Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 5 (2013) 243 – 248

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Parallel and Distributed Computing Systems

Bio-inspired Motion Attention for Gist Perception under Spatio-Temporal Dimension

Jiawei Xu[a*], Shigang Yue[b]

[a,b]School of Computer Science, University of Lincoln, Lincoln, LN6 7TS, United Kingdom

Abstract

This paper analysis the motion effects on human visual system.By using electroenc ephalogram (EEG) and gammo-amino butyric acid (GABA) we prove middle temporal area (MT) are sensitive on motion perception. It links the bridge between LGN(lateral geniculate nucleus), V1(Primary visual cortex) and MST (medial shlerior temporal area), the feedback between these area are paraelle and circular. Nearly all neurons in MT area show their preference on the specific motion direction and angle. The instantanous firing rate at the specific phase is 10 times higher than other phases at a certain neuron. The neurons that react similar response to a certain kind of features can compose a neuron cluster and work synchronously. The contribution of us is to simulate a computational motion field model, which can refelect the neuron activities in MT. This model can explain the attention selection mechanism and visual perception to some extent. However, further endeavor should be addressed on the interconnection and correlation of neurons.

Keywords: Motion field, multi-scale motion vector, spatio-temporal dimension

* Corresponding author. Phone: +44 (0) 7574339972
E-mail: jxu@lincoln.ac.uk or abyssecho@hotmail.com

doi:10.1016/j.aasri.2013.10.085

1. Introduction

The current researches acknowledge three criterions on human visual perception. They are sparse criteria, temporal slowness criteria and indepent criteria. This paper research the motion cues based on the sparse standard. The meaning behind the sparse criteria advice that most neurons shows 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 deducation and feature extraction to the visual system research.

Temopral 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.

In this paper, we mimic a computational model which cans reflect the human visual selection instantaneously. In order to represent the complex and irregular neuron activities, we simplify and quantify the problem then it can be simulated on the PC easily. The framework is illustrated in figure 1..

The paper is organized as the following, the section 2 elaborated on the model and algorithm which gives a detail description of algorithm. The experiments part is given in section 3. Section 4 shows the conclusion and comparsion with other state-of-the art models.

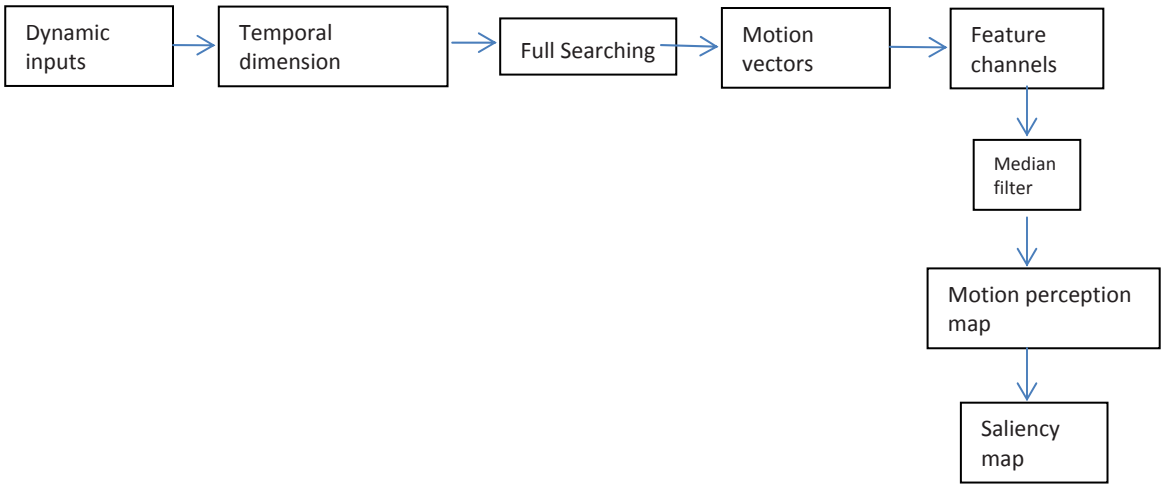

Figure 1. The framework of our proposed model

2. Model analysis and algorithm designing

The model designing concept is desciribed 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 orientations weight 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.

We select the uncompressed video samples to keep the information fidelity. In each frame, the spatial layout of motion vectors would compose a field called Dynamic Motion Perception Field [3]; the terminology can be looked as the photoreceptor of human eyes, all motion vectors can be replaced as the perceptual response of optic nerves [3]. The model is setted into 3 types of feature cues, motion intensity cues, spatial phase, and temporal phase, when the moving vector in dynamic motion perception field go through such cues, they will be transformed into three kinds of feature maps. We fuse the normalized output of cues into a motion cues map and gist perception by the 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. The results of our experiments are shown in figure 2 which gives the flowchart of the original inputs video, the motion cues map and the corresponding gist perception output. The first video clip is based on the 22 spatio-temporal dimensional, while the others are 18 25, 109 dimensions, respectively. Following on the prediction, we have three cues at every position of full searching block. Each 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 8*8 Marco block due to the computational burden. Then the intensity cues can be obtained by computing the magnitude of motion vector.

$$I(i,j) = \sqrt{dx_{i.j}^2 + dy_{i,j}^2} \,/\, MaxMag \tag{1}$$

here (dx_{ij}, dy_{ij}) indicate two component product of motion vector, and $MaxMag$ is the maximum value in dynamic motion perception field. The spatial coherence phase induces the spatial phase consistency of moving 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. Firstly, 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, the phase distribution is calculated by the full searching entropy as following:

$$\begin{aligned} C_s(i,j) &= -\sum_{t=1}^{n} p_S(t)\log(p_s(t)) \\ p_s(t) &= SH_{i,j}^m(t) / \textstyle\sum_{k=1}^{n} H_{i,j}^m(k) \end{aligned} \tag{2}$$

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

$$\begin{aligned} C_t(i,j) &= -\sum_{t=1}^{n} p_t(t)\log(p_t(t)) \\ p_t(t) &= TH_{i,j}^W(t) / \sum_{k=1}^{n} TH \sum_{i,j}^{W}(k) \end{aligned} \tag{3}$$

where C_t denotes temporal coherence, $TH_{i,j}^W(t)$ 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.

3. Experimental Results

To evaluate the efficiency of the gist perception, we have applied the approaches on several types of benchmark video sequences from the benchmarks. The detail of the testing results is given in table 2.We test our model on a Acer laptop, each consisting of a Quad-Core 2.1 GHz IntelXeon Processor, the benchmark video testing dataset is list as the table 1 and based on different genre.

Table 1. An example of a table

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

Table 2. The salient regions comparisons, hereby we set the ground truth as 1, the ground truth salient regions is labelled by 10 participants.

	ROC area
Itti& Koch saliency model [8,9]	0.7812
Achanta's saliency model [11]	0.7458
AIM [10]	0.7319
Our model	0.8316

4. Conclusion and Future Work

Our contribution can be listed out as the following. We propose a novel method on the motion effects via bio-inspired human visual system by a computational model. This is a novel and state-of-the-art way. Unlike psychological methods, the technique by using computer vision can describe the human attention selection more vividly. 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 index widely 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 [8][9], AIM [10] and Achanta's method [11].

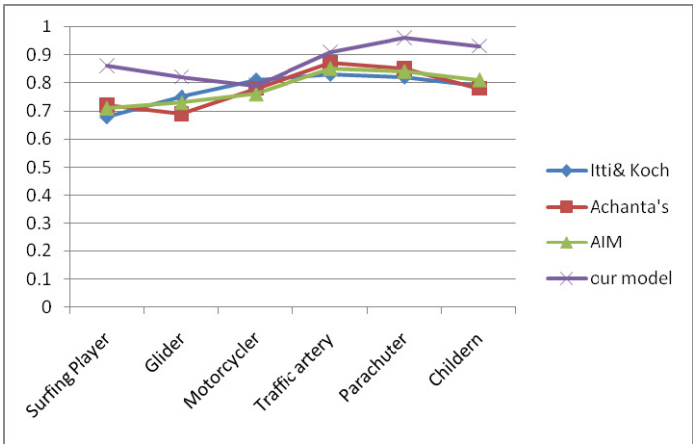

Figure 2. Evaluation of our proposed method

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:

$$P = \sum(S * A) / \sum(S)$$
$$R = \sum(S * A) / \sum(A)$$
$$F = 2 * P * R / (P + R) \quad (4)$$

Here S donates the proposed attention regions, A is the ground truth attention regions, S*A indicates the grayscale 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.

This paper focus on the motion cues and the effect bring to 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.

Acknowledgements

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]Liu, J and Newsome, WT (2005). Correlation between speed perception and neural activity in the middle temporal visual area. J. Neurosci. 25(3):711-722.
[2]Newsome, WT and Paré, EB (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). J. Neurosci. 8: 2201-2211.
[3]YF Ma, L Lu, HJ Zhang, M Li, "A user attention model for video summarization" ACM Multimedia'02..
[4]http://www.viratdata.org
[5] http://cim.mcgill.ca/~lijian/database.htm
[6]http://www-prima.inrialpes.fr/PETS04/caviar_data.html
[7]http://people.csail.mit.edu/tjudd/research.html
[8]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.
[9]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.
[10]L.Bruce, N.D.B., Tsotsos, J.K., Attention based on information maximization. Journal of Vision, 7(9):950a, 2007.
[11]R. Achanta and S. Süsstrunk, Saliency Detection for Content-aware Image Resizing, IEEE International Conference on Image Processing, 2009.
[12]L Chelazzi, M Corbetta, Cortical mechanisms of visuospatial attention in the primate brain,The new cognitive neurosciences, 667-686