Journal of Signal and Information Processing, **2013, 4, 30-35**
doi:10.4236/jsip.2013.43B006 Published Online August 2013 (http://www.scirp.org/journal/jsip)

Scientific Research

Research on Motion Attention Fusion Model-Based Video Target Detection and Extraction of Global Motion Scene

Long Liu, Boyang Fan, Jing Zhao

The Faculty of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China.
Email: liulong@xaut.edu.cn

Received April, 2013.

ABSTRACT

For target detection algorithm under global motion scene, this paper suggests a target detection algorithm based on motion attention fusion model. Firstly, the motion vector field is pre-processed by accumulation and median filter; Then, according to the temporal and spatial character of motion vector, the attention fusion model is defined, which is used to detect moving target; Lastly, the edge of video moving target is made exactly by morphologic operation and edge tracking algorithm. The experimental results of different global motion video sequences show the proposed algorithm has a better veracity and speedup than other algorithm.

Keywords: Target Detection; Attention Model; Global Scene

1. Introduction

Moving target detection and extraction has been a hot spot in the field of video analysis, which has extensive value in use. It can be roughly divided into two categories: one is that the lens is stationary, under the local motion scene, and the other is that the lens is moving, under the global motion scene. Under the local motion scene, the method of moving target detection is relatively mature, but under the global motion scene, due to the complexity of motion, moving target detection and extraction is always a difficult problem.

Video moving target detection algorithm is mainly based on spatio-temporal information such as texture, color and motion. Under the local motion scene, the typical methods are inter-frame difference [1-3] and background reconstruction [4-8]. The inter-frame difference method detects variable and invariable characteristics of frame to separate moving target from static background. The main idea of background reconstruction is to reconstruct background without foreground moving target in advance and then subtract the background frame from the current frame for target detection. The difficulty in moving target detection and extraction under global motion scene is that video motion characteristic is the result of the superposition of global motion and local motion. The most effective solution at present is the detection algorithms based on motion compensation [9-10]. Its main clue is to use six-parameter affine model to estimate global motion, then recursive least square is adopted to calculate model parameters, obtain the relative motion between moving target and background utilizing motion compensation and finally acquire target region (TR). The computation process for motion parameter model is complicated, at the same time the estimation accuracy will be affected by the moving target size and motion complexity. So in the case of bigger target area or complicated motion, these algorithms may not realize real time and accuracy.

In recent visual technology research, achievements on human physiological and psychological are gradually integrated into the visual perception, which play a significant role in promoting the development of the visual technology. Studies have shown that human visual process is characterized by a bottom-up combining with a top-down process. Bottom-up process belongs to early vision which has noting to do with the specific content of image while depends on visual contrast caused by constituent elements of the image. *i.e.*, the greater contrast the region is, the easier it is to attract the attention of the visual system. In 1998, Itti, Koch *et al*. [11-12] proposed the concept of attention region which introduces characteristics of human vision for observed image for the first time. Firstly, low-level features, such as intensity, color, orientation are extracted from the input image after linear filtering, and then local visual contrast is calculated by Gaussian pyramid and Center-surround operator. After fusion of visual contrast with different scales and features, a comprehensive visual saliency map is obtained. On this basis, Ma Yufei *et al*. [13] proposed a motion

attention model considering the energy of motion vector and the spatio-temporal correlation to analyze motion attention on the basis of analysis of the motion vector. Guironnet and Zhai [14] proposed an attention model based on spatio-temporal information fusing static and moving target model in 2005. Jing Zhang [15] and Seung-Hyun Lee [16] applied extraction of region of attention (ROA) to target segment on static image and achieved much effects. Junwei Han [17] took advantage of attention model to segment video target. The global motion estimation and compensation was used and static attention and dynamic attention fusion was carried out to get the final result, but this method is limited to the local motion scene.

In summary, human complex visual system possesses attention mechanism and the attention is caused by feature contrast (*e.g.* color, intensity and motion). Human visual system can commendably captures moving target under global motion scene. This paper holds that this is due to human visual attention caused by the moving target and global motion contrast and moving target its own motion contrast. Movement under the global motion video scene is caused by global motion superimposed on local motion, and tends to motion contrast. If a reasonable motion attention model can be constructed, then moving target detection under global motion scene will be better solved. According to the spatio-temporal characteristic of motion vector, this paper builds a motion attention fusion model which is used to detect motion vector field, and obtain ROA, then accurately extract target.

2. Pre-Processing of the Motion Vector Field

Motion vector field directly reflects motion information of video signal, and it is estimated based on Optical Flow Equation(OFE). Let the intensity of image pixel $r=(x,y)^T$ at time t be $I(r,t)$, and OFE is defined as follows:

$$\mathbf{v}\cdot\nabla I(\mathbf{r},t)+\frac{\partial I(\mathbf{r},t)}{\partial t}=0 \tag{1}$$

where v is defined as $\mathbf{v}=\left(v_x,v_y\right)^T=\mathbf{d}/\Delta t$. Horn. Schunck [18] solved the equation on the condition of smoothness constraint. Added in different constraint, there will be other different solutions.

Motion vector field estimated by adjacent frames with Optical Flow method presents a sparse and local mess motion characteristic. Because the moving degree of adjacent frames is not enough strong and video signal exists some noise at the same time. In this paper, the motion vector field is pre-processed by accumulation and median filter. Motion vector accumulation process is: Set the current frame for the n^{th} frame, the center of block (k,l), the corresponding motion vector $\left(v_x^{k,l}(n),v_y^{k,l}(n)\right)$, and accumulation of adjacent frames calculated by Equation (2). For denoising, median filter is utilized after the accumulation of motion vector, *i.e.* each nonzero motion vector is replaced by adjacent motion vector median.

$$\left(v_x^{kl},v_y^{kl}\right)=\sum_{i=n-c}^{i=n+c}\left(v_x^{kl}(i),v_y^{kl}(i)\right) \tag{2}$$

A compact and uniform motion vector field which is suitable for motion analysis will be obtained after accumulation of motion vector and denoising.

3. Moving Target Detection Based on Motion Attention Fusion Model

This paper holds that movement of target has motion contrast in time and space, which is the basis to make use of attention to solve the problem of target detection. Analyzing factors of motion attention caused by motion vector, this paper ultimately proposes a motion attention fusion model and applies it to target detection under global motion scene.

3.1. Motion Attention

Motion attention existing in time and space is caused by motion contrast. And it can be reflected by adjacent spatio-temporal correlation degree of motion vector. The weaker correlation degree is, *i.e.* the stronger motion contrast induced by motion vector and neighbors is, and then the more attention will be attracted, vice versa. **Figure 1** shows spatio-temporal motion contrast of motion vector.

The motion vector generally appears to have strong correlation in time dimension. Motion vector correlation degree is measured by the motion vector difference between two adjacent motion vectors in the time. Temporal correlation degree $L_{k,i,j}^T$ is defined as follows:

$$L_{k,i,j}^T=|\Delta V|=|\vec{V}_{k,i,j}-\vec{V}_{k-1,i,j}| \tag{3}$$

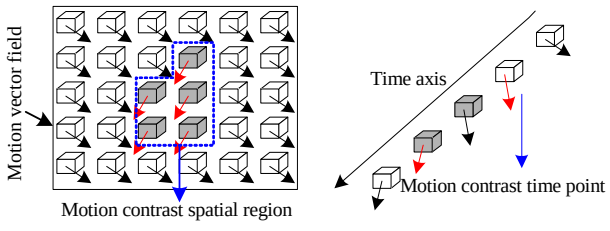


Figure 1. Spatio-temporal motion contrast of motion vector.

where the motion vector at (i,j) in the k^{th} and $(k-1)^{th}$ frame are denoted as $\vec{V}_{k,i,j}$ and $\vec{V}_{k-1,i,j}$ respectively.

The motion vector in different regions expresses

different correlation degree in spatial dimension. When the movement is caused by the global motion, motion vector correlation degree is strong. While caused by the global motion and local motion simultaneously, it is relatively weak.

A difference between a motion vector and its 8-connected boundary motion vector mean is utilized to define the local motion correlation degree. Suppose $MB_{k,i,j}$ is macro block centered at (i, j) in the k^{th} frame, i and j are horizontal and vertical coordinates of macro block; $S_{k,i,j}$ is a set of the macro block and its neighbors. Spatial correlation degree $L^S_{k,i,j}$ is defined as:

$$L^S_{k,i,j} = |\Delta V| = |\vec{V}_{k,i,j} - \vec{u}_{k,i,j}| \tag{4}$$

Here, $\vec{V}_{k,i,j}$ is motion vector at (i, j) in the k^{th} frame and $\vec{u}_{k,i,j} = \sum_{\{i,j|\vec{V}_{k,i,j} \in S^k_{k,i,j}\}} \vec{V}_{k,i,j} \Big/ 8$.

In conclusion, temporal motion attention is caused by change magnitude of motion vector, while spatial motion attention is caused by the distribution of motion vector; correlation in time and space is described as adjacent motion vector difference and the difference between neighbor average of motion vectors and itself.

3.2. Motion Attention Fusion Model

Motion attention and the correlation of motion vector in time and space are closely related. This paper considers quantifying motion attention with degree of correlation. According to section 3.1, temporal motion attention factor and spatial motion attention factor are defined as follows:

$$A^T_{k,i,j} = L^T_{k,i,j} \tag{5}$$

$$A^S_{k,i,j} = L^S_{k,i,j} \tag{6}$$

where (i, j) is the position of motion vector in the k^{th} frame, T is time, S is space.

Motion attention contains two factors: time and space, so a fusion model of those two factors is considered when modeling motion attention. Firstly, a linear fusion model is defined using a simple linear combination of temporal motion attention factor and spatial motion attention factor.

$$A_{k,i,j} = \alpha \cdot A^T_{k,i,j} + \beta \cdot A^S_{k,i,j} \tag{7}$$

Here, $\alpha > 0$ and $\beta > 0$ are the weight coefficients. As shown in Equation (7), linear operation is simple and efficient. But it is not enough to reasonably reflect the contrast changes of spatio-temporal motion attention from the perspective of spatio-temporal effect on motion attention. The paper holds that spatio-temporal biased effect to attention is different at different moments, which is due to changes of motion contrast in two aspects. In the attention model, a part of attention effect changes should be added. In this way, it can truly reflect objective changes and finally a motion attention fusion model is defined as:

$$\begin{aligned}\tilde{A}_{k,i,j} &= A_{k,i,j} + 1/2 \cdot \delta \cdot \max\left(A^T_{k,i,j}, A^S_{k,i,j}\right) \cdot \sigma \\ &= \alpha \cdot A^T_{k,i,j} + \beta \cdot A^S_{k,i,j} + 1/2 \cdot \delta \cdot \max\left(A^T_{k,i,j}, A^S_{k,i,j}\right) \cdot \left|A^T_{k,i,j} - A^S_{k,i,j}\right| \\ &= \alpha \cdot L^T_{k,i,j} + \beta \cdot L^S_{k,i,j} + 1/2 \cdot \delta \cdot \max\left(L^T_{k,i,j}, L^S_{k,i,j}\right) \cdot \left|L^T_{k,i,j} - L^S_{k,i,j}\right|\end{aligned} \tag{8}$$

where $\tilde{A}_{k,i,j}$ denotes the attention, δ is the bias controller and σ is deviation. The third part of Equation (8) denotes the spatio-temporal biased effect on attention, which reflects the stronger effect on attention when spatio-temporal attention effect is changing.

3.3. Determination of Moving Target Region

In a global motion scene, sometimes because of interference and inaccurate estimation, there will be a local and temporary movement contrast of motion vector field. This suggests motion vector field estimated by Optical Flow method is not accurate, which isn't beneficial to distinguish whether the motion macro block belongs to TR. The proposed model in section 3.2 can be sure to determine the motion vector macro blocks which draw attention in motion vector field, but to determine whether it is belongs to the TR needs further processing.

To be noticed, motion contrast generated by interference or inaccurate estimation of Optical Flow method is usually temporary, while generated by moving target is relatively continuous. This paper firstly calculates moving macro blocks attention average on adjacent time, this will greatly reduce misjudgment caused by interference and the inaccurate estimation. Average calculation as shown in Equation (9) and determine whether moving macro block belongs to the TR by Equation (10).

$$F_{k,i,j} = \frac{1}{n+1} \cdot \sum_{k=t-n}^{k=t} \tilde{A}_{k,i,j} \tag{9}$$

$$F_{k,i,j} = \begin{cases} \geq T & MB_{k,i,j} \in TR \\ < T & MB_{k,i,j} \notin TR \end{cases} \tag{10}$$

where the parameter $n > 0$ is a integer, $T > 0$ is a judging threshold, $MB_{k,i,j}$ is macro block.

4. Precision Extraction of Moving Target Region

4.1. Morphologic Operation

TR detected in section 3.3 is likely to produce hollows,

and this is because the motion contrast often exists in the boundary region of the target and background. The characteristic of binary image mathematical morphological closing operation is that the most basic morphological filter can effectively fulfill in the target holes, connect adjacent objects and smooth the boundary and at the same time does not obviously change the area of the original target. According to the results in section 3.3, this paper eliminates the inner cavity area of TR based on the morphological closing operation and obtains relatively complete TR.

4.2. Precision Target Region

In order to meet different application requirements, target boundary should be refined to obtain an accurate target region. Precision target contour relate to edge detection and tracking and a typical solution is track edge to connect of the edge of TR. What the main problem is how to determine the direction of tracking edge. This paper makes rough direction of edge as initial tracking direction and constantly adjusts the tacking direction when tracking as shown in **Figures 2(a)** and **(b)** . The process of refining edge is showed as follows:

Step 1: Use Canny operator to obtain texture edge binary image of coarse segmentation region.

Step 2: Casually select a center point of an edge pixel block as the initial tracking point and a direction from it to adjacent edge pixel block center as the initial tracking direction. If two adjacent blocks exist, then the following steps performed respectively.

Step 3: Judging whether the 8 pixels around the point as shown in **Figure 2(c)** are the edge pixels. If they are, a most close tracking direction pixel will be selected as edge pixel, or the point will be selected.

Step 4: Appoint an edge pixel determined in **step 3** as a new tracking pixel, the direction from it to adjacent edge pixel block center as a new tracking direction, perform **step 2** again. When next one adjacent block has already in the image edge and no other adjacent blocks, then end the operation.

When edge tracking is completed, a more accurate target contour is obtained, then fulfills the inner of contour, and finally get accurate motion target region.

5. Experimental Results

In this section, the proposed method was tested with a variety of standard video sequences. **Figure 3** shows the block diagram of the target detection method based on the motion attention fusion model. The global motion compensation method proposed in [9-10] was compared with the proposed method. Select parameter $\delta = 0.9$ and threshold $T = 5.6$ and MATLAB 2010.

Experimental sequence, such as "Foreman", "Stefan

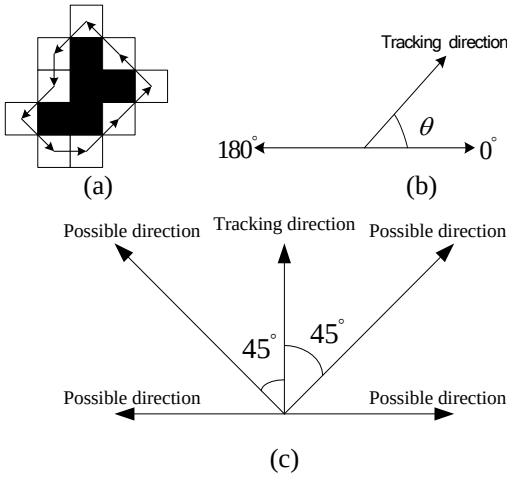


Figure 2. Edge tracking (a) alignment of boundary (b) tracking direction angle (c) possible tracking direction.

and "Coastguard" are tested, and above video sequences are global motion video scenes. "Foreman" sequence with characteristic that a moving target is relatively big, camera movement and the target motion shakes intensively; "Stefan" with characteristic of camera movement in a horizontal direction, target small and movement in a horizontal direction, target small and flexible variability in movement direction; For "Coastguard" sequence, camera and target motion remain in a horizontal direction, movement is slow, and there are two moving targets. The method proposed in [9-10] and the proposed algorithm in this paper is denoted by algorithm 1 and algorithm 2, respectively. **Figure 4** shows the results of "Foreman" (CIF) the 2^{nd} , 12^{st} frames,

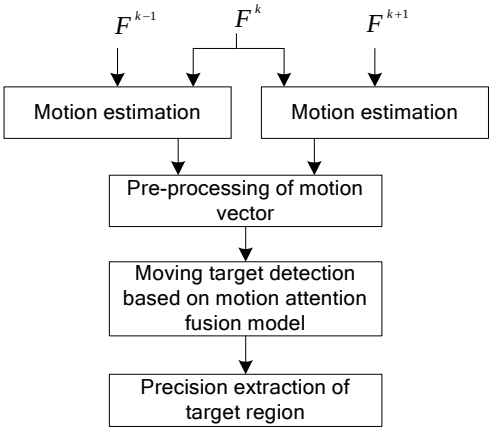
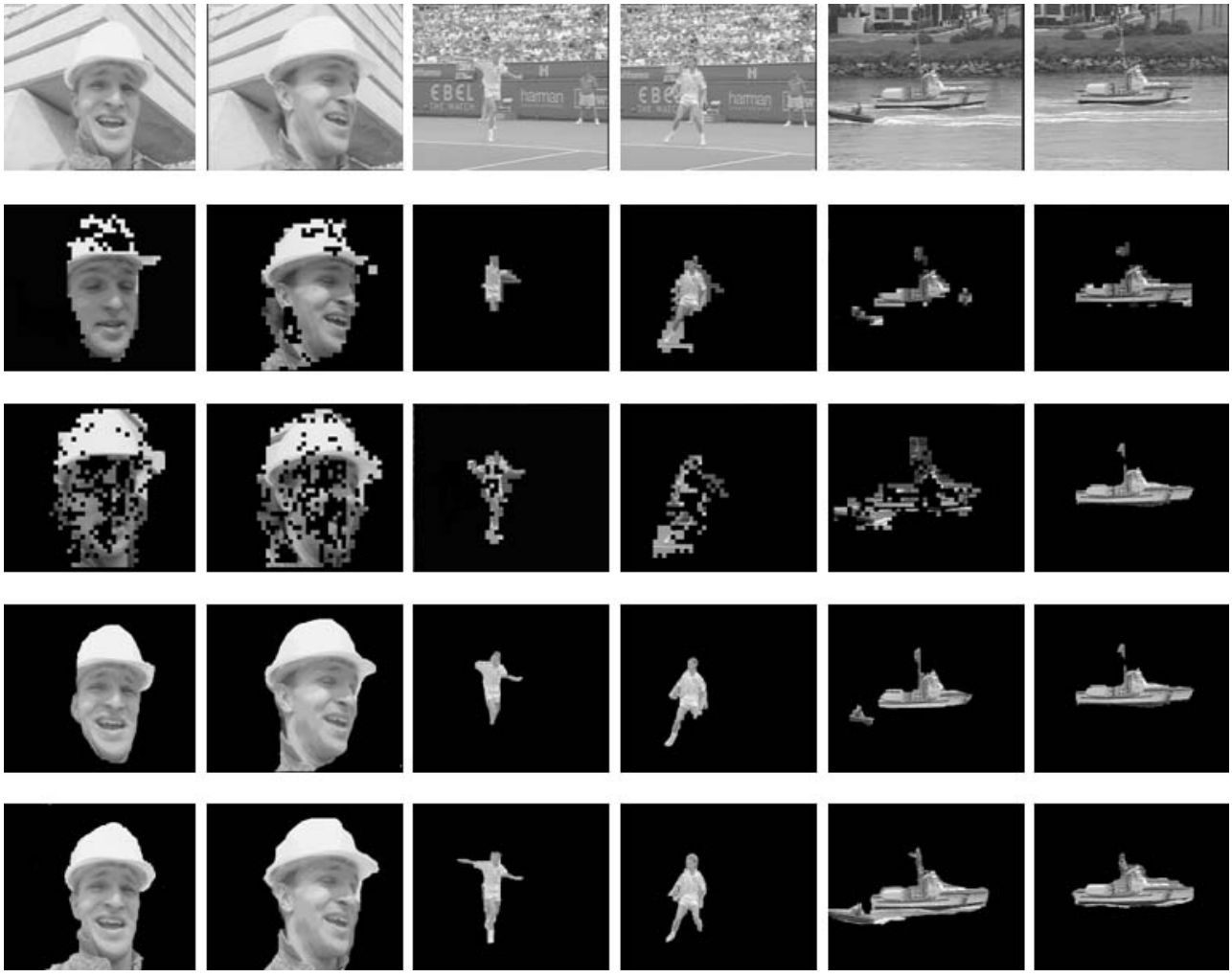


Figure 3. Block diagram of the target detection method based on the motion attention fusion model.

"Stefan"(CIF) the 26^{th} , 41^{st} frames, "Coastguard" the 101^{st} , 151^{st} frames. From row 1 to row 5, respectively as fol- lows: the original image frame of video sequence, the preliminary result of algorithm 1, the preliminary



(a) (b) (c)

Figure 4. Test results (a) "Foreman" (b) "Stefan" (c) "Coastguard".

result of algorithm 2, the finally test result of algorithm 1, the finally test result of algorithm 2. we can see that due to the global motion estimation inaccuracy, target detection and extraction error is higher than algorithm 2 in video sequence of big target, *e.g.* "Foreman"(**Figure 4(a)**); For small moving target "Stefan" sequence, the target movement is intense, the algorithm 1 error is bigger, while the relative target motion is smooth, algorithm 1 and 2 achieve the same effort(**Figure 4(b)**); For "Coastguard", the algorithm 1 and algorithm 2 results are the same because of target movement is smooth and keep in a horizontal direction(**Figure 4(c)**). **Table 1** shows the two algorithms Time Consumption (TC) statistic results comparison, testing data showed that in the same test environment, computing speed of algorithm 2 is significantly higher than algorithm 1, and this is because algorithm 2 avoids the computational cost brought by the global motion estimation, which greatly increases the operation speed.

Table 1. Two algorithms TC statistic results comparison.

Test video	Format	Frame	TC of algorithm 1	TC of algorithm 2
Foreman	CIF	1-125	350 ms/f	81 ms/f
	QCIF	1-125	127 ms/f	37 ms/f
Coastguard	CIF	5-75	283 ms/f	67 ms/f
	QCIF	5-75	92ms/f	23 ms/f
Stefan	CIF	20-125	293 ms/f	74 ms/f
	QCIF	20-125	110 ms/f	34 ms/f

In a word, the proposed algorithm based on motion attention fusion model using the motion vector in temporal and spatial attention factor can effectively detect and extract moving target under global motion scene, avoid the shortage of poor robustness and heavy computation

caused by global motion estimation and improve the veracity and real-time performance, and shows it has widespread application value.

6. Conclusions

This paper proposed a target detection and extraction method based on motion attention fusion model under global motion scene. Firstly, motion vector field generated by optical flow is pre-processed by accumulation and median filter; Then, according to the temporal and spatial character of motion vector, the attention fusion model is defined, which is used to detect moving target; Lastly, the target region is exactly extracted. The experimental results of different global motion video sequences show the proposed algorithm has a better veracity and real-time performance than other algorithms.

7. Acknowledgements

The work was supported by Education Department of Shannxi Industrialization Cultivation Project (2012JC19) and Xi 'an Technology Transfer to Promote Engineering Major Project (CX12126) for research.

REFERENCES

[1] J. Wang and E. Adelson, "Representing Moving Images with Layers," *IEEE Transactions on Image Processing*, Vol. 3, No. 5, 1994, pp. 625-638. doi:10.1109/83.334981

[2] H. G. Musmann, M. Hotter and J. Ostermann, "Object-oriented Analysis Synthesis Coding of Moving Images," *Signal Processing: Image Communication*, Vol. 1, No. 2, 1989, pp. 117-138. doi:10.1016/0923-5965(89)90005-2

[3] N. Diehl, "Object-oriented Motion Estimation and Segmentation in Image Sequences," *Signal Processing: Image Communication*, Vol. 3, No. 1, 1991, pp. 23-56. doi:10.1016/0923-5965(91)90028-Z

[4] C. Kim and J.-N. Hwang, "Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 12, No. 2, 2002, pp. 122-129. doi:10.1109/76.988659

[5] C. Stauffer and W. E. L. Grimson, "Adaptive Background Mixture Models for Real-time Tracking," *IEEE Computer Society Conference on Computer Vision and Pattern and Recognition*, Vol. 2, Fort Collins, CO, Jun 1999, pp. 246-252. doi:10.1109/CVPR.1999.784637

[6] D. Magee, "Tracking Multiple Vehicle Using Foreground, Background and Motion Models," *Image and Vision Computing*, Vol. 22, No. 2, 2004, pp. 143-155.

[7] C. R. Wren, A. Azarbayejani, T. Darrell and A. P. Pentland, "Pfinder: Real-time Tracking of the Human Body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 780-785. doi:10.1109/34.598236

[8] I. Haritaoglu, D. Harwood and L. Davis, "W4: Real-time Surveillance of People and Their Activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, 2000, pp. 809-830. doi:10.1109/34.868683

[9] Q. Bin, M. Ghazal and A. Amer, "Robust Global Motion Estimation Oriented to Video Object Segmentation," *IEEE Transactions on Image Processing*, Vol. 17, No. 6, 2008, pp. 958-967. doi:10.1109/TIP.2008.921985

[10] H. Xu, A. A. Younis and M. R. Kabuka, "Automatic Moving Object Extraction for Content-based Applications," *IEEE Transactions on Circuits and System for Video Technology*, Vol. 14, No. 6, 2004, pp. 796-812. doi:10.1109/TCSVT.2004.828338

[11] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, Vol. 2, No. 3, 2001, pp. 193-203.

[12] 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, pp. 1254-1259. doi:10.1109/34.730558

[13] Y. F. Ma and H. J. Zhang, "A Model of Motion Attention for Video Skimming," *IEEE International Conference on Image Processing 2002*, Vol. 1, New York, USA, 2002, pp. 129-132. doi:10.1109/ICIP.2002.1037976

[14] Guironnet and Mickael., "Spatio-temporal Attention Model for Video Content Analysis," *IEEE International Conference on Image Processing*. Vol. 3, 2005, pp. 1156-1159.doi:10.1109/ICIP.2005.1530602

[15] J. Zhang, L. Zhou and L. S. Shen, "Regions of Interest Extraction Based on Visual Attention Model and Watershed Segmentation," *IEEE international Conference Neural Networks & Signal Processing*, Zhenjiang, China, Jun 8-10, 2008, pp. 375-378. doi:10.1109/ICNNSP.2008.4590375

[16] S.-H. Lee, J. Moon and M. Lee, "A Region of Interest Based Image Segmentation Method Using a Biologically Motivated Selective Attention Model," *2006 international Joint Conference on Neural Networks*, Canada, July 16-21, 2006, pp. 1413-1420.

[17] J. W. Han, "Object Segmentation from Consumer Video: A Unified Framework Based on Visual Attention," *IEEE Transactions on Consumer Electronics*, Vol. 55, No. 3, 2009, pp. 1597-1605. doi:10.1109/ICNNSP.2008.4590375

[18] B. K. P. Horn and B. G. Schunck, "Determining Optical Flow," *Artificial Inteligence*, Vol. 17, 1981, pp. 185-203. doi:10.1016/0004-3702(81)90024-2

