Shot Boundary Detection Algorithm Based on Multi-Feature Fusion

[1]Kai Jin, [2]Hong-cai Feng
*1, First Author School of Mathematic & Computer Science, Wuhan Polytechnic University, Wuhan 430023, China, sir.jinkai@gmail.com
*2, First Author Modern Education Technology Center, Wuhan Polytechnic University, Wuhan 430023, China, fenghc@whpu.edu.cn

[3]Qi Feng, [4]Chi Zhang
[3]Business School, Newcastle University, Newcastle Upon Tyne, United Kingdom, NE2 4LB
[4] School of Mathematic & Computer Science, Wuhan Polytechnic University, Wuhan 430023, China
Electronic banking department, department of hubei branch, Agricultural Bank of China
xtfo119@163.com

***Abstract*—To establish a general and robust shot boundary detection algorithm, according to characteristics of lens conversion and the ideal of multiple video features fusion, a shot boundary detection algorithm is proposed based on YUV histogram, texture feature and edge orientation histogram in the paper. Besides, global and self-adaptive threshold are combined to use so as to control the process of shot boundary detection and enhance the accuracy of threshold selection. The experiment results show that the algorithm can effectively realize video shot boundary detection and strengthen the robustness of the detection.**

Keywords-shot boundary detection (SBD); histogram; feature fusion; threshold selection;

Ⅰ.INTRODUCTION

With the rapid development of multi-media and network technology, the consequence caused video information expansion has become an increasingly prominent problem. How quickly and efficiently find the useful information has become an urgent task. So, CBVR (Content-Based Video Retrieval) has become a hot research. Nevertheless, shot boundary detection is the basis of the CBVR, which impacts directly the performance of CBVR. Usually, shot boundary has two types which include cut and gradual transition. Cut refers to a lens switching to another lens directly, and gradual transition refers to a lens switching to another lens through the video transition frame, among them gradual transition contains fade [1], dissolve, and wipe, etc. In view of these two types of lens conversion, domestic and overseas scholars put forward a variety of different detection methods, including color histogram method, based on edge feature or motion feature detection method, etc. These methods have achieved some effects, but also there are some obvious inadequacies. As follows, [2] proposed a detection method based on color histogram, which can effectively detect the cut type, but can't detect gradual transition very well; [3] proposed a detection method based on motion vector, which has a good detection rate for various types of video, but there still exist the problem of insufficient detection capability when handling with some videos containing the complex object motion. Thus it can be seen, using a single feature for video shot detection, it's easy to appear the problem of insufficient detection capability, resulting in missed or false detection. Therefore, this paper proposes a shot boundary detection algorithm based on multi-feature fusion, which adopts color histogram and multiple features fusion[4] ideas for feature extraction, in this way, it can effectively avoid the missed or false detection to be caused by only using a single feature for shot detection. Besides, this paper also adopts the method of global and self-adaptive threshold combined to enhance the accuracy of threshold selection and avoid the arbitrariness and diversity to be caused by manually setting the threshold. The experimental results show that the algorithm can effectively realize shot boundary detection, and has a good robustness and higher recall ratio and precision ratio.

Ⅱ. VIDEO PHYSICAL FEATURES EXTRACTION

A. Color Feature Extraction

Ⅰ) YUV color space

YUV describes a color space through the luminance/chrominance. Among them, Y represents the luminance, namely gray value; U and V represents chrominance, which is used to specify the color of a pixel. The main reason of choosing YUV color space is two separate signals of luminance signal Y and chrominance signal U, V, which can reduce the lens detection calculation by only processing Y. In addition, the chrominance of the different frame images in the same scene is basically the same and single, even if the chrominance would not be much change in the adjacent frames of shot boundary. Therefore, this paper uses the change of Y component as the main basis to judge the color feature difference.

$$Y = 0.3R + 0.59G + 0.11B \quad (1)$$

Among them, R, G, B represents respectively red, green, blue component value of RGB color model.

Ⅱ) Color feature representation

Heterogeneous block weighted histogram method can effectively suppress the influence of inserted subtitles at the top or bottom of video segments [5]. In addition, in view of the method of identification image of the human eye is nonlinear, inhomogeneous, during the sampling of the retina, it has a higher resolution in the central area

than in the surrounding area. Therefore, this paper adopts the method of strengthening the central characteristics to adapt to human visual characteristics, in order to improve the accuracy of the video image retrieval.

As shown in Figure1, we divided a video frame image into 3×3 sub-blocks, vertical and horizontal sub-block ratio of 1:3:1, the four corners of the weight value is set to 0.

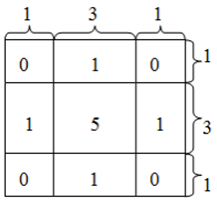

Figure1. Heterogeneous Block Weighted

This paper calculates frame difference based on the heterogeneous block weighted of Y component. Firstly, we divided the luminance space into Q interval, marked for 1,2,…,q,…,Q. To improve the accuracy of the threshold detection, we take S_k (the number of pixel in sub-block) as the size of current sub-block k, among them, $h_{ik}(q)$, $h_{jk}(q)$ represent respectively the histogram of luminance interval q of i-th frame and the j-th frame in the sub-block k, then the histogram distance of luminance interval q of i-th frame and the j-th frame in the sub-block k is

$$\overline{d}_{ijk} = \sum_{q=1}^{Q} | h_{ik}(q) - h_{jk}(q) | / S_k \tag{2}$$

Corresponding weight values of the sub-block are respectively w_1, w_2,…, w_9 weighted matrix marked as W, as the formula (3) shown.

$$W = \begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 5 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{3}$$

We can calculate the histogram different value of corresponding block between two frames' by formula (2), marked respectively as $\overline{d}_{ij1}, \overline{d}_{ij2}, ..., \overline{d}_{ij9}$, so the histogram different value of i-th frame and j-th frame is.

$$DF_{i,j} = \sum_{n=1}^{9} w_n \overline{d}_{ijn} / \sum_{n=1}^{9} w_n \tag{4}$$

B. Texture Feature Extraction

Texture feature refers to the change of the image gradation, the reaction of properties of the image itself. According to human visual system's simulation, Gabor decomposed a frame image into filtering image, to reflect the strength change of different frequency and direction [6]. Gabor filter is actually a group of wavelet, which can capture energy in a specific frequency and direction. Therefore, we can get a local frequency description by expanding the signal, so as to capture the local feature and energy of the signal and extract texture feature in the group of energy distribution [7].

Sine wave of two-dimensional Gabor function along the X axis is

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{-(x^2+y^2)}{2\sigma^2}} (\cos\frac{2\pi x}{l} + j\sin\frac{2\pi x}{l}) \tag{5}$$

Among them, (x,y) as the coordinate point of G(x,y), σ is standard deviation, $l = 6\sigma / f$ (f represents the center frequency).

By rotation and scale transformation of G(x,y), we can get a set of filter.

$$g_{mn}(x,y) = \alpha^{-m} g(x',y') \quad \alpha > 1 \tag{6}$$

Among them, m, n represent respectively scale and direction, α as decomposition coefficient, $x' = \alpha^{-m}(x\cdot\cos\theta + y\cdot\sin\theta)$, $y' = \alpha^{-m}(x\cdot\sin\theta + y\cdot\cos\theta)$, $\theta = n\pi / k, (\theta \in [0,k])$, k represents the number of overall direction (k for positive integer).

For a given image $I(x,y)$, its Gabor wavelet transform can be defined as

$$w_{mn}(x,y) = \iint I(x,y) g^{*}_{mn}(x - x_1, y - y_1) d_{x_1} d_{y_1} \tag{7}$$

Among them, * represents conjugate complex, (x_1, y_1) as the reference point. This paper selected mean value μ_{mn} and variance σ_{mn} as the similarity measure.

$$\mu_{mn} = \iint | w_{mn}(x,y) | dxdy \tag{8}$$

$$\sigma_{mn} = \sqrt{\iint (| w_{mn}(x,y) - \mu_{mn} |)^2 dxdy} \tag{9}$$

To describe the image texture, this paper adopts μ_{mn} and σ_{mn} as component to structure feature vector f. Usually scale m for 5, direction n for 6, namely feature vector expressed as f=(μ_{00}, σ_{00}, σ_{01},…, μ_{45}, σ_{45}).

C. Edge Feature Extraction

Edge, which exist in between target and background, objectives and goals, area and regional, is the basic feature of image, containing rich image information (such as direction, shape, etc.). The purpose of edge detection is to extract edge information of image, to eliminate unrelated information and reduce data analysis. This paper adopts the new model of [8] to calculate the level gradient value $G_{x_{i,j}}$ and vertical gradient value $G_{y_{i,j}}$ of pixel point $P_{i,j}$, thus draw the following value:

The gradient magnitude value which can reflect regional edge sharpness is:

$$Mag(P_{i,j}) = \sqrt{G^2_{x_{i,j}} + G^2_{y_{i,j}}} \tag{10}$$

The gradient direction value which can reflect edge direction of different pixel is:

$$D_{ir}(P_{i,j}) = arc\tan\sqrt{\frac{G_{x_{i,j}}}{G_{y_{i,j}}}},\ -\frac{\pi}{2} \le D_{ir}(P_{i,j}) \le \frac{\pi}{2} \qquad (11)$$

Divided $D_{ir}(P_{i,j})$ value range into 16 equal parts, each value interval is $[-\frac{9\pi}{16}+\frac{\pi}{16}k, -\frac{\pi}{2}+\frac{\pi}{16}k]$, k=1,2,…,16. Quantitative calculation of $D_{ir}(P_{i,j})$, each pixel $P_{i,j}$ in a frame image has two corresponding value $Mag(P_{i,j})$ and $\theta(P_{i,j}) = k$. Take $\theta(P_{i,j})$ as X axis, the edge direction histogram is

$$H = \sum_{i,j} Mag(P_{i,j})\delta[\theta(P_{i,j}) - t],\ \delta[x-t] = \begin{cases} 1, x = t \\ 0, x \neq t \end{cases} \qquad (12)$$

III. DESCRIPTION OF THE ALGORITHM

A. Feature Fusion

This paper proposes a shot boundary detection algorithm based on multi-feature fusion, which adopts color histogram and multiple features fusion ideas for feature extraction, in this way, it can effectively avoid the missed or false detection to be caused by only using a single feature for shot detection. For example, given two frame images, according to the color histogram difference value $DF_{i,j}$, texture similarity f and edge histogram H, the overall similarity of two image can be calculated. Further, by the Euclidean distance [9], the color histogram similarity d_1, texture similarity d_2 and edge histogram d_3 between the two frame images can be calculated.

In order to calculate more convenient, the paper takes the absolute value of the above feature. So, the color histogram similarity, texture similarity and edge histogram similarity between i-th frame and j-th frame is defined as follows:

$$d_1(i,j) = \| DF_{i,j} \|;\ d_2(i,j) = \| f_i - f_j \|;$$
$$d_2(i,j) = \| H_i - H_j \| \qquad (13)$$

This paper adopts the Multiplicative Fusion method to fuse the above three kinds of feature values in order to obtain the fusion feature value $D_{i,j}$.

$$D_{i,j} = d_1 \times d_2 + d_2 \times d_3 + d_1 \times d_3 \qquad (14)$$

B. Threshold Selection

Attributed to the tiny changes of two frame images, the current frame difference may be larger than the cut threshold. So, this paper sets a global threshold D to avoid the above situation. Under normal circumstances, the frame difference change is gentle in the same shot, to stabilize within a smaller range, and fluctuates around the average value of frame difference [10]. Therefore, this paper adopts the average value of frame difference multiplied by a threshold coefficient as a local adaptive threshold to judge the shot boundary.

From the formula (14) shows that the adjacent frame difference in a video segment are respectively $DF_{1,2},...,DF_{i,i+1},...,DF_{n-1,n}$, thus we can calculate the average value of adjacent frame difference, $D = \frac{1}{n-1}\sum_{i=1}^{n} DF_{i,i+1}$ (n=2,3 , …), as the global threshold.

By the local adaptive threshold method, we can calculate the average value of frame difference.

$$avg_i = \frac{1}{i-l}\sum_{j=l}^{i-1} DF_{j,j+1} \qquad (15)$$

Among them, avg_i represents the nearest cut frame to distance current frame (referred to as the i-th frame) or the average value of frame difference from the next frame (referred to as the $l-th$ frame) of gradient sequence end frame to the previous frame (referred to as the i-1th frame) of current frame.

The cut threshold, gradient threshold, flash threshold respectively set to:

$$TH_1 = \alpha \times avg_i,\ TH_2 = \beta \times avg_i,$$
$$TH_3 = \gamma \times avg_i \qquad (16)$$

Among them, α is a cut coefficient, β is a gradient coefficient, γ is a flash coefficient.

C. Algorithm Implementation

Take MPEG format videos for example, assume that the number of video frames contained in the whole video segment is n. By the formula (1), RGB space model is converted into YUV space model. Shot boundary detection algorithm is described as follows:

1) Assuming the initial value of cut set Cut, the start frame set Gras of gradient sequence, the end frame set Grae of gradient sequence, the start frame set Flas of flash sequence, the end frame set Flae of flash sequence, are all empty, and the frame number i is set to 1.

2) To calculate the global threshold D by the formula (14) and expression $D = \frac{1}{n-1}\sum_{i=1}^{n} DF_{i,i+1}$ (n=2,3,…).

3) To judge whether i less than n. If i<n and i is the first frame of current shot, set i=i+1, go on detection; if i<n and i is the 2th or 3th frame of current shot, take the frame difference mean of the current frame to the subsequent four frames (total of 5 frames) as the local frame difference mean avg_i; if i=n, skip to step 9), the end of shot detection.

4) To calculate local adaptive threshold TH1, TH2 and TH3 by formula (15) and (16).

5) To judge whether $D_{i,i+1}$ greater than the global threshold D. If $D_{i,i+1} > D$ and $D_{i,i+1} > TH_1$, show that the current frame is a cut preselected frame; if $D_{i,i+1} > D$ and $D_{i,i+1} < TH_1$, have to judge whether the

current frame is a gradient preselected frame, skip to step 7); if $D_{i,i+1} < D$, set i=i+1, return to step 3), go on to detect the next frame.

6) To calculate non-adjacent frame difference $DF_{i,i+2}, DF_{i,i+3}, ..., DF_{i,i+9}$ by the formula (14), and respectively compare with TH3. If $DF_{i,j} < TH_3$ in the non-adjacent frame difference, to stop comparison, show that (i+j)-th frame is the end frame of flash sequence, the video sequence between i-th frame to (i+j)-th frame is a flash sequence. Continue to judged j value, if j>3, to take i value put into the start frame set Flas of flash sequence, (i+j) value put into the end frame set Flae of flash sequence, set i=i+j+1, return to step 3); otherwise i-th frame is a cut frame, take i value put into set Cut, set i=i+2, return to step 3).

7) To judge $DF_{i,i+1}$ whether less than the gradient threshold TH_2 .If $DF_{i,i+1} < TH_2$, show that the current frame does not the boundary frame, set i=i+1, return to step 3); otherwise, go to step 8), for the gradient detection.

8) Assuming that the current frame is a start frame of gradient sequence, $DF_{1,2}, ..., DF_{i,i+1}, ...$, respectively compare with TH_2, if $DF_{i,i+j+1}, DF_{i,i+j+2}, DF_{i,i+j+3}$ are all less than TH_2, show that (i+j)-th frame is the end frame of gradient sequence. Usually, a gradient sequence is greater than 3 frames, if j<=3, show that the sequence is noise and must been removed from the gradient sequence, set i=i+1, return to step 3); otherwise, j>3, take i value put into Gras, (i+j) value put into Grae, set i=i+j+1, return step 3).

9) The end of the operation, output the values of Cut, Gras, Flas, Flae, to obtain the cut frames, gradient sequences and flash sequence.

Ⅳ. THE EXPERIMENTAL RESULTS AND ANALYSIS

This paper adopted four news video segments of MPEG format as experimental object, which are derived from shot boundary detection project in TRECVID2007. Cut type and gradual transition type account for approximately 57.6% and 36.4% of all video frames. Local threshold coefficient set to: α=5, β=2.5, γ=8.

NIST (National Institute of Standards and Technology) has proposed a standard evaluation method for SBD, which is used to measure the quality of SBD mainly by recall and precision [11].

$$\mathrm{Re}call = Correct\ Detection \Big/ Ground\ Truth$$

$$\mathrm{Pr}ecision = Correct\ Detection \Big/ All\ Detection$$

This paper conducted the simulation experiments on the four video segments by MATLAB_R2012a simulation software, the specific results as shown in Table1, Table2.

Table1. The simulation result of different video segments

Video	Frame	Actual		Missed		False		Recall（%）		Precision（%）	
		cut	gradual	cut	gradual	cut	gradual	cut	gradual	cut	gradual
Se1	392	226	143	5	21	4	15	97.74	83.54	98.19	87.70
Se2	375	216	136	4	18	5	14	98.10	85.24	97.64	88.13
Se3	383	220	139	5	19	7	14	97.65	84.80	96.74	88.33
Se4	395	227	144	3	20	6	16	98.64	84.37	97.32	87.09
Total	1545	889	562	17	78	22	59	98.03	84.49	97.47	87.80

Seen from Table1, the proposed algorithm has a higher recall rate and precision rate. The overall recall rate and precision rate of cut type are respectively 98.03% and 97.47%; the overall recall rate and precision rate of gradual transition are respectively 84.49% and 87.80%. For cut detection, the recall rate is higher than precision, which is due to the false detection to be caused by the rapid movement of the objects in a lens and the light changes; but for gradual transition, the precision recall is higher than recall, which is due to the missed detection to be caused by the tiny color difference between front and rear lens.

Table2. The performance comparison results of two algorithms

Algorithm	[4] algorithm				The proposed algorithm			
Video	Se1	Se2	Se3	Se4	Se1	Se2	Se3	Se4
Recall（%）	89.45	88.63	88.95	87.73	90.64	91.67	91.22	91.50
Precision（%）	91.48	90.75	91.03	90.35	92.94	92.88	92.53	92.20

Seen from Table2, the proposed algorithm has a higher recall rate and precision rate compared with [4] algorithm. The main reason is as follows, the proposed algorithm adds a flash detection, which can effectively detect the flash sequence; using the heterogeneous block weighted method for video frame image, this can effectively suppress the influence to be caused by inserting subtitles at the top or bottom of video segments; adopting to global and self-adaptive threshold method, this can effectively enhance to the accuracy of threshold selection.

Ⅴ.CONCLUTIONS

In this paper, to detect the shot boundaries by feature fusion and local self-adaptive threshold method, it can effectively reduce the impact of flash and subtitle, and strengthen the robustness and versatility of shot boundary detection algorithm. In addition, this paper uses the change of Y component as the main basis to judge the color feature difference in order to reduce the lens detection calculation. Heterogeneous block weighted histogram method is adopted in the paper, which can effectively suppress the influence of inserted subtitles at the top or bottom of video segments. What's more, the paper uses the Gabor filter, which can well capture energy in a specific frequency and direction, so as to get a local frequency description by expanding the signal and extract texture feature in the group of energy distribution. Besides, this paper also adopts the method of global and self-adaptive threshold combined to enhance the accuracy of threshold selection and avoid the arbitrariness and diversity to be caused by manually setting the threshold. However, the efficiency of the algorithm and adaptive threshold determine need to be further improved, which will also be the direction of future research.

ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation of Hubei Province (No.: 2009Chb008, 2010CDB06603), key scientific research projects of Hubei Provincial Education Department (No.: D20101703).

REFERENCES

[1] CERNEKOVA Z, PITAS I, NIKOU C. Information theory-based shot cut/fade detection and video summarization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006,16(1):82-91.

[2] Feng Hong-cai, Yuan Xiao-juan, Ming Wei. A shot boundary detection method based on color space[C].International Conference on E-Business and E-Government, 2010, pp.1647-1650.

[3] Wang Cheng-ru, Wang Hui-hui. MPEG video shot boundary detection based on motion vectors[J]. Journal of Computer Applications, 2012,32(5):1269-1271.

[4] Liu Qun, Jiang Wei, Wu Yu. Approach of shot-boundary detection based on multi-feature fusion[J]. Computer Engineering and Application, 2010,46(13):171-174.

[5] Jin-Wook Lee, Jae-Soo Cho. Effective Lane detection and tracking method using statistical modeling of color and lane edge-orientation[J]. Advanced Information Sciences and Service Sciences, 2010, 3(2):40-47.

[6] Sun Shi-ran, Ai Si Ka Er AMDL, Liu Wen-Hua. Image retrieval based on the entropy value and the Gabor filter[J]. Laser Journal, 2011,32(2):24-26.

[7] Qingshan Yang, Chengan Guo. Fuzzy ensemble of local Gabor sparse representation classifiers for face recognition[J]. Advanced Information Sciences and Service Sciences, 2011, 10(3):345-354.

[8] Jiang Wei, Chen Hui. New edge detection model based on fractional differential and Sobel operator[J]. Computer Engineering and Application, 2012,48(4):182-185.

[9] Miyazawa M, Peifeng Zeng, et al. A systolic algorithm for euclidean distance transform[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006,28(7):1127-1134.

[10] Ren Jinchang,Jiang Jianmin,Chen Juan.Shot boundary detection in MPEG videos using local and Global indicators[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009,19(8):1234-1238.

[11] Pan Chen-ming, Chuang Yung-yu, Wisnton H.Hsu. NTU TRECVID-2007 fast rushes summarization system[C]. International workshop on TRECVID video summarization, 2007, pp:74-78.