Automatic Video Region-of-Interest Determination Based on User Attention Model

Wen-Huang Cheng
Graduate Institute of Networking and Multimedia
National Taiwan University
wisley@cmlab.csie.ntu.edu.tw

Wei-Ta Chu
Dept. of Computer Science and Information Engineering
National Taiwan University
wtchu@cmlab.csie.ntu.edu.tw

Jin-Hau Kuo
Dept. of Computer Science and Information Engineering
National Taiwan University
david@cmlab.csie.ntu.edu.tw

Ja-Ling Wu
Graduate Institute of Networking and Multimedia
National Taiwan University
wjl@cmlab.csie.ntu.edu.tw

***Abstract*—This paper presents a framework for automatic video region-of-interest determination based on user attention model. In this work, a set of attempts on using video attention features and knowledge of applied media aesthetics are made. Three types of visual attention features we used are intensity, color, and motion. Referring to aesthetic principles, these features are combined according to camera motion types on the basis of a newly proposed video analysis unit, frame-segment. We conduct subjective experiments on several kinds of video data and demonstrate the effectiveness of the proposed framework.**

I. INTRODUCTION

With the amazing growth in the amount of multimedia documents, one of the key technologies of content analysis is the region-of-interest (ROI) determination [1]. An ROI is a portion of a multimedia document that audiences show more interest in or pay more attention to than others; it provides end users a more concise and informative representation of multimedia contents.

According to psychological findings about the primate visual system and eye fixation, quite a few vision models for still images have been developed to simulate the cognitive mechanism of human beings. One well-known approach is based on Itti's user attention model [2], in which several spatial visual features are combined into a single saliency map for representing local conspicuity in images. This model has been extensively studied in many fields and was shown to be robust in intelligent processing of digital images. However, due to the ignorance of temporal aspects, its extension to moving pictures needs to be explored. Some approaches for analyzing video ROI are then proposed. However, video's specific characteristics have not been taken into account properly till now. In summary, the difficulties associated with conventional video ROI determination, based on user attention model, can roughly be divided into two categories. The first one is the lack of temporal and motion information, and the second is that fixed or video-genre-based feature combining method seems to be problematic for practical use.

In this work, we consider the problem from the viewpoint of applied media aesthetics [5]. Our goal is to develop a framework that can be used to determine the video ROI using computable visual features. In addition to light and color, object motion is adopted as one visual feature in our attention model. Rather than a single frame, we choose a short video clip, i.e. a frame-segment, as the basic unit for conducting video analysis. A camera-motion assisted algorithm for combining visual features is developed and integrated to the framework. We conduct lots of experiments on kinds of video data and demonstrate the effectiveness of the proposed framework in video ROI determination.

The rest of this paper is organized as follows. Section 2 presents the proposed framework for video ROI determination. User attention representation and camera motion utilization are described in Section 3. Section 4 discusses the dynamic ROIs determination from a saliency map. Section 5 shows experimental results, and Section 6 presents our concluding remarks.

II. AN OVERVIEW OF THE PROPOSED FRAMEWORK

The block diagram of the proposed framework is illustrated in Figure 1. The input video is first segmented by a reliable shot boundary detection algorithm [6], which can correctly detect abrupt shot changes and gradual transitions. Further, each shot is partitioned into non-overlapped "frame-segment" (will be explained in Section 3). For each frame-segment, one camera motion type is registered. This camera motion information will be used to generate the saliency map later. Meanwhile, the corresponding feature maps generated from each feature models (e.g., including intensity contrast, color contrast, and motion) are computed. By taking account of the camera motion types, different kinds of feature maps are combined elaborately. Finally, the integrated saliency map is constructed. The video ROIs are then dynamically estimated according to the distribution of saliency values.

This work was partially supported by the CIET-NTU(MOE) and National Science Council of R.O.C. under the contract No. NSC93-2622-E-002-033, NSC93-2752-E-002-006-PAE and NSC93-2213-002-006.

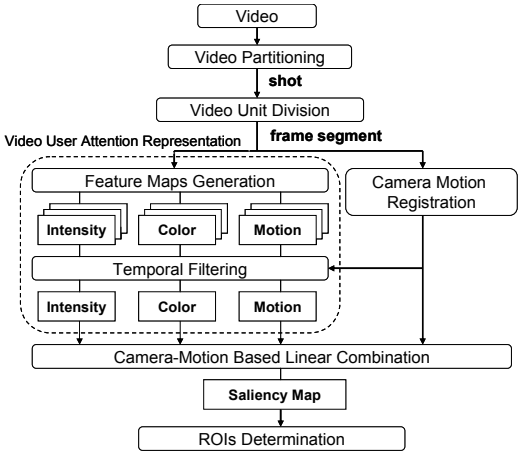

Figure 1. Block diagram of the proposed framework for conducting video ROIs determination.

III. User Attention Representation

A. User Attention Model

User attention refers to the ability of a viewer concentrating his attention on some visual objects or regions. Previous researches showed that this physiological process could be modeled by the so-called user attention model [2, 3]. In our work, three types of video-oriented visual features (i.e. intensity, color, and motion) are adopted to model the visual attraction of videos by using the same idea.

1) Contrast Based Intensity and Color Feature Model

In psychology, perceptual experiments have showed that some color pairs, such as red-green and blue-yellow, possess high spatial and chromatic opposition. The same characteristics exist in high difference lighting or intensity pairs. Based on these observations, we include three contrast based feature models, including intensity, red-green color, and blue-yellow color contrast models, into our user attention representation module.

2) Motion Feature Model

The motion of objects plays an essential role in a video. Two feature models, i.e. x-motion and y-motion models, are used to represent the motion information of a video frame. The x-motion and the y-motion refer to the horizontal and the vertical movements of a specific pixel within a frame, respectively. The 2-D structure tensor [8] is used to compute each pixel's motion magnitude in a specific direction.

B. Frame-segment

In previous researches, user attention is modeled and determined mostly for only a single, at most, for two consecutive frames. The collection of determined regions of each independent single frame composes the final ROIs of a video sequence. However, based on our previous observation [4], we found that the single- or two-frame based approach only generates acceptable results for image but not for video ROI analysis. For example, the focus point may swiftly tremble due to a slight difference between two consecutive frames. This unpleasant phenomenon does not exist in viewers' attention. If the estimated ROIs are applied to other extended applications, such as scalable coding and content adaptation, the prescribed defect will cause significant deficiency in both bit rate and quality.

Due to the fact that the content of a video would not change drastically in a short duration, we take a short video clip, called frame-segment, as the unit for conducting the video ROI analysis. The newly defined frame-segment takes both spatial and temporal correlations into account and can suppress noises caused by sudden luminance changes, such as flashlights. In our experiments, the length of a frame-segment is empirically set to 0.5 seconds.

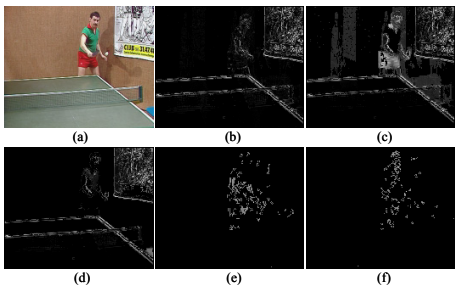

Figure 2. Example of feature maps. (a) original video frame, (b) intensity, (c) red-green color, (d) blue-yellow color, (e) x-motion, and (f) y-motion feature maps.

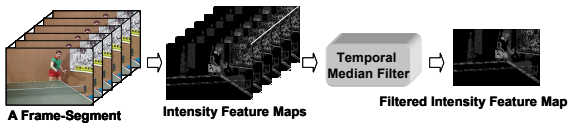

Figure 3. The procedure of generating the filtered intensity feature map.

C. Feature Map and Filtered Feature Map

For each frame-segment, the distributions of each of the features are calculated and constructed as five feature maps, as shown in Figure 2. A temporal median filter, as shown in Figure 3, is then applied to find the filtered feature map of each specific feature. Each filtered feature map represents the general characteristics of a specific feature in a frame-segment. In other words, both the spatial saliency distribution and the temporal saliency variation of a video are used to model the visual attraction of videos.

D. Camera Motion Based Saliency Map Generation

1) Relations between Camera Motion and User Attention

Nowadays, a large amount of videos are produced according to the principles of applied media aesthetics, especially the *expert-produced* videos [5]. From the viewpoint of video shooting, different camera motions have different impacts on the audience's reception. In other words, they influence the relative importance of each visual feature and reveal what and where the video-maker wants

viewers to see. The idea has been extensively used in film or TV show productions. Therefore, it is our belief that camera movement should be considered in the process of ROI determination. Figure 4 gives a real example. If you look at one of the video frames intensively, your focus will soon be attracted by the player who is controlling the basketball. However, if you look at these frames rapidly in succession, you will find that your eyesight moves right with the camera and will not concentrate yourself on the players anymore.

In our work, seven camera motion types are registered: zoom, left-pan, right-pan, up-tilt, down-tilt, static-with-no-motion, and static-with-object-motion. The spatio-temporal slices and tensor histograms based motion analysis techniques [8] are used to register the camera motion type of each frame-segment.

Figure 4. Demonstrations of user attention under right-pan camera motion. The video frames (a) to (d) are captured in an interval of 0.5 seconds from a TV sports program.

TABLE I. WEIGHTS FOR FILTERED FEATURE MAPS UNDER DIFFERENT CAMERA MOTION TYPES.

	Intensity	RG color	BY color	X-motion	Y-motion
Zoom	0.2	0.2	0.2	0.2	0.2
L/R-Pan	0.05	0.05	0.05	0.75	0.1
U/D-Tilt	0.05	0.05	0.05	0.1	0.75
Static	0.15	0.075	0.075	0.35	0.35
Motion	0.05	0.05	0.05	0.425	0.425

2) Saliency Map Generation

Weights of the filtered feature maps for combining the generic saliency map are decided according to the registered camera motion type. The generic saliency map is generated according to the following equation.

$$S(N) = \alpha_{c,1} \times FFM_1 + \alpha_{c,2} \times FFM_2 + \cdots + \alpha_{c,n} \times FFM_n, \quad (1)$$

where $S(N)$ is the generated generic saliency map of a frame-segment with length N. FFM_i is the i-th filtered feature map of that segment, and $\alpha_{c,i}$ is the weight of the corresponding FFM_i under given camera motion type c. Table I shows the weights for various camera motion types and filtered feature maps used in our framework. These weights are defined elaborately to present characteristics of different camera motion types. For example, when camera panning occurs, the horizontal motion should be emphasized.

IV. VIDEO ROI DETERMINATION

Although the saliency maps have showed the ability to characterize the visual attraction of a video, the generated ROIs still have the probability of failure in capturing the essence if they are not determined properly according to these maps. In this work, the appropriate position and size of an ROI is determined by the regular moments [7]. Since there may be multiple key objects in a frame-segment, a method for dynamically determining the number of ROIs will then be presented.

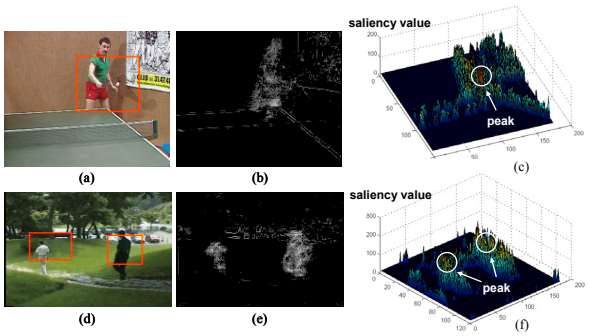

Figure 5. Examples of frame-segments with (a) one and (d) two ROIs. (b) and (e) are the corresponding saliency maps, and (c) and (f) are the 3-D profiles of the saliency maps of (a) and (d), respectively.

A. Saliency Weighted Regular Moment

We use the saliency weighted regular moments [7] to determine the centroid and the size of each ROI. It follows.

$$m_{pq} = \sum_{1}^{M}\sum_{1}^{N} x^p y^q s(x,y), \quad p,q = 0,1,2,..., \quad (2)$$

$$\eta_{pq} = \frac{\sum_{1}^{M}\sum_{1}^{N} (x-\bar{x})^p (y-\bar{y})^q s(x,y)}{m_{00}^{(p+q+2)/2}}, \ \bar{x} = \frac{m_{10}}{m_{00}}, \ \bar{y} = \frac{m_{01}}{m_{00}}, \quad (3)$$

where M, N, are the dimensions of the saliency map and $s(x, y)$ is the saliency value function corresponding to the pixel (x, y). In the saliency map, the centroid (x, y) of an ROI is determined as $(\bar{x}, \bar{y}) = (m_{10}/m_{00}, m_{01}/m_{00})$, and its region size $m \times n$ is set as $(k\sqrt{\eta_{20}}) \times (k\sqrt{\eta_{02}})$, where $k = 2$.

When determining the ROIs, we observe that if those saliency points are clustered around a concentrated area, they generate a small ROI. It implies that an obvious attentive region exists. Contrarily, if those saliency points are scattered across the saliency map, it implies that there is no obvious attentive region and the region size will be very large. In this case, even one ROI is claimed, actually, it implies no apparent region attracts the audience's attention.

B. Dynamic Determination of ROIs

Sometimes, there are more than one ROI in a frame-segment. For example, in a distance view of a tennis game, two players may form two different ROIs. We devised a method to resolve this problem explicitly. In a saliency map, each ROI usually consists of a set of saliency values peaked at the center of its 3-D profiles. For example, if a frame-segment has two ROIs, its saliency map usually has two separate peaked sets, as shown in Figure 5(f).

We assume that the saliency value ranges from 0 to R (in this work, R is 255). In each saliency map, if a pixel's

saliency value is greater than a predefined threshold, it is added to the peak set (*PS*). The pixels in the *PS* are further clustered basing on the Euclidean distances among them, and are then divided into several disjoint subsets. That is,

$$PS = \bigcup_{i=1}^{n} PS_i, \quad when \ PS_i \cap PS_j = \varnothing \ if \ i \neq j. \tag{4}$$

In this way, a saliency map is divided into *n* regions, and each region corresponds to a peak subset PS_i. One ROI is declared for each region. With this scheme, the number of ROIs can be determined dynamically and automatically for each frame-segment. However, normally, the number of ROIs in a frame will be no more than three. If there are more than three key objects in a frame, the viewer may be confused and lose his focus [5].

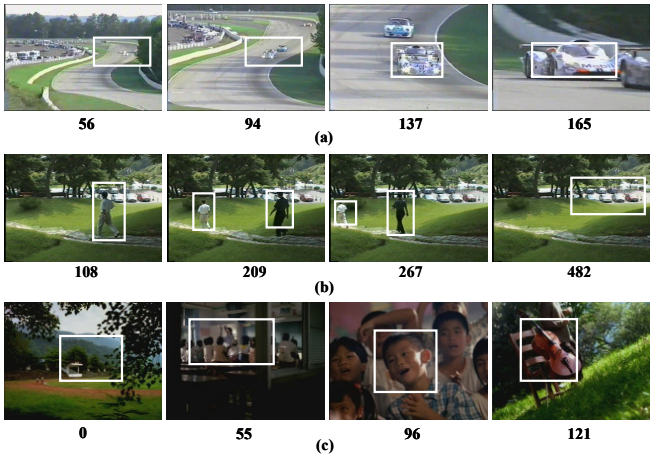

Figure 6. Sample results of ROI determination for the sequences (a) "car-racing", (b) "walking", and (c) "commercial". (The number below each frame is the time index in the corresponding video sequence)

V. EXPERIMENTAL RESULTS

The experimental videos are all *expert-produced* sequences, and are chosen from three kinds of categories: TV shows, sport programs, and TV commercials. Examples of ROI determination for some video sequences are shown in Figure 6. It is noted by exploiting the camera motion information, the true key subjects can be correctly identified. For example, the frame 137 in Figure 6(a) is under the "zoom-in" operation, and the car in the upper side of the frame is reasonably ignored. The same case is shown in the frame 96 of Figure 6(c). On the other hand, Figure 6(b) demonstrates the accuracy of the proposed dynamic algorithm in deciding the number of ROIs.

To further evaluate the performance of our approach, a subjective experiment is designed and accomplished. The experimental setting in our work is similar to that of [3]. Currently, we have two testing data sets. Data-Set-I contains videos from TV shows, films, and commercials. Data-Set-II contains clips from various sport games. Each data set includes about 15 sequences with determined ROIs, and the total length of them is approximately 60 minutes. Then, ten observers are invited to participate in the user study. They are requested to assign one subjective comment for each testing video. Three comments: GOOD, ACCEPTABLE, and FAILED are adopted in our experiments. The statistical results of our user study experiments are listed in Table II.

TABLE II. EVALUATION RESULTS OF DETERMINED ROI.

	GOOD (%)	ACCEPTABLE (%)	FAILED (%)
Data Set I	82	15	3
Data Set II	73	21	6
Avg.	78	18	4

It is reasonable that the overall performance of Data-Set-I is better than that of Data-Set-II, because the videos in Data-Set-I are produced more reliably according to the principles of applied media aesthetics than those of the Data-Set-II. The clips in Data-Set-II are sport related videos, and more semantic or game related rules are needed to facilitate the accuracy of ROI determination. Overall, the experimental results show that most of the observers (more than 95%) feel comfortable with the estimated video ROIs, which proves the effectiveness of the proposed framework.

VI. CONCLUSION

This paper presents an automatic video ROI determination framework, which provides an alternative way towards high-level semantic video analysis. The main contribution of this work is the investigation of a video-oriented fusion scheme for integrating visual features to facilitate the ROI determination. Both user attention model and applied media aesthetics are considered in the scheme. Experimental results show that the proposed framework is effective in video ROI determination. This work is very useful for a variety of vision systems and video content analysis.

REFERENCES

[1] N. Dimitrova, H.-J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," IEEE Multimedia, vol. 9, pp. 42-55, July-Sept. 2002.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Machine Intell., vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[3] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in ACM Multimedia Conf., 2002, pp. 533-542.

[4] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu, "A user-attention based focus detection framework and its applications," in Proc. of Pacific-Rim Conference on Multimedia, Singapore, 2003.

[5] H. Zettl, Sight, Sound, Motion: Applied Media Aesthetics. Belmont, CA: Wadsworth, 1990.

[6] W.-T. Chu, W.-H. Cheng, S.-F. He, C.-W. Wang, and J.-L. Wu, "A unified framework using spatial color descriptor and motion-based post refinement for shot boundary detection," in Proc. of Pacific-Rim Conference on Multimedia, Tokyo, Japan, 2004.

[7] R. Paramesan, P. Ramaswamy, and S. Omatu, "Regular moments for symmetric images," IEE Electronics Letters, vol. 34, no. 15, pp. 1481-1482, July 1998.

[8] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang, "Motion analysis and segmentation through spatio-temporal slices processing," IEEE Trans. Image Processing, vol. 12, no.3, pp. 341-355, March 2003.