**基于显著性视觉注意的快速场景分析模型**

Laurent Itti, Christof Koch, and Ernst Niebur

**摘要:**这是一种受到行为和早期灵长类视觉系统的神经元结构启发而提出的视觉注意系统。多尺度图像特征被组合成单个地形显著图。然后，动态神经网络根据显著性降低的顺序选择参加的位置。系统通过以计算有效的方式快速选择要详细分析的显著位置来分解场景理解的复杂问题。

**术语索引**——视觉注意，场景分析，特征提取，目标检测，视觉搜索

## 1 简介

尽管神经元硬件可用于这种任务的速度有限，但是PRIMATES具有很好的实时解读复杂场景的能力。在进一步处理之前，即时和更高的可视化过程[1]倾向于选择可用感知信息的一个子集，最有可能减少场景分析的复杂性[2]。 该选择似乎以视野的空间限定区域的形式实现，即所谓的“注意力集中”，其同时以快速，自下而上，显著驱动和任务独立的方式以及较慢，自上而下，意志控制和任务依赖的方式扫描场景[2]。

注意模型包括“动态路由”模型，其中只有视野的一个小区域的信息可以通过皮层视觉层次被捕获。 通过动态修改皮质连接或通过在自顶向下（依赖任务）和自下而上（依赖于场景的）控制下建立特定的时间活动模式来选择所参与的区域[3]，[2]，[1]。

这里使用的模型（图1）建立在Koch和Ullman [4]提出的第二生物可靠的架构上，并且结合了多个模型[5]，[6]。它与所谓的“特征整合理论”有关，用来解释人类视觉搜索策略[7]。视觉输入首先被分解为一组地形特征图。然后，不同的空间位置竞争每个地图中的显著性，使得只有从周围脱颖而出的位置才能保留。所有特征图以纯粹的自下而上的方式提供给主“显著图”，其在整个视觉场景中地形地识别本地显眼。在灵长类动物中，这样的地图被认为位于后顶叶皮层[8]以及丘脑脉脉核中的各种视觉图中[9]。模型的显著性图具有引起注意力转移的内部动力。因此，该模型代表了自下而上的显著性可以引导注意力的转移，并且不需要任何自上而下的引导来转移注意力。该框架提供了一种大规模并行方法，用于快速选择少量有趣的图像位置，以通过更复杂和时间消耗更长的对象识别过程进行分析。在“引导式搜索”中扩展这种方法，使用较高皮质区域的反馈（例如，关于要发现的目标的知识）来加重不同特征的重要性[10]，使得仅具有高权重的那些可以达到更高的处理水平。

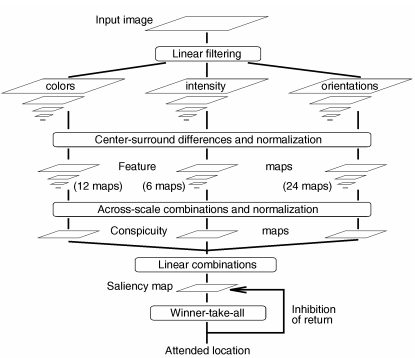


图1 模型框架

## 2 模型

输入以静态彩色图像的形式提供，通常以640 x 480分辨率数字化。 使用二元高斯金字塔[11]创建九个空间尺度，其逐渐低通滤波并对输入图像进行子采样，产生范围从1：1（刻度0）到1：256（刻度8）的水平和垂直图像降低因子， 包含八个尺度。

每个特征由一组类似于视觉感受场的线性“中心环绕”计算（图1）：典型的视觉神经元在视觉空间（中心）的小区域中是最敏感的，而在与中心（环绕）同心的地方更宽，更弱的对抗区域抑制神经元反应。对局部空间不连续性敏感的这种结构特别适合于检测显著性突出的位置，并且这是视网膜，外侧膝状核和初级视觉皮层中的一般计算原理[12]。中心环绕在模型中表现为细微尺度与粗尺度之间的差异：代表中心信息的图像尺度c取c∈{2,3,4}，代表周边背景信息的图像尺度s取，其中。通过将尺度较大的图像和尺度较小的图像进行跨尺度减操作获得两个地图之间的标度差，表示为“”。不仅对c使用多个尺度，而且对于取同样通过包括在中心和环绕区域之间包括不同的尺寸比率（与先前使用的固定比率[5]相反）来产生真正的多尺度特征提取。

**2.1 早期视觉特征提取**

用r，g，b分别表示红色、绿色和蓝色三个通道的值，那么亮度图像I可以表示为I=（r+g+b）/3。I用来构造高斯金字塔，其中尺度。r，g和b通道被I标准化，以便将色调与强度去耦。 然而，由于色调变化在非常低的亮度（并且因此不显著）下不可感知，因此归一化仅在I大于其整个图像的最大值的1/10的位置处施加（其它位置产生零r，g ，和b）。产生了四个宽调谐（broadly-tuned）的颜色通道，红色是R=r-(g+b)/2，绿色是G=g-(r+b)/2，蓝色是B=b-(r+g)/2，黄色是Y=(r+g)/2-|r-g|/2-b（负数设置为0）。通过这些颜色通道构造四个高斯金字塔，，和。

中心细尺度c和周围粗尺度s之间的中心围绕差值（之前定义的）产生特征图。第一组特征图涉及强度对比度，在哺乳动物中，由对明亮的环境中的暗中心或黑暗环境中的明亮的中心敏感的神经元检测到[12]。在这里，两种类型的敏感性是在一组六个I(c,s)图中同时计算的（使用整流），其中c∈{2,3,4}，

. (1)

类似地为颜色通道构建第二组地图，其在皮层中使用所谓的“双色对立”系统来表示：在其接收场的中心，神经元被一种颜色激活（例如， 红色）并禁止另一种颜色（例如，绿色），而在周围区域则是是相反的。在主视觉皮层中，红/绿，绿/红，蓝/黄，黄/蓝颜色对都存在这种空间和色彩竞争关系[13]。 因此，在模型中创建特征图以同时考虑红/绿和绿/红的双竞争以及构建反映蓝/黄和黄/蓝的双竞争关系：

(2)

. (3)

采用方向Gabor金字塔，方向，得到局部方向信息[11]。（Gabor滤波器，即余弦光栅和2D高斯包络的乘积，近似于原始视觉皮层中取向选择性神经元的接受场灵敏度分布（脉冲响应）[12]。）方向特征图作为一组，编码中心和周围区域尺度之间的局部方向对比度：

. (4)

总体来说，计算出来了42个特征图：6个亮度图，12个颜色图和24个方向图。

**2.2 显著性图**

显著性图的目的是通过显著性的空间分布来表示醒目度（conspicuity-saliency）或显著度：在视野中每个位置用一个定量向量表示，引导注意位置的选择。通过一个动态神经网络模型，特征图的组合可以向显著性图提供自下而上的输入。

组合不同特征图的难点在于，对不同的的特征图表示了不可比较的模态的先验信息，有不同的动态范围和提取机制。而且，42个特征图结合时，在一些特征图中表现非常强的显著目标可能被其它更多的特征图的噪声或不显著的目标所掩盖。

在缺少自上而下的监督的情况下，采用一个归一化操作算子，整体提升那些有部分强刺激峰值（醒目位置）的特征图，而整体抑制那些包含大量可比峰值响应。归一化计算方法包括（图2）：

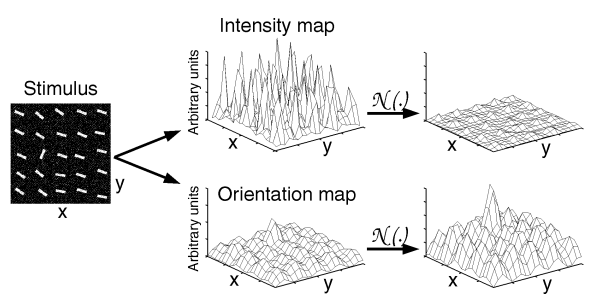


图2 归一化算子

1. 将所有图的取值固定为，消除依赖于模态的幅值差异；
2. 计算图中最大值和其他所有局部极值的平均值；
3. 整幅图像乘以。

我们只考虑局部活动的最大值，使归一计算只比较特征图中有意义的“激活点”相关的响应并忽略响应均匀的区域。将整个地图中的最大活动与平均总体激活度进行比较，测量最活跃位置与平均值的不同。 当这个差异很大时，最活跃的位置脱颖而出，特征图被加强。 当差异很小时，地图不包含任何显著点的并被抑制。设计背后的生物动机是粗略地复制皮层侧向抑制机制，其中相邻的相似特征通过特定的解剖学定义的连接相互抵抗[15]。

通过跨尺度相加，特征图被结合成3个醒目度图（conspicuity maps），表示灰度(5)，表示颜色(6)，表示方向(7)，显著性图的尺度。它们通过跨尺度加操作得到，该操作包括缩减每幅图至尺度四分之一然后逐点相加：

(5)

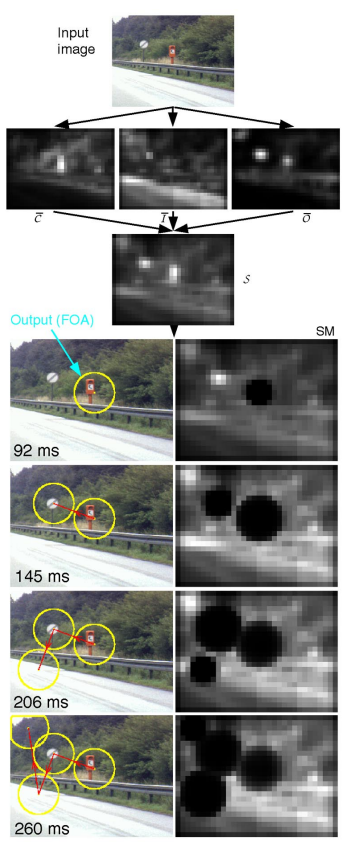
(6)

对于方向，首先通过组合给定的六个特征图来创建四个中间图，然后将其组合成单个定向显著图：

(7)

创建三个独立通道，和的动机，以及它们的个体归一化是一个假设，即相似的特征在显著性方面竞争更加明显，而不同的模态独立于显著性图。 三个显著图被归一化，并且归结为显着图的最终输入S：

(8)

任何时候，显著性图（SM）的最大值定义了最显著的图像位置，为注意焦点（FOA）定向。 我们现在可以简单地选择最活跃的位置来定义下一个模型应该参加的点。然而，在神经元似乎合理的实现中，我们将SM建模为尺度四的泄漏的积分和发射神经元的2D层。 这些模型神经元由单一的电容组成，其集成了突触输入递送的电荷，漏电流和电压阈值。当达到阈值后，产生一个原型尖峰，并且电容电荷分流到0[14]。SM以尺度进入一个生物可行的2D“赢家全胜”（WTA）神经网络[4]，[1]，其中单位之间的突触相互作用确保只有最活跃的位置保留，而所有其他位置被抑制。

**图3基于自然图像的模型的操作示例。并行特征提取产生颜色对比度()，强度对比度()和方向对比度()的三个醒目度图。这些被组合以形成显著图（SM）的输入S。最引人注目的地方是橙色的电话盒，在颜色对比度上表现的很强；它成为第一个被注意的位置（92 ms模拟时间）。在抑制反馈产生并抑制显著图中的这个位置之后，下一个最显著的位置被成功选择。**

SM中的神经元接收来自S的兴奋性输入，并且都是独立的。SM神经元在更突出的位置的潜力因此增加更快（这些神经元被用作纯粹的积分器，不会发射）。每个SM神经元激发其相应的WTA神经元。 所有WTA神经元也相互独立演化，直到一个（“获胜者”）首先达到阈值并触发。这触发三个同时的机制（图3）：

1）FOA转移到胜者神经元的位置;

2）WTA的全局抑制被触发并完全抑制（复位）所有WTA神经元;

3）具有FOA大小和新位置的区域局部抑制在SM中瞬时激活；这不仅使FOA的动态变化，允许下一个最显著的位置随后成为赢家，而且还防止FOA立即返回到先前参加的位置。

人类视觉心理物理学已经证明了这种“抑制回报”[16]。为了稍微偏向模型，随后跳到紧邻当前所在位置的空间位置的突出位置，在FOA（Koch和Ullman的“接近偏好”规则[4]）的附近的SM中短暂激活一个小激励。

由于我们不建立任何自上而下的注意组件，因此FOA是一个简单的磁盘，其半径固定为输入图像宽度或高度较小的一半。选择模拟神经元的时间常数，电导率和触发阈值（详见[17]），以便FOA在大约30-70毫秒（模拟时间）内从一个显著位置跳到下一个位置，已经参与过的区域被抑制约500-900毫秒（图3），正如在心理物理学上所观察到的[16]。 这些延迟的相对幅度的差异证明足以确保对图像进行彻底的扫描，并防止循环仅通过有限数量的位置。 所有参数在我们的实现[17]中是固定的，并且所有图像被研究后证明这个系统是稳定的。

**2.3 与空间频率范围模型比较**

Reinagel和Zador [18]最近使用眼睛跟踪装置来分析由人类产生的眼睛扫描路径的局部空间频率分布，同时自由观察灰度图像。他们发现固定位置处的空间频率含量显著高于随机位置的平均值。虽然眼睛轨迹可能与注意轨迹不同，但在视觉控制下[1]，视觉注意力通常被认为是一种预注意机制，对自由观察产生强烈的影响。因此，有趣的是验证我们的模型是否会重现Reinagel和Zador的发现。

我们构造了空间频率内容（SFC）的简单测量：在给定的图像位置，从每个I(2)，R(2)，G(2)，B(2)和Y(2)图里提取16x16的图像块，并且将2D快速傅里叶变换（FFT）应用于图像块。对于每个图像块，应用阈值来计算不可忽略的FFT系数的数量; 该阈值对应于刚刚感知的光栅的FFT幅度（1％对比度）。 SFC测量是五个相应图块中不可忽略的系数数的平均值。选择图块的大小和规模，使得SFC测量对于与我们的模型大致相同的频率和分辨率范围是敏感的; 此外，我们的SFC测量和模型一样在RGB通道以及灰度强度下计算。使用这一方法，创建尺度为4 的SFC图，并与显著图进行比较（图4）。

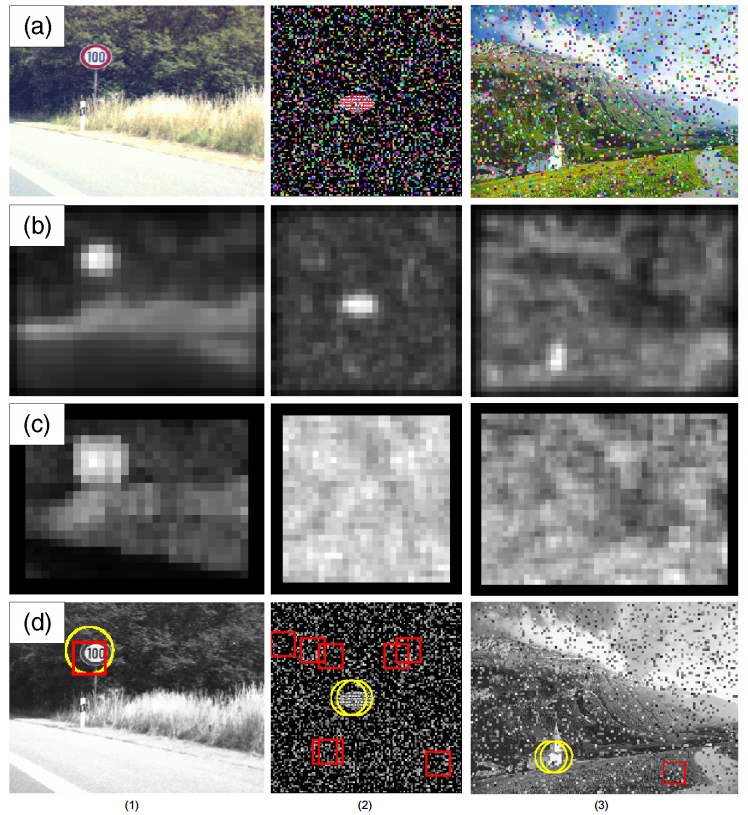


图4 （a）彩色图像的例子。（b）相应的显著图输入。（c）空间频率内容（SFC）图。（d）显著性图的输入高于其最大值（黄色圆圈）的98％以及SFC高于其最大值（红色方块）的98％的图像块的位置。显著性图对噪声非常强大，而SFC则不是。

## 3 结果和讨论

虽然显著图的概念已被广泛应用于FOA模型[1]，[3]，[7]中，关于其构造和动力学的细节却几乎没有。在这里，我们将研究前馈特征提取阶段，映射组合策略以及显著性映射的时间属性如何有助于整个系统性能。

**3.1 一般性能**

该模型经过人工图像广泛测试，以确保正常运行。例如，具有相同形状但与背景对比度不同的几个对象以降对比度的顺序被测试。该模型证明对于这种图像的噪声增加是非常鲁棒的（图5），特别是如果噪声的性质（例如其颜色）不直接与目标的主要特征相冲突。

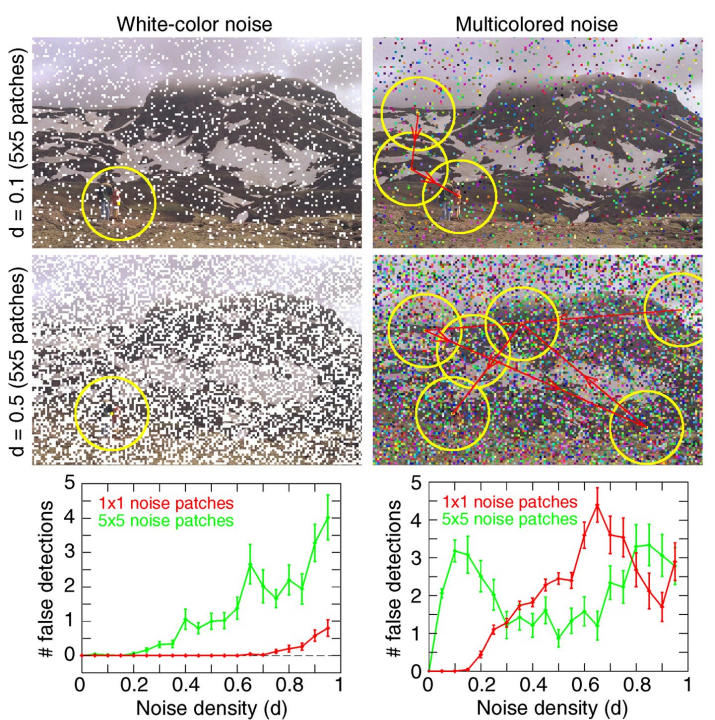


图5 噪音对检测性能的影响，以768x512的场景说明，其中一个目标（两个人）以其独特的色彩对比而显著突出。在发现目标之前的错误检测的平均值±S.E.表现为50个噪声实例的噪声密度的函数。该系统对于不直接干扰目标的主要特征（左；强度噪声和色彩目标）的噪声是非常鲁棒的。当噪声与目标具有相似的特性时，它会损害目标的显著性，系统首先突出显示其他特征物体（这里是强度的粗略变化）。

使用图2所示类型的图像，该模型能够重现许多弹出式任务[7]的人类表现。当目标以它自己的独特方向，颜色，亮度或尺寸和周围干扰物不同时（如图2），不管干扰物的数量如何，目标总是第一个被注意的。 相反，当目标与特征的结合不同时（例如，它是红色垂直和绿色水平条的混合阵列中唯一的红色水平条），寻找目标所需的搜索时间随着干扰物数量的增加而线性增长。这两个结果已经在人类中广泛观察到[7]，并将在3.2节中讨论。

我们还用真实图像测试了模型，从自然户外场景到艺术绘画，并使用来对特征图进行归一化（图3和[17]）。有许多这样的图像，很难客观地评估模型，因为没有客观的参考可用于比较，观察者可能不同意哪些位置是最显著的。然而，在所研究的所有图像中，大多数被注意的地点是感兴趣的对象，如面孔，旗帜，人物，建筑物或车辆。

在与Reinagel和Zador [18]类似的实验中，使用具有显著交通标志（90图像），红色苏打罐（104图像）的自然场景或车辆紧急情况三角形符号（64图像）做测试，将模型预测与局部SFC的测量值进行比较。与Reinagel和Zador的实验结果相似，被注意地点的SFC显著高于平均的SFC，而因子由首次注意点的2.5±0.05降至第八次被注意的地点的1.6±0.05。虽然这个结果并不一定表明人眼注意力固定点与模型的注意轨迹之间的相似性，但它表明，像人类一样，根据共同的假设，即更丰富的频谱内容的区域更具信息性，模型被吸引到“信息丰富”的图像位置。SFC特征图与大多数图像的显著图类似（例如，图4.1）。然而，对于具有强烈，延伸的光照或颜色变化（例如，由于斑点噪声）的图像而言，两种特征图的表现有根本上的不同：虽然这些区域表现出统一的高SFC，但由于它们的均匀性，它们具有低的显著性（图4.2和图4.3 ）。在这样的图像中，显著性图通常与我们主观的显著性更一致。数量上，对于这里研究的258幅图像，被注意点的SFC显著低于最高SFC，因子从第一次被注意点的0.90±0.02降至第八次注意点的0.55±0.05：模型注意到的高SFC的地方，不一定是SFC最高的地方。因此，似乎显著性的不仅仅是衡量局部SFC。实现空间内部特征竞争的模式比纯粹的局部SFC方法能更好地捕捉到主观显著性。

**3.2 优点与缺点**

我们提出了一个模型，其架构和组件模仿灵长类早期视觉的属性。尽管其架构简单且是前馈特征提取机制，该模型能够在复杂的自然场景下达到优秀的表现。例如，它能迅速检测到各种显眼的交通标志形状（圆形，三角形，正方形，矩形），颜色（红色，蓝色，白色，橙色，黑色）和纹理（字母标记，箭头，条纹，圆圈），虽然它并没有专门为此而设计。这种强大的表现加强了一个观点，即从早期视觉过程接收输入的独特的显著性图可以有效地引导灵长类动物中的自下而上的注意力[4]，[10]，[5]，[8]。从计算角度来看，这种方法的主要优点在于大量并行实现，不仅是计算昂贵的早期特征提取阶段，也包括注意力聚焦系统。比以前的模型广泛地基于松弛技术[5]，我们的架构可以轻松地在专用硬件上实时操作。

该模型可以预期的表现类型主要取决于一个因素：只有物体特征在至少一个特征图中显性表示才能产生弹出效应，即独立于分散对象数量的快速检测[ 7]。如果不对预注意特征提取过程进行修改，此模型不能检测联合特征。尽管当目标以它自己的独特方向，颜色，亮度或尺寸和周围干扰物不同时，目标能被迅速检测到，它不能检测未实现的特征类型的目标显著性（如T型连接或线段终点，因为特定的神经元的存在还是有争议的）。为了简单起见，也没有在特征图内实现任何循环机制，因此不能产生轮廓竞争和封闭现象，这些现象对某类人类弹出任务很重要[19]。另外，目前我们的模型并不包括任何已经在人类显著性中发挥重要作用的大细胞运动通道[5]。

关键模型组件是归一化，为任何情况下的显著性计算提供了一个通用的机制。由模型实现的显著性方法虽然经常与局部SFC相关，但更接近于人类的显著性，因为它实现了显著位置之间的空间竞争。 我们的前馈实现比先前提出的迭代方案更快更简单[5]。 在神经元中，在条纹和外来皮层细胞的非经典接受领域已经观察到类似于的空间竞争效应[15]。

总而言之，论文提出了一个简单的显著性驱动的视觉注意焦点计算模型。其基于生物学发现而构建的结构能复制灵长类动物视觉系统的许多特性。检测目标方法的效率主要取决于实现的特征类型。 因此，这里呈现的框架可以通过实现专用特征图轻松地适应任意任务。

## 感谢

我们感谢Werner Ritter和戴姆勒奔驰的交通标志图像和Pietro Perona以及两位评论人员的出色建议。

这项研究得到美国国家科学基金会，加州理工学院神经元工程中心和美国海军研究所的支持。

## 参考文献

[1] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, and F. Nuflo, “Modelling Visual Attention via Selective Tuning,” Artificial Intelligence, vol. 78, no. 1-2, pp. 507–545, Oct. 1995.

[2] E. Niebur and C. Koch, “Computational Architectures for Attention,” R. Parasuraman, ed., The Attentive Brain, pp. 163–186. Cambridge, Mass.: MIT Press, 1998.

[3] B.A. Olshausen, C.H. Anderson, and D.C. Van Essen, “A Neurobiological Model of Visual Attention and Invariant Pattern Recognition Based on Dynamic Routing of Information,” J. Neuroscience, vol. 13, no. 11, pp. 4,700–4,719, Nov. 1993.

[4] C. Koch and S. Ullman, “Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry,” Human Neurobiology, vol. 4, pp. 219–227, 1985.

[5] R. Milanese, S. Gil, and T. Pun, “Attentive Mechanisms for Dynamic and Static Scene Analysis,” Optical Eng., vol. 34, no. 8, pp. 2,428–2,434, Aug. 1995.

[6] S. Baluja and D.A. Pomerleau, “Expectation-Based Selective Attention for Visual Monitoring and Control of a Robot Vehicle,” Robotics and Autonomous Systems, vol. 22, no. 3-4, pp. 329–344, Dec. 1997.

[7] A.M. Treisman and G. Gelade, “A Feature-Integration Theory of Attention,” Cognitive Psychology, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[8] J.P. Gottlieb, M. Kusunoki, and M.E. Goldberg, “The Representation of Visual Salience in Monkey Parietal Cortex,” Nature, vol. 391, no. 6,666, pp. 481-484, Jan. 1998.

[9] D.L. Robinson and S.E. Peterson, “The Pulvinar and Visual Salience,” Trends in Neurosciences, vol. 15, no. 4, pp. 127–132, Apr. 1992.

[10] J.M. Wolfe, “Guided Search 2.0: A Revised Model of Visual Search,” Psychonomic Bull. Rev., vol. 1, pp. 202–238, 1994.

[11] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C.H. Anderson, “Overcomplete Steerable Pyramid Filters and Rotation Invariance,” Proc. IEEE Computer Vision and Pattern Recognition, pp. 222-228, Seattle, Wash., June 1994.

[12] A.G. Leventhal, The Neural Basis of Visual Function: Vision and Visual Dysfunction, vol. 4. Boca Raton, Fla.: CRC Press, 1991.

[13] S. Engel, X. Zhang, and B. Wandell, “Colour Tuning in Human Visual Cortex Measured With Functional Magnetic Resonance Imaging,” Nature, vol. 388, no. 6,637, pp. 68–71, July 1997.

[14] C. Koch, Biophysics of Computation: Information Processing in Single Neurons. New York: Oxford Univ. Press, 1998.

[15] M.W. Cannon and S.C. Fullenkamp, “A Model for Inhibitory Lateral Interaction Effects in Perceived Contrast,” Vision Res., vol. 36, no. 8, pp. 1,115–1,125, Apr. 1996.

[16] M.I. Posner and Y. Cohen, “Components of Visual Orienting,” H. Bouma and D.G. Bouwhuis, eds., Attention and Performance, vol. 10, pp. 531–556. Hilldale, N.J.: Erlbaum, 1984.

[17] The C++ implementation of the model and numerous examples of attentional predictions on natural and synthetic images can be retrieved from <http://www.klab.caltech.edu/~itti/attention/>.

[18] P. Reinagel and A.M. Zador, “The Effect of Gaze on Natural Scene Statistics,” Neural Information and Coding Workshop, Snowbird, Utah, 16-20 Mar. 1997.

[19] I. Kovacs and B. Julesz, “A Closed Curve Is Much More Than an Incomplete One: Effect of Closure in Figure-Ground Segmentation,” Proc. Nat’l Academy of Sciences, U.S.A., vol. 90, no. 16, pp. 7,495– 7,497, Aug. 1993.