Fusion of Saliency Maps for Visual Attention Selection in Dynamic Scenes

Jiawei Xu
School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, United Kingdom

Shigang Yue
School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, United Kingdom

***Abstract*—Human vision system can optionally process the visual information and adjust the contradiction between the limited resources and the huge visual information. Building attention models similar to human visual attention system should be very beneficial to computer vision and machine intelligence; meanwhile, it has been a challenging task due to the complexity of human brain and limited understanding of the mechanisms underlying the human attention system. Previous studies emphasized on static attention, however the motion features, which are playing key roles in human attention system intuitively, have not been well integrated into the previous models. Motion features such as motion direction are assumed to be processed within the dorsal visual and the dorsal auditory pathways and there is no systematic approach to extract the motion cues well so far. In this paper, we proposed a generic Global Attention Model (GAM) system based on visual attention analysis. The computational saliency map is superimposed by a set of saliency maps via different predefined approaches. We added three saliencies maps up together to reflect dominant motion features into the attention model, i.e., the fused saliency map at each frame is adjusted by the top-down, static and motion saliency maps. By doing this, the proposed attention model accommodating motion feature into the system so that it can responds to real visual events in a manner similar to the human visual attention system in a realistic circumstance. The visual challenges used in our experiments are selected from the benchmark video sequences. We tested the GAM on several dynamic scenes, such as traffic artery, parachuter landing and surfing, with high speed and cluttered background. The experiment results showed the GAM system demonstrated high robustness and real-time ability under complex dynamic scenes. Extensive evaluations based on comparisons with other approaches of the attention model results have verified the effectiveness of the proposed system.**

Keywords-Global Attention Model; Saliency Map Fusion; Motion Vector Field

I. INTRODUCTION

Human vision system can optionally process the visual information and adjust the contradiction between the limited resources and the huge visual information. Building attentions models similar to human visual attention system should be very beneficial to computer vision and machine intelligence; meanwhile, it has been a challenging task due to the complexity of human brain and limited understanding of the mechanisms underlying the human attention system.

The human visual system includes the eyes, the connecting pathways through to the visual cortex and other parts of the brain. The visual pathway is the part of the central nervous system which gives organisms the ability to process visual detail. A complete visual pathway is consisted of the binocular vision field, temporal retina, ganglion ciliare, chiasma opticum, tractus opticus and so forth [1], while the lateral geniculate nucleus (LGN) is the primary relay center for visual information received from the retina of the eye. The LGN is found inside the thalamus of the brain. Most classical models of visual processing emphasize the lateral geniculate nucleus (LGN) as the major intermediary between the retina and visual cortex. Previous research revealed that the primary visual pathways begins in the retina, continues through the lateral geniculate nucleus (LGN) of the thalamus, and enters the first cortical way station in primary visual cortex (V1) [2]. This geniculostriate pathway is a fundamental part of early visual processing.

Human vision system is the evolved optical organ that can optionally process the vast and complex visual information with limited resources. By adopting visual mechanism in the research fields such as image processing pattern recognition and machine vision, the amount of processed information and computational resources can be reduced and the efficiency of information processing can be increased effectively. Therefore, the research of computational model for human visual attention [3][4] is a very attractive research area.

There are two interactive ways of attentional competition: **bottom–up** competition performs contrast in terms of pre-attentive features, whereas **top–down** biasing modulates attention based upon task instructions. Bottom–up and top–down contributions are combined together to decide which item is attended. Recent psychophysical research has shown that top–down factors play the dominant role in attentional competition [5]. In contrast to the well-developed computational models of bottom–up competition [6], the alternatives of top–down biasing have not been fully exploited. Starting from the Feature Integration Theory of Treisman and Gelade [7], Treisman hypothesized that simple features were represented in parallel across the field, but that their conjunctions could only be recognized after attention had been focused on particular locations. Recognition occurs when the more salient features of the distinct feature maps are integrated. Subsequently, Einhäuser et. al. (2006) [8] and Krieger, Rentschler, Hauske, Schill, and Zetzsche (2006) [9] suggested incorporating higher order statistics to fill some of the gaps between the predictive powers of current saliency map models. One way of doing this is adding more feature channels such as faces or text into the saliency map. This

significantly improves the accuracy of prediction (Cerf, Frady, & Koch, 2009) [10]. The extent to which such bottom-up, task-independent saliency maps predict human fixation eye movements [11][12] under free-viewing conditions remain under active investigation (Donk & Zoest, 2008 [13]; Foulsham & Underwood, 2008; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009)[14]. Bottom-up saliency has also been adopted (Chikkerur, Serre, Tan, & Poggio, 2010)[15]; Navalpakkam & Itti, 2005 [16]; Rutishauser & Koch, 2007)[17] to mimic top-down searches.

However, these models have emphasized on the static images, in many real situations, there are moving objects in the visual scenes – the attention has to priorities and shift to those moving objects. In our daily life, it is hard to image the reality without motion exists. Motion is a property of objects by movement of a body of mass taken over time. Motion is natural or what man makes, or channels, to exert force, do work, travel, or to push and pull loads. In our paper we put our emphasis on the motion anaysis and simulation in GAM. As human perception are environment-centralized [18] and evolved to cope with dynamic challenges. In biological visual systems, visual pathways are specialized in dealing with the motion cues and optical 'changes' [19][20][21][22]. Hereby we consider fusion motion features onto the multi-channels saliency map to facilitate fast response of the attention model to motion cues in an interactive environment. Unlike the previous research on visual attention model [23][24][25][26], our proposed model will behave more like a human visual attention system in real scenes as we discussed below.

The key inspiration of the proposed attention model can be tracked to the medial temporal area in the biological fields. The medial temporal (MT) area [27] is involved primarily in the detection of motion. Motion features have been studied in the previous model [28][29][30][31], however, both of them are not well integrated or reflected in current models. Motion features extraction is a kind of specific form of dimensionality reduction. In this paper, we decomposed the motion into angular movements and scalar transformation as they are the basic and generic motion vector in the space. They are the simplest motion patterns in the real world. These vectors comprise the motion vector field and thus lead to the motion saliency maps when we set a temporal dimensional field beforehand. At the later stage, the motion saliency map is integrated into the static saliency map and top-down saliency map together to form a final fused saliency map. With the motion feature integrated into the saliency map, the proposed attention model will be able to respond to motion feature naturally. Furthermore, motion feature is a dominant factor in the complex dynamic scenes and we mimic the real circumstance after adopting the motion cues into the model.

The rest of the paper is organized as follows. In Section II we propose an efficient fusion attention framework. The top-down saliency map, static saliency map, motion saliency map and model fusion are described in Section III. The experimental results and performance evaluation are reported in Section IV. Finally, Section V concludes the paper.

II. The Proposed Attention Model

Studies of attention models can be categorized according to the two attributes mentioned previously. The first attribute is based on the spatial contrast analysis; the second attribute is based on the frequency domain analysis. The following sections elaborate on the visual attention model and describe our proposed attention model briefly. The details of the proposed model will be further illustrated in section 3.

A. Related work on visual attention model

Recent work on computational models of visual attention are generally classified into four parts with different aims, they are salience map, temporal tagging, emergent attention and selective routing, respectively. A majority of computational models of attention follow the structure adapted from the Feature Integration Theory (FIT) [7] and the Guided Search model. Koch and Ullman [32] proposed a computational architecture for this theory and Itti [33][34] were among the first ones to fully implement and maintain it. The main idea here is to compute saliency in each of several features (e.g., color, intensity, orientation; saliency is then the relative difference between a region and its surrounding) in parallel, and to fuse them in a scalar map called the "saliency map". Le Meur et al. adapted the Koch-Ullman's model to include features of contrast sensitivity, perceptual decomposition, visual masking, and center-surround interactions. Some models have added features such as symmetry [35], texture contrast [36], curvedness [36], or motion [37] to the basic structure. For example, a computational model of visual attention is an instance of a model of visual attention, and not only includes a formal description for how attention is computed, but also can be tested by providing image inputs, similar to those an experimenter might present to a subject, and then seeing how the model performs by comparison. We put our emphasis on the saliency region detection with the combination of prior knowledge and feature extraction to get a relatively realistic and logical situation.

In addition to the mentioned models, several motion models of visual saliency have been developed over the past years. In Hou's model [38], the motion saliency map is obtained through the multi reference frames, and enhanced by spatial saliency information. In Raj's model [39], he advised the statistical saliency model to include motion as a feature. These models do improve the motion features under complex scenes. However, these models consider the motion cues as a separate channel and do not fused it into the simulated human visual system well. In our model, we not only computed the motion cues as a part of saliency map before the fusion stage, and also consider the prior knowledge or task-driven stimuli, and static saliency map into the final weighted saliency fusion. This process can lead to a better accuracy to the ground-truth saliency map which is indicated by our human eyes compared with other previous models. We described our model as below.

B. The Proposed System Overview

Figure 1 illustrates the framework of our system. The system is composed of four modules, video preprocessing, feature pooling, saliency map acquisition and final saliency map fusion.

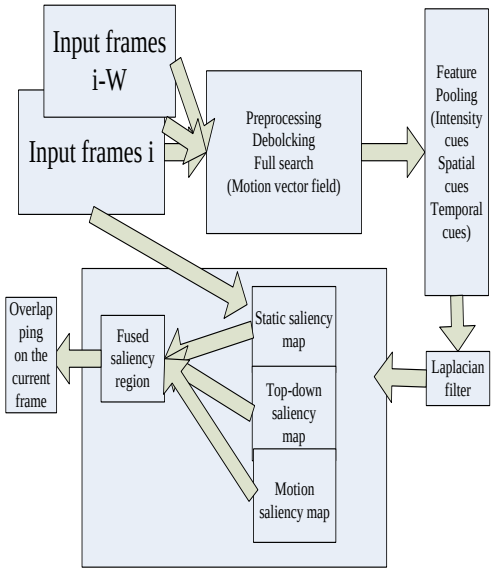


Fig. 1. System overview

In video preprocessing module, the input video is segmented into 8*8 blocks. Then the full search between the previous W frame, which W is setting by our different users, generate all the motion vectors by observing the moving components on the HSV channel, with RGB turning into HSV beforehand.

The feature pooling module chooses the selected cues from all W frames. Our proposed idea is composed by three channels, which are motion intensity, spatial cues and temporal cues, respectively. After using Laplacian filter, which carries the isotropic properties, we can smooth the processing images and get a refine results before getting the motion saliency map.

The saliency map acquisition module contains three layers, the static saliency map is obtained from the Itti and Koch bottom-up attention model, which the top-down saliency map is generated by the AIM model [40], our motion saliency map is an important role on the later stage of final fusion module.

The final saliency map module is to combine all the saliency maps by giving the fusion weight and variance as we describe in the later part. The attention region is detected and we overlapped it on the current frame, remember that the W is a human-supervised frame rate setting before our proposed model.

In the following sections, we will illustrate our attention model in details.

III. Attention Saliency Maps

A. Top-down Saliency Map

This section first presents an overview of the top-down saliency map and its underlying assumptions. It goes on to explain in the details of the different components and their functions. The processing architecture of the system is shown in Figure 2.

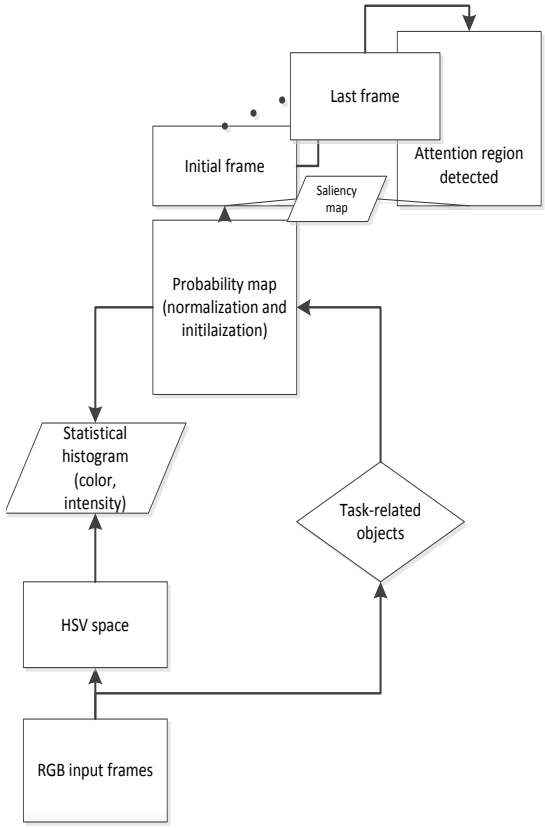


Fig. 2. The scheme of top-down saliency map

As shown in the Figure 2, the system consists of the several units which are all implemented in the systems. The web camera input might be produced by a motorized pan-and-tilt camera simulating a moving eye. Here we adopt a stationary camera image and only simulate the eye movements. The work space that the eye can explore is an area formed by 352*288 RGB pixels. We convert the input frames into HSV which can enlarge the color contrast in spatial dimension. After obtaining the statistical histogram and designating the attention region, we can get the probability map (PM) [41], we adopt normalization and initialization to search the attention regions on each frame. The saliency map is a key process of our flowchart, which combines a number of visual feature maps into a combined map that assigns a saliency to every location in the visual field (Itti, Koch and Neibur, 1998). Each feature map is typically the result of applying some simple visual operator to the input image. For example, a feature map could consist of an activity pattern that may code for intensity, color, motion or some other visual cues. The result of summing the different feature maps is that the saliency map will code for locations in the image with many features.

Here, the saliency map selects the region of interest (ROI) by evaluating the possible area in the initial frame. Alternatively speaking, our model is based on the task-related visual searching. Variants of the task are obtained by positioning a certain number of distractors. The saliency region is detected by the following algorithm:

Algorithm 1 Top-down Saliency Map
i. Initialize proposed algorithm: (a) select the target region; (b) calculate the color histogram of each frame; (c) obtain the probability model of the object's features; (d) set the initial position.
ii. Process next frame: (a) read next frame; (b) set the previous frame's target center to the current frame's one; (c) set the searching window's width and height by half of the minimum bounding rectangle's width and height; (d) calculate the joint histogram; (e) obtain candidate model
iii. Tuning the weights on the probability map.
iv. Compute the next location of the target candidate by assuming the 10 pixels bias.
v. Compute the deviation to the tolerance range, if the current frame iteration stops, the current position is the saliency center and the algorithm goes back to step ii. Otherwise, to adjust the target position and to go back to step iii in order to continue the current frame's iteration.

As the algorithm implements, we can obtain our attention area by giving the attention region we assumed before, which can be formally called a semi-supervised top-down saliency map. In next stage, we need the static saliency map as one of the components before the saliency map fusion.

B. Static Saliency Map

As we mentioned in the previous section, the first attribute is based on the spatial contrast analysis. Different spatial locations compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into saliency map, which topographically codes for local conspicuity over the entire visual scene. The second attribute is based on the frequency domain analysis. Recently, a simple and fast algorithm, called the spectrum residual (SR), was proposed based on the Fourier Transform [42]. The paper argues that the spectrum residual corresponds to image saliency. Following this, the Phase spectrum of the Fourier Transform (PFT) was introduced, which achieved nearly the same performance as the SR [43]. On the other hand, the choosing of perceptive unit differentiates the existing attention analysis methods. The perceptive unit can be chosen as pixel, image block, region or object. A pixel/block contains little perceptive information. Comparatively, an object contains much perceptive information but is difficult to be obtained because object detection is still an open problem in the area of computer vision. In color images, an object is composed of one or more regions. In other words, a region is a unit between a pixel/block and an object. It contains more perceptive information than a pixel/block and can be obtained by image segmentation, which is much easier than object detection. So, in our work we adopt region as perceptive unit. This choice also enables the proposed method to analyze visual attention at multi-scales for the adaptive size of region. Since our purpose is to obtain the image patches which can be used as perceptive unit, image segmentation is simplified by performing color quantization using K-Means. Then the neighboring pixels of same color are regarded as a region. The saliency of each region can be calculated by using equation (1):

$$Sal(k) = \log p(f(k)) \sum_{i=1}^{K} d(f(k), f(i)) * G_k(i,k) \quad (1)$$

where κ is the total number of regions in the image. G_k is the DoG function of region κ , of which the radius is the same with the region. $f(k)$ is the feature of region κ and $p(f(k))$ s its probability calculated by using the color quantization result. $d(f(k), f(i))$ is the distance between features. It is evaluated in our work by using Gaussian distance. An example of static saliency calculation is illustrated in Figure 4.

C. Motion Saliency Map

In this section, we introduce the architecture of motion attention model and hereby this is the key contribution of our work. We intergrated this element into our model as previous approaches [44][45][46] are not well considered or simplified this part. Here, we start our research based on AVI video stream. However, we only select the uncompressed video clips to keep the information fidelity. As shown in figure 1, we can obtain approximate motion information from the motion vectors. In each frame, the spatial layout of motion vectors would compose a field called Motion Vector Field (MVF)[47][48]. If we consider MVF as the retina of eyes, the motion vectors will be the perceptual response of optic nerves. We set 3 types of cues, intensity cues, spatial coherence cues, and temporal coherence cues, when the motion vectors in MVF go through such cues, they will be transformed into three kinds of feature maps. We fuse the normalized output of cues into a saliency map by linear combination, and it will be tuned by the weight as describe in section 3.4. Finally, the image processing methods are adopted to detect attended regions in saliency map image, where the motion magnitude and the texture are normalized to [0, 255]. The selection of texture as value, which follows the intuition that a high-textured region produces a more reliable motion vector,

provides this method a significant advantage that when the motion vector is not reliable for camera motion, the V component can still provide a good presentation of the frame. After transforming the RGB to HSV color space, motion saliency can be calculated using the segmentation result of section. An example of saliency map and motion attention is illustrated in Figure 3. Figure 3(a) is the corresponding motion saliency map based on 9 dimensional MVF, while figure 3(b) is the result provided on 2 dimensional MVF. According to our assumption, there will be three cues at each location of macro block $MB_{i,j}$. Marco block is a basic unit of motion estimation in video encoder and it is consisted by an intensity pixel and two chromatic pixel blocks. Hereby we adopt 16*16 Marco block due to the computational burden. Then the intensity cues can be obtained by computing the magnitude of motion vector

$$I(i,j) = \sqrt{dx_{i,j}^2 + dy_{i,j}^2} / MaxMag \tag{2}$$

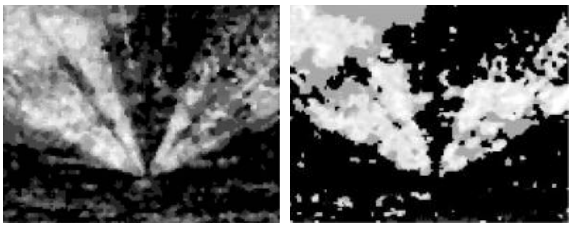
here (dx_{ij}, dy_{ij}) indicate two components of motion vector, and $MaxMag$ is the maximum magnitude in MVF. The spatial coherence cues induces the spatial phase consistency of motion vectors has high probability to be in a motion object. By contraries, the area with inconsistent motion vectors is possible to be located near the edges of objects or in the still condition. First, we calculate a phase histogram in spatial window with the size of $m*m$ pixels at each location of Marco block. The bin size of each is 10 degree, as we segment the 360 degree into 36 intervals, which means from 0 degree to 10 degree we regard it as a same angle. Then, we measure the phase distribution by entropy as following:

$$\begin{aligned} C_s(i,j) &= -\sum_{t=1}^{n} p_S(t)\log(p_s(t)) \\ p_s(t) &= SH_{i,j}^m(t) / \sum_{k=1}^{n} H_{i,j}^m(k) \end{aligned} \tag{3}$$

where C_s donates spatial coherence, $SH_{i,j}^m(t)$ is the spatial phase histogram whose probability distribution function is $p_s(t)$, and n is the number of histogram bins. Similarly, we define temporal phase coherence within a sliding window with the size of W(frames). It will be the output of temporal coherence cues as expressed below:

$$\begin{aligned} C_t(i,j) &= -\sum_{t=1}^{n} p_t(t)\log(p_t(t)) \\ p_t(t) &= TH_{i,j}^W(t) / \sum_{k=1}^{n} TH\sum_{i,j}^W(k) \end{aligned} \tag{4}$$

where C_t denotes temporal coherence, $TH\sum_{i,j}^W(t)$ is the temporal phase histogram whose probability distribution function is $p_t(t)$ and n is still the number of histogram bins. Moreover, we increase the frame number as a temporal dimension and the output is easier to distinguish the difference. The result indicates the attended region can be more precise if we elongate the frame number as shown in figure 5. The Laplacian filter is to remove the impulse noise generated by the input frames. Hereby we adopt the median filter can also preserve the edge information and sharpen the image details. We adopt 3*3, 7*7…, 25*25 window slides at the experiment stage, but finally we utilize 3*3 window as the convenience of later computation.



(a) (b)

Fig. 3. the saliency map results from video set 1, as we can see more information fidelity from (a) but poor information from (b), the left one donates n=1~10,2~11,...,10~20, which means 9 motion vector fields in summation, while the right one only aggregates i=1~3, 2~4,3~5, only computes 2 motion vector fields.

Generally, the intensity cue reveals the highly moving objects. The spatial cues indicate the different motion objects in spatial, while the temporal cues donates the variability of one object in the temporal dimensional. Also, the motion orientation weights the motion saliency map and affect the results on a critical extent. For example, when we capture a 135 degree motion on a motion saliency map consisted by most of 45 degree motion vector. This is quite singular and obvious to our human vision system, which means a high tuning weight on the next stage.

In figure 4, we demonstrate one of our experiment results, here W in equation 4 is set to 23, which means 23 frames constructs a motion vector field. The first column indicates the first frame and the W frame from the top to bottom, respectively. The second row illustrates the top-down saliency map by using AIM and the static saliency map from the Itti & Koch model, which the bottom figure shows the motion saliency map proposed by our model and the right most is our final fused saliency map.

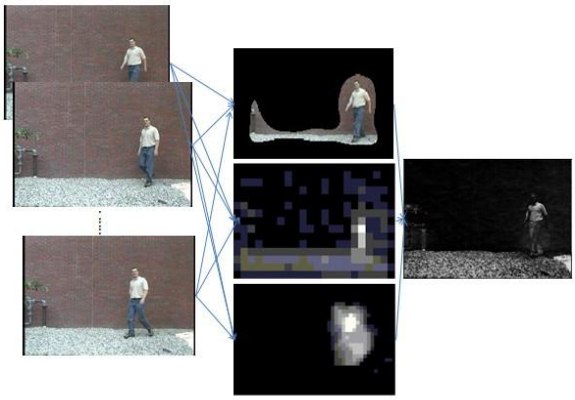


Fig. 4. The first column indicates the first frame and the W frame from the top to bottom respectively. The second row illustrates the top-down saliency map by using AIM and the static saliency map from the Itti& Koch model, which the bottom figure shows the motion saliency map proposed by our

model and the right most is our final fused saliency map

D. Saliency Map Fusion

After multi-channel analysis, we obtain three saliency maps including static saliency map, motion saliency map and top-down saliency map. Suitable fusion of the maps produces the final attention map. Map fusion can be performed with linear and nonlinear methods. In our work we adopt linear method for simplicity and with adaptive coefficients to fit different types of videos. Considering that our goal is to detect Region of Interest (ROI) from the saliency/novelty maps, we model this progress as binary classification and use the variance between classes to determine the fusion coefficients. Saliency map is first classified with the method of maximum variance between classes. Let Var_S be the result variance between classes:

$$Var_S = (n_1(\mu_1 - \mu)^2 + n_2(\mu_2 - \mu)^2) / n \quad (5)$$

where n_1 and n_2 are the number of samples of the two classes, n is the total number of samples. μ_1 and μ_2 are the means of two classes, μ is the mean of all the samples. In equation 6, let Var_s donates for the variance on static saliency maps, Var_{motion} and $Var_{top-down}$ be the variances between classes of motion saliency map and top-down saliency map respectively. Then the fusion weight for static saliency maps is

$$W_s = Var_s / (Var_s + Var_{motion} + Var_{top-down}) \quad (6)$$

The weights for motion saliency map and top-down saliency map are similarly calculated. We obtained w_M and w_N for the weight of motion saliency map and top-down saliency map. Hereby M_S, M_M and M_N indicate the static saliency map, motion saliency map and top-down saliency map respectively. Finally, the fused attention map is computed as

$$AM = w_S M_S + w_M M_M + w_N M_N \quad (7)$$

with the above equation, motion features will be integrated as a part of fused saliency map, as attention is mostly guided by motion cues in dynamic scenes. The top-down saliency map is also fused into the fusion stage. The static saliency map is calculated in the final saliency map based on the bottom-up computational attention model.

The motivation behind the evaluation is to find an efficient and robust fusion method that not only extracts all useful information, but also reduces the impact of false findings.

After applying the proposed maps fusion method, we selected the points higher than the threshold; the threshold is obtained from the subtraction from each consecutive block 3*3. The higher value we get, the more probability of attention region it is, since it implies the higher differences between foreground and background. After we get the labeled points on each frame, if the points occupy on a relatively concentrated area, we then assumed it as the region of interests. To indicate the region of interests, we will add a red circle with the radius of 64 pixels to indicate ROI on each frame. In the experiments and evaluation section, we will test the proposed attention model with various video sequences.

The effect of saliency fusion stage is to make our model more close to the real human visual system. As human visual system is affected by exogenous and endogenous stimuli, our model considers the **top-down** and **bottom-up** mechanism to link a bridge for both biological and artificial vision systems.

IV. Performance Evaluation

To demonstrate the effectiveness of the propose attention model, we have extensively applied the method on several types of video sequences from the benchmarks. The detail of the testing videos is given in Table 1.

A. Data Set and Experimental Results

We applied our system on different types of videos to verify its feasibility and generality. The dataset are from [49][50][51][52], as detailed in Table 1, includes surfing player, parachute landing, outdoor, traffic artery and other video sequences with high or low motion features.

TABLE I. Testing Videos

No.	Video	Subjects	Frames
1	Surfing player	Surfing player	25
2	Glider	Glider pilots	30
3	Traffic artery	Vehicles	108
4	Toll station	Tollbooth collector	258
5	Parachuter landing	Parachuter	33

The first testing set contains video sequences with prominent motions. The video are mainly focusing on the motion surfing player.

In this case, the motion saliency map plays an important role in the feature fusion, when the spatial and temporal cues do not take a large percentage on the equation (4). Hereby the value of the variables in our model is described as the following in the Table 2.

We apply our proposed algorithms to these video clips and the most relevant region of interest will be found out and highlighted with a red circle. All the results are shown in the figures below. As the image size of each frame is 640 by 480 pixels, the circled region of interest can be relatively small as you will notice, for example, in Figure 6.

TABLE II. PARAMETER SETTING

$n = 25$	$n_1 = 9$	$n_2 = 2$	$w_N = 0.3$
$\mu_1 = 4.5$	$\mu_2 = 1$	$\mu = 12.5$	$Var_S = 27.3$
$Var_{motion} = 29.1$	$Var_{top-down} = 24.6$	$w_S = 0.34$	$w_M = 0.36$

The second test video clip is about mostly dynamic scenes that recorded a parachute player descending in the atmosphere. Our model also proves its robustness and efficiency in this experiment. With the motion vector fusion in the attention model, our accuracy reaches 87% since it contains rich entropy and the weight factor is relatively high. The variables change a little due to different total frames and number of samples. We have not listed all the parameters as the table 2 because the effects of saliency maps are quite similar. And interest region is also detected and indicated using a red circle with radius of 64 pixels. The circle is relatively small due to higher resolution of these images (640*480).

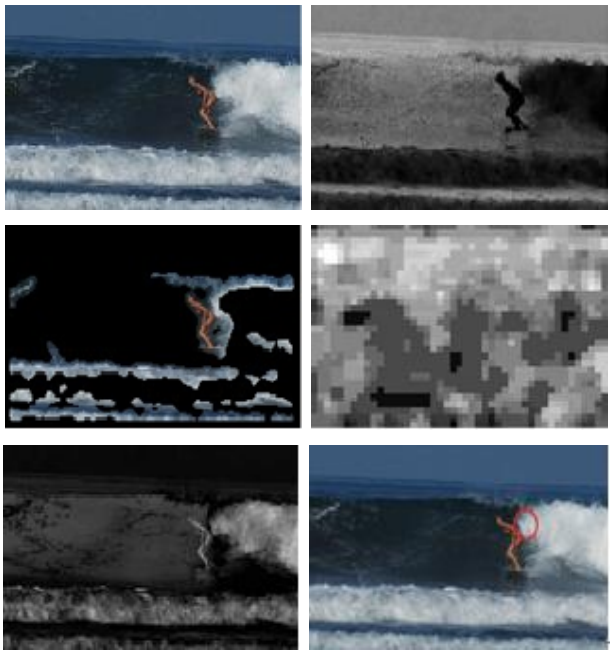


Figure 6. Testing Video 1: the illustration of one example results using our proposed model, the first left image indicates the original video, the right image shows the static saliency image; the left image on the second row shows the top-down saliency map while the right one is the motion saliency map. The third row of the left image is the fused saliency map while the right one is the attention region, red circle indicates our detected attention region by using the fusion of saliency map. The total detected frames are about 25 frames with nearly 16 frames accurately detected.

On the test video 3, the traffic artery with multiple objects contains in the frame. We can accurately detect the approaching vehicles in the video sequences – as shown in the example images in the following Figure 7. The interest region indicated by the red circle is very small, as the size of the images is 960 by 540. On video clip 4 and 5, the accuracy is 74.8% and 78.7%, respectively, which is a satisfactory result overall. The images size in video clip 4 and 5 are 352 by 288 and 640 by 480, respectively.

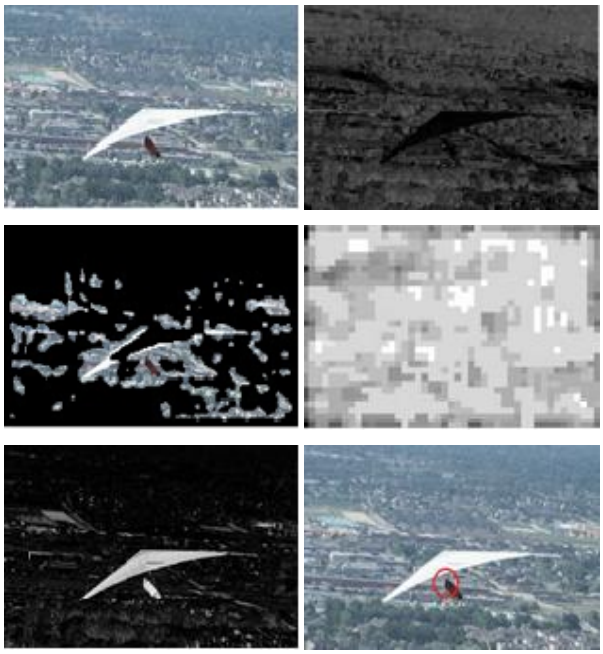
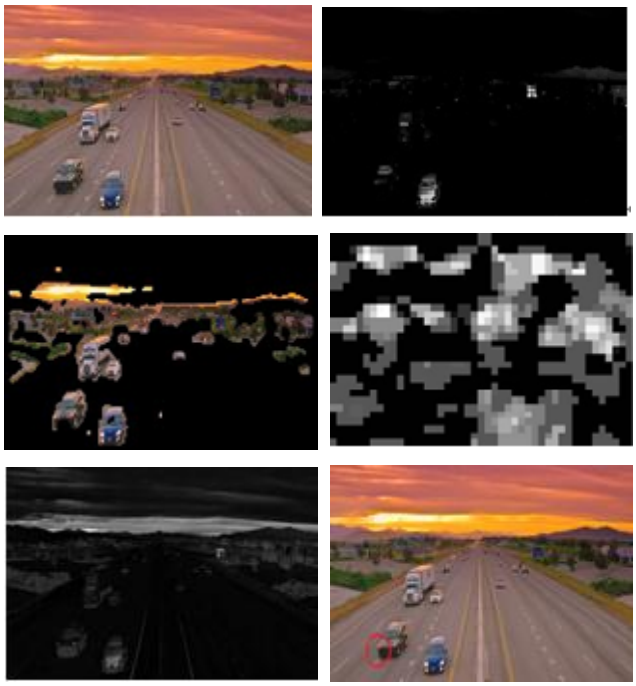


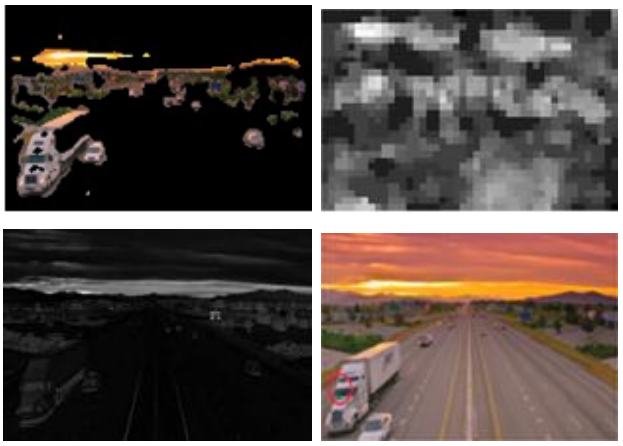
Figure 7. Testing Video 2: the illustration of one example results using our proposed model, the first left image indicates the original video, the right image shows the static saliency image; the left image on the second row shows the top-down saliency map while the right one is the motion saliency map. The third row of the left image is the fused saliency map while the right one is the attention region, red circle indicates our detected attention region by using the fusion of saliency map. The total detected frames are about 30 frames with nearly 19 frames accurately detected.



(Group A, selected from the video frame 7)

(Group B, selected from the video frame 31)

(Group C, selected from the video frame 78)



(Group D,selected from the video frame 93)

Figure 8. Testing Video 3: the illustration of four examples results using our proposed model, the first left image indicates the original video, the right image shows the static saliency image; the left image on the second row shows the top-down saliency map while the right one is the motion saliency map. The third row of the left image is the fused saliency map while the right one is the attention region, red circle indicates our detected attention region by using the fusion of saliency map. The total detected frames are about 108 frames with nearly 95 frames accurately detected. Hereby we randomly select 4 frames results to illustrate our results. They are the frames at 7,31,78 and 93, respectively.

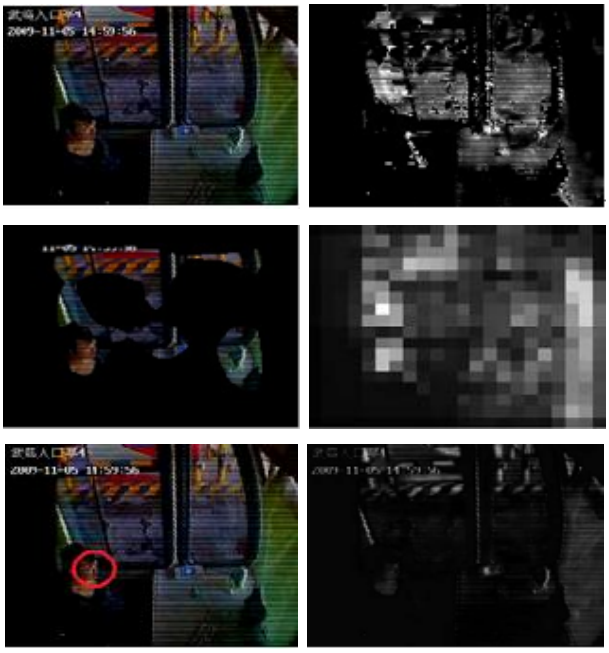


Figure 9. Testing Video 4: the illustration of one example results using our proposed model, the first left image indicates the original video, the right image shows the static saliency image; the left image on the second row shows the top-down saliency map while the right one is the motion saliency map. The third row of the left image is the fused saliency map while the right one is the attention region, red circle indicates our detected attention region by using the fusion of saliency map. The total detected frames are about 258 frames with nearly 193 frames accurately detected.

Figure 10. Testing Video 5: the illustration of one example results using our proposed model, the first left image indicates the original video, the right image shows the static saliency image; the left image on the second row shows the top-down saliency map while the right one is the motion saliency map. The third row of the left image is the fused saliency map while the right one is the attention region, red circle indicates our detected attention region by using the fusion of saliency map. The total detected frames are about 33 frames with nearly 26 frames accurately detected.

B. *Comparison and Evaluation*

In order to further verify the proposed method, we compared our approach with several state-of-the-art methods. We measured the overall performance of the proposed method with respect to precision, recall, and O-measure, and compared them with the performance of existing competitive automatic salient object segmentation methods, such as Itti & Koch's method [53][54], AIM [40] and Achanta's method [55].

According to the standard evaluation methods, precision is the percentage that the detected saliency map divided on the non-ground-truth saliency map as been predicted. Recall is a measure of the percentage provided that the detected saliency map divided on the ground-truth saliency map as been predicted. The highest percentage of precision indicates the real attention as the test participants assumes them as the attention region. The recall is similar as the false positive. O-measure is a special method which predicts the overall performance of the model. The O-measure used in this study is calculated from:

$$O_{\Omega} = \frac{(1+\Omega^2)\, precision * recall}{\Omega^2 * precision + recall} \qquad (8)$$

Here we use Ω =0.25 in our experiments to weigh precision than recall. For all three performance measures, a high percentage is indicating a better performance.

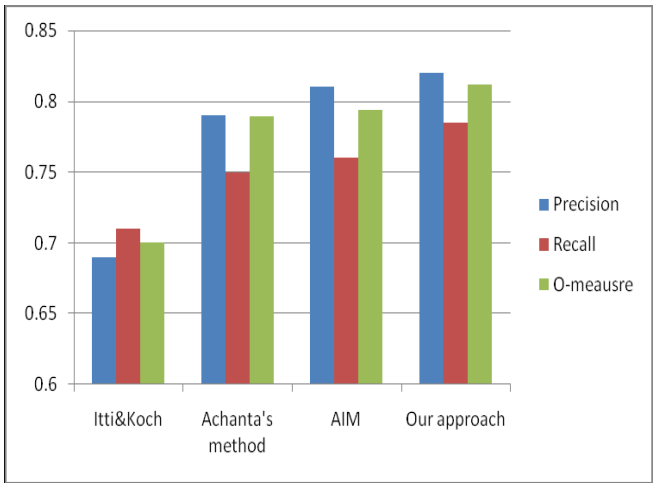


Figure 11. Performance comparison between the proposed method and the state-of-the-art methods. The vertical axe indicates the percentage of precision, recall and O-measure index, the horizontal axes is the approaches proposed and our model.

As shown in Figure 11, we compared the precision, recall and O-measure among the methods of Itti & Koch [53][54], Achanta [55], AIM [40] and proposed model. The proposed method outperformed the state-of-the-art attention models in all three performance measures- precision, recall, and O-measure.

V. Conclusion And Future Work

In this paper, we proposed a novel automatic approach to detect attention region from video sequences. We designed a motion saliency map by using the motion vector field, which shows a high entropy and motion vector on the specific region.

The saliency map fusion can extensively balance the low-level and high-level features and optimize them to the maximum extent as the experiments going. Our experiments demonstrated that the saliency model outperforms other models with respect to performance measures.

This paper has addressed the bottom-up cues into the saliency model, however, in human visual system, top–down attention and the combination of the bottom-up and top-down cues play significant role in the emergence of high level intelligence [56].

It is also believed that the ability to respond to motion cues is vital for not only low level animals such as insects [57], but also import in the emergence of complex human brains. We will further integrate more motion cues into the attention model, and will implement these models to robots for efficient human robot interaction in the future.

Acknowledgement

Thanks to all of the collaborators whose modeling work is reviewed here, and to the members of school of computer science, at the University of Lincoln, for discussion and feedback on this research. This work was supported by the grants of EU FP7-IRSES Project EYE2E (269118), LIVCODE (295151) and HAZCEPT(318907).

REFERENCES

[1] L. Ferman, H. Collewijn, A. Van Den Berg, A direct test of Listing's law-I. Human ocular torsion measured in static tertiary positions, Vision Research 27 (1986) 929–938.

[2] E. Proeto, U and Peter A.Tass,Timing of V1/V2 and V5+ activations during coherent motion of dots:An MEG study,Neuroimage 37(2007)1384-1395.

[3] A. Oliva, A., Torralba, A., Castelhano, M.S., & Henderson, J.M. Top down control of visual attention in object detection. IEEE Proceedings of the International Conference on Image Processing,2003, Vol I, 253-256.

[4] D. Parkhurst, Modeling the role of salience in the allocation of overt visual attention,Vision Research,2002.

[5] C. Rentschler, , Fixations in natural scenes: Interaction of image structure and image content, Vision Research 46 (2006) 2535–2545.

[6] G. Kreiman, T. Serre and T. Poggio. On the limits of feed-forward processing in visual object recognition. Cosyne, February 2007.

[7] A. Treisman, A M, Gelade, G. A feature-integration theory of attention. Cognitive Psychol, 1980, 12(1): 97-136.

[8] R.Einhäuser et al. Differences of monkey and human overt attention under natural conditions, Vision research, 46, 1194–1209,2006.

[9] C. Rentschler, , Fixations in natural scenes: Interaction of image structure and image content, Vision Research 46 (2006) 2535–2545.

[10] M. Cerf, P. Frady, C Koch, Faces and text attract gaze independent of the task: Experimental data and computer model, Journal of Vision, 9(12):10, 1-15,2009.

[11] A. Oliva, Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search,Psychological Review, 113, 766-786.

[12] A.Castelhano, M.S., & Henderson, J.M. Incidental Visual Memory for Objects in Scenes. Visual Cognition, 2005,12, 1017-1040.

[13] H. Donk, M. & van Zoest, W. Effects of salience are short-lived. Psychological Science, 19 (7), 733-739, 2008.

[14] CM Masciocchi , S Mihalas , D Parkhurst , Niebur E.Everyone knows what is interesting: salient locations which should be fixated, Journal of Vision, vol. 9 no. 11 article 25, 2009.

[15] S. Chikkerur, T. Serre, C. Tan and T. Poggio. What and where: A Bayesian inference theory of attention. Vision Research (Special issue on "Mathematical Models of Visual Coding"). 55(22), pp. 2233–2247, Oct 2010.

[16] V. Navalpakkam, L. Itti, Optimal cue selection strategy, In: Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005), pp. 987-994, Cambridge, MA: MIT Press, 2006.

[17] U. Rutishauser, C. Koch, Probabilistic modeling of eye movement data during conjunction search via feature-based attention, Journal of Vision, 7(6):5, 1–20, 2007.

[18] YF Ma, L Lu, HJ Zhang, M Li, "A user attention model for video summarization" ACM Multimedia'02.

[19] S Yue and Rind F. Claire, "Postsynaptic organizations of directional selective visual neural networks for collision detection," Neurocomputing (in press), DOI: 10.1016/j.neucom.2012.08.027.

[20] S Yue and Rind F. claire, "Visually stimulated motor control for a robot with a paire of LGMD visual neural networks,", IJAMechS, (2012)

[21] HY Meng , A Kofi, S Yue, H Andrew, H Mervyn, P Nigel, H Peter "Modified Model for the Lobula Giant Movement Detector and Its FPGA Implementation," Computer Vision and Image Understanding, 2010, vol.114(11), pp.1238-1247. ISSN 1077-3142, doi:10.1016/j.cviu.2010.03.017.

[22] S. Yue and Rind F. Claire, "Collision detection in complex dynamic scenes using a LGMD based visual neural network with feature enhancement," IEEE Transactions on Neural Networks, 2006, vol.17(3), pp.705-716.

[23] J. J. Clark, "Spatial attention and saccadic camera motion," in Proc. IEEE Int. Conf. Robot. Autom., Leuven, Belgium, 1998, pp.3247–3252.

[24] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," IEEE Trans. Image Process.,vol. 17, no. 5, pp. 633–644, May 2008.

[25] R Bodor, B Jackson, and N. Papanikoloupolos, ; "Vision-based human tracking and activity recognition", XI Mediterranean Conf. on Control and Automation, 2003.

[26] J. Tsotsos and A Rothenstein (2011),"Computational models of visual attention", Scholarpedia, 6(1):6201. doi:10.4249/scholarpedia.6201.

[27] K.Tanaka, H.Saito, G. Analysis of motion of the visual field by direction,expansion/contraction,and rotation cells clustered in the dorsal part of the medial superior temporal area of the macaque monkey. Journal of Neurophysiology,vol. 62 no. 3, 626-641,1989.
attention", Scholarpedia, 6(1):6201. doi:10.4249/scholarpedia.6201.

[28] C. Alais, D., & Blake, R. (1999). Neural strength of visual attention gauged by motion adaptation. Nature Neuroscience, 2(11), 1015–1018.

[29] A. Cyrus Arman, "Effects of feature-based attention on the motion aftereVect at remote locations", Vision Research, 46 (2006) 2968–2976.

[30] P. Anandan, "Measuring visual motion from image sequences," Ph.D.thesis, Univ. of Mass., May 1987.

[31] AC Huk & DJ Heeger , Pattern-motion responses in human visual cortex, Nature Neuroscience, 5:72-75, 2001.

[32] L. Itti, C. Koch, and E, Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 20, NO. 11, NOVEMBER 1998.

[33] L. Itti, C. Koch, Feature Combination Strategies for Saliency-Based Visual Attention Systems, Journal of Electronic Imaging, Vol. 10, No. 1, pp. 161-169, Jan 2001.

[34] [34] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20, No. 11, pp. 1254-1259, Nov 1998.

[35] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. BMVC, 2008.

[36] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. ICCV, 2009.

[37] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. SPIE, 2003.

[38] [38] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in IEEE Conf. Computer Vision and Pattern Recognition,2007.

[39] Raj, AlvinStatistical Saliency Model incorporating motion saliency and an application to driving Massachusetts Institute of Technology. Dept. of El.

[40] L.Bruce, N.D.B., Tsotsos, J.K., Attention based on information maximization. Journal of Vision, 7(9):950a, 2007.

[41] M. DeGroot, Optimal Statistical Decisions, McGraw-Hill, (1970).

[42] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in IEEE,Conf. Computer Vision and Pattern Recognition, 2008.

[43] J Li, Visual Saliency Based on Scale-Space Analysis in the Frequency Domain, IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 6, No. 1, January 2007.

[44] J. J. Clark, "Spatial attention and saccadic camera motion," in Proc. IEEE Int. Conf. Robot. Autom., Leuven, Belgium, 1998, pp.3247–3252.

[45] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," IEEE Trans. Image Process.,vol. 17, no. 5, pp. 633–644, May 2008.

[46] TZ Lauritzen , M D'Esposito, DJ Heeger , MA Silver . Top-down flow of visual spatial attention signals from parietal to occipital cortex, Journal of Vision, 9(13):18, 1-14, 2009.

[47] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in IEEE Conf. Computer Vision and Pattern Recognition,2007.

[48] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using hase spectrum of quaternion fourier transform," in IEEE,Conf. Computer Vision and Pattern Recognition, 2008.

[49] ftp://ftp.cs.rdg.ac.uk/pub/PETS2001/

[50] http://cim.mcgill.ca/~lijian/database.htm

[51] http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml

[52] http://people.csail.mit.edu/tjudd/research.html

[53] L. Itti, Quantitative Modeling of Perceptual Salience at Human Eye Position, Visual Cognition, Vol. 14, No. 4-8, pp. 959-984, Aug-Dec 2006.

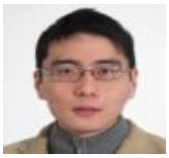
[54] R. J. Peters, L. Itti, Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention, In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2007.

[55] R. Achanta and S. Süsstrunk, Saliency Detection for Content-aware Image Resizing, IEEE International Conference on Image Processing, 2009.

[56] M. Begum and F. Karray, "Visual attention for robotic cognition: a survey," *IEEE Transactions on Autonomous Mental Development*, vol.3(1), pp92-105, 2011.

[57] S. Yue and F.C. Rind, "Redundant neural vision systems - competing for collision recognition roles," *IEEE Transactions on Autonomous Mental Developments*, 2013 (in press) (DOI 10.1109/TAMD.2013.2255050.

AUTHORS PROFILE



Jiawei Xu received the B.S. and M.S. degrees in computer engineering from Shanghai University of Engineering Science and Technology, Shanghai, China, 2007 and Hallym University, Korea, 2010, respectively. Now he is a PhD student in the School of Computer Science, University of Lincoln, UK. His research interests include computer vision, human attention models, and visual cortex modeling. He was a pattern classification engineer in JTV Co.Ltd, Beijing during the year of 2011.

Shigang YUE is a Professor of Computer Science in the Lincoln School of Computer Science, University of Lincoln, United Kingdom. He received his PhD and MSc degrees from Beijing University of Technology (BJUT) in 1996 and 1993, and his BEng degree from Qingdao Technological University (1988). He worked in BJUT as a Lecturer (1996-1998) and an Associate Professor (1998-1999). He was an Alexander von Humboldt Research Fellow (2000, 2001) at University of Kaiserslautern, Germany. Before joining the University of Lincoln as a Senior Lecturer (2007) and promoted to Reader (2010) and Professor (2012), he held research positions in the University of Cambridge, Newcastle University and the University College London(UCL) respectively. His research interests are mainly within the field of artificial intelligence, computer vision, robotics, brains and neuroscience. He is particularly interested in biological visual neural systems, evolution of neuronal subsystems and their applications – e.g., in collision detection for vehicles, interactive systems and robotics. He is the founding director of Computational Intelligence Laboratory (CIL) in Lincoln. He is the coordinator for several EU FP7 projects. He is a member of IEEE, INNS, ISAL and ISBE.