SPATIAL-TEMPORAL ATTENTION ANALYSIS FOR HOME VIDEO

Xuekan Qiu[1], Shuqiang Jiang[2], Huiying Liu[1,2], Qingming Huang[1,2,], Longbing Cao[3]*

[1]Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
[3]Faculty of Information Technology, University of Technology Sydney, Sydney, Australia
{rabqiu, sqjiang, hyliu, qmhuang}@jdl.ac.cn, lbcao@it.uts.edu.au

ABSTRACT

In this paper, by considering the multiple spatial-temporal characteristic of visual perception system, we propose a novel home video attention analysis method. Firstly, each frame of the video is segmented into regions which are more informative than pixels and image blocks. Then the saliency of each region is analyzed by combining static, motion and location attentions. Finally a region based saliency map is generated for each frame, and an attention score curve is obtained for the video clip by combining attention scores of all regions in each frame. Both of them can be utilized in wide applications. This method takes advantage of the properties of human visual perception and can well present the attention information of home videos. Experimental results show the effectiveness of this approach.

Index Terms— visual attention, video analysis, attention score curve, saliency map

1. INTRODUCTION

Nowadays, people often shoot video clips to record something using their digital capture devices on hand. With the rapid increasing of video content, the effective management, transfer and browsing of home video becomes a necessity. The aim of this paper is to present a spatial-temporal home video attention analysis approach which can generate both attention scores of regions and frames. This method can be applied in many fields, such as video coding, video summarization and video browsing on small screens.

A great deal of research has been done on analyzing attention in still images [1, 2], which mainly used static information. There also has been some work on video attention analysis. Ma et al. proposed a generic framework of user attention model and used it for video summarization [3]. You et al. improved the framework by utilizing some perceptive models [4]. To find regions of interest needs other methods such as those proposed in [5-8]. Zhai and Shah [5] and Guironnet et al. [6] utilized static and motion information as spatial and temporal factors to obtain the attended areas. Liu and Gleicher [7] also analyzed image and motion saliency and applied it for retargeting video to small screens. Abdollahian and Delp [8] combined static and location saliency maps to find regions of interest in key frames of home videos. [3-8] all utilized multiple spatial-temporal factors. Analysis results of [3, 4] are temporal attention score curves and [5-8] are spatial saliency maps.

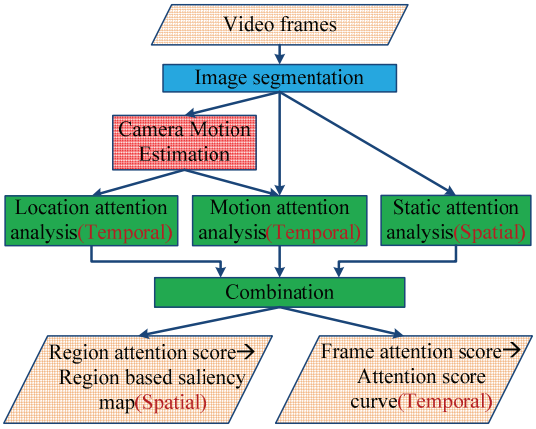


Fig. 1. Flow chart of video attention analysis.

In this paper we propose an attention analysis method which can generate both spatial and temporal attention information for home video clips. We choose region as perceptive unit by segmenting frame images. Then a fast and robust region based camera motion estimation method is utilized. After that, our attention analysis method is performed by combining static, motion and location factors. Contrast based method is used for static attention analysis. Local actual motion is calculated for each region and fuzzy growing method is utilized to extract moving foreground regions which will obtain higher motion effect scores. Moreover, we adopt the notion that viewers have the tendency to follow camera motion [8]. So location saliency map is generated according to camera motion. Finally the attention score of each region and each frame are calculated.

*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

And we can obtain a region based saliency map for each frame and an attention score curve for the home video clip. The flow chart of our approach is illustrated in Fig.1.

The rest of this paper is organized as follows. Section 2 presents this method in detail. Section 3 shows the results of our experiments. Section 4 concludes the paper.

2. VIDEO ATTENTION ANALYSIS

This section describes the detail of the proposed approach as illustrated in Fig.1.

2.1. Image segmentation

A region is more informative than a pixel and a block [9], so it is chosen as the perceptive unit in our method. The regions are obtained by using the segmentation method proposed in [10], which well accords with human perception. An example of image segmentation is shown in Fig. 2(b).

2.2. Camera motion estimation

In this paper, camera motion is important because it will be used to compute location saliency map and local motion. So we propose a fast and robust method to estimate camera motion. Four parameters camera motion model is adopted.

$$\begin{pmatrix} MV_x \\ MV_y \end{pmatrix} = \begin{pmatrix} zoom & rotate \\ -rotate & zoom \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} pan \\ tilt \end{pmatrix} \quad (1)$$

Motion vector computing method proposed in [11] is used. Pixels in the same macro block are assigned with same motion vectors. The mean (MMV) and the variance (VMV) of motion vectors of each region are computed as follows.

$$MMV_{i_x} = \sum_{j \in \mathrm{Region}_i} MV_{j_x} / N_i \quad (2)$$

$$VMV_{i_x} = \sqrt{\sum_{j \in \mathrm{Region}_i} \left(MV_{j_x} - MMV_{i_x}\right)^2} / N_i \quad (3)$$

where MV_{j_x} is the x-component of pixel j's MV, N_i is the number of pixels in region i. So MMV_{i_y} and VMV_{i_y} are computed in the same way. Fig. 2(c) shows the intensity of each region's MMV of the example frame.

To estimate the camera motion fast and robustly, we substitute MMV for MV and the coordinate of each region's center for $(x, y)^T$ in Eq. (1). Weighted binary linear regression is used to estimate the four parameters. For Eq. (1), we define the estimating error:

$$E = \sum_{i=1}^{N} w_i [(MMV_{i_x} - zoom \times x_i - rotate \times y_i - pan)^2 + (MMV_{i_y} + rotate \times x_i - zoom \times y_i - tilt)^2] \quad (4)$$

where $w_i = \theta_1(A_i) / (1 + \|VMV_i\|)$, $\theta_1(A_i)$ is the ratio of area of region i to the whole frame. To minimize E, partial of $zoom$, $rotate$, pan, $tilt$ are calculated and set to 0. We have

$$AX = B \quad (5)$$

where

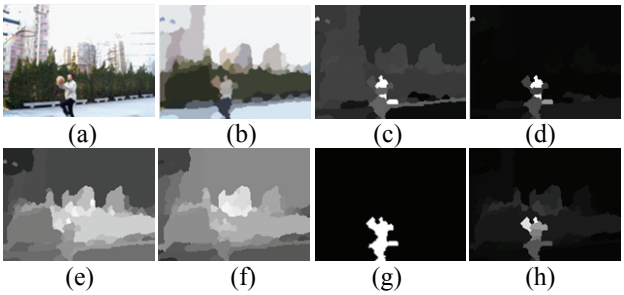


Fig. 2. Example of the saliency map calculation of a frame. (a) Original Frame, (b) Segmented image, (c) $\|MMV\|$, (d) $\|AMV\|$, (e) Static saliency map, (f) Location saliency map, (g) Moving foreground, (h) Final saliency map.

$$A = \begin{bmatrix} \sum w_i(x_i^2 + y_i^2) & 0 & \sum w_i x_i & \sum w_i y_i \\ 0 & \sum w_i(x_i^2 + y_i^2) & \sum w_i y_i & -\sum w_i x_i \\ \sum w_i x_i & \sum w_i y_i & \sum w_i & 0 \\ \sum w_i y_i & -\sum w_i x_i & 0 & \sum w_i \end{bmatrix}$$

$$X = \begin{bmatrix} zoom \\ rotate \\ pan \\ tilt \end{bmatrix} \qquad B = \begin{bmatrix} \sum w_i(x_i MMV_{i_x} + y_i MMV_{i_y}) \\ \sum w_i(y_i MMV_{i_x} - x_i MMV_{i_y}) \\ \sum w_i MMV_{i_x} \\ \sum w_i MMV_{i_y} \end{bmatrix}$$

By solving Eq. (5), we can obtain 4 parameters: camera zoom, camera rotate, camera pan and camera tilt in Eq. (1). For the example frame shown in Fig. 2, they are 0.001, 0.000, -5.322, 0.825.

2.3. Video attention analysis

After image segmentation and camera motion estimation, attention analysis is performed by combining static, location and motion attention analysis which are all region based. The details will be presented in the following subsections.

2.3.1. Static attention analysis

Contrast based static attention analysis method proposed in [9] is utilized. The static saliency of a region is calculated as:

$$S_i = \theta_0(P_i) \times \sum_{j=0}^{N-1} (FD_{i,j} \times \theta_1(A_j) \times \theta_2(SD_{i,j}) \times \theta_3(E_{i,j})) \quad (6)$$

The details can be found in [9]. In this paper, $\theta_0(P_i)$ is ignored, because it fuses with the factor of camera motion which will be presented in 2.3.2. The example of static saliency map is shown in Fig. 2(e).

2.3.2. Location attention analysis

Location is an important factor of attention. People always pay more attention to the central part of still images. And when watching videos, viewers have the tendency to follow the camera motion and look for the new objects those are about to enter the camera view [8]. The method proposed in [8] can generate camera motion based location saliency map. But it gives location saliency maps the same weight in any

(a) (b)

Fig. 3. Examples of location saliency maps. (a) that based on [8], (b) that based on this method.

situation. We add the weight and improve the saliency maps about camera panning and tilting as:

$$Map_P(i) = \max\left(0, 1 - \sqrt{\frac{(x_i - \mu_x)^2}{\sigma_x^2} + \frac{(y_i - \mu_y)^2}{\sigma_y^2}}\right) \quad (7)$$

$$Map_Z(i) = \begin{cases} |zoom| \times Z \times \left(1 - \frac{r}{r_{\max}}\right) & zoom \geq 0 (zoom\ in) \\ |zoom| \times Z \times \frac{r}{r_{\max}} & zoom < 0 (zoom\ out) \end{cases} \quad (8)$$

$$Map(i) = Map_P(i) + Map_Z(i) \quad (9)$$

In Eq. (7), $\mu_x = width/2 + pan \times P$, $\sigma_x = \sqrt{2} \times width/2$, $\mu_y = height/2 + tilt \times H$, $\sigma_y = \sqrt{2} \times height/2$. And in Eq. (8), $r_{\max} = \sqrt{width^2 + height^2}/2$, $r = \sqrt{(x_i - width/2)^2 + (y_i - height/2)^2}$. (x_i, y_i) is the center of region i. Map_P is the location saliency map about camera panning and tilting. Map_Z is about camera zooming. The difference between Map_P and location saliency map about camera panning and tilting in [8] is shown in Fig.3. And we add $|zoom| \times Z$ to Map_Z. It determines the weight of Map_Z in Map. Map_P is normalized to [0,1] while Map_Z not. This is useful for combing frame attention scores, because people always pay more attention to zooming in/out frames [3]. In our experiments, we set the value of P, H, Z to be 5, 5, 50 respectively. Fig. 2(f) shows the final location saliency map of the example frame.

2.3.3. Motion attention analysis

It has been pointed out [3] that moving objects in videos attract more attention. So it is useful for attention analysis to distinguish moving foreground from background. In our method, we extract the regions with drastic actual motion as foreground and the remaining ones as background.

The actual motion vector (AMV) of each region is first calculated by subtracting camera motion from MMV.

$$AMV_i = MMV_i - \begin{pmatrix} zoom & rotate \\ -rotate & zoom \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} pan \\ tilt \end{pmatrix} \quad (10)$$

where $(x_i, y_i)^T$ is the center of region i. Fig. 2(d) shows the intensity of each region's AMV of the example frame.

Fuzzy growing method [3] is performed to extract the foreground regions. We define the membership function of moving foreground regions as:

$$\mu_f = \begin{cases} 1 & \|AMV_i\| \geq f \\ \|AMV_i\|/f & 0 \leq \|AMV_i\| < f \end{cases} \quad (11)$$

f is set to be (width+height)/40 in our experiments. Firstly, for all regions, if region i satisfies $\|AMV_i\| \geq f$ and $\theta_1(A_i) \geq T_1$, we set it to be a foreground seed. T_1 is set to be width*height/480 in our experiments. Secondly, for each seed, fuzzy growing is performed. If region i is conjoint with a foreground region and $\|AMV_i\| \geq f/2$, region i is set to be a foreground region. Fig. 2(g) shows the extracted moving foreground of the example frame.

Then we assign moving foreground regions and background regions with different motion effect scores:

$$E_M(i) = \begin{cases} 2 & i \in motion\ foreground \\ \frac{1}{2} & i \in motion\ background \end{cases} \quad (12)$$

2.3.4. Combination

Because values of location saliency map are normalized and motion effect scores are fixed, for normalization purpose, the attention score of each region is generated by multiplying the corresponding values of static saliency map, location saliency map and actual motion effect scores, as shown in Eq. (13). Then all the region attention scores are combined together to obtain the attention score of frames, as shown in Eq. (14).

$$S_R(i) = E_M(i) \times Map(i) \times S_i \quad (13)$$

$$S_F(j) = \sum_i^N (\theta_1(A_i) \times S_R(i)) \quad (14)$$

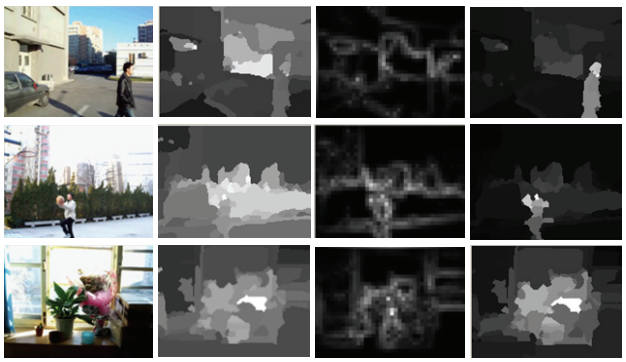
where $S_R(i)$ and $S_F(j)$ are the attention scores of region i and frame j respectively. $\theta_1(A_i)$ is the ratio of area of region i to the whole frame. Fig. 2(h) shows the final saliency map S_R of the example frame.

3. EXPERIMENTS

Home video clips are often shot to record something interesting and memorable. So home video attention is relatively explicit and objective than other kinds of videos. Experimental analysis is provided in this section to demonstrate the effectiveness of our method. The experimental videos are all recorded by a digital camera, resolution of which is 320*240.

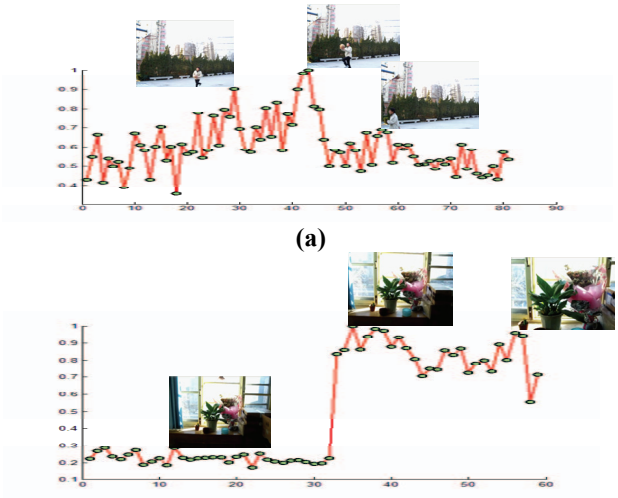
3.1. Saliency map

The saliency maps based on three different methods are illustrated in Fig.4. In the first line, camera is still and a person is walking to right. In the second line, a person is making a lay-up while camera pans a little left. In the third line, camera is zooming in while all objects are still. In the first and second lines, our maps show emphasis on the moving persons, while others not because they don't utilize local motion factor. In the third line, our map is brighter at the center regions as expected. Our three saliency maps all well show the spatial distribution of people's attention.



(a) (b) (c) (d)

Fig. 4. Examples of saliency maps. (a)Original frame, (b) Saliency map based on [9], (c) Saliency map based on [8] and (d) Saliency map based on our method.



(a)

(b)

Fig. 5. Examples of frame attention score curves.

3.2. Attention score curve

Fig.5 shows two frame attention score curves. In Fig. 5(a), camera pans a little left. Contrast and motion mainly decide the attention scores. A person with a ball is running to the left basketball stand in the left frame. He is making a lay-up in the middle frame. And in the right frame he has already completed the lay-up. The attention curve reaches its peak at the middle frame in which the person's motion is most drastic. In Fig. 5(b), camera zooms in the second half. The attention scores of the second half are higher because of $|zoom| \times Z$ in Eq. (8). Both two attention curves well demonstrate the temporal distribution of people's attention.

Attention curves can be used in many applications, such as video summarization. The frames shown in Fig.5 are key frames which are extracted using the method proposed in [3].

Experimental results on more other home video clips are also satisfactory.

4. CONCLUSIONS

In this paper we propose a video attention analysis method, which combines contrast based static attention, local motion attention, and camera motion based location attention analysis. With this method a region based saliency map and an attention score curve are generated for each frame and each home video clip respectively. Both of them can be used in many potential applications such as video coding, video summary, adaptive browsing on small screens, et al. The results of our experiments show that this approach accords with human visual perception system well. In future work, we will study special features of other kinds of videos and extend this method to wider applications.

5. ACKNOWLEDGEMENT

This work was supported by "Science100 Plan" of Chinese Academy of Sciences under Grant 99T3002T03, and the Knowledge Innovation Program of the Chinese Academy of Sciences under Grant No. 20076032.

6. REFERENCES

[1] L. Itti, C. Koch and 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, 1998

[2] Y.F. Ma and H.J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 374–381, 2003.

[3] 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, Oct. 2005.

[4] J.Y. You, G.Z. Liu, L. Sun, and H.L. Li, "A multiple visual models based perceptive analysis framework for multilevel video summarization," *IEEE Trans. on Circuits and Systems for Video Technology*, vol.17, no. 3, pp. 273-285, Mar. 2007.

[5] Y. Zhai and M. Shah. "Visual attention detection in video sequences using spatiotemporal cues," *Proceedings of the fourteenth ACM international conference on Multimedia*, pp: 815-824, October 2006.

[6] M. Guironnet, N. Guyader, D. Pellerin and P. Ladret, "Spatio-temporal attention model for video content analysis," *IEEE International Conference on Image Processing*, 2005.

[7] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," *Proceedings of the fourteenth ACM international conference on Multimedia,* pp:241-250, 2006.

[8] G. Abdollahian and E. J. Delp, "Finding regions of interest in home videos based on camera motion," *IEEE International Conference on Image Processing*, 2007.

[9] H.Y. Liu, S.Q. Jiang, Q.M. Huang, C.S. Xu and W. Gao, "Region-based visual attention analysis with its application in image browsing on small displays," *Proceedings of the fifteenth ACM international conference on Multimedia,* pp. 305-308, 2007.

[10] Q.X. Ye, W. Gao and W. Zeng, "Color image segmentation using density-based clustering," *International Conference on Multimedia and Expo*, vol. 2, pp. 401-403, 2003.

[11] R. Wang and T. Huang, "Fast camera motion analysis in MPEG domain," *IEEE International Conference on Image Processing*, vol.3, pp. 691-694, 1999.