Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **2**, March **2015**

Video Motion Characteristics Based Spatial-Temporal Salient Region Extraction Method

Ying Zhou[1,2], Ji-hong Zhang[1], Yong-sheng Liang[2], and Wei Liu[2]

[1]College of Information Engineering
Shenzhen University
Shenzhen 518060, P. R. China

[2]Shenzhen Key Laboratory of Visual Media Processing and Transmission
Shenzhen Institute of Information Technology
Shenzhen 518172, P. R. China

[1]Zhouying722@163.com

Received April, 2014; revised Octobe, 2014

Abstract. *The human eyes only observe the salient regions of the video. According to this, the motion characteristics based spatial-temporal salient region extraction method was proposed. Spatial saliency map was extracted by analyzing the log spectrum of each frame in the frequency domain. Temporal saliency map was obtained by global motion estimation, block matching and then Gaussian filter. According to the human visual characteristics and the subjective perception of different motion characteristics, the final saliency region was fused dynamically by spatial and temporal saliency map. The experiment was analyzed from both subjective and objective indicators. Visual observation and quantitative indicators show that the method proposed in this paper can reflect the human visual attention area more accurately than other classical extraction methods.*

Keywords: Video processing, Region of saliency, Visual attention model, Spatial saliency, Temporal saliency, Motion characteristics

1. **Introduction.** When the human visual system (HVS) observes a video or image, typically the amount of information into the visual field is great. These information can not be simultaneously received and processed by the human brain, and the importance of these information is also different. The HVS has the selectivity and only is interested in some part of the screen or area. These areas that can significantly attract the attention of the eyes are called region of saliency (ROS). Other areas are referred as non significant region or background area. ROS presents the psychological characteristics of HVS performance [1].

ROS is widely applied in target identification and tracking, video compression coding, scene classification, image retrieval, and so on. Scholars have conducted a lot of research work on how to efficiently and accurately obtain ROS of the video. In this paper, combined with the visual perception characteristics of HVS for video, a spatial-temporal fusing salient region extraction method was proposed.

2. **Visual Attention Mechanism.** When observing a complex scenepeople quickly focus on a few important regions and take advantage of the limited capacity process them priorly [2], which is the visual attention mechanism. The ROS extraction of video by visual attention model (VAM) is the more accurate extraction method, which matches

the physiological characteristics of the human eye better. VAM includes task-driven (Top-down) and data-driven (Bottom-up) two implementation modes.

The former combines subjective cognitive and visual scene analysis and extracts ROS based on statistical data of visual characteristics or a priori knowledge. The machine learning method is adopted. The latter simulates human vision and sensory stimulating process and extracts the different characteristics of scene (e.g. texture, color, orientation, density, etc.) as the saliency, and then models it and selects the significant attention region. The ROS extraction proposed in this paper is also based on the Bottom-up VAM.

The typical bottom-up VAM is proposed by Itti in 1998 at the University of California. Itti model [2] took brightness, color and direction features of image and gets saliency map by multi-scale integration. Harel proposed another similar GBVS model [3] (Graph based Visual Saliency). After extracting features, this model got saliency map by calculating and normalizing the activity diagram. Based on Itti model, Walther introduced competition mechanism of neural network to detect salient object [4]. Stentiford[5] proposed a context based model with random neighborhood concept. This model compared neighbor pixels in the remaining area with other randomly extracted pixels and saliency map was taken from it. [6] used gray-scale features in the frequency spectrum and got saliency map by inverse Fourier transform for remaining spectrum and phase spectrum of the input image. The above VAMs have achieved good results in extracting the saliency map of image. However, without considering the correlation of frames or the visual characteristics of human eyes, the ROS extraction of video by these models has significant difference from ROS observed by human eyes.

There are also some methods considering the spatial and temporal characteristics of the video. Itti extended static model to video by increasing the motion characteristics and flicker. Liu [7] proposed a saliency detection method fusing movement and space, which, on the basis of color contrast, introduced spatial depth and motion characteristics of the target. The temporal correlation and spatial saliency was combined [8]. Fang et al.[9] used DCT in the compressed domain to extract and fuse the brightness, color, texture of I-frame and motion characteristic of P-, B-frame.

It is proved that when viewing the video, the human eyes are sensitive to the moving object or target with greater contrast to the surrounding. In this paper, according to visual perception feature of HVS, spatial and temporal features reflecting human visual characteristics are extracted and dynamically weighted with different video content to ROS. Experimental results show that ROS extracted by our method is more consistent to the subjective perception of video content.

3. Video Motion Characteristics Based Spatial-Temporal Salient Region Extraction Method.

3.1. Spatial saliency calculation.

Hou notes that there are two stages of visual processing are involved while detecting and identifying the saliency map: first, the parallel, fast, but simple pre-attentive process; and then, the serial, slow, but complex attention process[10]. Some of the major objects are selected by integrating some primary features (such as orientation, the boundary, density, etc.) in the first stage. These objects are treated as candidates in the second stage and the ROS is determined from them. However, selecting the representative characteristics from image or video, and creating a visual attention model is a very complex process. Typicality of Feature selection as well as the complexity of the model design will directly affect the accuracy of the ROS extraction.

According to information theory, the image compression coding consists of two parts: one part is the novelty one representing feature information of image; another part is the priority knowledge, which can be compressed as redundant information. The log spectra of different images share similar trends with a simple linear relationship. In the existing models, the images are comprised of some unique and irregular salient regions and a large number of iterative non-salient regions with single content. The non-salient regions correspond to the peaks of amplitude spectrum. Gaussian kernel function can be used to filter these peaks, which is equivalent to suppress redundant information of the image and highlight the salient object. Therefore, we adopt method mentioned in reference [10].First we remove most of the redundant information by analyzing and processing log spectrum, and then get the spatial saliency map by inverse Fourier transform. The algorithm is shown as the following:

$$A(f) = \log|\hat{F}(f)| \tag{1}$$

$$I(A(f)) = A(f) * l(f,k) \tag{2}$$

$$R(f) = A(f) - I(A(f)) \tag{3}$$

$$S_f(f) = g(f) * \hat{F}^{-1}[\exp(R(f) + P(f))]^2 \tag{4}$$

Wherein, $l(f,k) = \frac{1}{\sqrt{2\pi}2^k\sigma} * \exp(-\frac{u^2+v^2}{(2^k\sigma)^2})$ is Gaussian kernel function$k = 1, 2, \cdots, K$ is the scale parameter; $K = \log 2\min\{H, W\}$ is determined by the image size; H and W denote the height and width of image separately. The accuracy of salient region extraction is depends on the scale parameter k. If k is too small, redundant information can not be inhibited effectively; if k is too large, the boundary information of salient region can be highlighted only. The optimal scale parameter k must meet the following equation:

$$k_e = \arg\min(-\sum_{i=1}^{n} S_S(k)_i \log(S_S(k)_i)) \tag{5}$$

Wherein, $S_S(k)_i$ is the salient map with the scale parameter k.

Assuming the input image is f, then $\hat{F}(f)$ denotes the Fourier transform format; $A(f)$ denotes the amplitude spectrum; then $I(A(f))$ denotes the Gaussian filtered output of $A(f)$ as above described; $R(f)$ denotes the spectral residual: $P(f)$ denotes the phase spectrum. With a Gaussian filter $g(f)$,we can get spatial saliency map $S_S(f)$.

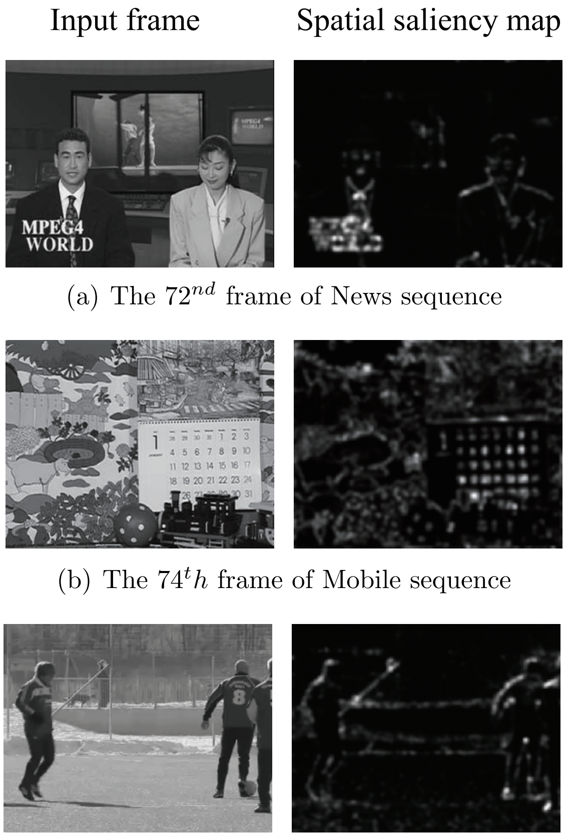
From H.264/SVC standard test sequences, we choose News, Mobile and Soccer three sequences (CIF) and extract the spatial saliency map of one frame using the above method. The results are shown in Figure 1:

As can be seen from the results of Figure 1, the ROS extraction based on spectrum can extract most significant regions in the image.

3.2. **Temporal saliency calculation.** Unlike image, the video contains motion information. Neuropsychological researches show that under the action of visual attention mechanism, HVS is more likely be attracted by moving objects or changing scenes while observing the video. And human eyes have a different perception of videos with different motion characteristics.

Actually there are three main types of moving scenes [11]:1) There is no change or little change in the background, only the foreground objects change. When watching this type of videos, human eyes focus on changing objects; 2) Background move intensely. If the total motion vector of frame is large, human eyes pay more attention to objects with little change; if most of the objects don't change very dramatically, the human eyes are more sensitive to the part of high-speed; 3) The entire scenes change intensely. Then the human eyes are difficult to distinguish the details of the video and more concerned about objects moving unobviously.

In this paper, we use the motion characteristics of the video to describe the human eyes' sensitivity to different changing video contents. In a video, the pixel motion generated



(a) The 72^{nd} frame of News sequence

(b) The 74^th frame of Mobile sequence

(c) The 70^{th} frame of Soccer sequence

FIGURE 1. The Spatial saliency map of current frame.

by the camera movement is called the global motion; the movement of the target objects is called local motion. Human eyes are mainly interested in change of target objects. Therefore, in order to describe the motion characteristics of the video content, we first need to separate the motion information of foreground and background based on the global motion estimation model and extract the target objects.

In this paper, the six-parameter model is adopted to estimate the global motion of the camera. This model can be used to model complex geometric motion, which is shown as following formula:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} * \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_5 \\ a_6 \end{bmatrix} \tag{6}$$

Where, $(x'\ y')$ denote the current coordinates of the center of current block relative to one of image; $(x\ y)$ denote the center coordinates of the matching block of the reference frame in the same position of the current frame; a_1, a_2, a_3, a_4 decide zooming and rotation; a5, a6 decide moving horizontally and vertically respectively. Error function of the sample points is defined as follows:

$$E(R) = \sum_{i=1}^{k} [(a_1 x_i - a_2 y_i + a_5 - x_i')^2 + (a_3 x_i + a_4 y_i + a_6 - y_i')^2] \tag{7}$$

According to the criteria of minimizing the error function, i.e.,

$$R_{opt} = \min_{R} E(R) \tag{8}$$

The optimal global motion estimation parameters can be obtained by the iterative least-square method:

$$\begin{pmatrix} a_1 \\ a_2 \\ a_5 \end{pmatrix} = \begin{bmatrix} \sum_{i=1}^{k} x_i^2 & \sum_{i=1}^{k} x_i y_i & \sum_{i=1}^{k} x_i \\ \sum_{i=1}^{k} x_i y_i & \sum_{i=1}^{k} y_i^2 & \sum_{i=1}^{k} y_i \\ \sum_{i=1}^{k} x_i & \sum_{i=1}^{k} y_i & k \end{bmatrix} * \begin{pmatrix} \sum_{i=1}^{k} x_i x_i' \\ \sum_{i=1}^{k} y_i x_i' \\ \sum_{i=1}^{k} x_i x_i' \end{pmatrix} \tag{9}$$

$$\begin{pmatrix} a_3 \\ a_4 \\ a_6 \end{pmatrix} = \begin{bmatrix} \sum_{i=1}^{k} x_i^2 & \sum_{i=1}^{k} x_i y_i & \sum_{i=1}^{k} x_i \\ \sum_{i=1}^{k} x_i y_i & \sum_{i=1}^{k} y_i^2 & \sum_{i=1}^{k} y_i \\ \sum_{i=1}^{k} x_i & \sum_{i=1}^{k} y_i & k \end{bmatrix} * \begin{pmatrix} \sum_{i=1}^{k} x_i y_i' \\ \sum_{i=1}^{k} y_i y_i' \\ \sum_{i=1}^{k} x_i y_i' \end{pmatrix} \tag{10}$$

We first use the block matching motion estimation algorithm to obtain the motion vector of the current frame, and then conduct a background map between the reference frame and the current frame with global motion estimation parameters. Finally the moving target is extracted.

Gaussian filter is used to filter the motion vector of moving target. Then according to three feature channels which are motion intensity, motion direction and motion density, bottom-up VAM is calculated by weighted sum of them. Finally we can get temporal saliency map of current frame $S_T(x)$. The algorithm processes as Figure 2.

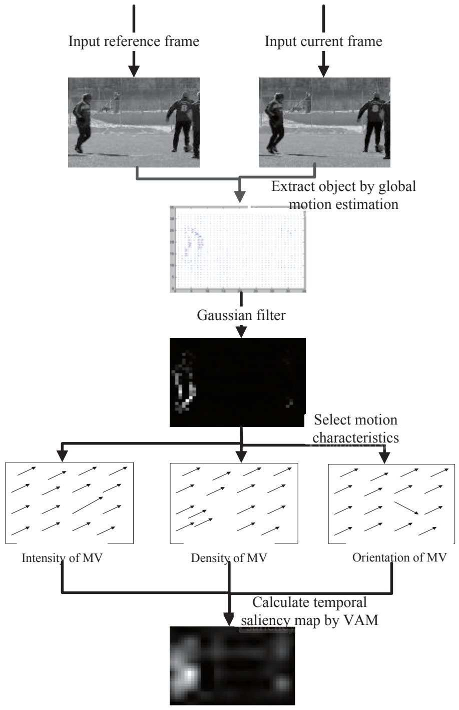

Input reference frame
Input current frame
Extract object by global motion estimation
Gaussian filter
Select motion characteristics
Intensity of MV
Density of MV
Orientation of MV
Calculate temporal saliency map by VAM

FIGURE 2. The Flow chart of the current frame temporal saliency map extraction

3.3. **Video motion characteristics based spatial-temporal salient region extraction.** According to subjective perception characteristics of HVS for the video scene, we propose a motion characteristics based spatial and temporal salient region extraction method. In this method, we first extract the spatial saliency map reflecting the video structure and the temporal saliency map reflecting motion characteristics of video, with a linear weighted combination way, then fuse adaptively temporal and spatial saliency map depending on the motion characteristics of video, as shown in the following formula:

$$S(x) = (1 - \omega_T) \times S_S(x) + \omega_T \times S_T(x) \tag{11}$$

Wherein, $S_S(x)$ denotes the spatial saliency map of video sequence x, $S_T(x)$ denotes the temporal saliency map, $S(x)$ denotes the fused saliency map. ω_T denotes the weight, which is dynamically adjusted according to the motion characteristics of the video. It takes into account three factors: the spatial distribution of motion ω_1, motion intensity 2 and motion complexity 3, which is shown as follows:

$$\omega_T = \omega_1 \times \omega_2 \times \omega_3 \tag{12}$$

$$\omega_1 = \frac{N_0}{N(s)} \tag{13}$$

Wherein, N_O denotes the number of non-zero macroblocks of moving targets' motion vector, $N(s)$ denotes the number of macroblocks of frame, then ω_1 represents the spatial distribution of motion.

$$\omega_2 = \frac{\sum\limits_{i=1}^{N_{kB}} (\|v_x\| + \|v_y\|)}{\sum\limits_{i=1}^{N(s)} (\|v_x\| + \|v_y\|)} \tag{14}$$

Where, v_x and v_y denote the horizontal and vertical coordinates of the motion vector of the target object; ω_2 denote motion energy, which is greater, the motion information is more abundant.

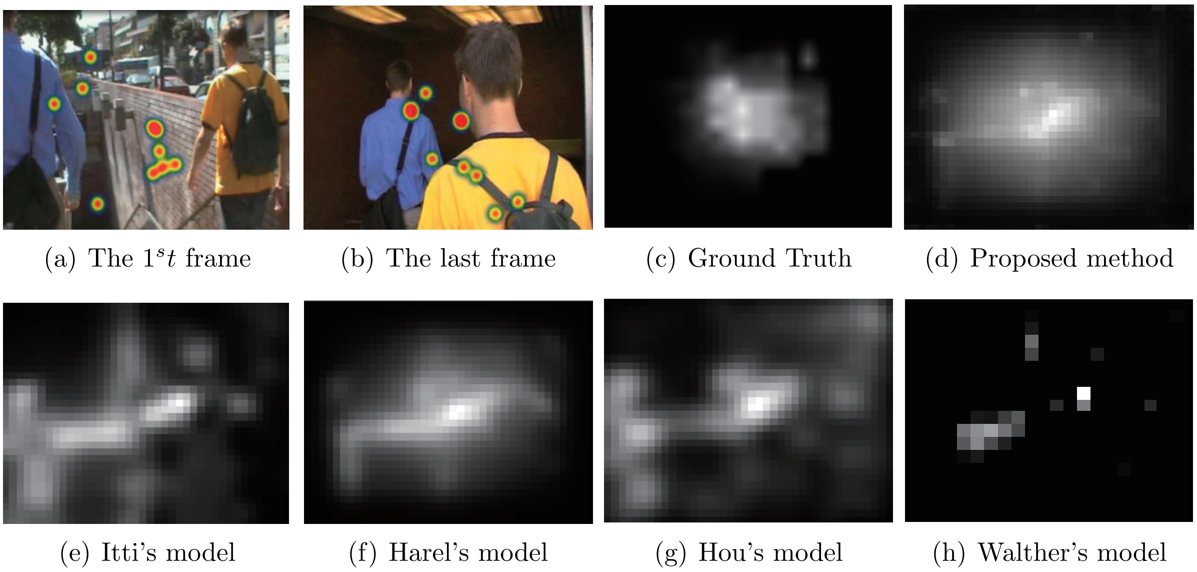
$$\omega_3 = \frac{-\left[\sum\limits_{i=1}^{m} \frac{N(s_i)}{N(s)} \times \log\left(\frac{N(s_i)}{N(s)}\right)\right]}{\lg(36)} \tag{15}$$

Where, si denotes the non-empty dimension of each direction histogram of target object motion vector; $N(si)$ denotes the number of macro blocks with non-zero motion vectors in each dimension; $i \leq 36$, the distribution in each dimension of motion vector of each macro block is obtained through the information entropy.

When the motion information contained in the video is abundant and preponderant, the weight of temporal saliency map will be increased; on the contrary, the weight will be decreased.

4. **Experimental Results and Analysis.** In order to verify the proposed motion characteristics based spatial-temporal salient region extraction method more in line with the subjective perception of HVS, the saliency map by our method is compared with other four by different spatial-temporal salient region extraction methods. These four saliency maps are calculated by Itti[2], Harel[3], Walther[4] and Hou[6] VAMs separately in the spatial domain and by our temporal saliency calculation method in the temporal domain. And the final saliency maps are the mean value of spatial saliency maps and temporal ones. Test videos are *Walk*, *Foreman*, *Tempete* and *Soccer* four sequences (CIF) selected from the H.264/SVC standard test sequences. The experimental results are evaluated

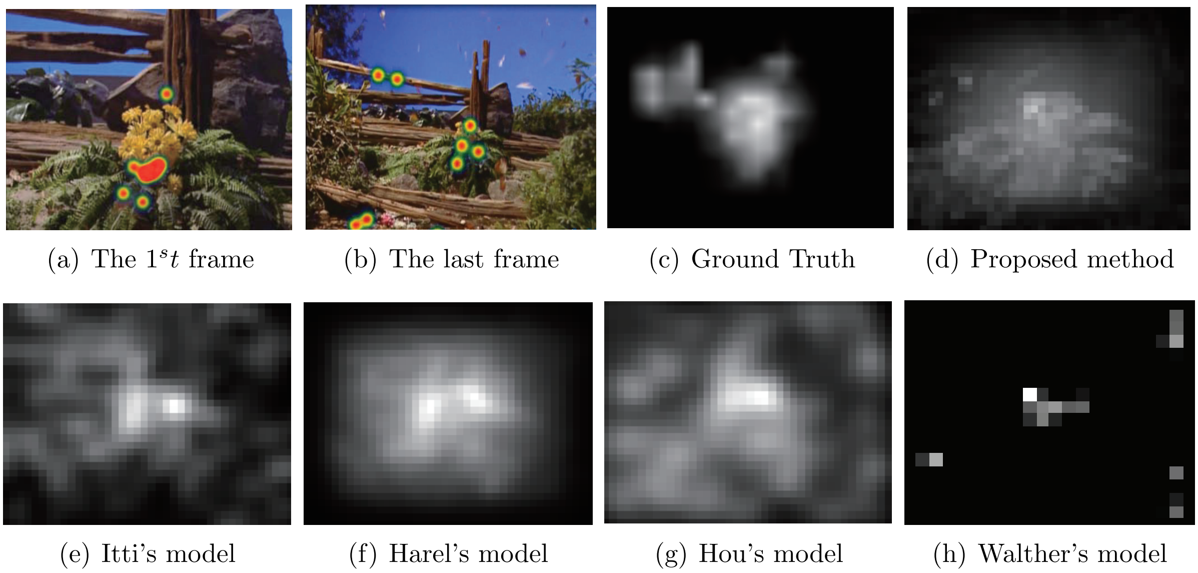
from the subjective and objective aspects. Firstly we use SMI eye tracker to get the attention region when human eye watching video and structure the Ground Truth. Then we compare Ground Truth with each saliency map extracted by the above mentioned methods. The results are shown in Figure 3 and Figure 4:



(a) The 1^st frame (b) The last frame (c) Ground Truth (d) Proposed method

(e) Itti's model (f) Harel's model (g) Hou's model (h) Walther's model

Figure 3. Saliency map comparison of *Walk* sequence.



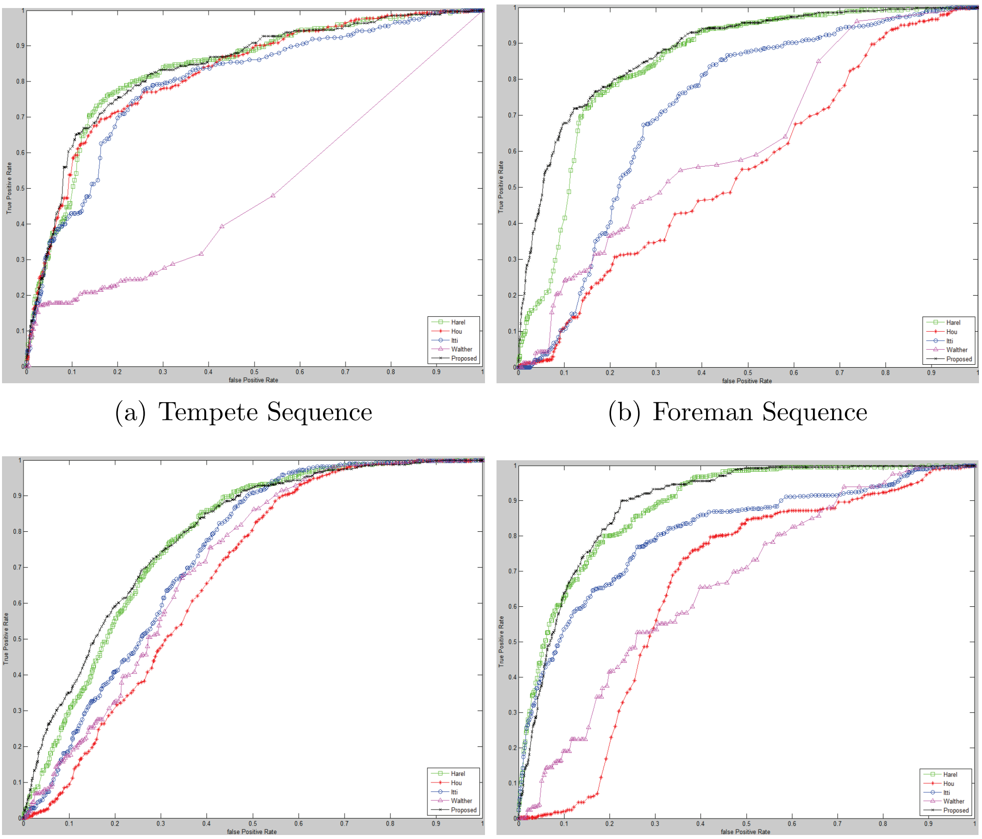
(a) The 1^st frame (b) The last frame (c) Ground Truth (d) Proposed method

(e) Itti's model (f) Harel's model (g) Hou's model (h) Walther's model

Figure 4. Saliency map comparison of *Tempete* sequence

Comparing Ground Truth and the above saliency maps, we can get the conclusion that compared to the others, the saliency map extracted by the method proposed in this paper is more in line with the perception of the human eyes.

As a measure of the effect of binary classifiers under different criteria, ROC is often used as a visual effect evaluation index. In the performance comparison, ROC curve and AUC value (the area under the ROC curve) are adopted to evaluate the quality of the extracted saliency map. The larger the AUC value, the better the performance of the algorithm described. ROC curves of each method are as shown in Figure 5, and AUC values are as shown in Table 1:



(a) Tempete Sequence

(b) Foreman Sequence

(c) Ground Truth

(d) Walk Sequence

Figure 5. ROC curve of each method.

Table 1. Comparison of AUC value

Model	Video Sequence			
	Tempete	Foreman	Soccer	Walk
Itti	0.7971	0.7203	0.7326	0.8105
Harel	0.8343	0.8458	0.7808	0.8873
Hou	0.8234	0.5544	0.6733	0.6543
Walther	0.4908	0.6256	0.7077	0.6647
Proposed	**0.8435**	**0.8801**	**0.7956**	**0.8948**

As can be seen from the results of Table 1 and Figure 5, in consideration of the HVS subjective perception for motion characteristics of the video, the motion characteristics based spatial-temporal salient region extraction method proposed in this paper can determine more accurately and efficiently the actual human eye's attention region. For the video sequences in which the contrast of moving object and background area is great (such as Foreman, Soccer), the effect is particularly evident.

5. **Conclusions.** While watching video, HVS is only interested in the moving objects or the region of great contrast with the surrounding. And for the videos with different motion characteristics, the subjective perception of the human eyes is different. According

to the visual characteristics of HVS, we propose a motion characteristics based spatial-temporal salient region extraction method. This method dynamically fuses temporal and spatial saliency map based on video motion characteristics (intensity, density and complexity). Experimental results show that, compared with ones by other classical methods, the saliency map extracted by the proposed method is more in line with human visual characteristics and subjective perception.

Acknowledgment. This work is partially supported by National Natural Science Foundation of China 61172165 and National Natural Science Foundation of Guangdong Province S2011010006113. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J. Zhang, L. Zhuo and X. G. Li, High efficiency Video Coding Techniques for Next Generation, *Posts & Telecom Press*, 2013.

[2] C. Koch, L. Itti 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, pp. 1254-1259, 1998.

[3] C. Koch, J. Harel and P. Perona, Graph-Based Visual Saliency, *Advances in Neural Information Processing Systems*, pp. 681-688, 2007.

[4] D. Walther and C. Koch, Modeling Attention to Salient Proto-objcets, *Neural Networks*, vol. 9, no. 9, pp. 1395-1407, 2006.

[5] A. Bamidele and F. W. M. Stentiford, An attention based similarity measure used to identify image clusters, 2005.

[6] X. Hou, J. Harel and C. Koch, Image Signature:Highlighting Sparse Salient Regions, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194-201, 2013.

[7] X. Liu, Z. jin, A. Zhao and et al, Salient region detection of interfusing motion and spatial relationships, *Journal of Huazhong University of Science and Technology*, vol. 41, no. 6, pp. 45-49, 2013.

[8] Y. Luo and Q. Tian, Spatio-temporal enhanced sparse feature selection for video saliency estimation, *IEEE Computer Society on Computer Vision and Pattern Recognition Workshops*, pp. 33-38, 2012.

[9] Y. Fang, W. Lin, Z. Chen and et al. A Video Saliency Detection Model in Compressed Domain, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27-38, 2014.

[10] X. Hou and L. Zhang, Saliency Detection: A Spectral Residual Approach, 2007.

[11] Y. Zhou, W. Liu and J. Zhang, Conetent aware based sorting approach of scalable video bitstream, *Journal of Signal Processing*, vol. 29, no. 8, pp. 1012-1018, 2013.