Optik 125 (2014) 6422–6427

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo

A biologically inspired spatiotemporal saliency attention model based on entropy value

Longsheng Wei*, Dapeng Luo

Faculty of Mechanical and Electronic Information, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 14 November 2013
Accepted 16 June 2014

Keywords:
Spatiotemporal saliency model
Attentional selection
Entropy value
Saliency map

ABSTRACT

A biologically inspired spatiotemporal saliency attention model based on entropy value is proposed 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. 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 the entropy value map. All the entropy maps are normalized and 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 model decides salient regions based on a spatiotemporal saliency map which is generated by integration of the dynamic and the static 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 Shi's model and Marat'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 [1–3] of visual attention are static and are inspired by the concept of feature integration theory [4]. The most popular is the one proposed by Itti et al. [5] and it has become a standard model of static visual attention, in which saliencies according to primitive features such as intensity, orientation and color are computed independently. There are also many models [6–8] bringing dynamic saliency to visual attention mechanism. Shi and Yang [9] proposed a model for motion detection in 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. Marat et al. [10] used an optical flow method to compute the dynamic saliency. The optical flow method does not require any prior knowledge of the scene; it can detect dynamic objects and 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. Ban et al. [11] also propose a dynamic visual attention model. Firstly, a static saliency map is obtained by a frame in a video. Secondly, an optimal scale is calculated for each pixel location and for each static saliency map. Thirdly, those optimal scales and static saliency maps are used to calculate the entropy to form entropy maps for every frame. At last, all the entropy maps are used to calculate a new entropy map, which is called dynamic saliency map. However, when the moving objects do not belong to the salient regions, Ban's model is very hard to attend to the moving regions.

In order to address the above problem, we propose a biologically inspired spatiotemporal saliency attention model based on entropy value 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 the local entropy map by the probability mass function in corresponding local region. All the

* Corresponding author.
E-mail address: weilongsheng@163.com (L. Wei).

http://dx.doi.org/10.1016/j.ijleo.2014.06.148
0030-4026/

local entropy maps are normalized and 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 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 spatiotemporal saliency map which is generated by integration of the dynamic and the static saliency map. At last, the size of each salient region is obtained by maximizing entropy of the spatiotemporal saliency map.

This paper is organized as follows. Section 2 presents the dynamic saliency model including feature extraction and dynamic saliency map. While attentional selection is described in Section 3, this part introduces how to acquire static saliency map, spatiotemporal saliency map and the size of salient region. Section 4 shows experimental results, and Section 5 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 correspond to the two main outputs of the retina [12]. 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 [10]. 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 the local entropy map by the probability mass function in corresponding local region. All the local entropy 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 [5]. Therefore, color features are divided into 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 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 [13].

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,

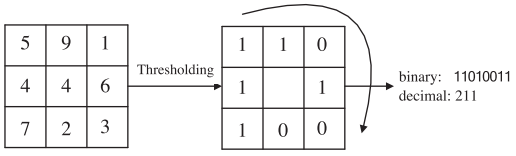

Fig. 1. The LBP operator.

simulate the receptive field structure of orientation-selective neurons in primary visual cortex [13]. A Gabor filter has the general form of:

$$h(x,y) = g(x',y')\ \cos(2\pi\omega_f x') \tag{1}$$

where

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

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

ω_f is the center frequency, which determines the scale of Gabor filter. σ_x and σ_y are the standard deviation of the Gaussian factor of the Gabor function in x and y directions, respectively. σ_x and σ_y have relation to frequency bandwidth B_f and orientation bandwidth B_θ:

$$\sigma_x = \sqrt{\frac{\ln 2}{2}}\frac{1}{\pi\omega_f}\frac{2^{B_f}+1}{2^{B_f}-1} \tag{4}$$

$$\sigma_y = \sqrt{\frac{\ln 2}{2}}\frac{1}{\pi\omega_f}\frac{1}{\tan(B_\theta/2)} \tag{5}$$

ϕ is the orientation of Gabor filter. In this paper, let $\omega_f = 0.12$, $B_f = 1.25$, $B_\theta = \pi/6$ and ϕ equal to 0°, 45°, 90° and 135°, respectively.

For texture feature, we consider local binary pattern (LBP) [14], which describes the local spatial structure of an image and has been widely used in explaining human perception of textures. 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. 1). The decimal form of the resulting 8-bit word (LBP code) can be expressed as follows:

$$\mathrm{LBP}(x_c,y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{6}$$

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:

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{7}$$

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. 2. 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 ith feature map F_i, we create a Gaussian pyramid of $F_{i,s}$, where $s \in \{1, 2, 3, 4\}$, according to the size of F_i. 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. In order to reduce the time of computation, we