Automatic region of interest determination in music videos

Wonjun Kim and Changick Kim
School of Engineering, Information and Communications University
119 Munji Street, Yuseong-Gu, Daejeon, 305-714, Republic of Korea
{ jazznova, ckim }@icu.ac.kr

***Abstract*-With the advent of third generation (3G) mobile devices, there is an increasing demand for intelligent display techniques to deliver better viewing experience to small mobile devices users. In this paper, we propose a novel framework for region of interest (ROI) determination in music videos. The most important issue is how to determine ROI under various illumination conditions. We employ the gray-world assumption to estimate and remove the light source color and reconstruct the color space by using hue information. Based on the reconstructed color space, we propose a new method to determine ROI by using the model of attractive force. The proposed method is tested on various music videos to confirm the efficiency and effectiveness of the proposed method.**

I. INTRODUCTION

This paper presents a novel framework for region of interest (ROI) determination in music videos. With the advent of new mobile services such as terrestrial and satellite DMB (Digital Multimedia Broadcasting), music videos have become one of popular contents consumed on mobile devices. Music video is composed of singers, band, and audience under various illuminations. When people watch the music video on their mobile devices, they tend to focus on the singer generally. However, if the size of the singer is very small on a mobile-device screen, it may be difficult to recognize the singer in the video and lead the viewers to uncomfortable experiences. Here this problem can be solved efficiently by using the intelligent display technique, which is based on ROI.

Determining ROI benefits in the applications of context-aware content adaptation, transcoding, intelligent information management, and so on. In addition, it can be a first step to semantic level interpretation of a video scene. There have been a number of approaches for efficient display by using ROI [1-5]. In [1], they define ROI based on the position of soccer ball. Through the simple method to find ball by using the intensity constraints, ROI is efficiently displayed on small screen. Kimura *et al.* [2] assume that the center of image is more important than other parts of image. They use the three features to extract ROI, which are brightness, contrast, and view angle. Cheng *et al.* [3] consider ROI for content recomposition on small display. They propose a high-level combination strategy to avoid the shape distortion of objects caused by changes in video aspect ratio. ROI is used popularly to analyze medical images. In [4], they use fuzzy and active contour models to extract ROI from motion affected MRI images. Liu *et al.* [5] present a local weak form geometric active contour to extract ROI from medical images, which is based on the multi-scale parametric and geometric deformable models. In addition, since coarse coding for non-ROI can be possible, ROI is widely used for efficient video coding [6-7].

In case of music video, skin color must play a significant role since a singer tends to get a main interest. It is observed that the skin color is very sensitive to illumination condition. In past years, many researchers have focused on color constancy problem [8-11]. Finlayson *et al.* [8] build a correlation matrix based on probability distribution with reference illuminations. In [9], authors distinguish color compensation with color constancy problem and show the comparison results of methods. In this paper, we employ the gray-world assumption that the average reflectance in a scene is achromatic [10]. Although more elaborate algorithms exist, gray-world is still widely used due to the low computational costs. The light source color is estimated and removed by using the gray-world assumption. Then we reconstruct the color space by using hue information and utilize the attractive force model based on the reconstructed color space to determine ROI.

The rest of this paper is organized as follows. In Section 2, the proposed method, which is composed of five steps, is explained in detail. The experimental results on various music videos are shown in Section 3, followed by conclusion in Section 4.

II. THE PROPOSED METHOD

Our goal is to define and magnify ROI for better viewing experience in music videos. The work flow is shown in Fig. 1. Each module in Fig. 1 is explained in the following subsections.

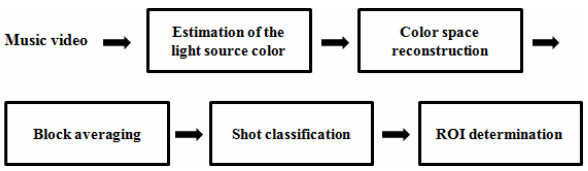

Fig. 1. System overview.

This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA (IITA-2007-C1090-0701-0017).

A. Estimation of the light source color

Since the illumination emitting color light in music stages shifts all scene colors toward the color of illumination, it is hard to apply useful clues such as skin color for ROI determination in music videos. In order to estimate the light source color, we employ the gray-world assumption that the average reflectance in a scene is achromatic. The image values for a Lambertian surface can be represented as follows [10]:

$$\mathrm{f} = \int_{\omega} e(\lambda) s(\lambda) \mathrm{c}(\lambda) d\lambda \tag{1}$$

where $e(\lambda)$, $s(\lambda)$ denote the light source and surface reflectance with the wavelength λ . $\mathrm{c}(\lambda)$ denotes the camera sensitivity functions, which is dependent on RGB components with the wavelength λ . ω is the visible spectrum. We estimate the light source color $e(\lambda)$ to remove various illuminations in music videos.

$$e = \begin{pmatrix} R_e \\ G_e \\ B_e \end{pmatrix} = \int_{\omega} e(\lambda) \mathrm{c}(\lambda) d\lambda \tag{2}$$

In the gray-world assumption, the average reflectance in a scene can be represented as (3) because it is achromatic.

$$\frac{\int s(\lambda, \mathrm{x}) d\mathrm{x}}{\int d\mathrm{x}} = k \tag{3}$$

Here x denotes the spatial coordinate in the image. Based on (2) and (3), we can simply estimate the light source color $e(\lambda)$ by computing the average pixel value as follows.

$$\begin{aligned} \frac{\int \mathrm{f}(\mathrm{x}) d\mathrm{x}}{\int d\mathrm{x}} &= \frac{1}{\int d\mathrm{x}} \int_{\omega} \int e(\lambda) s(\lambda, \mathrm{x}) \mathrm{c}(\lambda) d\lambda d\mathrm{x} \\ &= k \int_{\omega} e(\lambda) \mathrm{c}(\lambda) d\lambda = k\mathrm{e} \end{aligned} \tag{4}$$

B. Color space reconstruction

In this subsection, we use hue information to eliminate the light source color from the scene. The light source color can be eliminated by adding complements on the color circle to all pixels in the scene. However, since adding complements causes increment of the intensity of scene, we take only the ratio among RGB values from added results in each pixel. Then the ratio in each pixel is applied to original color space as follows:

$$\begin{aligned} R_R &= (R_O + G_O + B_O) \times R_{ratio} \\ G_R &= (R_O + G_O + B_O) \times G_{ratio} \\ B_R &= (R_O + G_O + B_O) \times B_{ratio} \end{aligned} \tag{5}$$

where R_R, G_R, B_R denote the RGB components of reconstructed color space while R_O, G_O, B_O denote the RGB components of original color space affected by various illuminations. $R_{ratio}, G_{ratio}, B_{ratio}$ denote the ratio between the modified each color component RGB, which can be represented by the sum of original component and complements on the color circle, and the sum of modified each color component RGB. Consequently, we can reconstruct the color space by removing the light source color from the scene. The result of color space reconstruction is shown in Fig. 2.

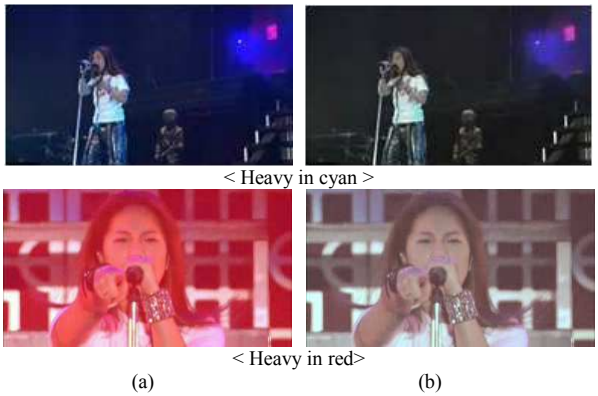

Fig. 2. Result of color reconstruction. (a) Original color. (b) Reconstructed color.

C. Block averaging

The goal of this module is to reduce the processing time by taking block-based approach each scene is divided into 20 × 15 blocks after color space reconstruction and then averaging process is applied to each block. To reduce the noise, a Gaussian filter is used before the block averaging.

D. Shot classification

Since people do not feel uncomfortable in close-up shots generally in spite of the small size of screen on mobile devices, ROI is used only for long shot. In this paper, the skin color detection algorithm is employed to classify the shot type because the number of skin color blocks in close-up shots and long shots are different. There have been many approaches on the skin color detection under varying illumination conditions. They use many algorithms to cope with various illumination conditions, such as lighting compensation [12], gamma correction [13-14], skin locus [15], and so on. Unlike previous work, we use the reconstructed color space proposed in the previous subsection to detect skin color under various illumination conditions. The HSI color space is employed, which is widely used for skin color detection [16-17]. The ranges used in this paper for skin color detection are (10, 60) for hue and (80, 255) for intensity. Early removal of non-skin color blocks is conducted by using the intensity range constraint to reduce the processing time, and then the hue constraint is applied to the remaining blocks only. We take

only three largest connected components of skin color blocks into account to minimize false shot classification. If the number of skin color blocks in the three largest connected components is larger than a pre-defined threshold value, the scene is classified into close-up shot. Otherwise, the scene is classified into long shot. The threshold value is set to 20 in this work.

(a) (b)

Fig. 3. Result of shot classification. (a) Long shot. (b) Close-up shot.

The result of shot classification by using skin color blocks is shown in Fig. 3. Detected skin color blocks on reconstructed color space are represented by green color.

E. ROI determination

Although skin color is very useful attribute in determining the shot type, it is not easy to determine the location of ROI in the long shot because a number of small skin color blocks can be scattered in the long shot frames. Since it is observed that ROI is highlighted by bright illumination, we can detect ROI candidates by using the intensity attribute. The proposed procedure for ROI determination is as follows. First, we define the ROI window which consists of 14 x 11 blocks with regard to the size of image in this paper. The example of ROI window is shown in Fig. 4.

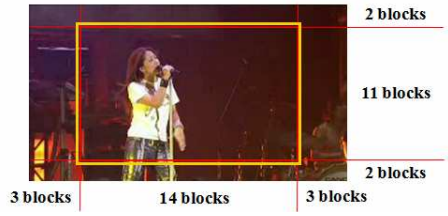

Fig. 4. The example of ROI window.

Then ROI candidate blocks (*CB*) are determined by using the average intensity of ROI window as follows:

$$\begin{cases} CB = 1, \ if \ ABI > m + \sigma \\ CB = 0, \ otherwise \end{cases} \quad (6)$$

where *ABI* denotes the average block intensity. m and σ denote average intensity of ROI window and corresponding standard deviation, respectively. ROI candidate blocks are represented as green color in Fig. 5. For the intelligent display by using ROI, the center of ROI needs to be matched with the singer position. The center of ROI moves along the horizontal direction only at the position shown in Fig. 4. We find the center of ROI window by using attractive force model. Attractive force between two objects can be obtained from (7),

$$F = \frac{m_1 \times m_2}{\alpha} \quad (7)$$

where m_1 denotes the mass at the center of image which is fixed 1 and m_2 denotes the number of ROI candidate blocks in each block column, respectively. α denotes the weight factor related to distance between the center of image and corresponding block column. The value of α decreases sharply as it is away from the center of image. With the distance between the center of image and ROI candidate blocks, the different weight is applied as follows.

Case 1) $d \leq 1 \times BW \rightarrow$ weight = 1.0
Case 2) $1 \times BW < d \leq 3 \times BW \rightarrow$ weight = 0.5
Case 3) $3 \times BW \leq d \rightarrow$ weight = 0.125

Here *BW* denotes the width of block. Finally, the center of block column with the strongest attractive force is determined as the center of ROI window. The result of ROI determination is shown in Fig. 5. We can see that most ROI block candidates are located at the left part of image. Although the dominant block column is far away from the center of image, it can be determined as the center of ROI window based on the fact that the number of ROI candidate blocks is much larger than other parts. The value of attractive force is shown in Table I.

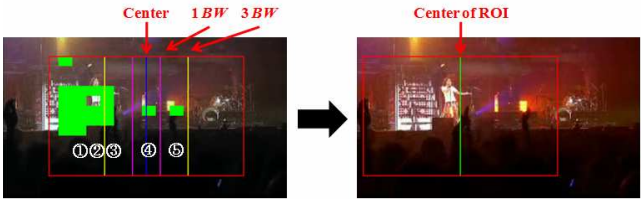

Fig. 5. The process of ROI determination.

TABLE I
THE VALUE OF ATTRACTIVE FORCE

Block column	①	②	③	④	⑤
Attractive force	0.625	0.375	**2.0**	1.0	0.5

We can see that the block column ③ is determined as the center of ROI window.

III. EXPERIMENTAL RESULTS

The framework for evaluating performance has been implemented by using Visual Studio 2003 (C++) under FFMpeg library, which has been utilized for MPEG decoding. All music videos used in this paper are encoded with image size of 288 x 160 and composed of more than 1000 frames,

respectively. Results of detected and magnified ROI are shown in Fig. 6. If the difference between previous position of ROI center and current position of ROI center is less than 2 block width, ROI window dose not move to prevent frequent change of ROI position. Although there are various illuminations in a scene, the center of ROI is matched with singer well.

(a) (b)

Fig. 6. Result of ROI determination. (a) Detected ROI. (b) Magnified ROI. ROI window and center are represented by using red rectangle and green line, respectively in (a).

The performance of proposed method is shown in Table II. The accuracy for ROI center and shot classification are estimated, which can be computed as follows:

$$ERC(\%) = \frac{\#\text{ of false detected ROI center}}{\#\text{ of total detected ROI center}} \times 100$$
$$ESC(\%) = \frac{\#\text{ of false detected shot}}{\#\text{ of long(or short) shots}} \times 100 \quad (7)$$

where *ERC* and *ESC* denote error for ROI center and error for shot classification, respectively. Since the illumination conditions change frequently in video 1, the value of *ERC* and *ESC* in video 1 is larger than that of video 2 slightly. The result of evaluating on processing time of two music videos is also shown in Table II. Since the ROI determination is not applied to the close-up shot after shot classification, the average processing speed of music video including more close-up shots is fast.

TABLE II
THE EVALUATION OF PERFORMANCE

Test videos	Video 1	Video 2
Total frames	1247 frames	1129 frames
ERC	48/568 x 100 = 8.45%	58/800 x 100 = 7.25%
ESC (Long shot)	14/568 x 100 = 2.46%	38/800 x 100 = 4.75%
ESC (Close-up shot)	53/679 x 100 = 7.81%	1/329 x 100 = 0.30%
Average processing time	28.36 frame/sec	27.55 frame/sec

IV. CONCLUSION

The efficient and robust method for ROI determination is proposed in this paper. The gray-world assumption is used to estimate the light source color on the basis that the average reflectance in a scene is achromatic. Then we use hue information of the light source color to reconstruct the color space. After color space reconstruction, block-based approach is taken to reduce the processing time. Based on the reconstructed color space, all shots in a music video are divided into long and close-up shot by using skin color detection. Finally, the center of ROI is determined by using the attractive force model, which is composed of distance from the image center and the number of ROI candidate blocks.

The proposed method is tested on various music videos. According to our analysis, the proposed method can be easily applied to various music videos for better viewing experience on mobile devices by using magnified ROI. Since the proposed algorithm runs almost real-time, this technique can be used for real-time mobile applications.

REFERENCES

[1] K. Seo, J. Ko, I. Ahn, and C. Kim, "An intelligent display scheme of soccer video on mobile devices," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 10, pp. 1395-1401, Oct. 2007

[2] M. Kimura and M. Yamauchi, "A method for extracting region of interest based on attractiveness," *IEEE Trans. Consumer Electronics*, vol. 52, no. 2, pp. 312-316, May 2006

[3] W. H. Cheng, C. W. Wang, and J. L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.* vol. 17, no. 1, pp. 43-58, Jan. 2007

[4] C. Weerasinghe, L. Ji, and H. Yan, "ROI extraction from motion affected MRI images based on fuzzy and active contour models," in *Proc. ICASSP'99,* vol. 6, pp. 3405-3408, Mar. 1999

[5] H. F. Liu, H. P. Ho, and P. C. Shi, "Local weak form geometric active contours for medical image segmentation," in *Proc. Biomedical Imaging 2004*, vol. 1, pp. 189-192, April 2004

[6] S. Fukuma, S. Ikuta, M. Ito, S. Nishimura, and M. Nawate, "An ROI image coding based on switching wavelet transform," in *Proc. ISCAS'03*, vol. 2, pp. 25-28, May 2003

[7] Y. Xie and G. Q. Han, "ROI coding with separated code block," in *Proc. Machine Learning and Cybernetics*, vol. 9, pp. 18-21, Aug. 2005

[8] G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: a simple, unifying, framework for color constancy," *IEEE Trans. Pattern anal. Mach. Intell.*, vol. 23, no. 11, pp. 1209-1221, Nov. 2001

[9] J. Kovac, P. Peer, and F. Solina, "Illumination independent color-based face detection," in *Proc. Image and Signal Processing and Analysis*, vol. 1, pp. 18-20, Sept. 2003
[10] J. V. Weijer and Th. Gevers, "Color constancy based on the grey-edge hypothesis," in *Proc. ICIP 2005*, vol. 2, pp. 11-14, Sept. 2005
[11] B. Funt, K. Barnard, and L. Martin, "Is machine colour constancy good enough ?," *Lecture Notes Comput. Sci*., vol. 1406, pp. 445-459, 1998
[12] R. L. Hsu, M. A. Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern anal. Mach. Intell*., vol. 24, no. 5, pp. 696-706, May 2002
[13] H. C. Do, J. Y. You, and S. I. Chien, "Skin color detection through estimation and conversion of illuminant color using sclera region of eye under varying illumination," in *Proc. Pattern Recognition*, vol. 1, pp. 327-330, 2006
[14] J. Yang, Z. Fu, T. Tan, and W. Hu, "Skin color detection using multiple cues," in *Proc. Pattern Recognition*, vol. 1, pp. 23-26, Aug. 2004
[15] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen, "Skin detection in video under changing illumination conditions," in *Proc. Pattern Recognition*, vol. 1, pp. 3-7, Sept. 2000
[16] M. J. Chen, M. C. Chi, C. T. Hsu, and J. W. Chen, "ROI video coding based on H.263+ with robust skin-color detection technique," *IEEE Trans. Consumer Electronics*, vol. 49, no. 3, pp. 724-730, Aug. 2003
[17] T. Sawangsri, V. Patanavijit, and S. Jitapunkul, "Face segmentation based on Hue-Cr components and morphological technique," in *Proc. ISCAS 2005*, vol. 6, pp. 5401-5404, May 2005