Detecting Shot Transitions Based on Video Content

E. Clua, M. S. Fonseca, A. Conci and A. Montenegro
Computer Science Department
Universidade Federal Fluminense
Rua Passos da Pátria, 156, Bl. E, s. 304, Niterói, RJ, Brazil – 24210-330
Phone: (55 21) 2629-5646 Fax: (55 21) 2629-5627 E-mail: esteban@ic.uff.br

Keywords: **content based segmentation, scene transition detection, video processing**

Abstract - Detection of scene transition is the first step on video segmentation, indexing and analysis. Although scene classification by human can be performed with visual or sonorous attributes at the same time, machine automatic classification usually relies on feature extraction of main visual characteristics. The use of color, shape, digital sound processing and voice signal altogether are investigated in this work. The color detection is based on the color histogram and shape detection is based on edge map histogram. Sound characteristics are resolved with the extraction of seven characteristics: short time average energy, zero-crossing rate, energy band ratio, delta spectral magnitude, root mean square of square sum of signals, high sounds and low value characteristics ratios. A Bayesian network is used on the decision for the transition. Finally, a new form of grouping frames is proposed. The results of the proposed method are summarized to show its efficiency.

1. INTRODUCTION

To help users find relevant information effectively, digital applications of audio, image and video indexing, retrieving and analysis have being growing in importance and volume, as a result of decreasing cost of storage devices and increasing network bandwidth capacities [1]. There are two widely-accepted approaches to characterize video in databases: shot based and object based approaches [10]. Shot based approaches are pointed as the best choice for high level video content indexing [3]. In these applications, scene transition detection or video segmentation is a basic step. This model also requires as a fundamental element the shot definition limits. Its identification must be based on the semantic of the acquired material, which is very complex due to the inherent complexity of multimedia contents [7].

Although image classification by human concerns several attributes, the use of color, shape and texture are the more frequently artifacts used on computational models [16]. On the other hand, research on digital sound and voice signal processing content in scene relies on extraction of several sonorous features [4]. Concerning the video detecting transition, the simplest way is detect the sharp transition (cut) that is simply a concatenation of the shots. Common approaches to cope with the cut detection are based on dissimilarity measure. The most popular cut detection approaches are based on pixel-wise comparison and histogram comparison [18]. However, cut detection can become complicated because of the presence of effects, like gradual transitions, flashes and fast camera and object motions. A well know approach for video detection transition is to transform the video into a 2D image and apply image processing methods in order to extract the different patterns regarding of each transition.

An interesting approach to detect video transition is through the visual rhythm [19], which concerns of getting a diagonal line from each frame. The image composed by lines represents the visual rhythm. Several image processing techniques are carried out in order to extract transitions. On this line, the spatial-temporal slice is also defined [20]. Both are related to the same video transformation and a sub-sampling of each frame, like the principal diagonal sub-sampling. However, when applied statistical measures to detect some patterns, the number of false detections is very high [19]. Markov models for shot transition detection have been also applied, but it fails when there is low contrast between textures of consecutive shots. A morphological and topological tool to detect cuts by analysis of the visual rhythm by sub-sampling, as suggested in [20], was used in this work.

This work proposes a semantic analysis for automatic detection of scene transition using a Bayesian network model to combine visual features with sound features in the scene. The shot identification considers the color aspects (based on color histogram), shape or texture aspects (based on edge map histogram) and sound aspects considering seven sound characteristics: short time average energy, zero-crossing rate, energy band ratio, delta spectral magnitude, root mean square of square sum of signals, high sounds and low value characteristics ratios. A comparison between manually and automatically obtained results four video analyzes illustrates the proposal performance.

2. THE PROPOSED METHOD

This section has the objective to present the developed methods for determination of the similarity between consecutives blocks of frame based on all attributes. The metric and the features extracted to represent the content of the video and the established way to divide sound and image frames are also discussed.

2.1 The Bayesian Models

This section considers the outline of Bayesian models in the way it is used on the problem of finding the scene transition, considering it's semantic meaning. Bayes theorem is a powerful probabilistic model widely used to represent relationship among probabilities a priori with probabilities a posteriori. The theorem also manage knowledge in domains which require reasoning under uncertainty [5]. Bayes theorem provides a concise representation of joint probability distribution among the variables, leading to an efficient inference process. With such characteristics, Bayesian models have been successfully applied to a wide range of problems, such as described in [11] and [13].

2.2 Used metric

The distance s between two features vectors can represent their similarity. This distance can be normalized such as $s \in [0, 1]$, so that it is a measure of the similarity between the images or sound frames from where they are extracted. In that way, the probability P of being similar can be expressed as $P = 1 - s$.

The distance function used is given by cosine between the two feature vectors on consideration or their correlation. Let $q = (c_{1q}; c_{2q}; : : : ; c_{vq})$, be the feature vector of a frame A, considering a specific kind of attribute in a *v-dimensional* space and $p = (c_{1p} ; c_{2p} ;... ; c_{vp})$, be the feature vector of a frame B on the same type of attribute and space, then:

$$sim(p,q) = \frac{p^t q}{\| p \| \| q \|} = \frac{\sum_{i=1}^{v} c_{ip} \times c_{iq}}{\sqrt{\sum_{i=1}^{v} c_{ip}^2} \times \sqrt{\sum_{i=1}^{v} c_{iq}^2}} \qquad (1)$$

2.3 Color Content

Several color spaces have been used for color representation based on the perceptual concepts. In the implemented system, the used color space is the *HSV*. An image pixel can be represented in a color space by a vector. Color quantization transforms a continuous tone into a discrete set of points. It maps each component of a continuous color signal into a series of colors. Through the quantization the dimension of the space is reduced, retaining the information of the color.

Uniform quantization in *HSV 162* (18x3x3) bits were used. Hue (H) is the attribute associated with the dominant wavelength. Axis H represents the more significant characteristic, where its values vary from 0° to 360° in *HSV* and it was quantized in 18 sections of 20° each. Axes S and V were partitioned in 3 cells each one. Saturation, given by S, and Value, given by V, have been limited by 0, 0.5 and 1 values [2]. Each frame, in term of its color, is characterized by its 1D normalized histogram with $m = 162$ cells:

$$h[_i] = \frac{1}{XY} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} f(x,y) = \begin{cases} 1 \text{ if } (x,y) \text{color} = i \\ 0 \text{ if } (x,y) \text{color} \neq i \end{cases} \quad i = 0,1,2,...m \quad (2)$$

where X, Y are the number of pixels in the frame on each direction, $f(x,y)$ is the color at position (x,y) and $m = 162$. In this context, color histograms are points in an m-dimensional space related "one-to-one" to the image frames [15].

2.4 Shape and Texture

The texture or shape model used in this work is represented by the angle of the edges detected on the image frame. If the image being analyzed presents no textural elements (or few textural elements, as a cartoon based video) the used approach characterizes the shape of the object in the scenery. But for real images, like movies or films, all textural elements are characterized by the methodology summarized in figure 1.

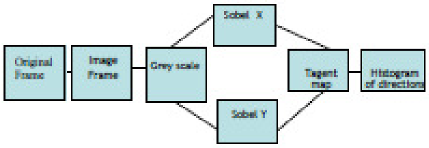

Fig. 1. Scheme for textural elements characterization.

In the same manner of the color content analysis, only the image frame are considered, but now the color information is discarded. The grey scale images are boundary detected using the directional Sobel filter. The edges in vertical and horizontal directions are used to compute the image angles. Angles are quantized in a histogram considering grouped in ten directions. These angular histograms are associated with each frame after their normalization.

2.5 Sonorous Features

Audio characteristics can be acquired by a single frame or by a group of sonorous frames. They are named shot time and long time characteristic respectively. For the second, a time window is used to encircle some frames. The interval or width of duration of this time window must be well defined in order to correctly characterize spoken signals [6]. Such window glides the frames for feature extraction but, like the shot time characteristics, the resulting feature is associated with the first frame. A preprocessing was realized for each video for adequate definition of this interval of time.

Some of the sound aspects considered use the signal directly in the time domain, namely: zero-crossing rate, root mean square of square sum of signals, short time average energy, high feature value and low values ratios. For others characteristics the signal are considered in the frequency domain by the use of a Discrete Fourier Transform (DFT) [12] [14].

The Zero Crossing Rate (ZCR) is related with the content and frequency of the waves and defined by the number of signals changes:

$$ZCR = \frac{1}{M-1} \sum_{m=0}^{M-1} | sign(x(m)) - sign(x(m-1)) | \quad (3)$$

where M is the number of samples in the window and *sign* $x(m) = 1$ if $x(m) > 0$ and *sig* $x(m) = 0$ if $x(m) < 0$. This feature, with and adequate window, can identify voice in a sound.

Delta Spectrum Magnitude (DSM) is used to characterize music signals [15] it is obtained from:

$$DSM = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log A(n,k\delta) - \log(A(n-1),k) + \delta)]^2 \qquad (4)$$

where N is total number of frames, K is the number of element on the DFT, δ is a value used to overflow prevention, and

$$A(n,k)=\left|\sum_{m=-\infty}^{\infty} x(m)w(nL-M)\,e^{-j\left(\frac{2\pi}{L}\right)km}\right| \quad (5)$$

where k is the frequency of frame n , $x(m)$ is the signal value, $w(m)$ is a function of the window of L length.

The Root Mean Square (RMS) value measures the signal energy, it is defined by:

$$RMS=\sqrt{\frac{1}{M}\sum_{m=0}^{M-1}x^2(m)} \quad (6)$$

where M is the number of samples and $x(m)$ is the signal value.

Short Time Average Energy (STE) is a simple and greatly used feature for audio segmentation [6] [8] [9]:

$$STE=\sum_{m=0}^{M-1}x^2(m) \quad (7)$$

High Feature-Value Ratio (HFVR) is also very used to differentiate voice and music in an audio [14] [17], it is defined by:

$$HFVR=\frac{1}{2N}\sum_{n=0}^{N-1}[\sin(ZCR(n)-1{,}5avZCR)+1] \quad (8)$$

where N is the total number of frames considered and $avZCR$ is the average ZCR in the considered window, and the other elements have the common meaning. To characterize silence and background sound the Low Feature-Value Ratio (LFVR) is considered:

$$LFVR=\frac{1}{2N}\sum_{n=0}^{N-1}[\sin(0{,}5avSTE-STE(n)+1] \quad (9)$$

where $avSTE$ is the average STE in the total number of frames considered.

The Energy Band Ratio (BER) consider the energy on a specify frequency band , f_i . It can be computed by Fourier Transform of the signal using different FFT approaches. Here the frequency spectrum is divided in 4 sub bands:

$$BER=\frac{\sum_{n}^{h}f_i(n)}{\sum_{n=0}^{N}f_i(n)} \quad (10)$$

where h = $N/4$ [11]. These seven normalized characteristic are used to form the feature vector related to sound and associated with each frame.

3. RESULTS AND PERFORMANCE ANALYSIS

Four types of videos are tested to illustrate the process and the results achieved with the proposed methodology. All these videos are real examples obtained from a local broadcast channel. They present great variety of sound and characters overlaid onto complex backgrounds, colors, illumination and textures. The videos, provided in VHS, are initially digitalized to AVI format. Then, the videos are separated in audio and still frames.

The still frames are stored in TIFF format while the audio frames are analyzed as WAV files. The TIFF files are blocked on groups of n=30 frames, to synchronize with the motion-picture film rate, in this case 30 frames per second. The WAV files are partitioned in block of 30 second each (900 elements). For a decision where there is transition, the probability obtained for each group of the Bayesian models are compared with a fixed and unique threshold value. The values 0.7 and 0.99998 are used for this work. Each of these blocks is manually analyzed by people. Discrepancy in manual identification is investigated conducting to a unique correct description of points of transition. Table 1 shows the number of scenes, blocks and frames of each video analyzed. Table 1 also presents the system performance to these videos but in percentage of identification of transitions and considering the two types of wrong answer: false positive and false negative. Note that comparing the automatic detected transition point with those at least 75% of correct detection was obtained.

	Video 1	Video 2	Video 3	Video 4
Number of scenes	14	5	27	7
Number of blocks	390	149	300	300
Number of Frames	11710	4496	9020	9003
Description	TV news	Show	Cartoon	Movie
Correct detection	76.92%	75%	76.92%	83.33%
False positives	53.85%	25%	38.46%	50%
False negative	23.08%	25%	26.92%	16.67%

Table 1. Videos characteristic, percentage of correct identification, false positive and negative.

4. CONCLUSION

This work presents a novel methodology to detect video transitions. Specifically, a similarity measure based on Bayesian network which combines traditional image processing techniques including also sound features is used. The proposed methodology can be considered an automatic tool for shot detection in video based on semantics of perceptually sounds, colors, texture or shapes. Since the proposed method is designed to locate scene transition, all tested videos are firstly analyzed by two persons in order to find these points. Comparing the automatic detected transition point with those at least 75%

of correct detection was obtained. Of course more elements could be added on the proposed approach. The Bayesian model is easily adapted to consider more elements as well as the used metric. It is also adequate the use of other forms of shapes and texture characterizations, specially associated with the content of the video on investigations. This means that others type of shape descriptors must be considered. Among these, the contour signature and Fourier descriptor could be adequate for cartoon like content. The DFT spectrum or co-occurrence matrixes approaches could be adequate for movies.

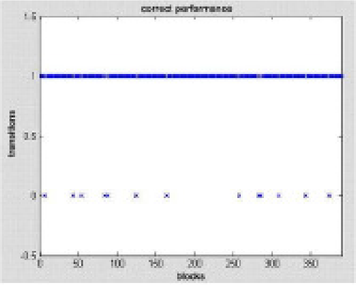

Fig. 2. Real transitions (function F(i)) for video 1: points marked on zero correspond to block where transition are presented

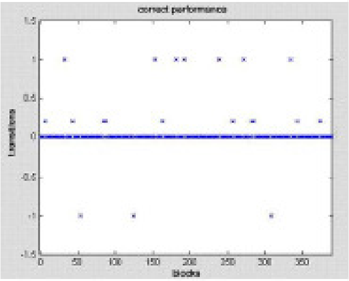

Fig. 3. Automatic detected transition (function G(i)) for video 1: points marked with zero correspond to block where transition are presented.

REFERENCES

[1] R. Brunelli, O. Mich, and C.M. Modena, "A Survey on the Automatic Indexing of Video Data," *J. of Visual Com. and Image Representation*, 10, 1999, pp.78-112.

[2] A. Conci and E. M. M. M. Castro, "Image mining by content", *Journal of Experts Systems with Applications*, Elsevier Science UK, 2002, pp. 377-383.

[3] J. Fan, A.K. Almagarmid, X. Zhu, W.G. Aref, and L.Wu, "ClassView: Hierarchical Video Shot Classification, Indexing, and Accessing", *IEEE Transactions on Multimedia*, Vol. 6, No. 1 2004, pp. 70-86.

[4] F. Gouyon, F. Pachet and O. Delerue. On the use of zero-crossing rate for an application of classification of percussive sounds. in *Proceedings of the COST G-6 Conference on Digital Audio Effects* (DAFX-00), 2000.

[5] F.V.Jensen. *Bayesian networks and decision graphs. Statistics for Engineering and Information Science.* Springer, 2001.

[6] D. Li. I. K. Sethi, N. Dimitrova and T. McGee. "Classification of general audio data for content-based retrieval", *Pattern Recognition Letters*, pp. 533-544, 2001.

[7] R. Lienhart, S. Pfeiffer and W. Effelsberg. "Video abstracting", Communications of ACM pp. 54-62. 1997.

[8] Z. Liu, Y. Wang and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification", *J. of VLSI Signal Processing Systems*, 1998.

[9] L. Lu, H. J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation", *IEEE Transaction on Speech and Audio Processing*, pp. 504-516, 2002.

[10] Maybury, M.T. (ed), Intelligent Multimedia Information Retrieval, AAAI Press/MIT Press, 1997.

[11] P. S. Rodrigues, A. A. Araújo and G. A. Giraldi. "Using Tsallis entropy a bayesian network for CBIR", in *Proceeding of International Conference on Image Processing (ICIP'05)*. Genova, Itália, 2005.

[12] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multi feature speech/music discriminator", in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1331-1334, 1997.

[13] I. Silva, B. Ribeiro-Neto, P. P Calado, E. S. Moura, and N. Ziviani, "Link-based and content-based evidential information retrieval in a belief network model", in *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 96–103, 2000.

[14] C. C. Tong and J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification", *IEEE Transaction on Speech and Audio Processing*, volume 9, pp. 441-454, 2001.

[15] H. W. Yoo, D. S. Jang, S. H. Jung, J. H. Park and K. S. Song, "Visual information retrieval system via content-based approach", *Pattern Recognition*, pp. 749–769, 2002.

[16] X. S. Zhou and T. S. Huang, "Edge-based structural features for content-based image retrieval", *Pattern Recognition Letters*, pp. 457–468, 2001.

[17] Y. Zhu and D. Zhou, "Scene change detection based on audio and video content analysis', in *IEEE International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 03)*, pp. 695 -702, 2003.

[18] Y. Wang, Z., Liu, and J.C. Huang, J.-C., "Multimedia content Analysis", *IEEE Signal Process*.pp. 12–36, 2000.

[19] M.G Chung , J. Lee, H. Kim, S.M.-H. Song and W.M Kim, "Automatic video segmentation based on spatio temporal features", *Korea Telecom* J. 4 , 4–14, 1999.

[20] S.J.F. Guimarães, M. Couprie, N.J. Leite, and A.A Araujo,."Amethod for cut detection based on visual rhythm" in: *Proc. of the XIV Brazilian Symposium on ComputerGraphics and Image Processing*, Brazil, Computer Society Press, pp. 297–304, 2001