A Computational Model of Extrastriate Visual Area MT on Motion Perception

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 are sensitive to motion perception under complex scenes. Building motion 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 vision system. This paper models the motion perception mechanisms in human extrastriate visual middle temporal area (MT) computationally. MT is middle temporal area which is sensitive on motion perception. This model can explain the attention selection mechanism and visual motion perception to some extent. With the proposed model, we analysis the motion perception under day time with single or multiple moving objects, we then mimic the visual attention process consisting of attention shifts and eye fixations against motion- feature-map. The model produced similar gist perception outputs in our experiments, when day-time images and nocturnal images from the same scene are processed. At last, we mentioned the future direction of this research.**

Keywords—*Motion perception; daytime and nocturnal scenes; spatio-temporal phase*

I. INTRODUCTION

The current research established three criterions on human visual perception. They are sparse criteria, temporal slowness criteria and independent criteria [1]. This paper researches the motion cues based on the sparse standard. The meaning behind the sparse criteria indicates that most neurons show a relatively low response to external stimuli, includes visual, auditory and olfactory signal, etc. Only a few of them yields a distinct activity. The response distribution of one neuron to the stimuli inputs has a property of sparse and discrete. These characters are of paramount importance and lead the dimensional deduction and feature extraction to the visual system research.

Temporal slowness criteria is described as following, the signal and environment are rapid change with the time, however, the features are slowly change with the time. Then, if we can extract slowly-changed features from the visual inputs, such as random motion, angular transformation or spin, the computational algorithm will be robust to the bio-inspired model.

The third criteria means the neuron are independent to external stimuli. The combination of independent feature subspace and multi-dimensional independent component analysis explain this criterion effectively.

This work is supported by the grants of EU FP7-IRSES Project EYE2E (269118).

Motion is a vector defined by direction and speed. In the primate visual system, motion is represented in a specialized pathway that begins in striate cortex (V1), extends through extrastriate areas MT (V5) and MST, and terminates in higher areas of the parietal and temporal lobes [2]. While the neural representation of direction in this pathway, and its relationship to perception, have been studied extensively. With the motion feature integrated into the saliency map, the proposed attention model will be able to respond to motion feature naturally. Motion feature is often a dominant factor in complex dynamic scenes. The model can mimic the visual process after adopting the motion cues into the model.

The middle temporal area (MT) is sensitive to visual motion, as discovered by neurobiologists using electroenc ephalogram (EEG) and gammo-amino butyric acid (GABA) [3]. It links the bridge between LGN (lateral geniculate nucleus), V1 (Primary visual cortex) and MST (medial superior temporal area), the feedback between these area are parallel and circular [4]. Nearly all neurons in MT area show their preference on the specific motion direction and angle. The instantaneous firing rate at the specific phase is 10 times higher than other phases at a certain neuron [5]. The neurons that react similar response to a certain kind of features can compose a neuron cluster and work synchronously [6]. These information may lead to attention shifts, eye fixations although the underlying neuronal mechanism has not been fully understood.

In this paper, we proposed a computational model which can mimic the human visual selection instantaneously. In order to represent the complex and irregular neuron activities, we model the visual motion via the topological way to cluster same response neurons in a cluster way.

This paper is organized as the following. Section 2 will briefly mention the previous work. The mathematical part and algorithm will be discussed in section 3. Experiments and evaluation combines section 4. And last section is conclusion and future work.

II. RELATED WORK AND MODEL FLOWCHART

The current research on MT is mostly based on the electrophysiological recording and micro stimulation experiments. In 2005, Jing Liu [7] studied Correlation between speed perception and neural activity in the middle temporal visual area. They trained rhesus monkeys on a speed discrimination task in which monkeys chose the faster speed of two moving random dot patterns presented simultaneously in spatially segregated apertures. Evidence from these

experiments suggests that MT neurons play a direct role in the perception of visual speed. Comparison of psychometric and neurogenetic thresholds revealed that single and multi-neuronal signals were, on average, considerably less sensitive than were the monkeys perceptually, suggesting that signals must be pooled across neurons to account for performance. The initial research on MT can be traced back to 1988, W T. Newsome [8] found the selective impairment of motion perception following lesions of MT. The injection of the ibotenic acid into MT caused striking elevations in motion thresholds; however, had little or no effect on contrast thresholds. The results indicate that neural activity in MT contributes selectively to the perception of motion.

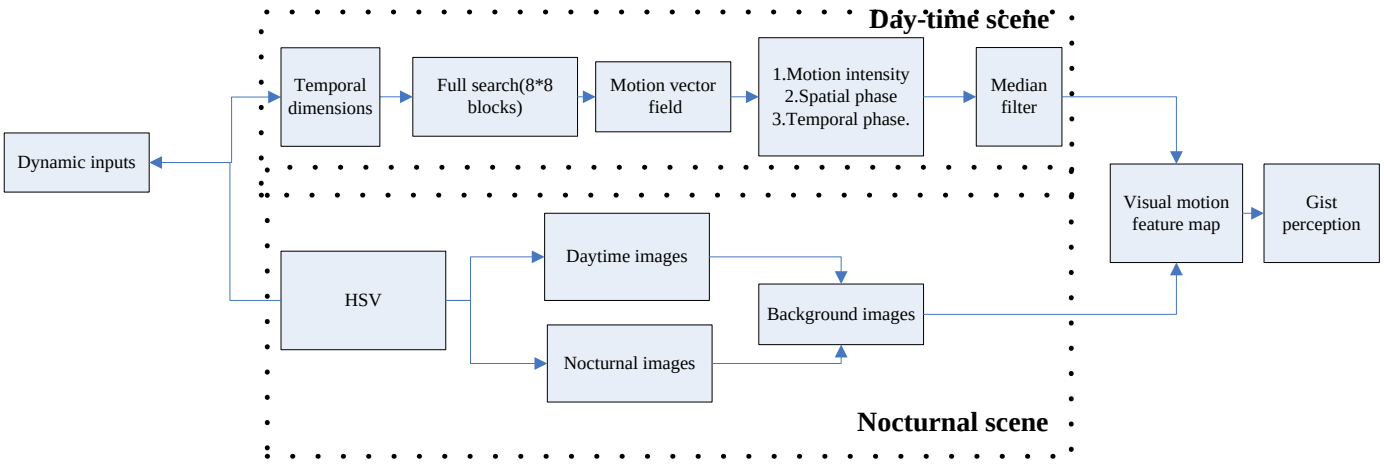


Fig.1. The sketch map of our designed model.

We divide our model into daytime and nocturnal directions, respectively. As many state-of-the art models focus on the daytime images and neglect the night scenes currently, this paper also compare the experimental results with the daytime scenes on the same situation, which further improve the robustness and rationality of proposed model. The framework of our model is represented in figure 1. The system consists of four parts, (1) Motion cues extraction under daytime, (2) objects based segmentation under nocturnal vision, (3) motion perception map, (4) gist perception. In the following section, we will explain the model and algorithm in detail.

III. Computational Model And Algorithm

The model designing concept is described as the following. The motion 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 affects 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.

A. Motion perception under Daytime video clips

In this section, we introduce the architecture of motion attention model under daytime scenes. We integrated this element into our model as previous approaches [9] [10] 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.

TABLE I. Motion perception map

```
%% %%   motion perception map   %%%%
for i=1:f
   if i==f
   break
   end
      frame1=C(i).data;
   frame2=C(i+1).data;
[px,py]=FullSearch(frame1,frame2);
PxPy(i).Mx=px;
PxPy(i).My=py;
MotAtt(i).Intensity=MotionIntensity(PxPy);
MotAtt(i).SpatialPhase=SpatialPhase(PxPy);
end
MotAtt(i-1).TemporalPhase=TemporalPhase(PxPy);
FeatureMap=MedianFilter(MotAtt);
for j=1:i
   if j==i
      break;
   end
SaliencyMap(j).data=FeatureMap(j).Intensity+FeatureMap(j).
TemporalPhase+FeatureMap(i-1).TemporalPhase;
img = 255*mat2gray(SaliencyMap(j).data)
 figure(1)
set(gcf, 'position', [0 0 1366 768]);
subplot(1,1,1)
image(img);
colormap(gray(255));
 axis image off;
end
```

In each frame, the spatial layout of motion vectors would compose a field called Motion Vector Field (MVF) [11]. 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 feature cues, motion intensity cues, spatial phase, and temporal phase, 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. 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:

$$C_s(i,j) = -\sum_{t=1}^{n} p_S(t)\log(p_s(t)) \tag{3}$$

and $$p_s(t) = SH_{i,j}^m(t) / \sum_{k=1}^{n} H_{i,j}^m(k) \tag{4}$$

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:

$$C_t(i,j) = -\sum_{t=1}^{n} p_t(t)\log(p_t(t)) \tag{5}$$

and $$p_t(t) = TH_{i,j}^W(t) / \sum_{k=1}^{n} TH\sum_{i,j}^W(k) \tag{6}$$

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. The detail code is given as the following.

TABLE II. Temporal phase cues

```
function TemporalPhaseVect=TemporalPhase(PxPy)
[p,q]=size(PxPy(1).Mx);
Timesize=length(PxPy);
for m=1:p
   for n=1:q

      for s=1:36
         Num(s)=0;
      end

      for i=1:Timesize
         TimeWinx(i)=PxPy(i).Mx(m,n);
         TimeWiny(i)=PxPy(i).My(m,n);
      x=TimeWinx(i);
      y=TimeWiny(i);
            if x>0&&y>0
   Phase=pi/2*0+asin(abs(y)/sqrt(x^2+y^2));
elseif x<0&&y>0
   Phase=pi/2*2-asin(abs(y)/sqrt(x^2+y^2));
elseif x<0&&y<0
   Phase=pi/2*2+asin(abs(y)/sqrt(x^2+y^2));
elseif x>0&&y<0
```

```
   Phase=pi/2*4-asin(abs(y)/sqrt(x^2+y^2));
elseif x>0&&y==0
   Phase=0;
elseif x<0&&y==0
   Phase=pi;
elseif x==0&&y>0
   Phase=pi/2;
elseif x==0&&y<0
   Phase=pi*3/2;
else Phase=0;
      end
       Angle=Phase/2/pi*360;
Data(i)=Angle;
      for t=0:35
      if (Data(i)-t*10)<10&&(Data(i)-t*10)>=0
         Num(t+1)=Num(t+1)+1;
      end
      if Data(i)==360
         Num(1)=Mum(1)+1;
      end
      end
    TemporalPhaseVect(m,n)=EntropyMethod(Num);
      end
   end
end
end
```

TABLE III. Motion intensity

```
%%%%   motion intensity  & spatial phase  %%%%%%%%
function Intensity=MotionIntensity(PxPy)
MotionVectX=PxPy.Mx;
MotionVectY=PxPy.My;
[m,n]=size(MotionVectX);
[p,q]=size(MotionVectY);
for i=1:m
   for j=1:n
      a=MotionVectX(i,j);
      b=MotionVectY(i,j);
      Vect(i,j)=sqrt(a^2+b^2);
   End
end
LargestVect=max(max(abs(Vect)));
for i=1:p
   for j=1:q
      c=MotionVectX(i,j);
      d=MotionVectY(i,j);
      Intensity(i,j)=sqrt(c^2+d^2)/LargestVect;
   end
end
end
```

TABLE IV. Spatial phase cues

```
%%%%   spatial pahse%%%%
 function TemporalPhaseVect=TemporalPhase(PxPy)
[p,q]=size(PxPy(1).Mx);
Timesize=length(PxPy);
for m=1:p
   for n=1:q

      for s=1:36
         Num(s)=0;
      end

      for i=1:Timesize
         TimeWinx(i)=PxPy(i).Mx(m,n);
         TimeWiny(i)=PxPy(i).My(m,n);

      x=TimeWinx(i);
      y=TimeWiny(i);

      if x>0&&y>0
   Phase=pi/2*0+asin(abs(y)/sqrt(x^2+y^2));
elseif x<0&&y>0
   Phase=pi/2*2-asin(abs(y)/sqrt(x^2+y^2));
elseif x<0&&y<0
   Phase=pi/2*2+asin(abs(y)/sqrt(x^2+y^2));
elseif x>0&&y<0
   Phase=pi/2*4-asin(abs(y)/sqrt(x^2+y^2));
elseif x>0&&y==0
   Phase=0;
elseif x<0&&y==0
   Phase=pi;
elseif x==0&&y>0
   Phase=pi/2;
elseif x==0&&y<0
   Phase=pi*3/2;
else Phase=0;
      end
```

B. Motion perception model under nocturnal video clips

The previous survey confirmed these facts. Cone-shaped and rod cells contain $6*10^6$ and $1.2*10^6$ on human retina, respectively [12]. The former one distributed on the center of retina, however, the later one are located on the periphery of retina. On the day time, human vision and motion perception are completed by the cone-shaped cells. However, rod cells activate its function under night vision. Cone-shaped cells, conversely, need high light intensities to respond and have high visual acuity. Different cone cells respond to different colors (wavelengths of light), which allows an organism to see color [13].

Rod cells are highly sensitive to light, allowing them to respond in dim light and dark conditions. These are the cells that allow humans and other animals to see by moonlight, or with very little available light (as in a dark room). However, they do not distinguish between colors, and have low visual acuity (measure of detail) [14].

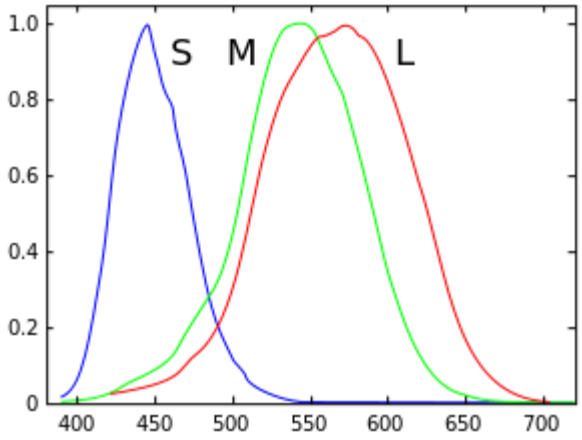


Fig.2. Normalized response spectra of human cone cells, S, M, and L types. Vertical axis: Response [15]. Horizontal axis: Wavelength in nanometers.

Generally, the difficulties of night image problem mainly contain two aspects. The first is that the obtained night image appears much noise, due to reasons of sensor noises or very low luminance. The second is the high light or dark areas in which the scene information cannot be seen clearly by the observers.

To mimic the biological process, we convert the videos from RGB to HSV color space for the convenience of process, and enhance the contrast of video inputs, thus lead to motion estimation at the later stage.

The enhancement of contrast can be classified into 3 steps. The first is calculate contrast *c*, then using the nonlinear transformation to get *c'*, which means x_l to *x* ,then last step is compute the pixel grayscale value using *c'*. The mathematical equation is:

$$c = \frac{|x - x_l|}{x + x_l} \tag{7}$$

$$c' = \psi(c) \tag{8}$$

$$x' = \begin{cases} \dfrac{x_l(1-c')}{(1+c')}, x < x_l \\ L_{\max} - \dfrac{(L_{\max} - x_l)(1-c')}{(1+c')}, x \geq x_l \end{cases} \tag{9}$$

where x_l is the average gray-scale value of attended pixel, $L_{\max}$ is the maximum gray-scale value, while ψ is convex transformation as $\psi(0) = 0, \psi(1) = 1, \psi(c) \geq c$.

Considering the background images of daytime and night are the images of the same scene captured under different illumination. Both objects, such as road, building, cars and player are extracted and the remaining part is classified into background. To distinguish the night vision and daytime vision, we assume if the luminance values of night images background are larger than the luminance of daytime images background, we classify the videos into night videos, vice verse.

After background model estimate, the background image of day and night (DB and NB) are transformed from RGB color space to HSV (Hue-Saturation-Value) color space. An illumination segmentation map $L_{(x,y)}$ can be computed as (10),

$$L_{(x,y)} = \begin{cases} 1(NB_{(x,y)}(V) - DB_{(x,y)}(V)) > 0 \\ 0(NB_{(x,y)}(V) - DB_{(x,y)}(V)) < 0 \end{cases} \tag{10}$$

Where $DB_{(x,y)}(V)$ and $NB_{(x,y)}(V)$ denote the luminance value of background image DB and NB separate at position (x,y) .

To achieve real-time and accurate moving objects segmentation, we first use illumination histogram equalization in the night video $N_{(x,y)}(V)$. Pixles will be classified into *M* levels according to their illuminance. After that different thresholds will be assigned for different classes in the background subtraction. Let $p(i)$ denotes the ratio of pixels, which luminance equals to *i* in $N_{(x,y)}(V)$, G denotes the equalized images,. and it can be computed through the equation (11):

$$G_{(x,y)} = M * f(m), m = 1, ...M \tag{11}$$

Where $f(m) = \sum_{x=0}^{x=m} p(i)$ and $G_{(x,y)}$ will be modified to nearest integral number. For the high light area has already

Been exacted. The motion map M can be computed by

$$M_{(x,y)} = \begin{cases} |N_{(x,y)}(R) - NB_{(x,y)}(R)| \\ |N_{(x,y)}(G) - NB_{(x,y)}(G)| > T(m)\}, or \\ |N_{(x,y)}(B) - NB_{(x,y)}(B)| \end{cases} \tag{10}$$

$$= 0$$

where $T(m)$ represents the threshold at luminance level *m* and $m = G_{(x,y)}$.

The final fusion rule we used is choosing the maximum value of the coefficients of the night input image and daytime reference background image for the high frequency band. For the low frequency band, the coefficients of the images are weighted according to the motion and illumination map.

C. Gist perception under dynamic scene

Recently, situation awareness (SA) [16] has been developed as a theoretical mental model for the gist perception under dynamic scenes.

It includes three levels: perception with focalized attention,comprehension of the current situation, and projection of future status. One interesting point of SA is that it proposes a goal-directed task analysis method to determine what aspects of the situation are important for perception

From the biological review, psychophysical experiments first demonstrated that humans are sensitive to average or centroid position. More recent work by Alvarez and Oliva [17][18] suggests that selective attention may play a minimal role in this process.

Using a multiple object tracking task found that even when observers were unable to identify individual unattended objects, they could localize the centroid of salient objects.

While Chong and Treisman [19] demonstrated that distributed attention could improve an estimate of the mean, this work showed that a summary might be derived even in the absence of attention. Consistent with this, Demeyere and colleagues found that a patient with simultanagnosia could perceive ensemble color in an array of stimuli despite being unaware of the array.

After obtaining the motion cues maps and fixation points, we selected the most gathering fixation points than other regions. After we get these points on each frame, if the points occupy on a relatively concentrated area, we then assumed it as the regions of attention. To indicate the region of interests, we will add a red circle with the radius of 64 pixels to indicate gist perception on visual scenes.

The computational results are elaborated in Section IV. We implement 4 groups of experiments and made the performance evaluation to compare our model's effectiveness with other standard models.

IV. EXPERIMENTS AND RESULTS

To demonstrate 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 results is given in table 3.

D. Benchmark Datasets

We applied our model on different types of videos to verify its feasibility and generality. The dataset are from [20][21][22][23], as detailed in Table 2, includes surfing player, parachute landing, outdoor, traffic artery and other video sequences with high or low motion features.

By implementing two kinds of experiments, we are intended to verify two predictions. The first one is to measure the motion effects on the judgment of human visual attention selection. We prove this predication by comparing the static attention selection model and the results generated by our model is more close to the ground-truth results, pointed out by the participants with normal or corrected vision. The second one is the potential eye fixations on video clips, we are trying to verify the predication that eye saccade yields simultaneous fixations in a millisecond time; however, human eyes are inclined to select the most dense regions with the fixation numbers. This predication matches the result as we can see from experiment 2.

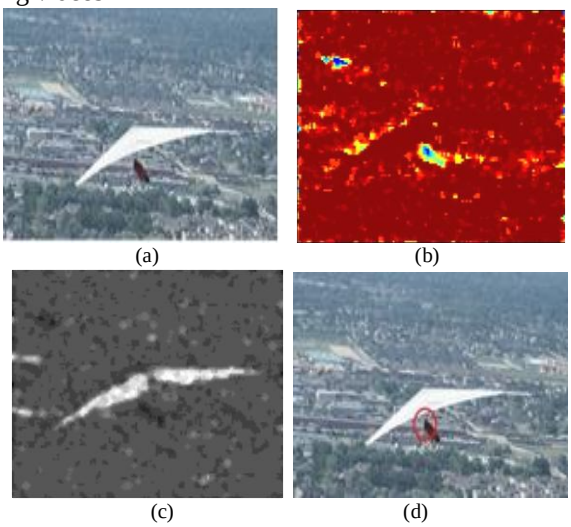
E. Experiments on the motion perception under daytime

The first group experiments are based on the single object moving on the video clip. The tests are short movies with AVI format and 1366*768 frame sizes, 15 fps.

TABLE V. Benchmark testing datasets of daytime vision

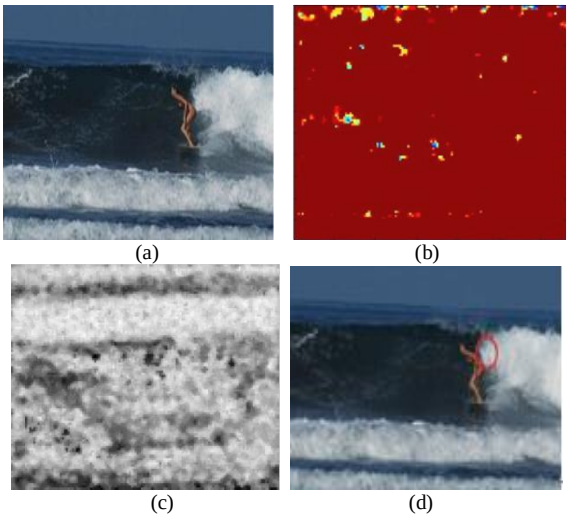
No.	Video Subjects	Temporal dimension
1	Surfing player	22
2	Glider	18
3	Moto cycler	25
4	Traffic artery	109

Testing videos



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

Fig.3. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 22 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.



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

Fig.4. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 18 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

The following figures emphasis multiple moving objects on the testing videos, we need to verify the model's robustness under more complex background.

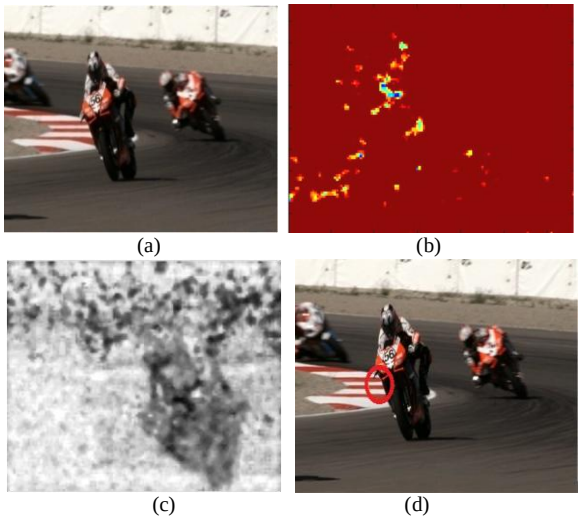


Fig.5. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 25 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

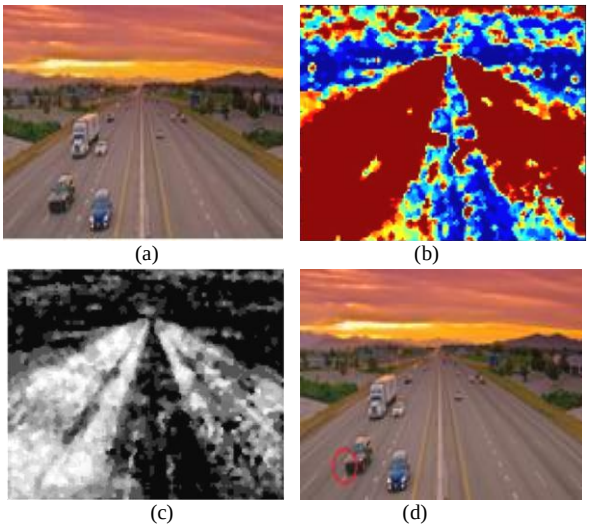


Fig.6. From top to bottom row, we illustrate a sample frame of one testing video clips shown on (a), (b) is the entropy map obtained by computing 109 temporal dimensions, (c) is the visual motion-feature-map, (d) is the gist perception.

From these experiments in figure 5 and figure 6, we conclude these common features as the following. First, the computational burden increase exponentially with the temporal dimension, our testing platform is based on a Windows 7 Intel Core i5 laptop using Matlab 2010b software. The shortest time is 5.73s; the longest time is 24.62s, respectively. Second, visual motion-feature-map in figure 5 (c) and 6 (c) indicate the dynamic motion vectors by computing the pre-setting temporal dimensions, the whiten area indicates higher entropy and motion activity area; however the darker area is relatively low-motion area. Third, the gist perception is based on the weight competition based on the maximum motion cues. Every weight competition computes for one fixation and the maximum value will be selected as the gist perception which represented by red circle for the saliency output. This is discussing in experiment 2.

F. Experiments on the eye fixations and motion perception under daytime

In this experiment, we analysis the relationship between the eye fixation and motion cues map. As we can find in figure 3, the potential eye fixations are representing by the symbol "+". We test on a new video clip with the genre "parachutes landing", each frame corresponds to a motion cues map as we show in figure 4.

Frame 30:

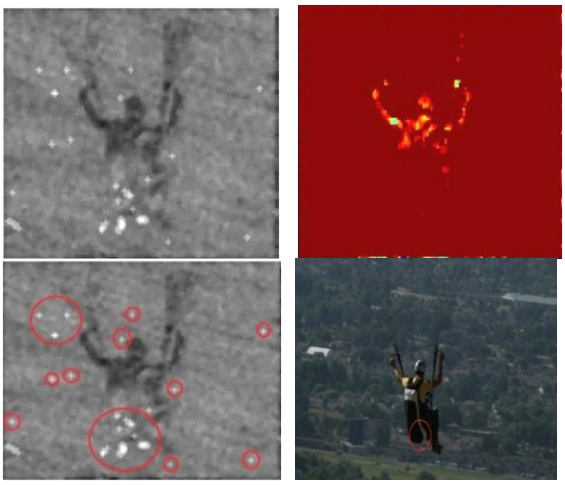


Fig.7. From left to right on first row, the left image is the motion cues map composed by 33 frames, while the right one is the corresponding entropy response with red setting background.

Frame 11:

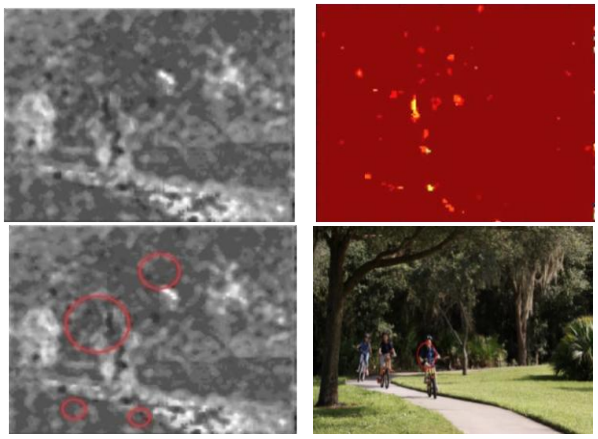


Fig.8. Another testing video with same methods in figure 4, only difference is 19 frames in total.

As shown in figure 7 and figure 8, in this group, we detected the salient regions on the center of the map and white "+" symbols are mostly scattered on the middle-bottom and left-center parts of the image. The white "+" symbols indicate the eye fixation regions; we can find the distinct result that most eye fixation regions are located on the parachutes with a larger circle. We also find other fixations with relatively small circle on the other part of images; however, these points will be selected as the sub-salient region according to WTA

(Winning-take-all) and IOR (Inhibition of return) mechanisms. The right image of bottom row indicates the saliency.

G. Experiments on the nocturnal motion perception

In this part, we illustrated the results by using the algorithm from part B of section 3. The experiments are based on the capture the same position scenes during daytime and night, using the high illumination to get the motion maps.

The figure 9 represents the images after contrast enhancement. Figure 10 shows the images under daytime and night background, then we computes the motion perception map by using the equation (7) (8) (9).

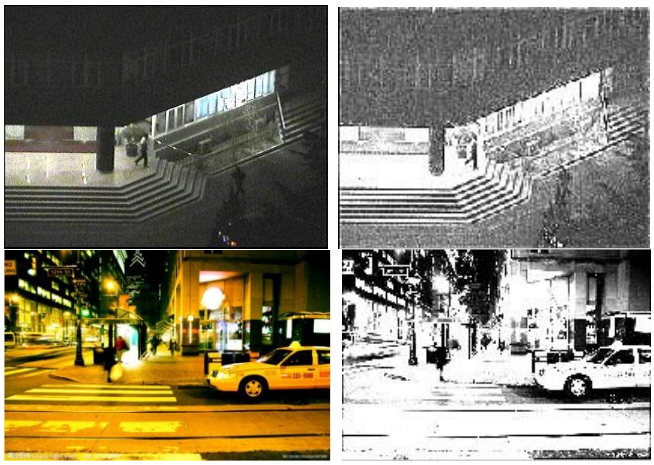


Fig.9. Frames enhancement examples by using the histogram equalization

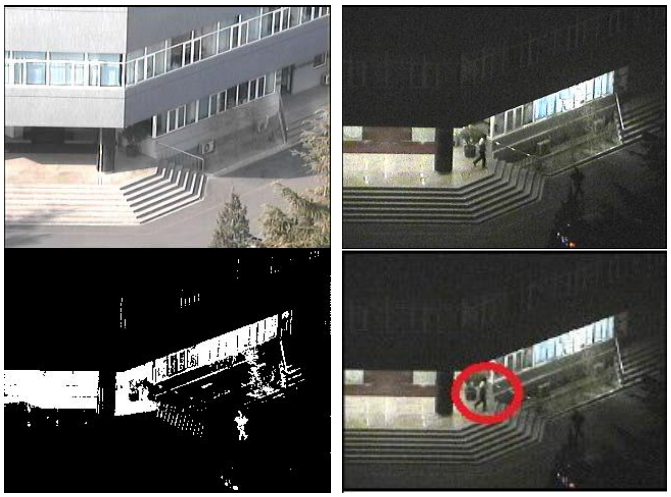


Fig.10. motion perception under nocturnal scenes. Top row, from left to right, daytime input video and night input video. Bottom row, motion perception map and gist perception of scenes.

H. Performance Evaluation

In order to further verify the proposed method, we compared our approach with several state-of-the-art methods. A lot of measure standard have been proposed since the attention models pop out.

Generally, there are 2 criteria adopted in the evaluation, the salient information is well displayed, quantify the attention models to sticking out the salient region. We measured the overall performance of the proposed method with respect to precision, recall, and F-measure, and compared them with the performance of existing competitive automatic salient object segmentation methods, such as Itti & Koch's method [24] [25], AIM [26] and Achanta's method [27].

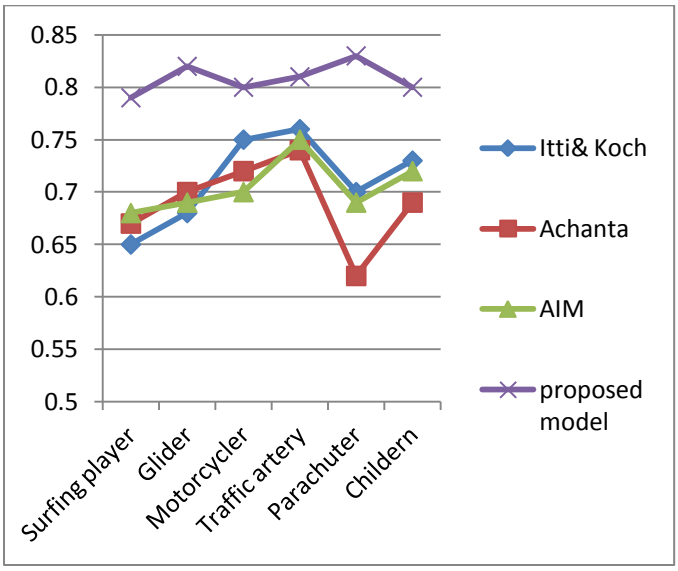


Fig.11. Evaluation of our proposed method under daytime

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. F-measure is a special method which predicts the overall performance of the model. Precision (P), recall (R), F-measure used in this study is calculated from:

$$P = \sum(S * A) / \sum(S) \quad (11)$$

$$R = \sum(S * A) / \sum(A) \quad (12)$$

$$F = 2 * P * R / (P + R) \quad (13)$$

Here S donates the proposed attention regions, A is the ground truth attention regions, $S*A$ indicates the gray-scale image by the gray value of pixel wise multiplication. $\sum(....)$ is the summation of the gray value of each pixel. Obviously, a larger value F means a better effect result.

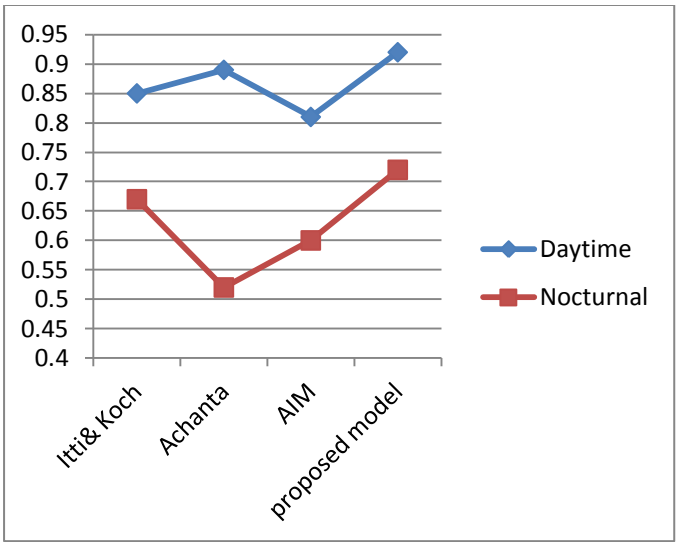


Fig.12. Evaluation of our proposed method under daytime and nocturnal scenes.

In figure 13, the horizontal axes are the proposed model by our model and other state-of-the-art models. We proposed three kinds of performance standards as the motion perception, eye fixations and nocturnal vision were compared with the ground-truth data (best result as 1), the vertical axes shows our results improved overall performance on these evaluation standards.

V. DISCUSSION AND CONCLUSION

In this paper, we proposed a new method to estimate the visual motion process on human visual attention and eye fixations by constructing a computational model. This is a novel and state-of-the-art way. Besides, a serial of comprasions are implemented to test the robustness on the model via the day-time and nocturnal scenes. Unlike psychological methods, the technique using computer vision explains the human attention selection more vividly. This model can explain the attention selection mechanism and visual motion perception to some extent. With the proposed model, we analysis the motion perception under day time with single or multiple moving objects, we then mimic the visual attention process consisting of attention shifts and eye fixations against motion- feature-map. The model produced similar gist perception outputs in our experiments, when day-time images and nocturnal images from the same scene are processed. At last, we mentioned the future direction of this research.

We focus on the motion cues and the effects on the human visual system. Generally, the results are satisfactory and we are trying to simulate the motion effects in the top-down and bottom-up pathway. As they will leads to different outputs if we consider individual agents in the real world. The daytime and nocturnal vision is also compared via different approaches.

This paper has addressed the motion cues into the human visual model, however, in real life, motion perception are mostly irregular and abrupt. The video clips are selected from benchmark and normalized before the experiments. The robustness of algorithm needs improvement in next stage. Also, it is also believed that the visual neurons to respond to motion cues is vital for not only low level animals such as insects, but also import in the emergence of complex human brains [28][29][30][31][32].

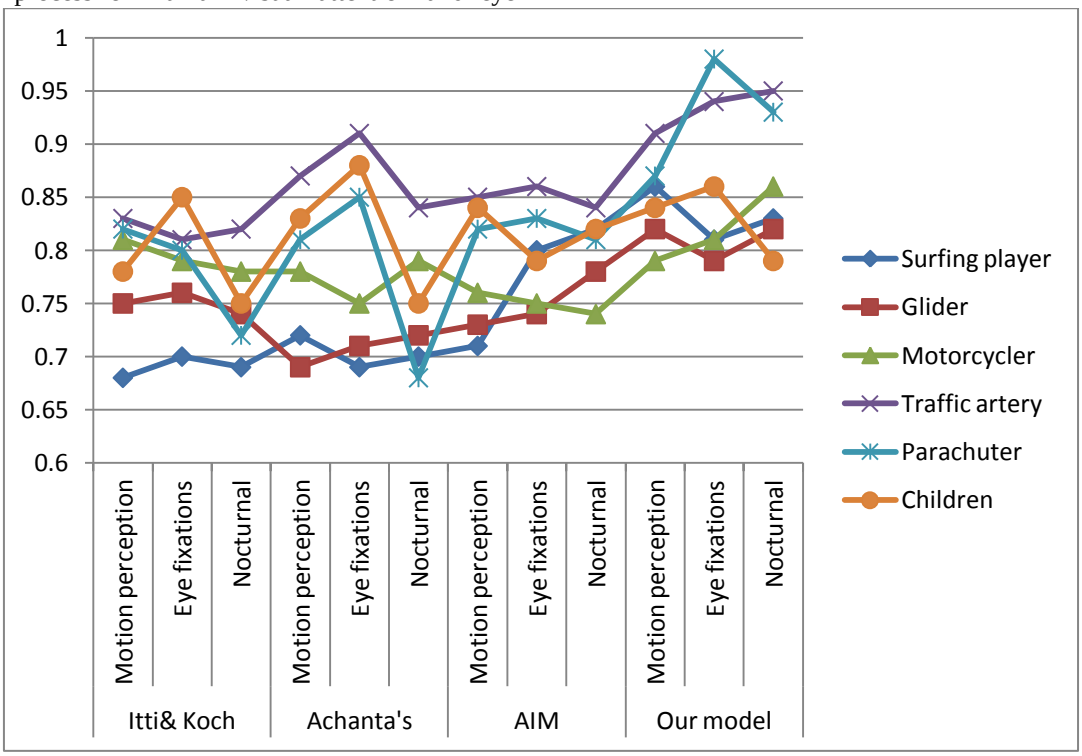


Fig.13. This figure indicates the precision (P), recall (R), F-measure comparions between the proposed method and other state-of-the-art methods under various testing standards, such as motion perceptionm eye fixations and nocutrunal vision.

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. Another important factor is the top-down cues will affect our visual decision largely during the daily life, this issue has been proved by Yang [33] and other scholars [34]. The later stage is to intergate motion cues and top-downs cues together which can reflect the visual processing and enhance the model's robustness in the future work.

VI. 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] C. Koch. "The quest for consciousness", Roberts & Company Publishers, 2004.
[2] J. Maunsell, D. Essen "Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation.". J Neurophysiol 49 (5): 1127–47, 1983.
[3] C. Rodman, "Afferent basis of visual response properties in area MT of the macaque. I. Effects of striate cortex removal". J Neurosci 9 (6): 2033–50, 1988.
[4] W.Desimone, ""Selective Attention Gates Visual Processing in the Extrastriate Cortex". Science 229(4715), 1985.
[5] M.Mercier,Sophie, "Motion direction tuning in human visual cortex", European Journal of Neuroscience,Vol29,pp424-434,2009
[6] YS Bonneh, , Motion-induced blindness and microsaccades: cause and effect, Journal of Vision, 10(14):22, 1-15, 2010.
[7] J. Liu and Newsome, WT. Correlation between speed perception and neural activity in the middle temporal visual area. J. Neurosci. 25(3):711-722, 2005.
[8] W. Newsome and EB Paré,. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). J. Neurosci. 8: 2201-2211, 1988.
[9] J. Tsotsos and A Rothenstein ,"Computational models of visual attention", Scholarpedia, 6(1):6201. doi:10.4249/scholarpedia.6201, 2011.
[10] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in IEEE Conf. Computer Vision and Pattern Recognition,2007.
[11] YF Ma, L Lu, HJ Zhang, M Li, "A user attention model for video summarization" ACM Multimedia, 2002..
[12] C.W Oyster,"The human eye: structure and function". Sinauer Associates, 1999.
[13] E.Strettoi, "Complexity of retinal cone bipolar cells". Progress in Retinal and Eye Research, 29 (4), pg. 272-283, 2010.
[14] A.Roorda, D.R.Williams, "The arrangement of the three cone classes in the living human eye". Nature 397 (6719): 520–522, 1999.
[15] R. W. G. Hunt . "The Reproduction of Colour" (6th ed.). Chichester UK: Wiley–IS&T Series in Imaging Science and Technology. pp. 11–12, 2004.
[16] 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.
[17] 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.
[18] Alvarez, G.A., & Oliva, A. Spatial Ensemble Statistics: Efficient Codes that Can be Represented with Reduced Attention. *Proceedings of the National Academy of Sciences*, 106, 7345-7350, 2009..
[19] SC Chong, A Treisman, Statistical processing: computing the average size in perceptual groups.Vision Research 45, 891–900, 2005.
[20] ftp://ftp.cs.rdg.ac.uk/pub/PETS2001/
[21] http://cim.mcgill.ca/~lijian/database.htm
[22] http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml
[23] http://people.csail.mit.edu/tjudd/research.html
[24] L. Itti, Quantitative Modeling of Perceptual Salience at Human Eye Position, Visual Cognition, Vol. 14, No. 4-8, pp. 959-984, Aug-Dec, 2006.
[25] 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.
[26] L.Bruce, N.D.B., Tsotsos, J.K., Attention based on information maximization. Journal of Vision, 7(9):950a, 2007.
[27] R. Achanta and S. Süsstrunk, Saliency Detection for Content-aware Image Resizing, IEEE International Conference on Image Processing, 2009.
[28] S. Yue and F.C. Rind, "Redundant neural vision systems - competing for collision recognition roles," *IEEE Transactions on Autonomous Mental Developments*, 2013 (in press).
[29] 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.
[30] S Yue and Rind F. claire, "Visually stimulated motor control for a robot with a paire of LGMD visual neural networks,", IJAMechS, 2012.
[31] 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, vol.114(11), pp.1238-1247, 2010.
[32] 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, vol.17(3), pp.705-716, 2006.
[33] J.Yang, M.Yang: Top-down visual saliency via joint CRF and dictionary learning. CVPR 2012: 2296-2303
[34] C.Kanan, M. H., Zhang, G. W. SUN: Top-down saliency using natural statistics. Visual Cognition, 17: 979-1003, 2009.

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.