

摘 要

人类视觉系统以大约每秒8.96Mbits的速度接收外部视觉信息，同时我们发现人类视觉系统的计算能力和存储能力虽然都有限，但却能有效的从这些信息中检测、过滤出重要的内容，譬如突然出现的危险物等，并对此做出及时的反应，这种高效的选择注意机制让当代具有强大计算和存储能力的计算机也望尘莫及，所以从生理物理学上揭示选择注意机制的内在机理，以及用计算机去模拟选择机制显得尤其重要。本文的目标是在现有的选择性注意机制的心理生理学实验依据的基础上设计视觉注意计算模型，使其能在现实场景中模拟人眼视觉系统的静态和动态特性。

首先，为了便于分析眼动过程的局部抖动行为和全局随机行为，本文利用高速眼动仪采集了24个被试在一个彩色自然场景图像集上的眼动数据，一方面应用在本文后续的实验中，另一方面发布在网络上供其他研究者使用(已有研究者索取)。同时本文还提出用延时嵌入(time-delay embedding)方法对动态视觉注意模型生成的扫视路径进行评估，相对于传统方法所采用的编辑距离，提出的方法需要调节的参数少，而且调节简单直观。

其次，鉴于早期的视觉注意计算模型不能精确的预测场景中的显著区域，本文提出了一个静态视觉注意模型：点熵率(Site Entropy Rate)。该模型以相关的心理生理学事实为依据，譬如视觉初级皮层中简单细胞的稀疏编码特性，神经网络中存在着不同空间尺度的连接，以及神经元的响应受到周围神经元的调制。提出用稀疏编码滤波函数去提取早期视觉特征，全连接图去模拟神经元之间的连接，并以人眼视觉系统最大化获取信息的特性为基本原理，提出用点熵率去度量显著性。此模型分别在常用的心理学刺激、彩色图像集、灰度图像集和视频图像集上进行了测试，并与当前的几种视觉注意模型进行了对比实验，结果显示提出的模型预测精度是最高的。

最后，视觉注意是人眼扫视行为的内部驱动机制，而扫视行为是一个动态过程，在这个动态过程中，不仅能指出哪些位置比较显著，还包含了这些显著位置的访问顺序，这两者的结合就是人眼扫视路径，大量的心理学实验表明，直接对静态显著度图的值进行排序不能解释真实的人眼扫视路径。基于信息最大化原理，本文提出了一个动态的视觉注意计算模型，它模拟人眼在观看自然

场景图像时的扫视路径。此模型集成了与视觉注意相关的三个重要要素：参考感知响应、中央凹-外周分辨率差别、视觉工作记忆，并为此三要素建立了统一的表示：多频带滤波响应图。模拟这三个要素之间的相互作用可以计算出一个多频带残差滤波响应图，然后用前面提出的点熵率去度量多频带残差滤波响应图上的残余感知信息(residual perceptual information)，残余感知信息随着眼动过程动态的变化，选择最大的残余感知信息处即为下一个关注点。通过与该领域两个重要工作的实验比较，证明提出的模型无论是在静态关注点还是动态扫视路径的预测都取得了最好的结果。

综上所述，本论文结合心理、生理、神经科学在视觉注意机制已有研究成果的基础上，从信息论的角度对视觉注意的静态和动态特性分别进行建模。在静态特性方面，提出了基于点熵率度量的视觉显著性模型；在动态特性方面，提出了一个模型去模拟人眼在自然图像上的扫视路径。然而我们知道，扫视行为不仅包括关注点的空间分布、关注点的顺序，还包括关注点的停留时间，扫视行为的个体差异性，影响第一个关注点的因素，上下文信息在扫视行为中的影响等，这些都是我们未来要研究的内容。

关键词： 视觉注意，扫视路径，信息最大化原理，稀疏编码，点熵率，全连接图，参考感知响应，中央凹成像，视觉工作记忆

Abstract

Human visual system is exposed to the outside world and receives a large amount of visual information at the speed of 8.96 Mbits per second. Although it has limited computing and storage capacity, the most important contents, such as the emergency of dangerous animals, can be efficiently captured and filtered from the mass information, and immediate responses can be performed. It largely outperforms modern computer with powerful computing and storage capacity in the aspects of low energy consumption and high efficiency. So it is very necessary to uncover the inner selective mechanism from the view of psychophysics and simulate this kind of mechanism by electronic computer. The aims of this thesis are to design computational models of selective visual attention based on known psychophysics evidences. These models can accurately simulate the static and dynamic properties of selective visual attention on natural scenes.

First, in order to analyze the local vibrating behavior and global stochastic behavior of eye movements, we use a high-speed eye-tracker to collect eye movement data from 24 subjects with a set of color natural images. This dataset can be used in our latter experiments and also for other researchers' work (There have been researchers asking for it). We also propose to employ time-delay embedding to evaluate the scanpaths generated by dynamic attention models. This method needs less parameter to be tuned and the parameter is easily tuned compared with classical edit distance.

Second, although many computational models of visual attention have been proposed in recent years, they generally have low prediction accuracy compared to human visual system. In this thesis, we propose a new computational model for visual saliency: Site Entropy Rate. The model is inspired by a few well acknowledged biological facts, such as sparse representation in primary visual cortex, two different spatial scales of cortical neuron connectivity, a neuron's activities are driven by the total synaptic input from its neighbors. First, the model extracts a number of sub-band feature maps to be early visual features using learned sparse

codes. Then, it adopts a fully-connected graph representation for each feature map, and runs random walks on the graphs to simulate the signal/information transmission among the interconnected neurons. Finally, Site Entropy Rate is proposed as a new visual saliency measure based on the principle of information maximization. To evaluate the proposed model, we do extensive experiments on psychological stimuli, two well known image data sets, as well as a public video dataset. The experiments demonstrate that the proposed model achieves the state-of-the-art performance of saliency detection.

Finally, as we know, visual attention is the driven force of human visual saccade which in fact is a dynamic search process. Human saccade contains not only fixation location but fixation order. The combination of these two aspects is the saccadic scanpath. Large amounts of psychology evidences show that it is unsuitable to explain human saccadic behavior with the scanpath by ranking static saliency values. Based on the principle of information maximization, we propose a computational model to simulate human saccadic scanpaths on natural images. The model integrates three related factors as driven forces to guide eye movements sequentially - reference sensory responses, fovea-periphery resolution discrepancy, and visual working memory. For each eye movement, we compute three multi-band filter response maps as a coherent representation for the three factors. The three filter response maps are combined into multi-band residual filter response maps, on which we compute residual perceptual information (RPI) at each location with Site Entropy Rate. The RPI map is a dynamic saliency map varying along with eye movements. The next fixation is selected as the location with the maximal RPI value. On a natural image dataset, we compare the saccadic scanpaths generated by the proposed model and the other two visual saliency-based models against human eye movement data. Experimental results demonstrate that the proposed model achieves the best prediction accuracy on both static fixation locations and dynamic scanpaths.

In summary, we propose two computational models to simulate the static and dynamic properties of visual attention, one for fixation distribution and the other for fixation order. In fact, human saccadic behavior contains several other factors, such as fixation duration, individual differences of human scanpaths, the

influential factors of the first fixation, the influences of contextual information on human scanpath, which are our focus in the future.

Keywords: Visual attention, Visual scanpath, Information maximization, Sparse coding, Site Entropy Rate, Fully-connected graph, Reference sensory response, Foveal imaging, Visual working memory