Visual Cognitive Mechanism Guided Video Shot Segmentation

Chenzhi Shao, Haifeng Li(✉), and Lin Ma

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, People's Republic of China
iamscz354@163.com, {lihaifeng,malin_li}@hit.edu.cn

Abstract. Shot segmentation of video sequences is one of the key technologies in video information processing, especially video retrieval. Traditional shot segmentation methods have low detection rate for the gradient shot and the abrupt shot, especially in a single scene. To deal with this problem, this paper proposes a video segmentation method based on visual cognition mechanism. This method proposes a block granularity color histogram to strengthen the visual salient area, and a highlight measure to describe the difference between the front and back frames. This brings great improvements to the accuracy of detecting shot switching in a single scene. In addition, based on the brightness visual perception in video, the difference between adjacent multi-frames in the sliding window is used to capture the brightness change for the gradient shots. Comparing with traditional methods, the proposed algorithm achieves better segmentation effect and has higher precision and recall rate.

Keywords: Video shot segmentation · Visual cognition mechanism · Block color histogram · Long time difference

1 Introduction

In the field of video retrieval, the accuracy of shot segmentation will directly affect the performance of video retrieval systems. Therefore, how to improve the accuracy of shot segmentation is one of the difficult problems in video analysis [1]. In multimedia information, videos can be divided into four different levels of frames, shots, scenes, and video streams according to different granularities. The shot refers to a set of consecutive frames of a camera that is continuous in time and space [2]. So the shot segmentation task is to divide a complete video into segments based on the shots. The shots can be mainly divided into the following two types: abrupt shots and gradual shots. An abrupt shot means that there is no obvious transition between two discrete frames; A gradual shot refers to a transition between two discontinuous frames, such as fade in, fade out, and dissolve. Therefore, our goal is to accurately find the boundaries of the two types of shot switching.

R. Xu et al. (Eds.): ICCC 2019, LNCS 11518, pp. 186–196, 2019.
https://doi.org/10.1007/978-3-030-23407-2_16

2 Related Work

The common shot segmentation method is mainly based on the difference value between two adjacent frames. Jinlai [3] used multi-feature fusion for video shot segmentation. Biswas [4] combined local similarity and global features, and used the method of matrix cosine similarity to detect the shot boundary. Mohanta [5] used the local feature-based frame transition parameters and frame estimation errors to achieve shot segmentation. In order to solve the copyright protection problem of multimedia video, Shang [6] proposed a motion vector based shot segmentation algorithm, and embedded the watermark in a suitable location to better protect the video. Chongke [7] proposed a shot boundary detection framework based on dynamic mode decomposition, which reduced the error detection rate. Baraldi [8] used hierarchical clustering for broadcast video for shot and scene detection. The above method has a high detection accuracy for shot switching in different scenarios, but the detection effect of the shot switching in the same scene is poor. In addition, the above methods have a low detection accuracy for the gradual shots, therefore, it is necessary to improve the accuracy of the shot segmentation.

As one of the most important channels of multimedia video, vision is also the main way of human cognition [9]. Under the constraints of visual physiology and visual sensitivity, the visual sensitivity of human eye to distinguish the details of the object is basically same for the similar things. The vision system is insensitive to absolute brightness, but is sensitive to color contrast. In the process of cognizing images, color plays an important role that can be directly perceived and conveyed by viewers [10]. According to the principle of visual continuity [11], vision tends to perceive continuous forms rather than discrete pieces. People use the same color to make visual recognition of continuous images. When the shot is switched, the color distribution of the video frame has a large difference, so we employ color as an important factor in the determination of the shot segmentation. In addition, according to cognitive psychology research, humans have different degrees of interest in different regions of the video, namely the visual attention mechanism [12]. In order to simulate the degree of interest of the biological vision system in different regions of the video image, this paper studies the influence of the video frame blocking strategy on the detection of the abrupt shots. For the detection of the gradual shots, a significant change occurs in the brightness of the video frame since the post-processing (fading, etc.) is added between the two discontinuous frames. Therefore, this paper studies the gradual shot detection method based on brightness information in view of human visual perception of brightness information [13].

The research content of the paper has the following aspects: 1. Research on the detection method of abrupt shot based on visual cognition mechanism. As for abrupt shot detection, this paper proposed a visual color block histogram detection method, which effectively solved the problem of low accuracy caused by shot switching in the same scene. 2. Research on the detection method of gradual shot based on visual cognition mechanism. In the aspect of gradual shot detection, the paper proposed a long-time difference detection method based on brightness information, which effectively improved the detection accuracy of the gradual shot.

3 Detection Method of Abrupt Shot Boundaries

3.1 Color Histogram Method

When detecting the abrupt shot, we can emply the difference of two continuous frames since the different shots are discontinuous in space and time, and there is no post-processing (fading, etc.) between the two frames of the abrupt shot. The abrupt shots in the video are shown in Figs. 1 and 2.

Fig. 1. Giagram of the abrupt shot A

Fig. 2. Giagram of the abrupt shot B

When we describe the difference between adjacent video frames, the common metric is the color histogram method [14], which is insensitive to camera shake and motion of objects within the shot. The color histogram is a statistical table reflecting the color distribution of an image pixel. The abscissa indicates the interval of each different size, and the vertical axis indicates the percentage of the total number of pixels in the image in a certain interval. It describes the proportion of different colors in the entire image, and does not care about the spatial location of each color. The HSV color space [15] is closer to the way humans feel color, encapsulating information about colors. Therefore, we first convert the original RGB color space to the HSV color space. The HSV model has three parameters: Hue, Saturation, and Value. The hue is the basic property of the color. The saturation refers to the purity of the color. The higher the value, the purer the color. The value means the brightness of the color. The conversion method from RGB color space to HSV color space is as follows:

$$V \leftarrow \max(R, G, B)$$

$$S \leftarrow \begin{cases} \dfrac{V - \min(R,G,B)}{V} & \text{if } V \neq 0 \\ 0 \quad \text{otherwise} \end{cases}$$

$$H \leftarrow \begin{cases} \dfrac{60 * (G - B)}{V - \min(R,G,B)} & \text{if } V = R \\ \dfrac{120 + 60 * (B - R)}{V - \min(R,G,B)} & \text{if } V = G \\ \dfrac{240 + 60 * (R - G)}{V - \min(R,G,B)} & \text{if } V = B \end{cases}$$

$$\text{if } H < 0 \text{ then } H \leftarrow H + 360.$$

$$\text{Output: } 0 \leq V \leq 1, \quad 0 \leq S \leq 1, \quad 0 \leq H \leq 360$$

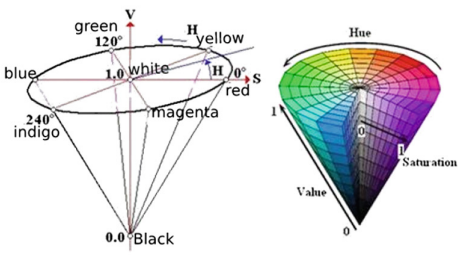

Fig. 3. HSV color model diagram

The formula for calculating the color histogram of an image is shown as follows:

$$H(k) = \frac{n_k}{N}, k = 0, 1, \ldots, 255 \tag{1}$$

In the above formula, n_k represents the number of pixels whose pixel value is k in the image, N refers to the total number of pixels in an image, and H(k) means the distribution of color histograms (Fig. 3).

3.2 Visual Color Block Histogram Strategy

Because of the frame images under different shots are similar in color in a single scene, people usually judge the situation by the region of interest rather than the overall color distribution of the image. Therefore, taking the color histogram of the entire image as a feature will result in a great error. Aiming at the situation, this paper proposed a segmentation method based on visual cognition mechanism, which is block color histogram. The block color histogram means that we divide the video frame firstly, and then compute the color histogram of each block. Finally, each color histogram is weighted to obtain the color distribution of the entire image. In order to highlight the main content of the image and reduce the influence of the background in the image, a larger weight is given to the main body region in the middle of the image, and the remaining background regions are given a smaller weight. In an image, since the four corners and the upper boundary are in the background area, a lower weight is given to the area. Sometimes the characters may occupy a vertical space in the multimedia video, so the left and right sides should be paid more attention. Therefore, the weights are assigned according to the scale shown in Fig. 4. The numbers in the rectangle represent the weight of each small block, while the 1:4:1 in the outer box represents the proportional relationship between the length and width.

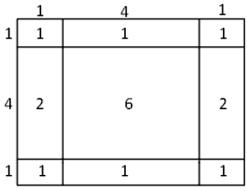

Fig. 4. Video frame block diagram

After the image frames in video are segmented, the color histogram difference of a small block of images in two adjacent frames on a single channel is calculated, as shown in the following formula:

$$d_m(i,j) = \frac{1}{2}\sum_{k=0}^{255} \left|H_{im}(k) - H_{jm}(k)\right| \tag{2}$$

In the above formula, $H_{im}(k)$ represents color distribution of the i-th frame image on the m-th block, $H_{jm}(k)$ refers to color distribution of the j-th frame image on the m-th block. Then, we calculate the color histogram difference of the adjacent two frames on a small block:

$$D_m(i,j) = \frac{1}{3}(d_{Hm}(i,j) + d_{Sm}(i,j) + d_{Vm}(i,j)) \tag{3}$$

Finally, the color histogram difference of the adjacent two frames is shown:

$$T(i,j) = \frac{\sum_{m=1}^{9} w_m * D_m(i,j)}{\sum_{m=1}^{9} w_m} \tag{4}$$

When the color histogram difference value $T(i, j)$ reaches a threshold, it is determined to be an abrupt shot. The choice of the threshold q is adaptive according to different types of video. The specific calculation method is as follows:

$$q = \alpha * \frac{T(1,2) + T(2,3) + \ldots + T(i,i+1) + \ldots T(N-1,N)}{N-1} \tag{5}$$

In the above formula, $T(i, i + 1)$ is the color histogram difference between video frame i and next frame. N refers to total number of frames, α is the coefficient and experiments show that α is suitable for $5 \sim 6$.

Algorithm 1. The Abrupt Shot Detection Algorithm

```
Input: A video to be detected containing N frames
Output: The abrupt shot frame number
1: Read original video frames
2: Convert each frame of RGB image to an HSV image, and
   scale the image proportionally
3: for i = 1 to N-1
    Compute T(i,i+1) according to formula 1~4
    sum = sum + T(i,i+1)
   end
4: q = sum/(N-1)*α
5: for i = 1 to N-1
    if T(i,i+1)>q, print i
   end
```

4 Detection Method of Gradual Shot Boundaries

In the detection of the gradual shot, since the transition forms such as fade in, fade out and dissolve are added between the two frames, the difference between their adjacent frames becomes smaller, which brings challenges to the shot segmentation. The gradual shots in the videos are shown in Figs. 5 and 6.

Fig. 5. Giagram of the gradual shot A

Fig. 6. Giagram of the gradual shot B

Since the brightness information changes between adjacent video frames are small when the gradual shot occurs, it is difficult for us to judge by the difference between the brightness information of adjacent frames. However, since the human visual system perceives regular changes in brightness information within a continuous number of frames, we can capture the change in brightness information. Based on this, this paper proposed a long-time difference method based on brightness to detect gradual shots. In this paper, a sliding window with a length of 6 frames is constructed, which slides from the front to the rear to observe the change of brightness difference of the video frames in the sliding window.

Similarly, we calculate the color histogram difference between adjacent frames:

$$d(i,j) = \frac{1}{2}\sum_{k=0}^{255} \left|H_i(k) - H_j(k)\right| \tag{6}$$

Since we only use the information on the brightness channel when performing gradual shot detection, the difference between adjacent frames can be calculated as:

$$T(i,j) = d_v(i,j) \tag{7}$$

It is assumed that successive frames in a sliding window in the video are frames $i-5, \ldots, i-2, i-1, i$, respectively. This paper first calculates the HSV color histogram of each frame image, and then computes the difference value $T(i, i-1)$, $T(i, i-2)$, $T(i, i-3)$, $T(i, i-4)$, $T(i, i-5)$. If the brightness of the image frames satisfied $T(i, i-1) < T(i, i-2) < T(i, i-3) < T(i, i-4) < T(i, i-5)$, and $T(i, i-1)$ is greater than a certain threshold, it is determined to be a gradual shot. The selection method of the threshold is as shown in the formula 5.

Algorithm 2. The Gradual Shot Detection Algorithm

```
Input: A video to be detected containing N frames
Output: The gradual shot interval (i, j)
1: Read original video frames
2: Convert each frame of RGB image to an HSV image, and
   separate the V channel
3: for i = 2 to N
    Compute T(i, i-1) according to formula 6~7
    sum = sum + T(i, i-1)
   end
4: q = sum/(N-1)*α
5: for i = 6 to N
    if T(i, i-1)<T(i,i-2)<T(i,i-3)<T(i,i-4)<T(i,i-5) &
       T(i,i-1)>q, print (i-5, i)
   end
```

5 Experiments and Results

5.1 Dataset and Evaluation Criteria

Videos to be detected are derived from TRECVID [16], which is the internationally authoritative dataset in the field of video detection. This paper used some of the videos published on the Open Video website. The following is the video information to be detected (Table 1):

Table 1. Test dataset information

Video name	Duration	Frames	Resolution	Shots	Abrupt shot	Gradual shot
anni005.mpg	6'19	11363	320*240	65	38	27
BOR08.mpg	28'07	50568	320*240	531	380	151
anni009.mpg	6'50	12307	320*240	103	39	64

In the task of shot segmentation, the evaluation benchmark is generally Precision and Recall. Precision refers to the percentage of the correct number of detected shots in the total number of detected shots, and the recall is the percentage of the number of correctly detected shots in the actual total number of shots. They are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

In the above formula, *TP* represents the number of correctly detected shots boundaries, *FP* means the number of incorrectly detected shots boundaries, and *FN* refers to the number of undetected shots boundaries. In addition, *F1* value is defined as follows, which is the average of recall and precision.

$$F1 = \frac{2 * P * R}{P + R} \tag{10}$$

5.2 Results and Analysis of Abrupt Shot Detection

When detecting the abrupt shots, the paper compared the traditional color histogram method, the method of other paper and the histogram method based on visual color block proposed in this paper. The experiment results are shown in Tables 2, 3 and 4:

Table 2. Anni005 video abrupt shots results

anni005.mpg	P	R	F1
Traditional color histogram method	0.86	0.89	0.88
Other paper [17]	0.91	0.85	0.88
Proposed	**0.97**	**0.95**	**0.96**

Table 3. BOR08 video abrupt shots results

BOR08.mpg	P	R	F1
Traditional color histogram method	0.84	0.89	0.86
Other paper [17]	0.92	0.88	0.90
Proposed	**0.98**	**0.96**	**0.97**

Table 4. Anni009 video abrupt shots results

anni009.mpg	P	R	F1
Traditional color histogram method	0.64	0.64	0.64
Other paper [17]	0.84	0.76	0.80
Proposed	**0.83**	**0.90**	**0.86**

As can be seen from the above experimental results, video frames are segmented and given a large weight to the main area in the abrupt shot detection, which increase the difference between discontinuous frames and improves the effect of shot segmentation. Especially in the block before and after, that is, the traditional color histogram method and the method in the paper on the comparison of experimental results, we can clearly see the advantages of the block color histogram proposed in this paper.

5.3 Results and Analysis of Gradual Shot Detection

When detecting the gradual shots, the paper compared the traditional color histogram method, the method of other paper and the long-difference method based on brightness information proposed in this paper. The experiment results are shown in Tables 5, 6 and 7.

Table 5. Anni005 video gradual shots results

anni005.mpg	P	R	F1
Traditional color histogram method	0.62	0.48	0.54
Other paper [17]	0.64	0.53	0.58
Proposed	**0.82**	**0.67**	**0.73**

Table 6. BOR08 video gradual shots results

BOR08.mpg	P	R	F1
Traditional color histogram method	0.52	0.43	0.47
Other paper [17]	0.46	0.64	0.53
Proposed	**0.70**	**0.71**	**0.71**

Table 7. Anni009 video gradual shots results

anni009.mpg	P	R	F1
Traditional color histogram method	0.45	0.43	0.44
Other paper [17]	0.59	0.62	0.60
Proposed	**0.83**	**0.63**	**0.72**

As can be seen from the above experimental results, in the detection of gradual shots, due to the small difference between the two frames, it is difficult to detect shot boundaries and recognition rate is lower than that of the abrupt shots. However, the long-time difference strategy based on brightness information proposed in the paper has a significant improvement over the traditional color histogram method. In addition, the employment of brightness information can also have a good tolerance for camera shake and motion of objects within a shot. However, in the detection of gradual shots, the scenes (sunset, etc.) in which the brightness gradually changes are hard to detect, so there is still room for improvement.

6 Conclusions

Shot segmentation task is the premise of later video information retrieval, so the accuracy of shot segmentation will directly affect the performance of video retrieval system. The paper first introduced the visual cognitive mechanism and human perception of color. Then, based on the visual cognitive mechanism, a histogram method

of visual color segmentation was proposed. This method strengthens the region of interest of biological vision system, weakens the influence of background, and improves the accuracy of abrupt shot detection. On the other hand, a long-time difference method based on brightness information was proposed to detect the gradual shots. The method improves the detection accuracy by capturing the perception rule of human vision for brightness. In addition, the paper used the method of window sliding, which is easy to achieve. Compared with other shot segmentation methods, the precision and recall are improved. The method can be used as an effective shot segmentation strategy. Due to the high computational complexity when comparing the difference between adjacent frames, variable step size method will be considered to speed up the shot segmentation in the future.

Acknowledgements. Our thanks to supports from the National Key Research and Development Program of China (2018YFC0806800), National Natural Science Foundation of China (61671187), Shenzhen Foundational Research Funding (JCYJ20150929143955341), Shenzhen Key Laboratory of Innovation Environment Project (ZDSYS201707311437102), Open Funding of MOE-Microsoft Key Laboratory of Natural Language Processing and Speech (HIT.KLOF.20150xx, HIT.KLOF.20160xx). The authors are grateful for the anonymous reviewers who made constructive comments.

References

1. Kumar, G.S.N., Reddy, V.S.K., Srinivas Kumar, S.: Video shot boundary detection and key frame extraction for video retrieval. In: Bhateja, V., Tavares, J.M.R.S., Rani, B.P., Prasad, V. K., Raju, K.S. (eds.) Proceedings of the Second International Conference on Computational Intelligence and Informatics. AISC, vol. 712, pp. 557–567. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8228-3_51
2. Xing, Y., et al.: Multi-scale shot segmentation based on weighted subregion color histogram. J. Inf. Hiding Multimed. Sig. Process. **6**(3), 622–628 (2015)
3. Lv, J., Bai, H.: Research on shot detection algorithm of self-adaptive dual thresholds based on multi-feature fusion. In: Pan, Z., Cheok, A.D., Müller, W., Zhang, M. (eds.) Transactions on Edutainment XIII. LNCS, vol. 10092, pp. 247–261. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-662-54395-5_21
4. Biswas, S.K., Milanfar, P.: One shot detection with laplacian object and fast matrix cosine similarity. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 546–562 (2016)
5. Mohanta, P.P., Saha, S.K., Chanda, B.: A model-based shot boundary detection technique using frame transition parameters. IEEE Trans. Multimed. **14**(1), 223–233 (2012)
6. Shang, G., et al.: Video watermark algorithm study of shot segmentation based on motion vector. In: Proceedings of the 2018 International Conference on Information Hiding and Image Processing. ACM (2018)
7. Bi, C., et al.: Dynamic mode decomposition based video shot detection. IEEE Access **6**, 21397–21407 (2018)
8. Baraldi, L., Grana, C., Cucchiara, R.: Shot and scene detection via hierarchical clustering for re-using broadcast video. In: Azzopardi, G., Petkov, N. (eds.) CAIP 2015. LNCS, vol. 9256, pp. 801–811. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23192-1_67
9. Monty, R.A., Fisher, D.F., Senders, J.W.: Eye Movements: Cognition and Visual Perception. Routledge, Abingdon (2017)

10. Schloss, K.B., et al.: Color inference in visual communication: the meaning of colors in recycling. Cogn. Res. Principles Implications **3**(1), 5 (2018)
11. Peteranderl, Sonja, Oberauer, Klaus: Serial recall of colors: two models of memory for serial order applied to continuous visual stimuli. Mem. Cogn. **46**(1), 1–16 (2018)
12. Li, N., Zhao, X., Ma, B., Zou, X.: A visual attention model based on human visual cognition. In: Ren, J. (ed.) BICS 2018. LNCS (LNAI), vol. 10989, pp. 271–281. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00563-4_26
13. Xing, D., et al.: Brightness–color interactions in human early visual cortex. J. Neurosci. **35**(5), 2226–2232 (2015)
14. Jeong, S., Won, C.S., Gray, R.M.: Image retrieval using color histograms generated by Gauss mixture vector quantization. Comput. Vis. Image Underst. **94**(1–3), 44–66 (2004)
15. Chen, T.-W., Chen, Y.-L., Chien, S.-Y.: Fast image segmentation based on K-means clustering with histograms in HSV color space. In: 2008 IEEE 10th Workshop on Multimedia Signal Processing. IEEE (2008)
16. Smeaton, A.F., Over, P., Doherty, A.R.: Video shot boundary detection: seven years of TRECVid activity. Comput. Vis. Image Underst. **114**, 411–418 (2010)
17. Li, Z., Liu, X., Zhang, S.: Shot boundary detection based on multilevel difference of colour histograms. In: 2016 First International Conference on Multimedia and Image Processing (ICMIP). IEEE (2016)