LOWER ATTENTIVE REGION DETECTION FOR VIRTUAL CONTENT INSERTION IN BROADCAST VIDEO

Huiying Liu[1, 2, 3], *Shuqiang Jiang*[1, 2], *Qingming Huang*[1, 2, 3, *], *Changsheng Xu*[4]

[1] Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS)
[2] Institute of Computing Technology, CAS, Beijing 100190, China
[3]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[4]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore
Email: {hyliu, sqjiang, qmhuang }@jdl.ac.cn, {xucs}@i2r.a-star.edu.sg

ABSTRACT

Virtual Content Insertion (VCI) is an emerging application of video analysis. For VCI the spatial position is very important as improper placement will make the insertion intrusive. To choose the spatial position, we propose the notation of Lower Attentive Region (LAR) and provide a generic framework of LAR detection for broadcast video. An LAR is defined, from the cognition point of view, as a region of the video frame which attracts less audience's attention. It can be changed with little interruption to the main content of the original video. The proposed LAR detection framework includes both bottom-up and top-down modules and can be adapted to all types of videos. Finally we apply the proposed LAR detection approach to broadcast sports video by integrating domain knowledge. The Experiments on LAR detection and VCI in broadcast video demonstrate the effectiveness of the proposed method.

Index Terms—Lower Attention Region, Virtual Content Insertion, Visual Attention, Information Theory

1. INTRODUCTION

Virtual Content Insertion (VCI) is an emerging application of video analysis and has been studied for several years. The task of VCI is to make the inserted content attractive to the viewers and meanwhile not intrusive. Thus the time point and spatial position of VCI must be chosen carefully to meet the requirement. To improve the attractiveness, Virtual Content (VC) can be inserted into video highlights or frames with little camera motion.

Compared with time point, spatial position is even more important as improper placement will make the insertion intrusive. Several works has been done to detect the suitable position. In [2] static region, goalmouth, central circle and boundary line of soccer video are detected to identify suitable locations for VCI. More generic works include VRM [1] and Lower Informative Region (LIR) [3]. Both of the two works don't need any domain knowledge and can be extended in all types of videos. However, only information theory is considered. Visual attention, an important mechanism of human visual system, is neglected. To detect the spatial position for VCI, a notation of Lower Attentive Region (LAR) is proposed in this paper. An LAR, from the cognition point of view, is defined as a region of the video frame which attracts less audience's attention. It will be presented and detected using visual attention and information theory in this paper.

In this paper we provide a generic LAR detection framework which can be applied to general video. With the detected LAR, VC can be inserted into videos with little intrusion. And as an application, an LAR detection method is also proposed for VCI in sports video using domain knowledge. The contribution of this paper can be concluded from the following aspects. Firstly, we propose a notation of LAR from the cognition point of view. Secondly, we provide a generic framework for LAR detection which includes both bottom-up module and top-down module. Finally, we propose a bottom-up LAR detection method which needs little prior knowledge.

The rest of this paper is organized as follows. In section 2, a generic LAR detection framework is presented. In section 3, we present the top-down LAR detection method with sports video playfield view shots as an example. In section 4, VCI methods are briefly presented. In section 5, experimental results on TV play series and on sports video are reported. We conclude the paper with future work in section 6.

2. LOWER ATTENTIVE REGION DETECTION

2.1. Framework

Shot is a basic unit of video. Within a shot, the content is of continuum both spatial and temporal. So in our work we

[*] Corresponding author

This work was supported in part by National Natural Science Foundation of China under Grant 60702035 and 60773136, in part by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117 and 2006AA010105.

perform LAR detection with shot as the spatio-temporal context. As illustrated in Figure 1, the LAR of each shot can be obtained by both bottom-up and top-down mechanism. In the bottom-up module visual attention and information theory are used to detect the LAR without any domain knowledge, which will be detailed in the following section. And in the top-down module, specific object is detected by using domain knowledge to identify LAR. An example of top-down LAR detection can be seen in section 3.

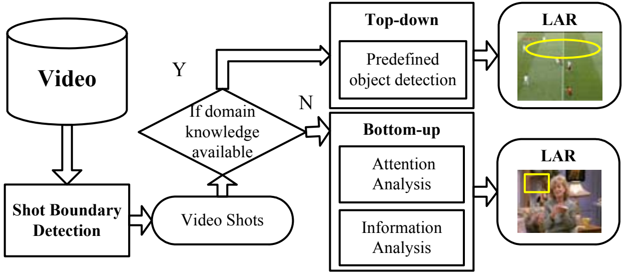

Figure 1. The framework of LAR detection

2.2. Attention Analysis

There are already some works to perform computational visual attention analysis, include static attention to detect Region Of Interest (ROI) of images [4, 5], and motion attention to detect ROI in spatio-temporal cues [6]. Performing attention analysis generates the saliency map of each frame. The LAR should be of lower saliency. In this paper, we will propose a new motion attention analysis method by integrating motion information into the region-based method [5]. Image segmentation is first performed in the method to obtain the regions. Then the saliency of each region is evaluated by combining both global contrast and contextual difference. It is calculated as follow:

$$S_i = \theta_0\left(P_i\right) \times \sum_{j=0}^{N-1}\left(FD_{i,j} \times \theta_1\left(A_j\right) \times \theta_2\left(SD_{i,j}\right) \times \theta_3\left(E_{i,j}\right)\right) \quad (1)$$

where $FD_{i,j}$ is the feagure contrast between two regions. $\theta_1\left(A_j\right)$ is region j 's area normalized to $[0,1]$. $\theta_2\left(SD_{i,j}\right)$ is a function of the spatial distance between the two regions and $\theta_3\left(E_{i,j}\right)$ is a factor related with the adjacent degree of the two regions.

Motion attracts much human attention and plays an important role in video analysis. Motion information can be obtained by several methods such as optical flow. However, a critical issue is that the motion estimation under moving camera is still a challenging problem and the motion information obtained is not so reliable. To alleviate the unpleasant impact, a cone-shaped Motion Vector Space (MVS) is adopted to represent motion vector [8]. This method transforms the MVS to HSV color space as follow:

$$\begin{aligned} &\mathit{Angle} &&\to H \\ &\mathit{Magnitude} &&\to S \\ &\mathit{Texture} &&\to V \end{aligned}$$

where the motion magnitude and the texture are normalized to $[0,255]$. The selection of texture as value follows the intuition that a high-textured region produces a more reliable motion vector. With this method the motion information can be presented intuitively. And the most significant advantage of this method is that when the motion vector is not reliable for camera motion, the V component can still provide a good presentation of the frame.

To analyze visual attention in spatio-temporal cues, we perform region based attention analysis on both the original image and the HSV image. Since the HSV image is feature of the original image, it can share the segmentation result of the original image. The example of attention analysis result can be seen in Figure 2, (a-i).

2.3. Information Analysis

The region outside the ROI attracts less attention but is still necessary to fully understanding of the scene. The necessity of a region is evaluated by its information and entropy in our work. The information of a region is a representation of its importance as a subset of the shot, while its entropy is the amount of information it contains as an independent image. An LAR must supplies less information and contains less entropy at the same time. The related works can be found in [1, 7]. In [1] block based entropy is used to detect the region of lower Viewer Relevance (VR). In [7] information of each image block is used to evaluate its attentiveness. The information of each spatio-temporal block is calculated by its conditional probability.

$$SSS(x,y,t) = -\log\left(p\left(B(x,y,t) \mid V(x,y,t-1), F(t)\right)\right) \quad (2)$$

where $B(x,y,t)$ is an image block on frame t , $V(x,y,t-1)$ is its temporal context and $F(t)$ is frame t ,which is used here as the spatial context. This method is concise and elegant but difficult to be implemented due to its high feature dimension.

In this paper a simple and effective method is adopted to calculate the information and entropy of each region and the result is used to evaluate the region's necessity. We adopt the entire shot as the spatio-temporal context and the pixels are assumed to be independent to each other. Let H be the normalized accumulative histogram of a shot, the information of a pixel of intensity l will be:

$$I(l) = -\log H(l) \quad (3)$$

Let R be a region of current frame, here a region is referred as a subset of the frame. Under the independent assumption its information is the sum of the information supplied by all of its pixels. Let h be its normalized histogram, then its information can be calculated as:

$$I = -\sum_{(i,j)\in I_{sub}} \log H\left(l(i,j)\right) = -A\sum_{l=0}^{bin} h(l)\log H(l) \quad (4)$$

where *bin* is the number of histogram bins and *A* is the region's area. The entropy of the region can be calculated as:

$$E = -\sum_{l=0}^{bin} h(l) \log h(l) \tag{5}$$

The final LAR should be of less information and less entropy. An example of information analysis can be found in Figure 2, in which (j) and (k) are the information map and the entropy map of the frame respectively.

Attention analysis generates the static saliency map and motion saliency map for each frame. Information analysis produces the information map and entropy map. Fusing the maps generates the frame attention map (See Figure 2, (l)). Finally, the shot attention map can be obtained by average the attention maps of the shot.

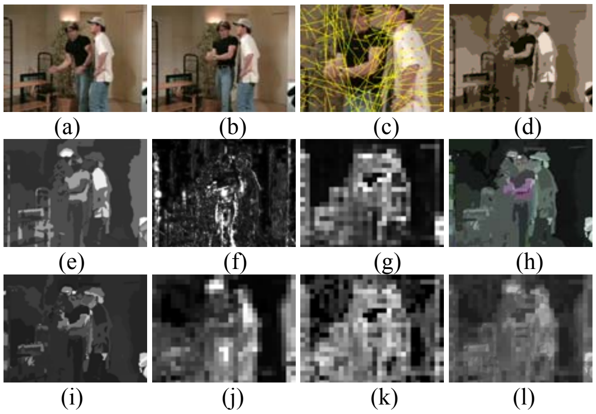

Figure 2. Example of bottom-up LAR Detection. (a): The previous frame. (b): Current frame. (c): Motion field. (d): Segmentation result. (e): Static saliency map. (f): Motion intensity. (g): Texture by DCT AC energy. (h): HSV image of motion vector. (i): Motion saliency map. (j): Information map. (k): Entropy map. (l): Final attentive map

3. LAR DETECTION FOR SPORTS VIDEO

Sports video VCI has been researched for its huge audience and commercial potential. In this section we will adapt the proposed LAR detection framework to sports video by using domain knowledge.

3.1. Method Overview

In sports video the shots can be classified into several types according to content. In [8] the shots are predefined into 8 semantic categories including 6 out of play segments and 2 in play ones. However, a much coarser classification is enough for the purpose of LAR detection. In our work the shots are classified into field view, player view (coach, referee) and audience view. Filed view and player view shots take up more than 90% of the video. So in our work, we perform LAR detection and VCI only on filed view and player view shots. This classification can be used to most types of sports including soccer, tennis and so on. Generally speaking, the field view shots are long shots of the field. The players attract most attention and the LAR can be defined and detected using the information of the playfield. Field view LAR detection will be presented in section 3.2 as an example of top-down LAR detection. Player view shots are player close up and the player may be running fast. For player view shots LAR can be detected by using the bottom-up method presented in section 2.

3.2. Field View LAR Detection

All the frames of filed view shots have the similar dominant color. For example, field view of soccer and baseball has the dominant color of green, while basket ball yellow. On the playfield there are usually many lines and/or circles, which can be used in LAR detection and in camera calibration. Also there are usually some particular objects that appear at the fixed position, such as the goal mouths in soccer video and the baskets in basketball video. They can be used to detect LAR in a top-down manner. So in field view shots, playfield detection, line and curve detection, as well as object detection [2] can be combined to detect the LAR.

Field detection can facilitate LAR detection by performing detection on the field region. Playfields may vary heavily between different types of sports, and vary lightly within same type of sport. The GMM based method [9] is adopted to detect the playfield. It is adaptive by sampling training data automatically from the video and robust by modeling the playfield with GMM. Detecting the lines and circles on the playfield helps LAR detection by supplying potential LAR which will be chosen using domain knowledge. In Figure 3, images (a)-(d) are examples of field view LAR detection. In the figure, the detected lines and circles are marked as yellow and red respectively. The center circle in (b) and the nearest rectangular in (d) are chosen as LAR by prior knowledge.

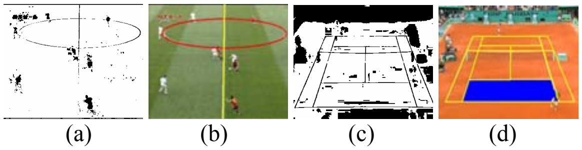

Figure 3. Examples of soccer video and tennis video LAR detection. (a, c): Playfield detection result; (b, d) The correspongding line and circle detection result.

4. VIRTUAL CONTENT INSERTION

With the detected LAR the VC can be inserted into the video without disturbing the audience. The methods for VCI include static and dynamic insertion. By static insertion the content floats over the original video, while dynamic insertion may be less intrusive by merge the content into the background. For dynamic insertion the camera parameter must be reconstructed and the content is adapted to the camera metric. For sports video the structure of playfield/court is stable. Thus the structure information is used in camera calibration [10]. In our work we use dynamic insertion in sports video playfield view shots and static insertion in other shots/videos.

5. EXPERIMENT

To verify the effectiveness of the proposed method we performed experiments on LAR detection and VCI on broadcast TV play series (22 minutes) and sports video including soccer and (45 minutes) tennis videos (26 minutes). The data are of MPEG format.

With the proposed bottom-up LAR detection method, the regions of lower saliency and lower necessity are detected correctly on player view shots and the TV play series. An example is illustrated in Figure 4. In the attention map (Figure 4, (b)), the regions of the player are of higher attention value than others, which is consistent with our common sense. With the attentive map the left-bottom corner of the frame is chosen for VCI, as illustrated in Figure 4 (a). And Figure 4 (d)-(h) are examples of TV play.

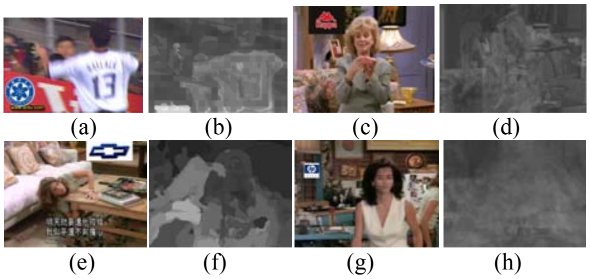

Figure 4. Results of LAR detection and VCI insertion. (a, b): Result of player view; (c-h): Result of TV play series.

For the field view shots, we insert the virtual content using dynamic method. Some of the results are illustrated in Figure 5. In the figure, the yellow/blue background is used to demonstrate the insertion effect. The images (a)-(c) are sampled from a soccer video field view shot. It can be seen that the VC is fused into the center circle as if it were on the real playfield. In the shot the camera is moving. The VCI method presented in Section 4 estimates the camera motion and adapts the VC to the camera metric. In the result, the boundary of the center circle is not very satisfying and the precision will be improved in our future work. Compared with soccer video, VCI in tennis video is easier for its less camera motion and the compactness of the court's structure. An example can be seen in Figure 5, (d)).

6. CONCLUSION

In this paper, an LAR detection method is proposed using visual attention analysis and information theory. As an application we proposed a generic method of LAR detection for VCI in sports video. In the future we will construct an integrated system of VCI and the temporal position for VCI will be researched too.

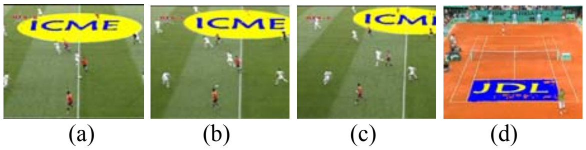

Figure 5. Results of VCI in sports video. (a-c): VCI result of soccer video: the 7th , 14th and 21st frames; (d): VCI in tennis video

7. ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of K. C. Wong Education Foundation, Hong Kong and "Science100 Plan" of Chinese Academy of Sciences under Grant 99T3002T03.

8. REFERENCES

[1] K. Wan, C. Xu, "Automatic Content Placement in Sports Highlights", *IEEE International Conference on Multimedia & Expo*, pp: 1893-1896, 2006

[2] C. Xu, K. W. Wan, S. H. Bui, Q. Tian, "Implanting Virtual Advertisement into Broadcast Soccer Video", *PCM*, 2004

[3] Y. Li, K. Wah Wan, X. Yan, C. Xu, "Real Time Advertisement Insertion in Baseball Video Based on Advertisement Effect", *Proceedings of the 11th ACM international conference on Multimedia*, pp: 343-346, 2005

[4] L. Itti, C. Koch, E. Niebur. "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *IEEE Trans on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11, pp: 1254-1259, 1998

[5] H. Liu, S. Jiang, Q. Huang, C. Xu, and W. Gao. "Region-Based Visual Attention Analysis with Its Application in Image Browsing on Small Displays". *15th ACM International Conference on Multimedia*, pp.305-308, 2007

[6] Y-F. Ma, X-S. Hua, L. Lu, H-J. Zhang. "A Generic Framework of User Attention Model and Its Application in Video Summarization", *IEEE Trans on Multimedia*, Vol. 7, No. 5, pp: 907- 919, 2005.

[7] G. Qiu, X. Gu, Z. Chen, Q. Chen and C. Wang, "An Information Theoretic Model of Spatiotemporal Visual Saliency", *IEEE International Conference on Multimedia & Expo*, pp: 1806-1809, 2007

[8] L-Y. Duan, M. Xu, Q. Tian, C-S. Xu, J. S. Jin, "A Unified Framework for Semantic Shot Classification in Sports Video", *IEEE Trans on Multimedia*, Vol. 7, No. 6, pp: 1066-1083, 2005

[9] S. Jiang, Q. Ye, W. Gao, T. Huang. "A New Method to Segment Playfield and Its Applications in Match Analysis in Sports Video". *12th ACM International Conference on Multimedia*, pp: 292-295, 2004

[10] X. Yu, X. Yan, T. T. P. Chi and L. F. Cheong, "Inserting 3D projected virtual content into broadcast tennis video", *Proceedings of the 14th ACM international conference on Multimedia*, pp: 619-622, 2006

[11] G. Zhang, X. Qin, W. Hua, T-T. Wong, P-A. Heng and H. Bao, "Robust metric recontruction from challenging video sequences", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007