A New Decision Making Criteria of ROI Evaluation in Video sequences and Comparison with Human Evaluation Psychology

Md. Rokunuzzaman, Kosuke Sekiyama, Toshio Fukuda
Dept. of Micro-Nano Systems Engineering & Nagoya University
Furo-cho, Nagoya-shi & 464-8603
Japan

Abstract:
This paper introduces a new decision making criteria by which the Region of Interest (ROI) is selected and evaluated. This criterion is based on Human psychology of relevance in visual perception. The experimental results are validated by comparing with Eye tracker system and Human evaluation measured by Subjective Correlation value for each ROI.

1. INTRODUCTION

To interpret a scene or video there need a selection of region of interest which should include important objects in it. However, the decision making about what objects are to be included and what objects should be excluded from the ROI is very difficult. This is due to the fact that, there is no strict rule how people will select the ROI as human has different psychologies of interest. Moreover, in real time there is less time available to make a decision of the important objects to include in the ROI. Furthermore, the evaluation of ROI is also important as it best describes the scene.

2. RELEVANT RESEARCHES

The ROI research is related with some concepts of psychological terms which are mentioned in the literatures of [1], [2] and [3]. The determination of ROI is closely related with attention and interest which are described in the literatures of [4], [5] and [6].This research implements the ROI selection in real time, therefore it is necessary to quote some relevant researches [7], [8]and[9] in this regard. The evaluation of ROI relevant with this research is reported in [10] and [11].

3. OUR CONTRIBUTION

In the existing researches, individual attention objects are selected and there is no boundary selection of ROI. The selection is based on pixel process which is mathematically complex and computationally expensive. Moreover, most of the existing methods evaluate the ROI by quantifying entropy of selected area which is totally mathematical aspect, not psychological aspect. We have developed a new decision making criterion which selects the boundary of ROI. We formalize a simple attention mechanism which scores the attention objects. We score the relevant objects based on this new criterion which is inspired from human psychology of relevance in visual perception. Finally we evaluate the ROI by quantifying the information content of ROI based on human interest and importance which is key requirement of ROI selection and evaluation.

4. ROI EVALUATION SYSTEM

4.1 Visual System Configuration

We develop our ROI selection and evaluation system by using the system as depicted in Table I.

Table-I: Specification of the System

Item	Specification
Vision Processor	Intel Core 2 Duo, 2.2 GHz, 2GB of RAM
Vision Sensor	Canon Pan-Tilt-Zoom Camera
Development Platform	Microsoft Visual Studio 2005
Language Used	C++, Visual C++
Code development	Intel's Open Computer Vision Library

4.2 Overview of the Method

In our method first we acquired the video and then subtract the background assuming moving background. For object based perception, we further detect objects with a blob filter. We compute the object properties in real time and use to calculate saliency of each object. We score the saliency according to value and select the most salient object with maximum score. We compute the relative distances of each surrounded objects from the most salient object. We select the nearest object as most relevant object and determine the ROI. The ROI is evaluated by summing up the saliency score of the most salient and most relevant object based on new decision making criteria, Chance factor. After evaluating the scene in real time we compare the ROI detection results with eye tracker. A subjective evaluation is carried offline by human evaluators and compared with ROI evaluation by our method based on subjective correlation. The details of each process are described in the following paragraphs. The overview of the system is shown in Fig.1.

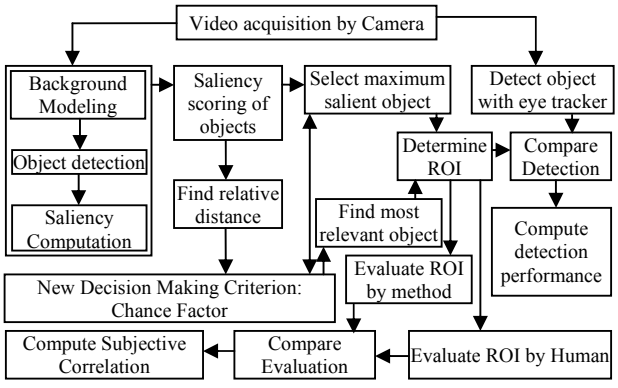

Fig: 1 Overview of the Method

978-1-4244-5095-4/09/$26.00 ©2009 IEEE

5. MATHEMATICAL FRAMEWORK

Mathematical framework consists of the attention objects determination, ROI determination and ROI evaluation.

5.1 Attention objects determination

This step consists of the following processes:

5.1.1 Background Modeling

To determine objects of interest, we need to separate the background. In visual surveillance, background cannot be fixed, therefore there need to model the background to achieve realistic performance. The most popular method is to use Stauffer's model [12]. According to his theory, background is assumed as the distribution which has higher evidence (ω) and lower variance (σ) in visual information. Therefore, Gaussians are ordered by the value of ω/σ. The highest value ordered at the top and updated by new distributions. The first distribution is chosen as background model, where

$$B = \arg\min_b \left(\sum_{k=1}^{b} \omega_k > T\right) \quad (1)$$

Where T is a measure of the minimum portion of the data that should be accounted for by the background and ω_k is the weight for k th distribution. The weights are adjusted at time t as

$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha(M_{k,t}) \quad (2)$$

Where α is the learning rate and $M_{k,t}$ is 1 for the model which matched and 0 for the remaining models. A match is defined as a pixel value within 2.5 standard deviations of a distribution.

Due to slow adaptation problems of this model an improved version is developed in [13]. In the improved version Gaussian mixture model is estimated by expected sufficient statistics update equations and then switch to L recent window version when the first L samples are processed. The weight is updated by equation 3 and 4 as

$$\omega_{k,t} = \omega_{k,t-1} + \frac{1}{t}(M_{k,t} - \omega_{k,t-1}) \quad (3)$$

$$\omega_{k,t} = \omega_{k,t-1} + \frac{1}{L}(M_{k,t} - \omega_{k,t-1}) \quad (4)$$

The L-recent window update equations gives priority over recent data therefore the tracker can adapt to changes in the environment.

5.1.2 Object detection

To mimic human like cognition, perceptual objects are detected by using a blob filter. After separating foreground from updated background, the foreground is again filtered by a blob filter and necessary information is extracted.

5.1.3 Saliency Computation

Saliency is the ability to make the perceptual objects to separate from its neighboring elements by some attributes, namely color, orientation and motion. In our method we define saliency value of a perceptual object is a combination of color, orientation and motion saliency value weighted by its information density. Each of the terms is explained as follows:

1. Color Saliency Value

Color is a very important factor for attention. Color saliency in this method is the difference in mean intensity of each object from their minimum value among all objects. Let $\overline{\mu_R}$, $\overline{\mu_G}$ and $\overline{\mu_B}$ be the mean values of R, G and B components of the object respectively, then the mean intensity is

$$\overline{I} = (\overline{\mu_R} + \overline{\mu_G} + \overline{\mu_B})/3 \quad (5)$$

Therefore, color saliency is

$$C_{sv} = (\overline{I} - \overline{I_{\min}}) \quad (6)$$

Finally after normalization,

$$C'_{sv} = C_{sv} / \max(C_{sv}) \quad (7)$$

2. Orientation Saliency Value

Orientation is calculated by fitting an ellipse in each object and calculating its angle between horizontal axis and the first side (i.e. length) in degrees. Then the Orientation Saliency Value is computed as:

$$O_{sv} = (O_i - O_{\min}) \quad (8)$$

Where O_i is the orientation in terms of angle of ith object, O_{min} is the minimum among n objects in the frame at that instant. Finally after normalization we get,

$$O'_{sv} = O_{sv} / \max(O_{sv}) \quad (9)$$

3. Motion Saliency Value

To determine motion saliency value, we need to estimate the motion. In this method we estimate the motion of each blob by their center to center Euclidean distance in two consecutive frames. Let the center position of ith blob at frame n be Cx_n^i , Cy_n^i and at frame n+1 be Cx_{n+1}^i , Cy_{n+1}^i respectively. Therefore the x component of motion of the ith blob can be formulated as

$$dx_n^{n+1}(i) = \left|(Cx_{n+1}^i - Cx_n^i)\right| \quad (10)$$

and for y component

$$dy_n^{n+1}(i) = \left|(Cy_{n+1}^i - Cy_n^i)\right| \quad (11)$$

Then the resultant motion of the ith blob is

$$M = \sqrt{(dx_n^{n+1}(i))^2 + (dy_n^{n+1}(i))^2} \quad (12)$$

Motion Saliency value of an object is the difference of motion of object and minimum of the surrounded object

$$M_{sv} = (M - M_{\min}) \quad (13)$$

After normalization we obtain

$$M'_{sv} = M_{sv} / \max(M_{sv}) \quad (14)$$

5.2 ROI determination

ROI determination consists of steps of saliency scoring, relevant object determination by new decision making criteria and ROI size determination.

5.2.1 Saliency Scoring

Each perceptual object are scored according to their total saliency value which is formalized as

$$T_{sv} = (C'_{sv} + O'_{sv} + M'_{sv}).I_D \quad (15)$$

Where, C'_{sv}, O'_{sv} and M'_{sv} are defined with eq. 7, 9 and 14 respectively and I_D is the information density. The reason to introduce this term is: pixels convey information about an object. Another aspect of human vision system is that it attains objects with larger area as it covers most of the portion of the retina. We define this term as

$$I_D = A_O / A_{OBB} \quad , 0 \le I_D \le 1 \quad (16)$$

Where, A_O is the area measured by number of pixels inside the object and A_{OBB} is the rectangular area of the box which fits the periphery of the object. We have further normalized to obtain the total saliency value between 0 to 1. Finally Eq. 15 becomes

$$T'_{sv} = T_{sv} / \max(T_{sv}) \quad (17)$$

The most salient object is selected which has maximum T'_{sv}.

5.2.2 Relevant object determination by new decision making Criterion: Chance Factor

According to human psychology of relevance, in selection of ROI, human tries to relate relevant objects with salient objects. Usually relative distance between the objects is a choice of relevant objects. If there are many salient objects, then it is very difficult to make decision which is more relevant. To solve this problem we introduce a term Chance Factor as a ratio of information density to relative distance between objects. Since it determines which salient objects will get chance to include in ROI, we name it as Chance Factor. Mathematically, it is expressed by eq. 18 as

$$CF = I_D / D_R \quad (18)$$

Where I_D is defined in eq.16 and D_R is the Relative distance calculated as the Euclidean distance of the surround objects from the maximum salient object and formulated as

$$D_R = \sqrt{(Cx_{MaxSalObj} - Cx_i)^2 + (Cy_{MaxSalObj} - Cy_i)^2} \quad (19)$$

Where, $Cx_{MaxSalObj}$ is the center x coordinate of the maximum salient object and Cx_i is the ith object's center x coordinate except maximum salient object. The second term applies for y coordinate.

5.2.3 ROI size determination

Human cannot pay attention to more than two salient objects at a time. Usually human tries to attend which is most salient and more relevant among surrounded salient objects. To reflect this behavior we assume two objects which make a ROI. For simplicity, we assume ROI as rectangle which covers most salient object in a rectangle R_1 and most relevant object in a rectangle R_2. The size of ROI is illustrated by Fig. 2 and based on minimum bounding rectangle whose sides are calculated as:

$$W_{ROI} = \{(Max\ x \mid x \in R_1, R_2) - (Min\ x \mid x \in R_1, R_2) \quad (19)$$

$$H_{ROI} = \{(Max\ y \mid y \in R_1, R_2) - (Min\ y \mid y \in R_1, R_2) \quad (20)$$

$$S_{ROI} = W_{ROI} \times H_{ROI} \quad (21)$$

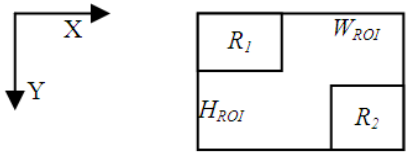

Fig.2 Illustration of ROI size determination

5.3 ROI Evaluation

ROI evaluation leads to quantification of salient region in terms of important information. The important information is quantified through saliency value at maximum attention and maximum relevance. Information also depends on size of image and ROI size. Based on this concept, lets formulate an evaluation function which quantifies the perceptual information of ROI. For formulation we need to define some variables as Image size S_I, most important information, I_{MI} and size ratio, S_r.

The definitions are:

$$S_I = \mathrm{Im}_w \times \mathrm{Im}_H \quad (22)$$

Where Im_w, Im_H are image width and image height in pixels respectively. The maximum relevancy value of object is determined by its saliency value corresponding to maximum CF denoted by $T_{sv}|_{\max(CF)}$. Making the summation we obtain,

$$I_{MI} = \max(T_{sv}) + T_{sv}|_{\max(CF)} \quad (23)$$

$$S_r = S_I / S_{ROI} \quad (24)$$

Where S_{ROI} is the size of ROI defined in eq. (21). Evaluation of ROI is quantified by EF_{ROI} which can be expressed as

$$EF_{ROI} = I_{MI} / S_r = (\max(T_{sv}) + T_{sv}|_{\max(CF)}) \times S_{ROI} / S_I \quad (25)$$

6. Experimental Results

The experiments are carried out to determine the boundary of ROI using new decision making criterion. The detected boundary is shown by blue rectangle in the following video results where real time videos are taken at different contexts.

6.1 ROI selection results using new decision making criterion

Fig.3 Amigobot is moving in colorful Premises (Indoor Video)

In Fig.3 an Amigobot is moving inside a room with different colorful objects. When it starts moving, the ROI contains Amigobot since other colorful objects are less important than the moving object. In the second image the ROI contains several objects along with Amigobot. This is due to

the fact that at that instant Amigobot is moving very slowly and it is considered as another colorful object with other object. Therefore, our new decision making criterion determines which objects are to be included and what objects are to be excluded.

Similarly other outdoor videos are examined with the same criterion and found very promising results shown by Fig. 4(a), 4(b) and 4(c).

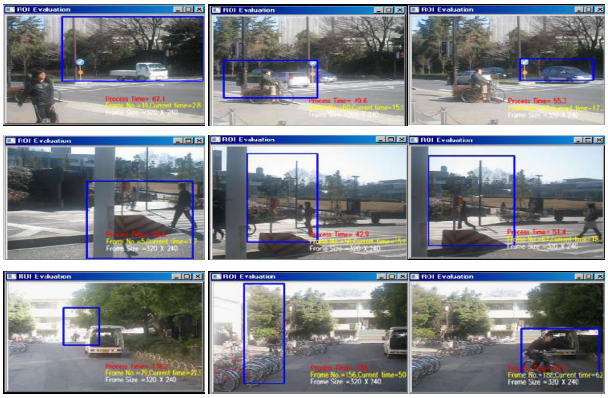

Fig.4 ROI selection in Outdoor videos: Upper image is for a, middle is for b and below is for c. 4(a) is entitled as Outdoor Video 1, 4(b) as Outdoor Video 2 and 4(c) as Outdoor Video 3.

6.2 ROI Evaluation results

The selected ROIs are evaluated in real time by the Evaluation Factor of the algorithm. One of the video results are shown by Fig. 5

Fig.5 ROI Evaluation in Outdoor Video 1: In this video the evaluation factor for second frame is higher than the first, third and fourth. This is due to the fact that ROI of second frame covers more salient and relevant objects than other frame sequences.

The evaluation factor quantifies the perceptual information at each frame sequence for each selected ROI instantly. From eq.25, the Evaluation Factor becomes high when the information content is high and size ratio is small. This means that Evaluation of ROI is high when it is stored in a compact region. Extraction of useful information is very important and it should have smaller size if we consider its easy communication.

6.3 Comparative Analysis with Eye Tracker

Prior to ROI evaluation, the appropriateness of ROI selection is necessary. This is because the selection of ROI boundary or location is very important prior to evaluation. In other words, which region we should analyze or evaluate is a crucial issue. However, how can be we sure that our algorithm selects the ROI which reflects human behavior of visual perception? To compare our ROI selection results, we choose to use Eye Tracker. Since eye search is related to human visual perception, we check the eye track point with our ROI center point. The comparative ROI selections for indoor video are shown by Fig 6.

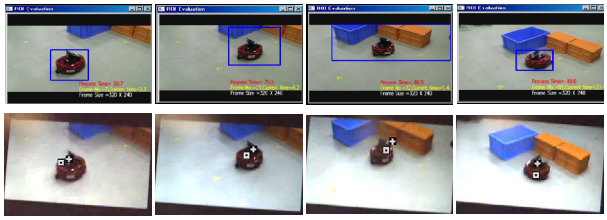

Fig.6 ROI selection comparative analysis: The upper video frames are obtained by using our ROI selection algorithm where as the below video frames are obtained by using Eye tracker system.

From the fig we can see our results are consistent with eye tracker result. To validate the results we used a statistical technique which proves how close our result to the eye track results. Fig.7 illustrates the comparison and Table II shows statistical analyses.

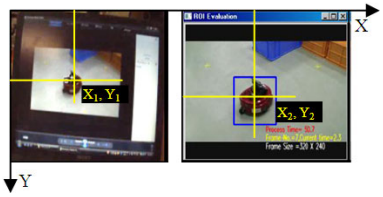

Fig.7. Indoor video is compared after eye track experiment by maintaining a same aspect ratio to see difference in attention point

Table-II: Comparative Statistical analysis of ROI location

Mean		Standard Deviation	
$\overline{\Delta x}$	$\overline{\Delta y}$	$\sigma_{\overline{\Delta x}}$	$\sigma_{\overline{\Delta y}}$
11.42	10.75	9.49	6.36

6.4 EF_{ROI} at different frames and corresponding ROIs

We have taken some sample frames from the video sequences and plotted in Fig. 8 (a). We are particularly interested in the Evaluation of the ROI and its corresponding ROI as shown in Fig. 8 (b) for the sampled frames. The numbers in the graph are used as link for the corresponding ROIs that are shown immediately below the graph. As we know the information is encoded in the pixels of image. In our research, we define ROI as the region which covers most of the salient and relevant information of image. ROI is evaluated based on this information content. Small ROI is good as it contains less byte as quality to store digital data in memory. Consequently, the small ROI is less informative due to its less evaluation value. Looking at the equation of evaluation factor, it can be seen that evaluation value is multiplied by the area ratio. If the ROI is sufficiently small size then usually it will contain less information. Referring to fig.8 (a) we can see the evaluation factor drops down as indicated by point 4 is due to this low information content. Our algorithm is intended to select and evaluate the ROI. By the value of evaluation factor, the system responses and makes decision by itself. In frame 1, we can see that there

are many salient objects like white van, road sign (yellow), road sign (blue), traffic signal light (red), man walking on sidewalk etc. Each salient object contains useful information. Any ROI which contains all of these objects, the information content will be higher. If we compare frame 1 with frame 3, we can see that frame 3 contains more salient objects than frame 1. That's why information content in the ROI of frame1 is low and that of frame3 is high. Then one might think that why the system chooses the ROI of frame 1 as it contains less information? The answer is: as the video is temporal based, the ROI should contain the object which has more temporal attribute (motion) than other salient objects at that instant.

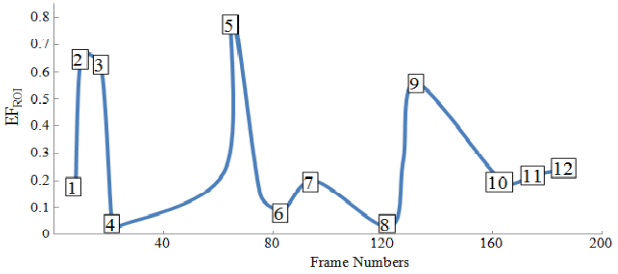

Fig.8 (a) Evaluation of ROI at different frames

Fig.8 (b) Evaluation of ROI at different frames

6.5 Subjective Correlation of ROI Evaluation

ROI evaluation or justification of its selection is entirely subjective. This is because, the human have different psychologies of interest in ROI selection. The same ROI can be evaluated by different scores by different people. Then how can we examine the effectiveness of this method of ROI evaluation? As ROI selection or evaluation is based on human psychology, we have to compare our results with evaluation of ROIs by Human evaluators. To make this correlation, we first design a scoring system for evaluation of each ROI at each frame as shown by Table IV. To compare different assessments (A) we need a correlation between assessments. This we define as Correlation value on assessment denoted by CV_{assess}. Assuming a 100 scale we assign its value for comparative assessment. To evaluate ROI we also assign some point range from 0~1 so that it can be compared with our EF_{ROI}.

Table III: Scoring System for ROI evaluation

Point range	Assessment(A)	CV_{assess}
0.0	Unjustified	0
0.1~0.4	Poor	25
0.5~0.55	No discrimination	35
0.56~0.65	Good	60
0.66~0.75	Very good	70
0.76~0.85	Best selection	80
0.86~0.95	Excellent	90
0.96~1.0	Fully Justified	100

Subjective Correlation is a quantitative measure of comparative assessment between human evaluation and systems evaluation of ROI. Let Subjective Correlation is denoted by CV_{subj}, assessment by Method is A_M and assessment by Human is A_H and difference in correlation value correspond to assessment is ΔCV_{assess}, then

Subjective correlation, CV_{subj} **(%)** = 100-(ΔCV_{assess}) (26)

If A_H=A_M, then the subjective correlation is 100. Otherwise it is calculated by eq.26 using the correlation values on assessment from Table IV. To compare ROI evaluation by the method, 20 human evaluators evaluates the same video sequences and give a justification value for each ROI of each sequence based on attention and relevance. The Evaluation Factors are not provided to them so that their decision is not influenced by it.

Each Justification values are ranges from 0~1. After evaluations from 20 evaluators, the justification values are averaged and plotted against each ROI of each video sequences indicated as ROI index in fig.7. This figure shows subjective evaluations of ROI for 4 different video sequences.

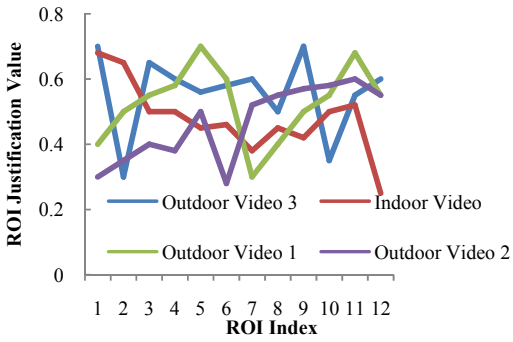

Fig. 9 ROI evaluation by Human Evaluators

To determine the ROI evaluation performance, we compare ROI evaluation by the method (*EFROI*) with this human justification value (JV) and compute a subjective correlation as an assessment performance. The reason behind this comparison is to see how our method is imitating human decision.

We have tabulated the evaluation results for Outdoor video2 as a sample result, by examining several selected video sequences in Table IV.

Table-IV Subjective Correlation as ROI Evaluation Performance

ROI Index	ROI Evaluation		Subjective Correlation CV_{subj} (%)	Average CV_{subj} (%)
	By Human JV	By Method EF_{ROI}		
1	0.40	0.20	100	85
2	0.50	0.60	75	
3	0.53	0.59	75	
4	0.55	0.10	90	
5	0.68	0.71	100	
6	0.61	0.10	65	
7	0.28	0.21	100	
8	0.42	0.10	100	
9	0.51	0.55	100	
10	0.58	0.30	65	
11	0.66	0.35	55	
12	0.52	0.40	90	

The average subjective correlation of ROI evaluation is found 85%. This means that our method is imitating human decision making most of the time. Therefore, we can conclude that the system is consistent with human evaluation psychology when it is evaluating ROIs in real time.

7. Conclusions And Future Work

We have successfully formulated a new decision making criterion for proper selection of ROI from a real scene. This new decision making criterion reflects the human behavior in visual perception. Human psychologies are properly implemented in our algorithm for making decision of selecting the ROI and its evaluation. Unlike other existing algorithms we consider human psychologies of relevance in selecting relevant objects without considering its context. The advantage of our new decision making criterion is it don't need any prior knowledge or online training. This criterion is developed by our own experience of everyday life when we observe the scene and try to concentrate if it is interesting or important to attend. Our method is imitating human decision in evaluating ROIs. Human observes the ROI and takes some time for making decision whereas our system decides it in a fraction of second. This method helps autonomous systems to respond to some unstructured environment and give an ability to make decision in choosing ROI or extracting perceptive information from visual scenes. Besides all these capabilities, our method needs some improvement. Camera motion should be taken into account to remove noise and false positives can be reduced if the threshold is adjusted both in background parameters and object detection filters. As saliency values are very small, the evaluation factor is not significant. This gives wrong assessment. Therefore, it should be improved by introducing weight for computation of saliency values. In future we will extend our algorithm to evaluate the ROI based on interaction between objects and share the information of ROI among autonomous agents.

References

[1] D. J. Levitin, "Foundations of cognitive psychology: core readings" pp.363-398, MIT Press, 2002

[2] Elizabeth A. Styles, "The Psychology of Attention", pp. 87-112, Psychology Press, 1997

[3] D. Sperber et al, "Relevance theory explains the selection task", Cognition, Vol. 57. pp. 31-95, 1995

[4] L. Itti et al, "A model of saliency-based visual attention for rapid scene analysis", Pattern Analysis and Machine Intelligence, Vol.20, No. 11, pp.1254–1259, 1998

[5] L. Elazary et al, "Interesting objects are visually salient", Vision, Vol. 8, No.3 pp.1-15, 2008

[6] N. Butko et al, "Visual Saliency Model for Robot Cameras", Proc. of the Int. Conf. of Robotics and Automation, ICRA, pp. 2398-2403, 2008

[7] H. Cheng et al, "Automatic video region-of-interest determination based on user attention model", Circuits and Systems, Vol. 4, pp. 3219 – 3222, 2005

[8] T. Sevilmis, M. Bastan et al, "Automatic detection of salient objects and spatial relations in videos for a video database system", Image and Vision Computing, Vol. 26, pp. 1384–1396, 2008

[9] I. Haritaoglu et al, "Real-Time Surveillance of People and Their Activities", Pattern Analysis and Machine Intelligence, Vol.22, No. 8, pp.809-830, 2000

[10] M. Clauss et al, "A Statistical Measure for Evaluating Regions-of-Interest Based Attention Algorithms", Pattern Recognition, Vol.3175, pp.383-390, 2004

[11] R. H. Huesman, "A new fast algorithm for the evaluation of regions of interest and statistical uncertainty in computed tomography" , Phys. Med. Biol., Vol. 29, No. 5, pp. 543-552, 1984

[12] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," in Proc. of the IEEE Int. Conf. of Computer Vision and Pattern Recognition, pp. 246-252,1999

[13] P. KaewTraKulPong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection," Proc. of the 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS01. pp.135-144, 2001