

In the next sections we describe all experiments and in Sec. 6.2.1 we present a global analysis of their results.

**Experiment 1** This experiment uses the difference between pixel (defined in Sec. 3) as the dissimilarity measure. A 1D signal is created from the dissimilarity values calculated on the video. According to [5], we apply a mathematical morphology operator, called inf top-hat operator, on this signal, and finally, we use a threshold to detect the cuts, i. e., if the result of the inf top-hat operator is greater than a threshold, then a cut is detected.

**Experiment 2** This experiment uses the histogram intersection (defined in Sec. 3) 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, we realize a subdivision in each frame, according to [12]. So, each frame contains 9 subframes, and the dissimilarity measure is applied to all correspondent subframes in consecutive frames, being realized the mean between these measures. The results of this experiment when compared to previous experiment produce worse results for robustness, falseless recall and gamma measures, but better results for missless error.

**Experiment 3** This experiment uses visual rhythm for cut video detection based on statistical method as described in [2]. Here, the parameters are different from those used in other methods, in particular the threshold. While in this method the threshold is locally adaptive and related to a parameter that vary from 1 to 10, in the other methods the threshold is fixed and global.

This method presents the best values of falseless recall in these experiments, but other quality measures of the proposed methods are better. In particular, this method has a very bad missless error rate.

**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. Step i) eliminates noise on the 1D signal, and step ii) reduces the number of false detection according to the quality measures. This method can be visualized as an hybrid between the method described in the experiment 1 and the proposed method described in Sec. 4.

The quality measures of this method has best results when compared to the previous experiments, with exception of the falseless recall rate of the experiment 3.

**Experiment 5** In this variant of our method introduced in Sec. 4, instead of applying the summation of the number of maxima points in each vertical line, we use the filtered maxima image as a mask to verify the grayscale value associated with each maxima point. Afterwards, we find the mean of these grayscale values in each vertical line. Then, a thresholding is applied to these results, and if the mean is greater than a threshold, then a cut is detected. We verify that the falseless recall presents the second best result of these experiments, but the other measures are worse when compared to the next experiment.

**Experiment 6** This experiment is related to the method defined in Sec. 4. In general, the robustness, the missless error and the gamma have the best results when compared to the others experiments, and the falseless recall present the third best value of all experiments.

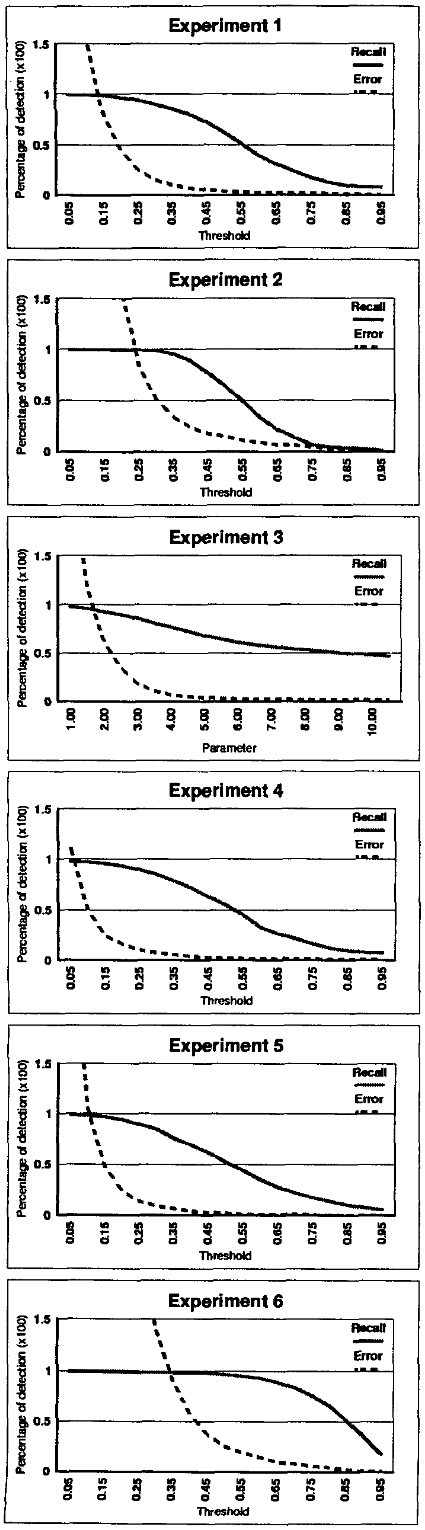


Figure 8: Experimental results.

### 6.2.1 Analysis of the results

In Fig. 8, we show graphically the experimental results for each experiment previously described. These graphics relate the threshold (except for experiment 3) to recall and error rates. From the functions illustrated in these graphics, it is possible to find the robustness, missless error rate, falseless recall rate and gamma measure, that are outlined in Table 2.

| | $\mu$ | $E_m$ | $R_f$ | $\gamma$ |
|---|---|---|---|---|
| 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** |

Table 2: Quality measures $\mu(0.10, 0.30)$, $E_m(0.03)$, $R_f(0.01)$ and $\gamma$.

From these experiments, we can verify that the proposed method generally produces the best results, mainly according to the robustness and the missless error rate. The result of the robustness means that the proposed method is a not very sensitive to small variations around an "optimal value". Another good point of our method is related to the missless error rate because generally, we want results without miss and with a smallest percentage of false detections, so that we can eliminate them posteriorly. Indeed, a post-processing is essential to increasing the quality of 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 with visual rhythm is significantly lower than for the experiments applied directly to the video.

## 6.3 Experiments for flash detection

In these experiments, we apply the methods described in Sec. 5.1 and in Sec. 5.2. In Fig. 9, we illustrate some experimental results. Considering two statistical measures, mean and median, we compute a component tree for each measure. The quality measures for the intersection of the filtering of the component tree and for top-hat filtering are outlined in Table 3.

| | $\mu$ | $E_m$ | $R_f$ | $\gamma$ |
|---|---|---|---|---|
| Top-hat | 0.05 | **0.61** | 0.26 | 0.56 |
| Component tree | **0.11** | 0.67 | **0.43** | **0.69** |

Table 3: Quality measures $\mu(0.40, 0.30)$, $E_m(0.05)$, $R_f(0.01)$ and $\gamma$.

## 7 Conclusions

In this work, we transform the video segmentation problem into a 2D image segmentation problem, and we propose

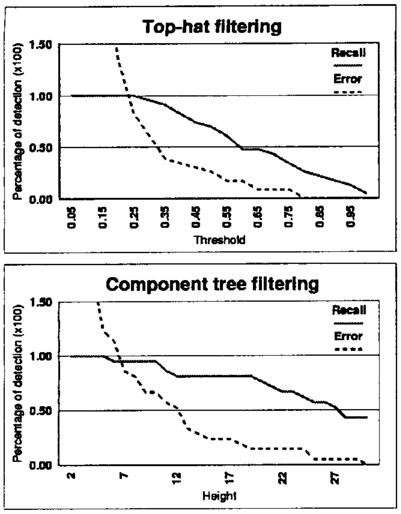


Figure 9: Experimental results for flash detection.

a method for cut detection from a video content simplification, called visual rhythm. Its main originality consists is the thinning step that decreases the number of false detections, with respect to the number of correct detections. This method is sensitive to filtering step due to the size of the structuring element that can eliminate some small regions, i. e., if the shot size (number of frames of the shot) is smaller than the size of the structuring element, then a miss occurs. To realize a comparative analysis between different methods for cut detection, we defined four quality measures: robustness, missless error, falseless recall and gamma. According to these quality measures, we verified that the proposed method have the best values of robustness, missless error and gamma measure, when compared experimentally to the other methods. Except for two methods, it has also the best falseless recall.

Another problem that we studied is related to the flash presence. In fact, due to the dissimilarity values, the flash can be confused with a sharp video transition, and with the aim of eliminating the choice of a dissimilarity measure, we proposed two methods for flash detection. A method is a variant of our cut detection method that uses a white top-hat by reconstruction and the another is related to a statistical measure filtering.

From this work, we observed that the visual rhythm presents an adequate simplification of the video content, which can be basis for future developments: i) identify some video effects, like pan, zoom, camera motion, from the detection of their correspondent patterns; ii) modify the proposed method to detect gradual video transitions, using the Canny's [13] filter to compute the horizontal gradient.

We can also remark that considering the video sequence as three-dimensional images, we could apply a variant of our method directly on the video data. We have to verify that the additional 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

[1] A. Hampapur, R. Jain, and T. E. Weymoth. Production model based digital video. *Multimedia Tool and Aplications*, 1:9–46, 1995.

[2] M. G. Chung et al. Automatic video segmentation based on spatio-temporal features. *Korea Telecom Journal*, 4(1):4–14, 1999.

[3] C. W. Ngo, T. C. Pong, and R. T. Chin. Detection of gradual transitions through temporal slice analysis. In *IEEE CVPR*, pages 36–41, 1999.

[4] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann, 1999.

[5] C.-H. Demarty. *Segmentation et Structuration d'un Document Vidéo pour la Caractérisation et l'Indexation de son Contenu Sémantique*. PhD thesis, École Nationale Supérieure des Mines de Paris, Janvier 2000.

[6] J. Serra. *Image Analysis and Mathematical Morphology: Theoretical Advances*. Academic Press, 1988.

[7] P. Soille. *Morphological Image Analysis*. Springer-Verlag, 1999.

[8] G. Bertrand, J.-C. Everat, and M. Couprie. Image segmentation through operators based upon topology. *Journal of Electronic Imaging*, 6:395–405, 1997.

[9] T. Y. Kong and A. Rosenfeld. Digital topology: Introduction and survey. *CVGIP*, 48:357–393, 1989.

[10] P. Salembier et al. Antiextensive connected operators for image and sequence processing. *IEEE Trans. on Image Processing*, 7(4):555–570, 1998.

[11] E. J. Breen and R. Jones. Attribute openings, thinnings and granulometries. *Computer Vision and Image Undestanding*, 64(3):377–389, 1996.

[12] A. Del Bimbo et al. Retrieval of commercials based on dynamics of color flows. *Journal of Visual Languages and Computing*, 11:273–285, 2000.

[13] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.