Digital Signal Processing 22 (2012) 760–767

Contents lists available at SciVerse ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A dynamic saliency attention model based on local complexity

Longsheng Wei [a,b], Nong Sang [a,*], Yuehuan Wang [a], Qingqing Zheng [a]

[a] Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, 430074, China
[b] Faculty of Mechanical and Electronic Information, China University of Geosciences, Wuhan, 430074, China

ARTICLE INFO

Article history:
Available online 4 May 2012

Keywords:
Dynamic saliency model
Attentional selection
Local complexity
Saliency map

ABSTRACT

A dynamic saliency attention model based on local complexity is proposed in this paper. Low-level visual features are extracted from current and some previous frames. Every feature map is resized into some different sizes. The feature maps in same size and same feature for all the frames are used to calculate a local complexity map. All the local complexity maps are normalized and are fused into a dynamic saliency map. In the same time, a static saliency map is acquired by the current frame. Then dynamic and static saliency maps are fused into a final saliency map. Experimental results indicate that: when there is noise among the frames or there is change of illumination among the frames, our model is excellent to Marat's model and Shi's model; when the moving objects do not belong to the static salient regions, our model is better than Ban's model.

1. Introduction

The human visual system can effortlessly detect an interesting region or object in natural scenes through the selective attention mechanism. Motion is clearly involved in visual attention based on the fact that people's attention is more easily directed to a motive stimulus in a static scene. Therefore, the human visual system interprets not only a static input scene but also a dynamic input scene with the selective attention mechanism.

Most computational models [1–7] of visual attention are static and are inspired by the concept of feature integration theory [8]. The most popular is the one proposed by L. Itti et al. [9] and it has become a standard model of static visual attention, in which salience according to primitive features such as intensity, orientation and color are computed independently. There are also many models [10–16] bringing dynamic salience to visual attention mechanism. Marat et al. [17] used an optical flow method to compute the dynamic salience. The optical flow method does not require any prior knowledge of scene to detect dynamic objects, and it can also deal with the instance of background motion. However, the optical flow method relies on the assumption of luminance constancy, so the result is easy to be affected by illumination and noise. Shi and Yang [18] proposed a model for motion detection in a video, in which dynamic part is obtained by frame difference. This model is very simple and it can obtain dynamic saliency map quickly. However, the result is easy to be affected by threshold and noise. Ban et al. [19] also proposed a dynamic visual selective attention model. Firstly, a static saliency map was obtained by a frame in a video. Secondly, an optimal scale was calculated for each pixel location and for each static saliency map. Thirdly, those optimal scales and static saliency maps were used to calculate the entropy to form an entropy map for every frame. At last, all the entropy maps were used to calculate a new entropy map, which was called dynamic saliency map. However, when the moving objects do not belong to the salient regions, Ban's model is very hard to attend the moving objects.

In order to address the above problem, we propose a dynamic saliency attention model based on local complexity in this paper. This model includes a dynamic attention phase and a static attention phase. In the dynamic attention phase, low-level visual features are extracted from current and some previous frames in a short video. Every feature map is resized into some different sizes. The feature maps in same size and same feature for all the frames are used to calculate a local complexity map. These complexity maps in same feature and different size are normalized and are fused into a dynamic map. All the dynamic maps in different feature are fused into a dynamic saliency map. In the static attention phase, same features are extracted and form multi-scale feature maps by center-surround differences in current frame, and then those feature maps are transformed into conspicuity maps, which are linearly combined into a static saliency map. Our proposed model decides salient regions based on a final saliency map which is generated by integration of the dynamic and the static saliency map. At last, the sizes of each salient region are obtained by maximizing entropy of the final saliency map. Our proposed model is shown in Fig. 1. The contents in the gray boxes are discussed in this paper.

The remainder of this paper is organized as follows. Section two presents the dynamic saliency model including feature extraction

* Corresponding author.
E-mail address: nsang@hust.edu.cn (N. Sang).

1051-2004/$ – see front matter
http://dx.doi.org/10.1016/j.dsp.2012.04.017

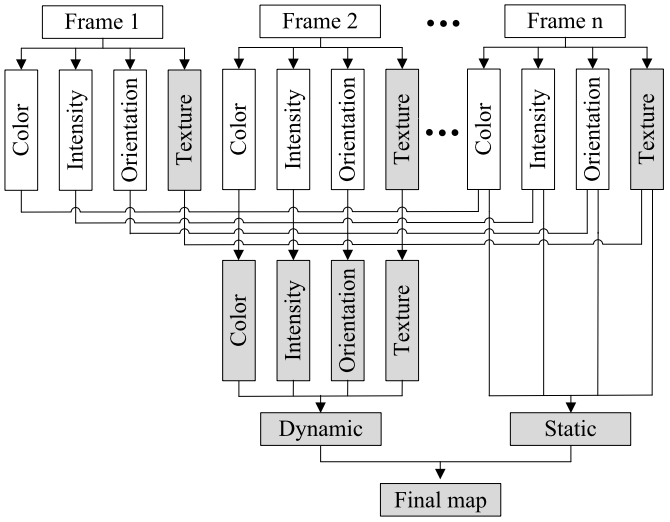

Fig. 1. Our model: Firstly, all the visual features are extracted from current and some previous frames; every feature map is resized into some different sizes. Secondly, a local complexity map is obtained by combining the feature maps in same size and same feature; all the local complexity maps are normalized and are fused into a dynamic saliency map. Thirdly, same features are extracted and form multi-scale feature maps by center-surround differences in current frame, and then through across-scale combinations, those feature maps are transformed into conspicuity maps, which are linearly combined into a static saliency map. At last, dynamic and static saliency map are fused into a final saliency map, which guides human visual attention.

and dynamic saliency map. While attentional selection is described in section three, this part introduces how to acquire static saliency map, final saliency map and the size of salient region. Section four shows experimental results, and section five concludes this paper.

2. Dynamic saliency model

Our proposed model is inspired by the human visual system from the retina cells to the complex cells of the primary visual cortex. The retina extracts two signals from each frame that corresponding to the two main outputs of the retina [20]. Each signal is then decomposed into elementary features by a bank of cortical-like filters. These filters are used to extract both dynamic and static information, according to their frequency selectivity, providing two saliency maps: a dynamic and a static one. Both saliency maps are combined to obtain a spatiotemporal saliency map [17]. Our model decomposes the input short video into different frequency bands: a lower spatial frequency one to simulate the dynamic output and a high spatial frequency one to provide a static output.

In this part, basic visual features are extracted from every frame in a short video. Every feature map is resized into some different sizes, which are transformed into lower gray-scale level. The feature maps in same size and same feature for all the frames are used to calculate local complexity map. All the local complexity maps are normalized and fused into a dynamic saliency map.

2.1. Feature extraction

For every frame in a short video, ten low-level visual features including two color contrast features, two intensity contrast features, four orientation features and two texture features are extracted in this passage. Let r, g and b are three color channels of input image, four broadly-tuned color channels are created: $R = r-(g+b)/2$ for red, $G = g-(r+b)/2$ for green, $B = b-(r+g)/2$ for blue, and $Y = (r+g)/2-|r-g|/2-b$ for yellow (negative values are set to zero). $RG = |R-G|$ is red/green contrast; $BY = |B-Y|$ is blue/yellow contrast. Therefore, color features are divided into

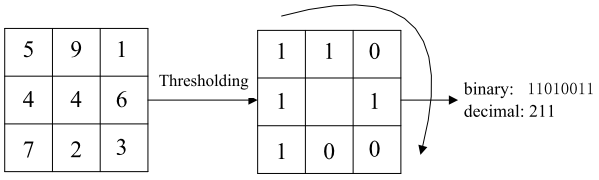

Fig. 2. The LBP operator.

red/green contrast and blue/yellow contrast two parties. Intensity feature includes intensity on (light-on-dark) and intensity off (dark-on-light). We convert the color object image into gray-scale image to obtain an intensity image and let center/surround contrast be intensity on, surround/center contrast be intensity off. The reason is that the ganglion cells in the visual receptive fields of the human visual system are divided into two types: on-center cells respond excitatory to light at the center and inhibitory to light at the surround, whereas off-center cells respond inhibitory to light at the center and excitatory to light at the surround [21]. There are four orientations in our model: 0°, 45°, 90° and 135°. The orientations are computed by Gabor filters detecting bar-like features according to a specified orientation. Gabor filters, which are the product of a symmetric Gaussian with an oriented sinusoid, simulate the receptive field structure of orientation-selective neurons in primary visual cortex [21]. A Gabor filter centered at the 2-D frequency coordinates (U, V) has the general form of

$$h(x, y) = g\big(x', y'\big)\exp\big(2\pi i\big(Ux' + Vy'\big)\big), \tag{1}$$

where

$$\big(x', y'\big) = \big(x\cos(\phi) + y\sin(\phi), -x\sin(\phi) + y\cos(\phi)\big), \tag{2}$$

$$g(x, y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right). \tag{3}$$

σ_x and σ_y are the scale parameter. In this paper, let $\sigma_x = 3.8274$, $\sigma_y = 5.8279$ and ϕ equal to 0°, 45°, 90° and 135°, respectively. For texture feature, we consider Local Binary Pattern (LBP) [22], which describes the local spatial structure of an image and has been widely used in explaining human perception of textures. Ojala et al. [23] first introduced this operator and showed its high discriminative power for texture classification. At a given pixel position (x_c, y_c), LBP is defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its eight surrounding pixels (Fig. 2). The decimal form of the resulting 8-bit word (LBP code) can be expressed as follows:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{4}$$

where i_c corresponds to the gray value of the center pixel (x_c, y_c), i_n to the gray values of the 8 surrounding pixels, and function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1 & x \geqslant 0, \\ 0 & x < 0. \end{cases} \tag{5}$$

Two LBP operators are used in this paper, one is original LBP operator and the other is extended LBP operator with a circular neighborhood of different radius size. The extended LBP operator can keep size and rotation invariance and its pixel values are interpolated for points which are not in the center of a pixel. The two LBP operators are illustrated in Fig. 3. Therefore, ten features are considered in this paper.

2.2. Dynamic saliency map

For every frame in the video, ten feature maps are extracted above. For i-th feature map S_i, we create a Gaussian pyramid of $S_{i,s}$, where $s \in \{1, 2, 3, 4\}$. In this way, each feature map has four different sizes, which equal to one second, one fourth, one eighth and one sixteenth respectively of the size of the feature map S_i. In order to reduce the time of computation, we work with 256 gray level feature maps for each size and transform them into a lower number of gray levels. Generally, good results are usually obtained with eight levels in normal illumination indoor and outdoor scenes. A higher value rarely gives better results, whilst lower values (say, two or four) may be used for night vision [24].

In this paper, we transform every feature map into eight gray levels. Let the maximal saliency value of all feature maps is M. There are n frames in the short video. For each coordinate (x, y) at successive k ($k \in \{1, 2, \ldots, n\}$) frames for i-th feature and the s-th scale, we normal this saliency maps $S_{i,s}(x, y, k)$ divided by M to a fixed range $[0, 1]$. After dividing the $[0, 1]$ range into some eight equal parts, we let the values in different parts be different integers, whose range is $[0, 7]$. Those integers are defined by Eq. (7). Fig. 4 shows an example of transforming a feature map (a) into eight gray level bands (b) and four gray level bands (c).

$$f_{i,s}(x, y, k) = S_{i,s}(x, y, k)/M, \tag{6}$$

$$g_{i,s}(x, y, k) = \begin{cases} 0 & 0 \leqslant f_{i,s}(x, y, k) \leqslant 1/8, \\ 1 & 1/8 < f_{i,s}(x, y, k) \leqslant 1/4, \\ \ldots & \ldots \\ 7 & 7/8 < f_{i,s}(x, y, k) \leqslant 1. \end{cases} \tag{7}$$

For the feature maps of the i-th feature and s-th scale for all frames, we calculate the local complexity [25]. Let $L(x, y)$ be a local round region whose center is (x, y) and radius is s. The local round region histogram $h_{i,s}(.)$ (i denotes the i-th feature and$s$denotes scale) is calculated by Eq. (8).

$$h_{i,s}(x, y, l) = \sum_{(x', y') \in L(x,y)} \sum_{k \in \{1,2,\ldots,n\}} \delta\big(l - g_{i,s}\big(x', y', k\big)\big), \tag{8}$$

where

$$\big(x', y'\big) \in L(x, y) \tag{9}$$

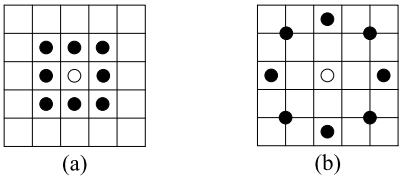

Fig. 3. (a) The original LBP operator; (b) The extended LBP operator.

$l \in \{0, 1, \ldots, 7\}$ and $\delta(.)$ is unit impulse function:

$$\delta(x) = \begin{cases} 1 & x = 0, \\ 0 & x \neq 0. \end{cases} \tag{10}$$

In order to avoid calculating the same pixel in gray level, we define the sign function $sign_{i,s}(.)$ as:

$$sign_{i,s}(x, y, l) = \begin{cases} 1 & h_{i,s}(x, y, l) \neq 0, \\ 0 & h_{i,s}(x, y, l) = 0. \end{cases} \tag{11}$$

The local complexity map of the i-th feature and s-th scale in coordinate (x, y) is:

$$C_{i,s}(x, y) = \sum_{l=0}^{7} sign_{i,s}(x, y, l). \tag{12}$$

Let $N(.)$ be normalization operator and $\oplus$ be point-by-point addition. The feature response map is formed by combining response map of different spatial scales and the same feature:

$$C_i(x, y) = \bigoplus_{s=1}^{4} N\big(C_{i,s}(x, y)\big). \tag{13}$$

We use spatial competition function f to combine all features to form a dynamic saliency map. For details regarding implementation of this feature combination strategy, please see Section 2.4 in [26]

$$M_d(x, y) = f\left(\sum_i N\big(C_i(x, y)\big)\right). \tag{14}$$

3. Attentional selection

In this part, dynamic and static saliency maps are fused into a final saliency map and the sizes of each salient region are obtained by maximizing entropy of the final saliency map.

3.1. Static saliency map and final saliency map

A static saliency map indicates how conspicuous every spatial location based merely on image itself. The static saliency map part in our proposed model is an extension of the model proposed by Itti et al. [9] since our model considers texture feature.

In order to be consistent with dynamic salience map, we still use color contrast, intensity contrast, orientation, and texture features. Ten feature maps are implemented, sensitive to color contrast (red/green and blue/yellow), intensity contrast (light-on-dark and dark-on-light), orientation (0°, 45°, 90°, 135°) and two texture features, as previously described. The extent to which the low-level features used here attract attention in humans has been previously investigated in details [27]. Center and surround scales are obtained using Gaussian pyramids with nine scales (from scale 0,

Fig. 4. (a) A feature map; (b) Eight gray level bands; (c) Four gray level bands.

the original image, to scale 8, the image reduced by a factor 256). Center-surround differences are then computed as point-wise differences across pyramid scales, for combinations of three center scales ($c \in \{2, 3, 4\}$) and two center-surround scale differences ($\delta \in \{3, 4\}$); thus, six maps are computed for each of the ten features, yielding a total of sixty feature maps. Each feature map is endowed with internal dynamics that operate a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition and linear combinations. The feature maps are fused step by step, thereby strengthening important aspects and ignoring others. Resultingly, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings and form conspicuity maps. All conspicuity maps are then linearly combined into a static saliency map M_s.

The dynamic saliency map and static saliency map are described above. The final saliency map is their weighted sum. Both maps compete for salience: the dynamic saliency map emphasizing its temporal salience; the static salient map showing regions that are salient because of its spatial conspicuities. To make the maps comparable, M_d is normalized in advance to the same range as M_s. When fusing the maps, it is possible to determine the degree to which each map contributes to the sum. This is done by weighting the maps with a dynamic saliency map factor t $(0 \leqslant t \leqslant 1)$

$$M = t \times N(M_d) + (1 - t) \times N(M_s). \tag{15}$$

After the computation of the final saliency map, the most salient region is determined and the focus of attention is directed there. Thus, we obtain salient regions, and then, we obtain salient sizes by maximizing entropy.

3.2. The size of salient region

In order to acquire the size of salient target object, we maximize the entropy of salient region. The entropy maximum is considered to analyze the sizes of salient regions [28]. The most appropriate scale x_s for each salient region centered at location x in the final saliency map is obtained by Eq. (16) which aims to consider spatial dynamics at this location:

$$x_s = \arg\max_s \left\{ H_D(s, x) \times W_D(s, x) \right\}, \tag{16}$$

where D is the set of all descriptor values which consist of the intensity values corresponding the histogram distribution in a local region with size s around an attended location x in final saliency map, $H_D(s, x)$ is the entropy defined by Eq. (17) and $W_D(s, x)$ is the inter-scale measure defined by Eq. (18).

$$H_D(s, x) = -\sum_{d \in D} p_{d,s,x} \log_2 p_{d,s,x}, \tag{17}$$

$$W_D(s, x) = \frac{s^2}{2s - 1} \sum_{d \in D} |p_{d,s,x} - p_{d,s-1,x}| \tag{18}$$

where $p_{d,s,x}$ is the probability mass function, which is obtained by normalizing the histogram generated using all the pixel values in a local region with a scale s at position x in the final saliency map, and the descriptor value d is an element in a set of all descriptor values D, which is the same set of all the pixel values in a local region.

4. Experimental results

We apply four schemes of dynamic visual attention model such as Marat's approach [17], Shi's approach [18], Ban's approach [19]

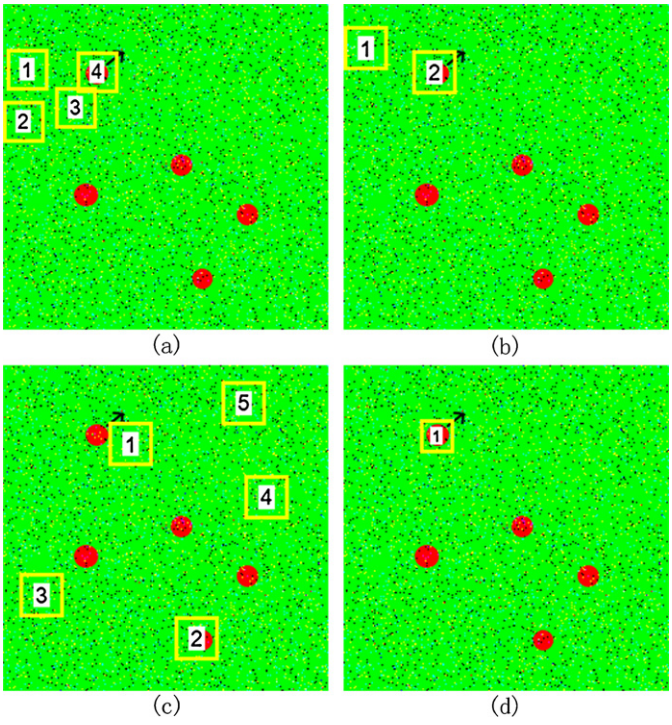

Fig. 5. The experimental results of single object synthetic scene, (a) Marat's model, the object is found in the 4th time; (b) Shi's model, the object is found in the 2nd time; (c) Ban's model, the object is not found within the first 5 times; (d) Our proposed model, the object is found in the 1st time.

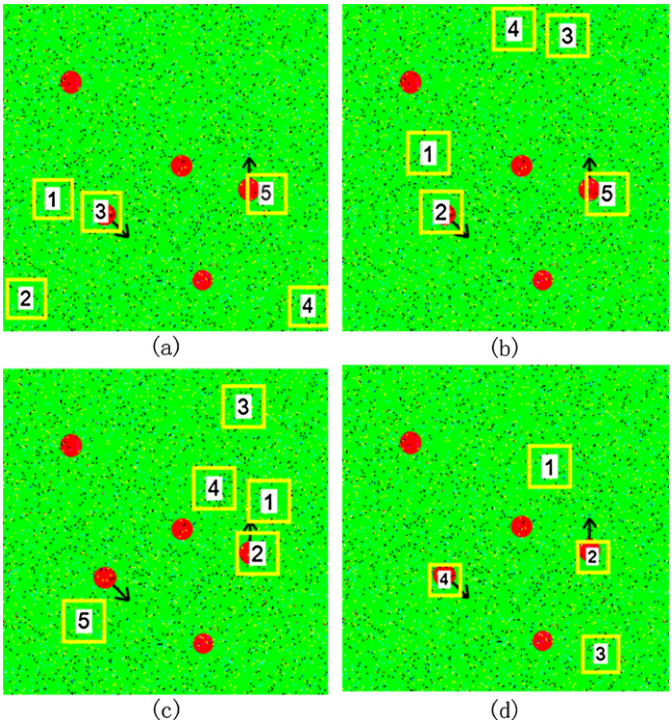

Fig. 6. The experimental results of multi-object synthetic scene, (a) Marat's model, two objects are found in the 3rd time and 5th time, respectively; (b) Shi's model, two objects are found in the 2nd time and 5th time, respectively; (c) Ban's model, one object is found in the 2nd time, but the other object is not found within the first 5 times; (d) Our proposed model, two objects are found in the 2nd time and 4th time, respectively.

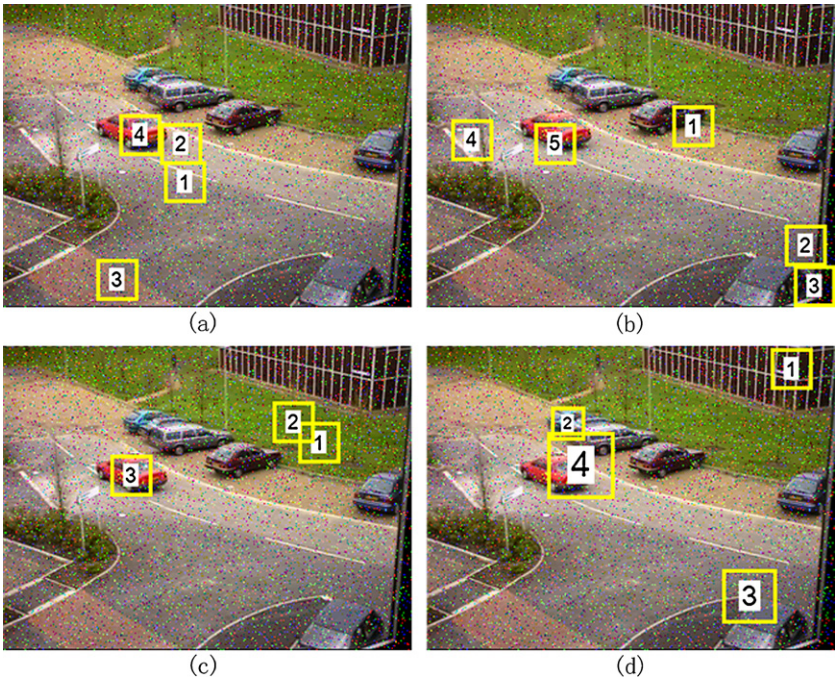

Fig. 7. The experimental results of single object nature scene, (a) Marat's model, the red car is found in the 4th time; (b) Shi's model, the red car is found in the 5th time; (c) Ban's model, the red car is found in the 3rd time; (d) Our proposed model, the red car is found in the 4th time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

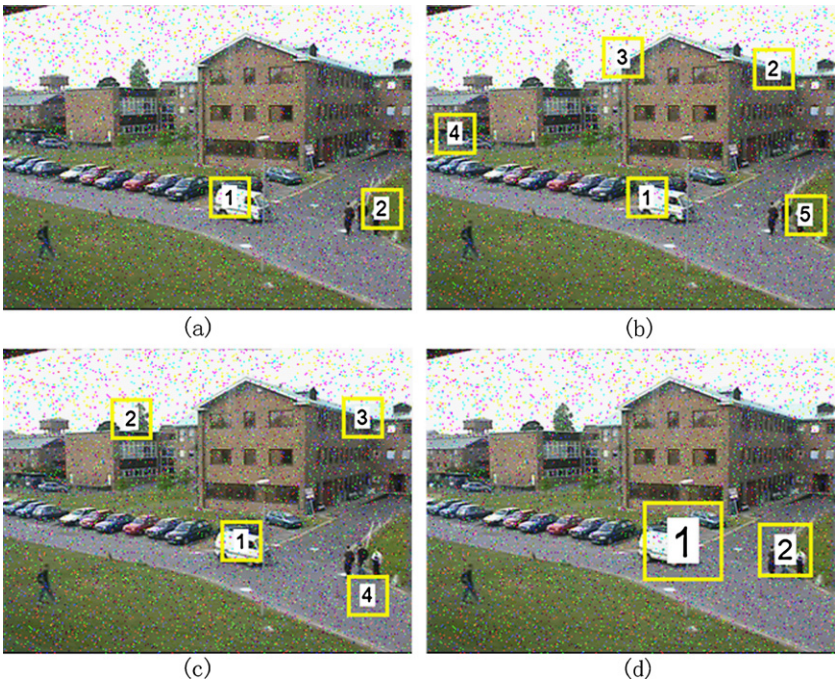

Fig. 8. The experimental results of multi-object nature scene. There are three motion objects in the scene, but the left pedestrian is not found within the first 5 times in all those four models, (a) Marat's model, two objects are found in the 1st time and 2nd time, respectively; (b) Shi's model, two objects are found in the 1st time and 5th time, respectively; (c) Ban's model, two objects are found in the 1st time and 4th time, respectively; (d) Our proposed model, two objects are found in the 1st time and 2nd time, respectively.

and our proposed approach on same short video sequences. Each video sequence includes one or more motion objects. We have done 50 group experiments and each group experiment is done by these four approaches. Those 50 experiment scenes are divided into 30 nature scenes and 20 synthetic scenes. The 30 nature scenes include 15 single object scenes and 15 multi-object scenes; the 20 synthetic scenes also include 10 single object scenes and 10 multi-object scenes. Each single object scene just contains one motion object, while each multi-object scene contains two or more motion objects. According to the different motion object, the experimental scenes are divided into single object synthetic scenes, multi-object synthetic scenes, single object nature scenes and multi-object nature scenes four part. We compare the four dynamic visual attention approaches in each part, respectively.

Fig. 5 provides the experimental results of single object synthetic scene. There are five red dots in the green background, just

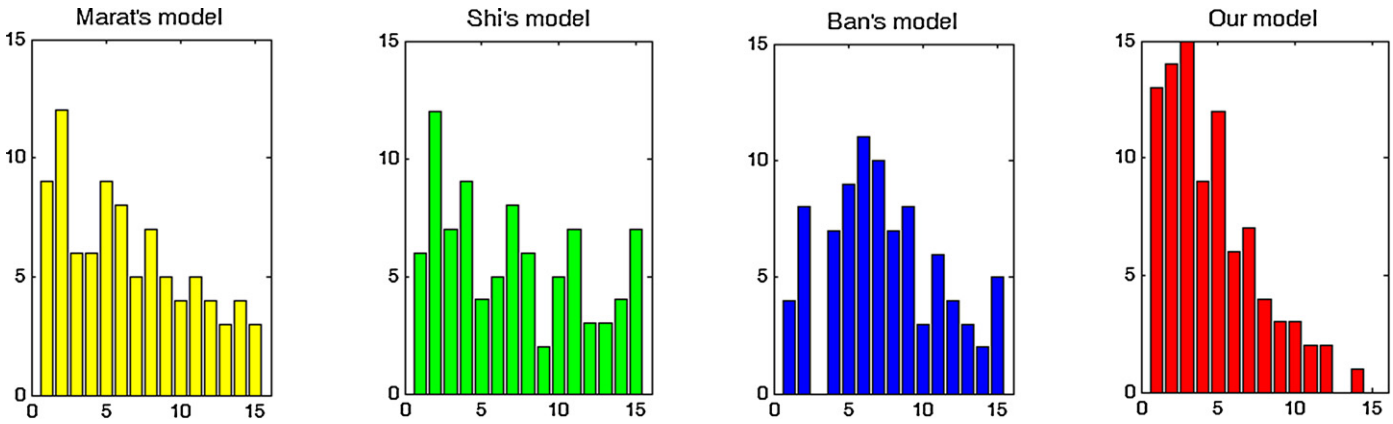

Fig. 9. All object scenes experimental results of four models (x-axis expresses the how many times the target object is found; y-axis expresses the total number of emerged in this time).

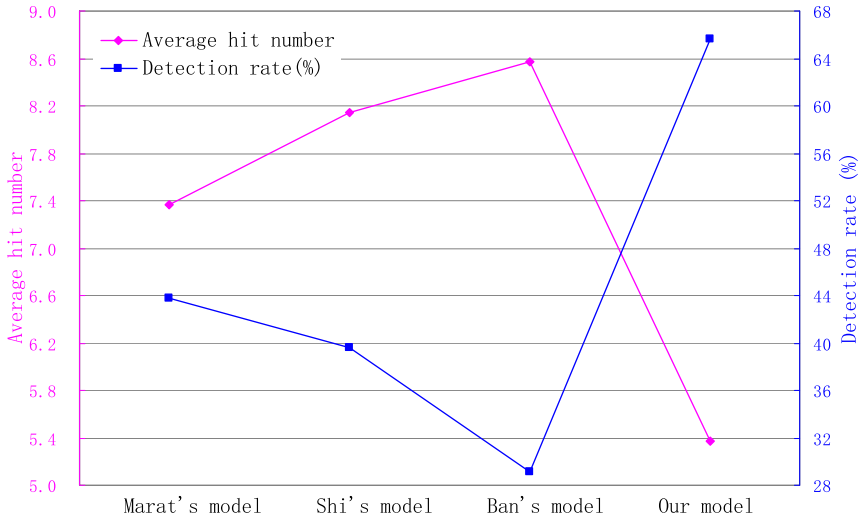

Fig. 10. The comparison of average hit number and detection rate for four different models in all scenes.

a dot is moving in 5 successive frames and this dot is marked by an arrow. Fig. 6 provides the experimental results of multi-object synthetic scene. The experimental results of single object nature scene and multi-object nature scene are shown by Fig. 7 and Fig. 8, respectively. In experiment, we take $t = 0.9$ which expresses that the dynamic map is more important than static map for the final saliency map. In those four models, our model included scale information in salient region, which was represented by a scaled box on the corresponding salient region; other models did not include any scale information and the boxes just expressed the location of salient region.

In order to evaluate the effect of each model, we introduce three definitions: hit number, average hit number [29] and detection rate. The hit number on an image for one target is the rank of the focus that hits the target in order of saliency. For example, if the 2nd focus is on the target, the hit number is 2. The lower the hit number, the better the search performance. If the hit number is 1, the target is immediately detected. The average hit number for an image set is the arithmetic mean of the hit numbers of all images. For example, if an image set just has three images and the hit number is 2, 3 and 5, respectively, then, the average hit number is 3.33. If target object is found within first 5 attention foci [30], then visual attention is regarded as a success. We define detection rate as the percentage of the number of scenes that targets detected within the first 5 attention foci to all the number of scenes.

Table 1
All object scenes average hit number and detection rate of four models.

	Marat's model	Shi's model	Ban's model	Our model
Av. hit number	7.3646	8.1458	8.5729	5.3750
Detection rate (%)	43.7500	39.5833	29.1667	65.6250

All the scenes experimental results of four models are shown in Fig. 9, corresponding average hit number and detection rate of those models are calculated in Table 1. The comparison of average hit number and detection rate for four different models in all scenes is expressed in Fig. 10.

According to above experimental results, synthetic scenes experimental results are better than nature scenes experimental results. The reason is that the backgrounds of synthetic scenes are simple, but the backgrounds of nature scenes are easy to be influenced by others, such as noise, illumination and clutter. Multi-object scenes experimental results are better than single object scenes experimental results. There are more dynamic regions in multi-object scenes, so there are more chances to be found within first several attention foci. When there is noise among the frames or there is change of illumination among the frames, our model is excellent to Marat's model and Shi's model. When the moving objects do not belong to the static salient region, our model is better than Ban's model. The reason is that our model uses the dynamic

information between feature maps while Ban's model uses the dynamic information between static saliency maps.

5. Conclusion

This paper presents a dynamic saliency attention model based on local complexity. The main process is described as following. Firstly, low-level visual features are extracted from current and some previous frames in a short video; every feature map is resized into some different sizes. Secondly, the feature maps in same size and same feature for all the frames are used to calculated the local complexity map. All the local complexity maps are normalized and are fused into a dynamic saliency map. Thirdly, same features are extracted and form multi-scale feature maps by center-surround differences in current frame, and then through across-scale combinations, those feature maps are transformed into conspicuity maps, which are linearly combined into a static saliency map; our proposed model decides salient regions based on a final saliency map which is generated by integration of the dynamic and the static saliency maps. At last, the sizes of each salient region are obtained by maximizing entropy of the final saliency map. Experimental results indicate that: when there is noise among the frames or there is change of illumination among the frames, our model is excellent to Marat's model and Shi's model; when the moving objects do not belong to the static salient regions, our model is better than Ban's model.

The key contribution of this paper is the local complexity method is used to calculate dynamic salience for all feature maps in a short video. Our proposed model for a dynamic scene can play an important role as an initial vision process for a more human-like robot system. This model presents a new approach for predicting the position of human gaze. There are lots of applications for dynamic visual attention model. For example, It can be used to direct foveated image and video compression [31–33] and levels of detail in non-photorealistic rendering [34]. It can also be used in advertising design, adaptive image display on small devices, or seam carving [35]. In our future works, we will extend our dynamic visual attention model to work in top-down environment by adding some prior knowledge [36].

Acknowledgments

This work was supported by the Chinese National 863 Grant No. 2009AA12Z10 and National Natural Science Foundation of China under contracts 60736010.

References

[1] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vision Res. 40 (2000) 1489–1506.
[2] L. Itti, C. Gold, C. Koch, Visual attention and target detection in cluttered natural scenes, Opt. Eng. 40 (9) (2001) 1784–1793.
[3] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vision Res. 45 (2005) 205–231.
[4] N.D. Bruce, J.K. Tsotsos, Saliency based on information maximization, presented at the Neural Information Processing Systems, 2005.
[5] O.L. Meur, P.L. Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 802–817.
[6] X. Hou, L. Zhang, Saliency detection: A spectral residual approach, in: Computer Vision and Pattern Recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 10, 2007, pp. 1–8.
[7] L. Wei, N. Sang, Y. Wang, A biologically inspired object-based visual attention model, Artificial Intell. Rev. 34 (2) (2010) 109–119.
[8] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognit. Psychol. 12 (1980) 97–136.
[9] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[10] R. Rosenholtz, A simple saliency model predicts a number of motion popout phenomena, Vision Res. 39 (19) (1999) 3157–3163.
[11] T. Veit, F. Cao, P. Bouthemy, Probabilistic parameter-free motion detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2004, pp. 715–721.
[12] T.M. Gersch, E. Kowler, B. Dosher, Dynamic allocation of visual attention during the execution of sequences of saccades, Vision Res. 44 (2004) 1469–1483.
[13] M.T. López, A. Fernández-Caballero, M.A. Fernández, J. Mira, A.E. Delgado, Motion features to enhance scene segmentation in active visual attention, Pattern Recognit. Lett. 27 (5) (2006) 469–478.
[14] O.L. Meur, P.L. Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vision Res. 47 (2006) 2483–2498.
[15] M.T. López, M.A. Fernández, A. Fernández-Caballero, J. Mira, A.E. Delgado, Dynamic visual attention model in image sequences, Image Vision Comput. 25 (2007) 597–613.
[16] A. Bur, Computer models of dynamic visual attention, Ph.D. thesis, Université de Neuchâtel, Switzerland, 2009.
[17] S. Marat, T. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin, Modelling spatio temporal saliency to predict gaze direction for short videos, Int. J. Comput. Vision 82 (2009) 231–243.
[18] H. Shi, Y. Yang, A computational model of visual attention based on saliency maps, Appl. Math. Comput. 188 (2007) 1671–1677.
[19] S.W. Ban, I. Lee, M. Lee, Dynamic visual selective attention model, Neurocomputing 71 (2008) 853–856.
[20] W.H. Beaudot, The neural information in the vertebra retina: a melting pot of ideas for artificial vision, Ph.D. thesis, Tirf Laboratory, Grenoble, France, 1994.
[21] S.E. Palmer, Vision Science, Photons to Phenomenology, MIT Press, Cambridge, MA, 1999.
[22] T. Mäenpää, M. Pietikäinen, Texture analysis with local binary patterns, in: C. Chen, P. Wang (Eds.), Handbook of Pattern Recognition and Computer Vision, 3rd edition, World Scientific, 2005, pp. 197–216.
[23] T. Ojala, M. Pietikäainen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognit. 29 (1) (1996) 51–59.
[24] M.T. López, A. Fernández-Caballero, M.A. Fernández, J. Mira, A.E. Delgado, Visual surveillance by dynamic visual attention method, Pattern Recognit. 39 (2006) 2194–2211.
[25] C. Yan, N. Sang, T. Zhang, K. Zeng, Image transition region extraction and segmentation based on local complexity, J. Infrared Millimeter Waves 24 (4) (2005) 312–316 (in Chinese).
[26] L. Itti, C. Koch, Feature combination strategies for saliency-based visual attention systems, J. Electron. Imaging 10 (1) (2001) 161–169.
[27] J.M. Wolfe, Visual memory: What do you know about what you saw? Curr. Biol. 8 (9) (1998) R303–R304.
[28] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2) (2001) 83–105.
[29] S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, Lecture Notes in Artificial Intelligence (LNAI), Springer, Berlin/Heidelberg, 2006.
[30] G.G. Ryu, I.H. Suh, S. Lee, Covert visual attention by object-based selective visual features and their saliency map, in: International Conference on Image Processing, Computer Vision, and Pattern Recognition, 2009, p. 6689.
[31] W.S. Geisler, J.S. Perry, A real-time foveated multiresolution system for low-bandwidth video communication, Proc. SPIE 3299 (1998) 294–305.
[32] Z. Wang, L. Lu, A.C. Bovik, Foveation scalable video coding with automatic fixation selection, IEEE Trans. Image Process. 12 (2003) 243–254.
[33] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (1) (2010) 185–197.
[34] D. DeCarlo, A. Santella, Stylization and abstraction of photographs, ACM Trans. Graphics 21 (3) (2002) 769–776.
[35] M. Rubinstein, A. Shamir, S. Avidan, Improved seam carving for video retargeting, ACM Trans. Graphics (SIGGRAPH) 27 (3) (2008).
[36] V. Navalpakkam, L. Itti, An integrated model of top-down and bottom-up attention for optimal object detection speed, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 2049–2056.

Longsheng Wei received the BE in information and computing science from Anhui University in 2005, Hefei, China. He received MS degree in theory of probability and statistics from Huazhong University of Science and Technology in 2007, Wuhan, China. He received PhD degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 2011, Wuhan, China. He is currently an Assistant Professor in the Faculty of Mechanical and Electronic Information, China University of Geosciences, Wuhan, China. His research interests involve visual attention, image processing, and artificial intelligence.

Nong Sang graduated from Huazhong University of Science and Technology and received his BE degree in computer science and engineering in 1990, MS degree in pattern recognition and intelligent control in 1993, and PhD degree in pattern recognition and intelligent control in 2000. He is currently a professor at the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China. His research interests include pattern recognition, computer vision, and neural networks.

Yuehuan Wang received the BE in electronic precision machinery from University of Electronic Science and Technology of China in 1993, Chendu, China. He received MS degree in computer architecture from Huazhong University of Science and Technology in 1996, Wuhan, China. He received PhD degree in pattern recognition and intelligent control from Huazhong University of Science and Technology in 2001, Wuhan, China. He is currently a professor at the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China. His research interests include biological vision principle of image analysis method, real-time automatic target recognition methods and systems, and computer vision.

Qingqing Zheng received her BE and MS degree in communication engineering in 2001 and 2004 respectively from Xi'an Jiaotong University. She is an instructor of Wuhan University of Science and Technology now. She is currently pursuing her PhD in the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China. Her research interests involve image processing and pattern recognition.