Silkman Statistical Journal

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

自然图像类别统计

Antonio Torralba[*] **Aude Oliva**[†]

翻译：林绪虹[‡]

摘要

在本文中，我们研究了属于不同类别的自然图像的统计特性及其与场景和对象分类任务的相关性。我们讨论了二阶统计量如何与图像类别、场景规模和对象相关联。我们提出了如何以前馈方式计算场景分类，以便在视觉处理链的早期提供自上而下的上下文信息。结果表明，直接基于低级特征的视觉分类（无需分组或分割阶段）可以有利于对象定位和识别。我们展示了如何在探索图像之前使用简单的图像统计数据来预测场景中对象的存在和不存在。（本文中的一些图片仅在电子版中为彩色）

Keywords: Statistics, Image Processing.

1. 导言

Figure 1 显示了通过平均来自同一语义类别的图片而创建的平均图像集合。根据 Rosch 及其合作者 (1976) 的开创性工作，人们在同一抽象级别识别大多数对象：基本级别（例如汽车、椅子）。已经证明，同一基本级别类别的对象具有相似的共同特征，并且通常具有相似的形状。这在 Figure 1 中显示的平均图像或“原型”中得到了说明。在每个原型图像中，有关局部特征（例如彩色区域分布和强度模式）之间存在的空间相似性水平的信息通过清晰度来展示。面部和行人等对象类别在像素强度分布方面比其他对象组（例如椅子）更为规则。

与物体类似，描绘环境场景的自然图像也可以归类为基本类别（Tversky 和 Hemenway 1983），因此，预计它们具有共同的特征（Jepson *et al* 1996）。尽管环境场景中部分和区域的组织比物体中少得多，但生成的原型图像在空间上并不是静止的。场景图片的类别与图像中彩色区域的分布之间存在很强的关系。类似地，物体周围背景场景的结构和颜色模式的分布也受到限制。Figure 1（场景中的物体）中显示的第三行平均原型是通过对数百张图像进行平均而创建的，这些图像被限制为在图像中以某一比例呈现特定物体（在平均之前，图像被平移以使感兴趣的物体位于中心）。由于物体与其上下

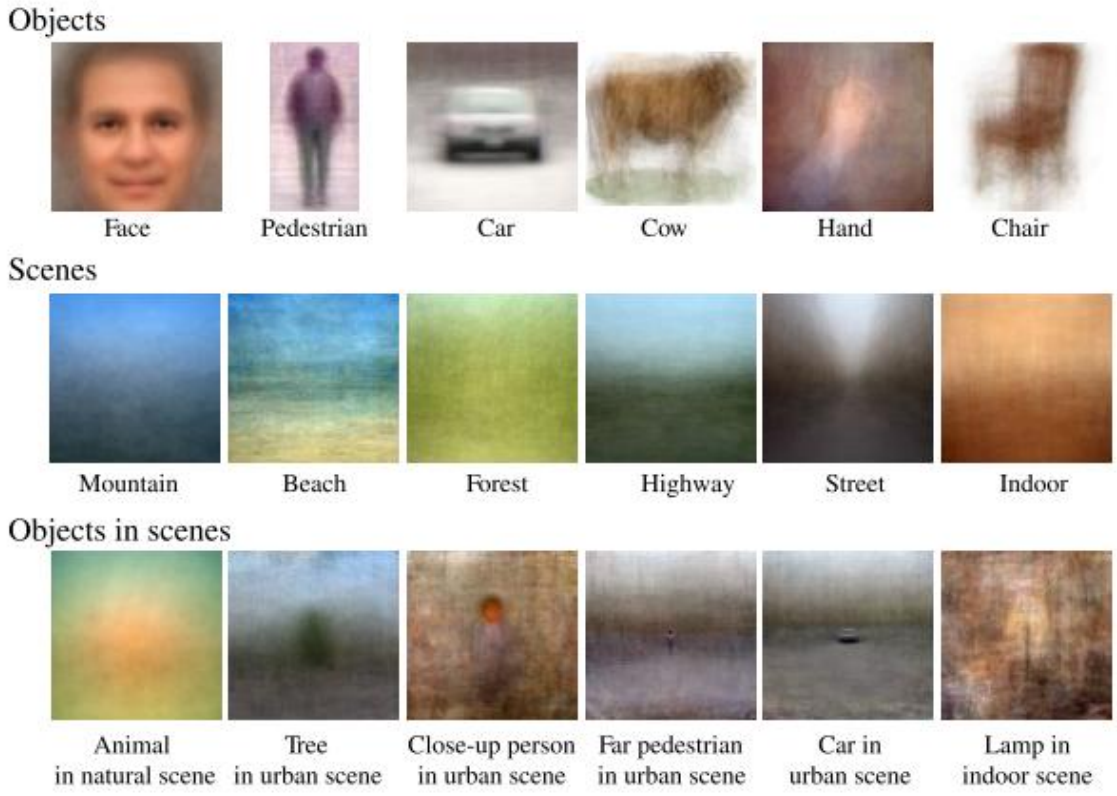

图 1: 物体、场景和场景中物体类别的平均图片，每个类别使用 100 个或更多样本计算得出。样本被选择为与观察者具有相同的基本水平和视点。场景中的物体组（第三行）代表图像中心物体周围平均外围信息的示例。

文之间存在相关性，因此背景不会平均为均匀场（Torralba 2003）。相反，背景展现出在通常发现特定物体的所有环境中所共有的纹理和颜色图案。

Figure 1 说明了在考虑生态观察条件时自然图像类别中发现的规律性程度。自然图像类别的统计数据不仅取决于视觉世界如何构建以服务于特定功能，还取决于观察者所采用的视角。由于不同的环境类别由不同的物体和材料构建而成，并且观察者的视角施加了其自身的限制（例如其大小和运动），我们预计会在图像信息的统计分布中发现强烈的偏差。这些统计数据可能对不同的动物物种有不同的偏差。

在本文中，我们说明了自然图像的简单统计数据如何随着观察者与世界之间的相互作用而变化。本文的结构如下：第 2 节描述了每个场景类别的自然图像的统计特性。第 3 节和第 4 节分别介绍了自然图像的光谱主成分和场景调整滤波器，并描述了如何使用这些方法执行简单的场景分类任务。第 5 节总结了场景分类的计算方法，第 6 节介绍了自然图像的简单统计方法。显示了简单统计数据在执行对象检测任务中的稳健性。

2. 自然类别的统计特性

2.1. 自然图像的 1/F 光谱

已发现自然图像的统计数据遵循特定的规律。开创性研究（Burton 和 Moorhead 1987、Field 1987、1994、Tolhurst *et al* 1992）发现，自然图像的平均功率谱呈 $1/f^{\alpha}$ 形式，其中 $\alpha \sim 2$（或为 $\alpha \sim 1$ 如果考虑到振幅谱，参见 Figure 2 (a)）。

相关研究发现方向分布存在偏差，如 Figure 2 的功率谱所示。在现实世界的图像中，包括自然景观和人造环境，垂直和水平方向比倾斜方向更常见（Baddeley 1997 年，Switkes 等人 1978 年，van der Schaaf 和 van Hateren 1996 年，Oliva 和 Torralba 2001 年）。一个更完整的平均功率谱模型（使用极坐标）可以写成

$$E[|I(f,\theta)|^2] \simeq A_s(\theta)/f^{\alpha_s(\theta)} \tag{1}$$

其中光谱的形状是方向的函数。函数 $A_x(\theta)$ 是每个方向的振幅缩放因子，$\alpha_s(\theta)$ 是频率指数，是方向的函数。这两个因素都对功率谱的形状有影响。在分别考虑人造和自然场景图像的功率谱时，需要方程 (1) 的模型（参见 Figure 2 和 Table 1，以及 Baddeley1996、1997）。Table 1 显示，斜率 α 和振幅 A 的值随方向和环境类型而变化[1]。方向的各向异性分布也与

氢	氧	V
自然		
一个		
1.98 (0.58)	2.02 (0.53)	2.22 (0.55)

[1]本研究中使用的数据库包含约 12 000 张场景和物体图片。图像大小为 256×256 像素。它们来自 Corel 图片库、数码相机拍摄的照片和从网络下载的图像。

氢	氧	V
一		
0.96 (0.40)	0.86 (0.38)	1 (0.35)
人造		
一个		
1.83 (0.58)	2.37 (0.45)	2.07 (0.52)
一		
1 (0.32)	0.49 (0.24)	0.88 (0.29)

表 1: 代表人造和自然环境的图像的平均 α 和 A 值。A/f^{α} 模型在三个方向（水平方向，f_x；倾斜方向和垂直方向，f_y）对每幅图像的功率谱进行拟合，从而得到 α 和 A 值。拟合在频率区间 [0.02, 0.35] 周期/像素内进行。对振幅因子 A 进行了归一化处理，使最大平均值归一。平均值是根据每个类别 3500 多张图像计算得出的（参见图 2(b)、(c)）。如果对平均功率谱进行拟合，也会得到类似的值。括号中的数字表示标准偏差。

神经生理学数据显示，早期皮质阶段的细胞数量随着空间尺度调节和方向的变化而变化（例如，中央凹的垂直和水平调节细胞比倾斜细胞多，DeValois 和 DeValois 1988）。

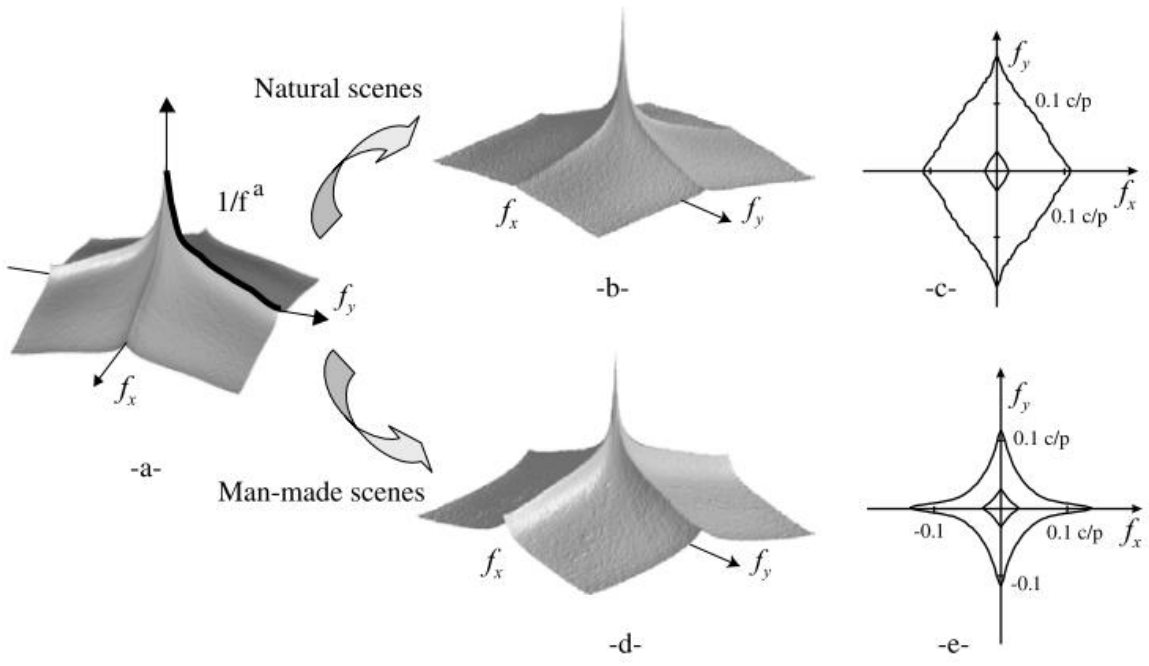

图 2: (a) 12,000 张图像的平均功率谱（纵轴采用对数单位）。用 6,000 张人造场景图片 (b) 和 6,000 张自然场景图片 (d) 计算的平均功率谱；(c) 和 (e) 是它们各自的光谱特征。轮廓图代表光谱特征能量的 50% 和 80%。选择轮廓是为了使部分内成分的总和代表总量的 50%（和 80%）。单位为每像素周期数（另请参阅 Baddeley 1996）。

2.2. 图像类别的光谱特征

不同类别的环境还表现出不同的方向和空间频率分布，这些分布在平均功率谱中得到体现（Baddeley 1997、Oliva *et al* 1999、Oliva and Torralba2001）。Figure 3 显示，各种人造类别之间的区别主要在于不同尺度的水平和垂直轮廓之间的关系，而自然环境的光谱特征具有更广泛的变化形状。

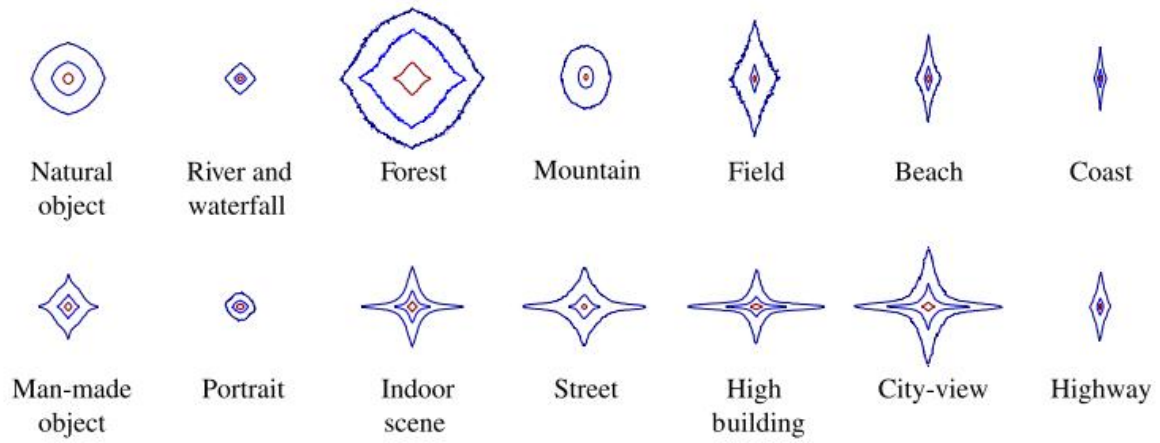

图 3: 14 个不同图像类别的光谱特征。每个光谱特征都是通过对每个类别的几百幅图像的功率谱取平均值而获得的。轮廓图代表光谱特征能量的 60%、80% 和 90%（能量是通过将傅里叶分量的平方相加而获得的）。光谱特征的大小与斜率 (α) 相关。较大的 α 值会导致高空间频率下的能量快速衰减，从而产生较小的轮廓。整体形状是 $\alpha(\theta)$ 和 $A(\theta)$ 的函数。

当考虑街道、高速公路、建筑物、森林等环境场景的基本类别时，每个场景类别的特定光谱特征更加引人注目。从 Figure 3 的轮廓图中，我们可以看到，主要的空间尺度和主要的方向是代表不同体积或深度范围的场景类别的典型特征。大规模场景（例如海滩、海岸、田野）图片的光谱特征以地平线为主导。当场景背景越来越靠近观察者时（从山脉到封闭的景观和自然物体），光谱特征在高空间频率下变得各向同性且更密集。光谱特征的形状与图像主要成分应位于的尺度（例如大小）相关（例如森林中的纹理更细，瀑布中的纹理更粗）。

2.3. 场景比例和图像比例

考虑不同尺度的场景时，图像统计数据也会有所不同。Figure 4 显示了具有相似深度范围的场景的光谱特征。这些特征是从图像数据库中获得的，其中要求四个受试者提供图像中所代表环境的平均深度或体积（Torralba 和 Oliva 2002）。场景尺度以米为单位。每个光谱特征都是通过平均相似距离范围内的场景图片的功率谱来计算的。

当考虑较大的尺度变化（大于 10 倍）时，描绘不同尺度场景和物体的图片的空间统计和光谱统计之间存在显著差异。至少有两个因素可以解释结构和深度范围之间的依赖关系。首先，任何特定观察者对特定场景所采取的视角都受到场景体积的限制。只要观察者能够直接操纵物体，就可以从无数个视点观察许多现实世界的物体。然而，随着距离和尺度的增加，人类观察者的可能视点变得越来越有限和可预测。例如，高层建筑通常是从地面或另一栋建筑的窗户观察的。其次，由于功能限制以及

塑造每个尺度空间的物理过程，构成一个场景的部分或物体在不同尺度上存在很大差异。

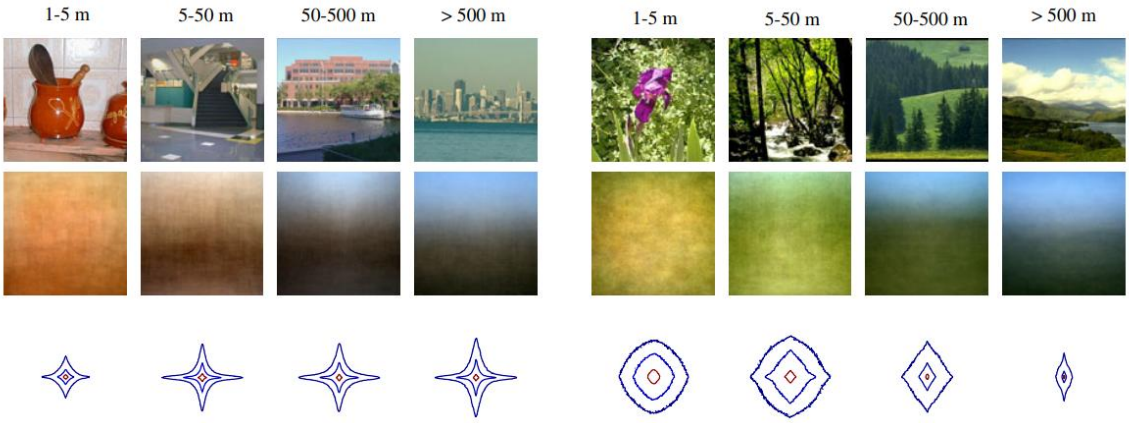

图 4: 作为场景尺度函数的平均空间图像和光谱特征。场景比例是指观察者与组成场景的主要元素之间的平均距离。每幅图像的平均值和光谱特征都是通过 300-400 幅图像计算得出的。

Figure 4 强调了不同尺度范围内人造环境和自然环境之间的差异。近距离观察人造物体往往会产生由平坦光滑的表面组成的图像。因此，近距离观察的功率谱能量主要集中在低空间频率中。随着观察者与场景背景之间的距离增加，视野会涵盖更大的空间，可能包含更多物体。人造场景的感知图像看起来像是分解成较小部分（物体、墙壁、窗户等）的表面集合。因此，随着视野覆盖的区域随距离增加而增加，场景变得更加混乱，与高空间频率相对应的光谱能量也会增加。相比之下，自然环境的光谱特征在增加深度时表现不同。

Figure 4 显示，当观察者与背景之间的距离增大时，自然结构会变得更大、更平滑（由于图像的空间采样，小颗粒会消失）。因此，平均而言，随着距离的增加，杂波水平会降低，高空间频率中的能量也会降低。此外，方向模式会随尺度而变化。近距离观察自然结构时，其方向趋向于各向同性（观察者的视角不受约束）。随着距离的增大，对垂直和水平方向的偏向会增加，同时观察者的视角也会受到更多限制。随着距离的不断增加，能量主要集中在垂直空间频率中，因为非常大的环境场景是沿着水平层组织的。为了识别场景或浏览这种全景环境，面对视角限制，观察者可能会考虑朝地平线看去，以在视觉上拥抱整个场景。

一些研究已经检验了自然图像统计的尺度不变性（例如 Field 1987、Ruderman 1997 等）。这些研究集中于不同图像尺度下小波输出统计数据之间的相似性。结果表明，一些图像统计数据是尺度不变的。在这里，我们区分了*图像尺度*（指空间频率尺度）和*场景尺度*（指观察者与构成场景的元素之间的平均距离）。请注意，对于我们考虑的距离范围（从 1 米到几公里），距离问题不能建模为缩放因子。随着距离的每次数量级变化，感知的图像也属于不同的场景语义类别（单个物体、房间、地点、大型户外和全景场景）。因此，我们可以预期图像的统计数据可能会在改变场景尺度时发生变化，并提供有关场景可能深度范围的有用分类信息（参见第 5 和第 6 节，以及 Torralba 和 Oliva 2002）。

Figure 5 显示了在考虑图像尺度和场景尺度时定向小波的输出能量如何变化。所使用的小波是定向 Gabor 滤波器，在 12 个不同方向上以 1/4 周期/像素的径向频率进行调整。通过将图像从 256^2 像素下采样到 32^2 像素，以两个因子获得图像尺度的变化。通过对具有不同深度范围的场景图片（从

物体和纹理的特写视图到全景视图和自然景观）获得的输出进行平均，获得场景尺度的变化。在场景尺度上观察到了光谱特征最重要的修改。当修改场景尺度时，极坐标图的形状会通过改变每个方向上的能量而演变。然而，在图像尺度上，变化很小。这种观察对于自然环境比人造场景更引人注目。

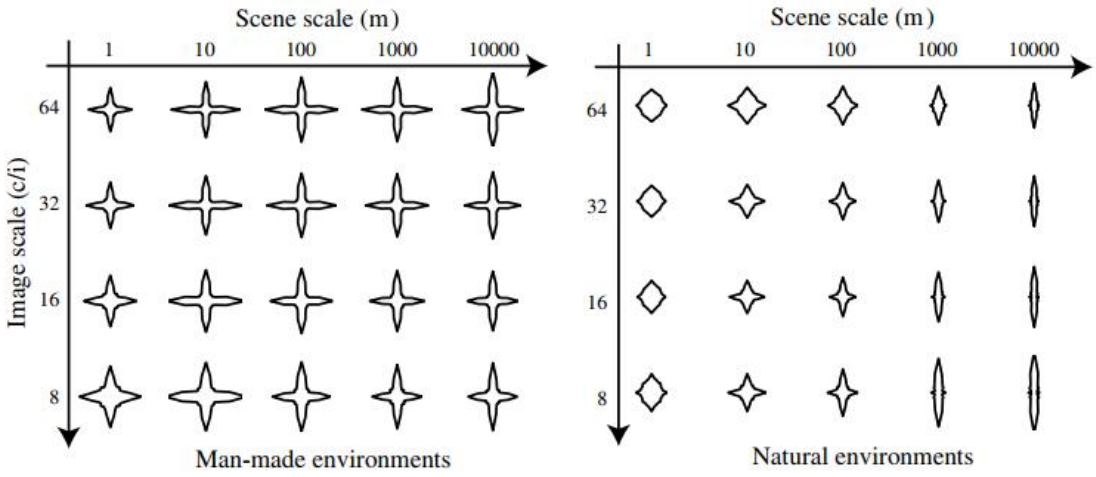

图 5: 多尺度定向 Gabor 滤波器响应的极坐标图。每个方向的幅度对应于整个图像的平均总输出能量。能量在不同图像尺度上的归一化是通过乘以一个常数来实现的，因此具有 $1/f$ 振幅谱的噪声在所有图像尺度上都具有相同的极坐标图。

2.4. 非平稳统计数据

自然图像的另一个重要特征是图像统计数据如何随空间位置而变化。当考虑眼睛相机的所有可能方向时，自然图像的统计数据预计是尺度不变的（Field 1994, 1999, Ruderman 1994, 1997）和静止的（特征在位置方面均匀分布，Field (1994), (1999)）。对于物体近距离视图的图像统计数据，情况确实如此，这些物体平均而言是静止的，因为相机没有首选的视点（参见 Figure 6 ，场景比例为 1 米）。然而，对于包含大体积的场景的图像，由于其高度和可能的位置（在地板上），人类观察者可能采用的视点会变得更加受限。如果观察者的任务是识别大空间场景的身份，那么大多数有用信息将在看向地平线时给出。因此，不同的图像统计数据将表征图像的上半部分和下半部分（例如，天空的平滑纹理、顶部的天际线的长垂直轮廓、底部的杂乱形状）。Figure 7 显示了图像反转如何影响对场景绝对深度的感知的示例。请注意，不仅是场景的相对形状被误解，而且绝对比例也被误解。对于大多数观察者来说，左侧图像比右侧图像看起来更接近结构。

由于图像统计数据是观察者的功能，因此从鸟瞰视角拍摄的大规模场景图像应能得出几乎平稳的特征分布，因为视点在感知图像的可能方向上完全不受限制。对于站立的人类观察者来说，视点受到很大限制，从而产生具有非平稳统计数据的图像，如 Figure 6 所示。非平稳性与感官和认知系统非常相关，因为它可以提供特定类型环境的不变特征。

3. 真实世界图像的主要成分和功率谱

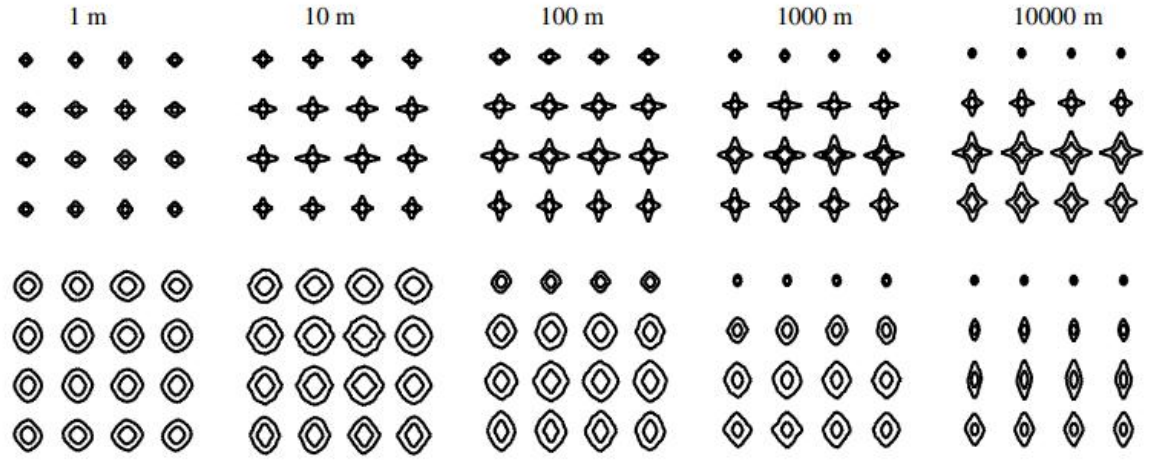

图 6: 不同深度尺度（从左到右，从近景到全景）的人造环境（上图）和自然环境（下图）组中图像统计数据的非稳态性说明。光谱特征是通过对图像中 4×4 位置的窗化功率谱进行平均而获得的。随着场景尺度的增加，图像统计变得非稳态。

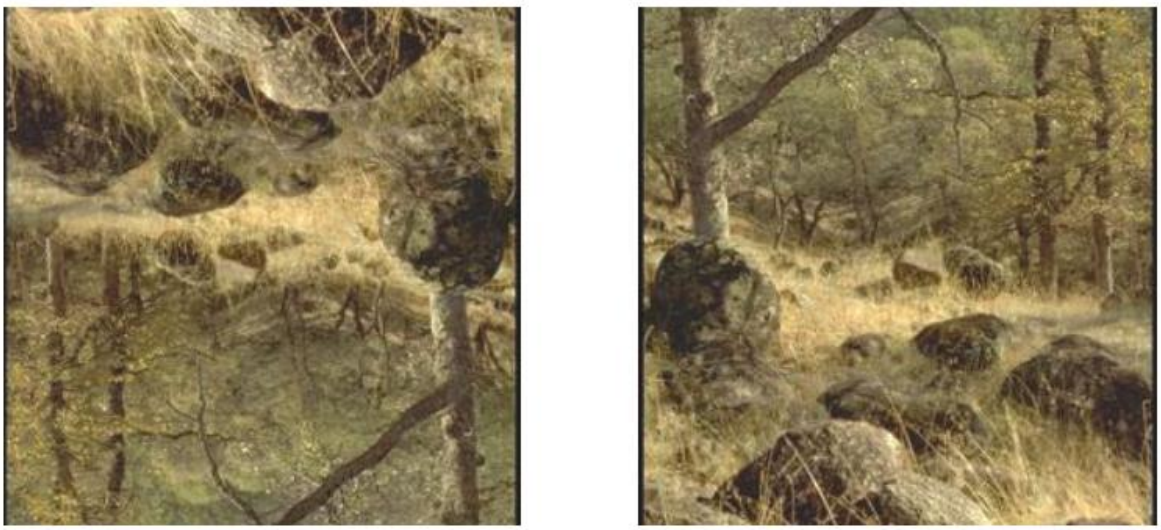

图 7: 自上而下对深度判断的影响。左边的图像一般被认为是灌木丛的特写，顶部可能还有蜘蛛网。右边的图像被归类为森林内部，对应的距离比左边的图像大。左侧图像对应的是颠倒和左右反转后的右侧图像。由于假设光线来自上方，倒置反转会影响对凹凸的感知，因此会改变感知到的场景相对三维结构。此外，错误的识别还会影响感知空间的绝对比例。

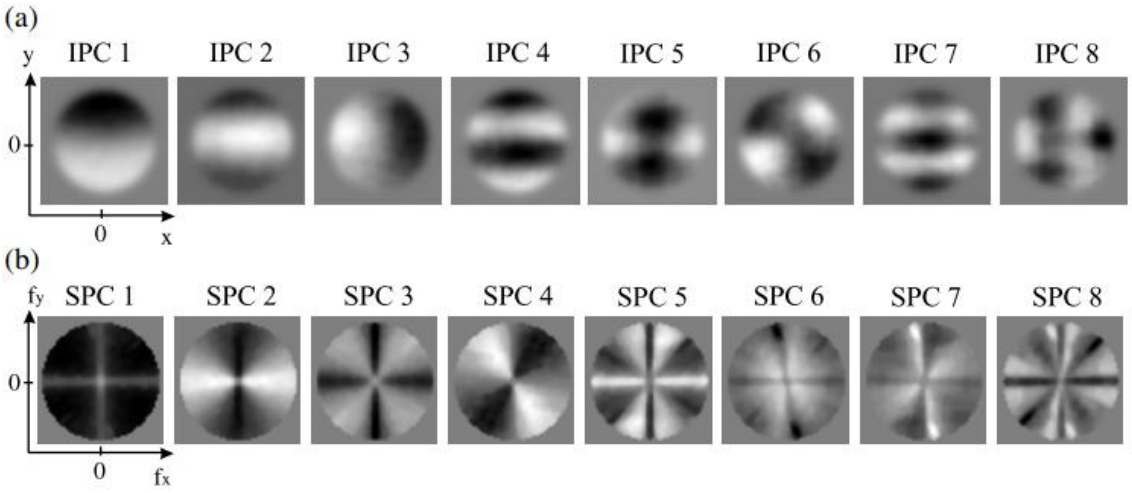

图 8: (a) 图像的前八个主成分（IPC）和 (b) 图像的能谱（SPC）。频率 $f_x = f_y = 0$ 位于每个图像（SPC）的中心。频率区间为 [-1/2, 1/2] 。在应用 PCA 之前，每个频率的振幅都根据其标准偏差进行了归一化处理。

主成分分析 (PCA) 已广泛应用于编码和识别的视觉问题。在人脸识别领域，当人脸正确对齐和缩放时，PCA 会提供一组简化的正交函数，能够重建人脸（Craw 和 Cameron 1991，Turk 和 Pentland 1991）。此操作有助于在低维空间中执行识别过程（Sirovich 和 Kirby 1987，Swets 和 Weng 1996）。PCA 还用于获取视觉输入的有效代码，以适应自然刺激的统计数据（例如 Hancock *et al* 1992，Liu 和 Shouval 1994）。图像主成分（IPC）将图像分解为

$$i(x,\, y) = \sum_{n=1}^{p} v_n \mathrm{IPC}_n(x,\, y) \tag{2}$$

其中 $i(x,\, y)$ 是图像沿空间变量 x 和 y 的强度分布。$P \leqslant N^2$ 是 IPC 总数，$N2 = 2562$ 是图像像素数。IPCn(x, y) 是协方差矩阵的特征向量：T = E[(i−m)(i−m)T]，其中 i 是重新排列在列向量中的图像像素。E 是期望算子。$m = E[i]$ 是图像的平均值。vn 是描述图像 i(x, y) 的系数。Figure 8 (a) 显示了根据 5000 张真实场景图片计算出的 IPC。正如 Field (1994) 所讨论的那样，自然图像的平稳性决定了 IPC 形状（Figure 8 (a)），它对应于傅里叶基。

在这里，我们通过取图像离散傅里叶变换（DFT）的平方来计算图像的功率谱：

$$\Gamma(k_x, k_y) = \frac{1}{N^2} |I(k_x, k_y)|^2$$

其中

$$I(k_x, k_y) = \frac{1}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} i(x, y) \exp\left(-\frac{\mathrm{j}2\pi}{N} (x k_x + y k_y) \right)$$

$f_x = k_x/N$ 和 $f_y = k_y/N$ 是离散空间频率。功率谱[2]，$\Gamma(k_x, k_y)$，对每个空间频率和方向的能量密度进行编码整個图像。

PCA 应用于功率谱给出了考虑图像之间的结构变异性。首先，对功率谱进行归一化对于每个空间频率的方差:$\Gamma'(k_x, k_y) = \Gamma(k_x, k_y)/\text{std}[\Gamma(k_x, k_y)]$,其中 $\text{std}[\Gamma(k_x, k_y)] = \sqrt{E[(\Gamma(k_x, k_y) - E[\Gamma(k_x, k_y)])^2]}$. 这种归一化补偿了 $1/f^\alpha$ 功率谱形状。谱主成分 (SPC) 分解图像的归一化功率谱为

$$\Gamma_s^{'}(k_x, k_y) = \sum_{n=1}^{P} u_n \text{SPC}_n(k_x, k_y). \tag{5}$$

P 是 SPC 的数量。

Figure 8 (b) 显示了生成的 SPC。根据第 2 节的结果，前三个主成分表现出垂直和水平光谱成分优势。Figure 9 显示了一组场景图像，这些图像根据其功率谱沿第二和第三光谱主成分的投影进行组织。沿着 SPC2 轴，具有主导地平线的场景（例如开阔景观和郊区开阔场景）位于图的顶部，与定义封闭和封闭环境的场景（具有各向同性的功率谱）相对。沿着 SPC3 轴，具有十字形功率谱（主要是城市地区）的图像位于一个极端，而自然景观位于另一个极端。出现了广泛的环境类别组织，表明自然图像二阶统计量的可变性可能与自然场景分类任务有关。前三个 SPC 的线性组合能够以 80% 的准确率将人造场景与自然场景区分开来。

4. 场景识别的接受场

SPC 所达到的组织水平表明，自然图像二阶统计量的变异性可能与分类目的相关。如图 2-6 所示，我们观察到二阶图像统计量随自然度维度（人造景观与自然景观）和开放度维度（与深度相关，Baddeley 1997，Oliva 和 Torralba 2001）而变化。这表明，环境视图沿这两个感知维度的分类状态可以以前馈方式从一组低级检测器计算出来，这些检测器编码的信息类似于功率谱提供的信息（另见第 6 节）。

在本节中，我们寻找最佳光谱统计数据，以区分人造、自然、开放和封闭场景类别。使用归一化图像功率谱 $\Gamma'(k_x, k_y)$ ，可以实现两个场景类别之间的线性区分，如下所示

$$w = \sum_{k_x=0}^{N-1}\sum_{k_y=0}^{N-1} \Gamma'(k_x, k_y)\text{DST}(k_x, k_y) = \sum_{k_x=0}^{N-1}\sum_{k_y=0}^{N-1} \Gamma(k_x, k_y)\text{DST}'(k_x, k_y)$$

$\text{DST}'(k_x, k_y)$ 是区分两个类别所需的光谱成分的权重（DST 代表判别光谱模板）。w 是最具判别性的特征，是图像功率谱的加权和。由于 SPC 定义了一个完整的正交基来描述归一化功率谱，我们可以将 DST 写成

$$\text{DST}(k_x, k_y) \simeq \sum_{n=1}^{P} a_n \text{SPC}_n(k_x, k_y).$$

[2]虽然公式（4）中没有反映出，但为了计算先例断面的光谱特征，我们对图像应用了环形汉明窗，以避免边界伪影。

图 9: 将图像投影到功率谱第二和第三主成分所代表的空间中。图像根据光谱特性排列：SPC_2 将能量主要集中在 f_y 轴的图像放在图的上方，而能量主要集中在 f_x 轴的图像则放在图的下方。SPC_3 将能量位于 f_x 和 f_y 轴（十字形）的图像与能量位于斜方向的图像对立起来。这就出现了一种粗略的场景组织：人造场景与自然场景，开放环境与封闭环境。

系数 a_n 表示如何对每个 SPC 进行加权，以构建特定的模板 DST。这里，我们使用了前 $P = 16$SPCs。系数 a_n 由监督学习阶段确定。在学习阶段，每幅图像由特征列向量 $u = \{u_n\}$ 表示，u_n 是图像功率谱在第 n 个 SPC 中的投影。然后，我们定义两组不同场景类别的图像（例如，人造环境和自然环境的图片，与第 3 节中描述的图像数据库相同）。可以通过应用 Fisher 线性判别分析（例如 Ripley 1996、Swets 和 Weng 1996）来学习 DST 的参数 a_n，该分析寻找给出最佳分类率的系数 a_n。

经过训练后，人造场景与自然景观的分类率达到 93%（而仅使用 SPC 时为 80%）。训练和测试分别在不同的一组图像上进行，每组图像有数千个样本。当将判别分析应用于开放和封闭环境之间的区分时，正确分类率达到 94%。

虽然可以使用更复杂的分类器，但线性分类器可以进行简单的分析。由于方程 (6) 的线性，在光谱域 (DST) 中执行的区分可以写在空间域中。然后可以计算调整到全局场景统计数据的感受野，以区分自然场景的类别。

具有传递函数 $H\left(k_x, k_y\right)$ 的离散滤波器的输出能量可以计算为

$$E = \sum_{k_x=0}^{N-1} \sum_{k_y=0}^{N-1} \Gamma(k_x, k_y) |H(k_x, k_y)|^2. \tag{8}$$

该表达式类似于用于计算结构特征 w 的公式 (6)。但是，由于滤波器传递函数的平方幅度不能为负值，因此可以通过计算两个滤波器输出能量之间的差异来实现 DST。在这种情况下，我们可以将 w 计算为两个能量之间的差异，即 $w = E_+ - E_-$，其中 E_+ 和 E_- 分别是传递函数为 H_+ 和 H_- 的两个滤波器的输出能量。在这种情况下，我们得到

$$w = \sum_{k_x=0}^{N-1} \sum_{k_y=0}^{N-1} \Gamma(k_x, k_y) (|H_+(k_x, k_y)|^2 - |H_-(k_x, k_y)|^2). \tag{9}$$

使用此表达式，可以获得 DST 的正值和负值。几个函数 H_+ 和 H_- 给出相同的 DST 结果。这里，我们使用

$$|H_+(k_x, k_y)|^2 = \text{rect}[\text{DST}'(k_x, k_y)] \tag{10}$$

和

$$|H_-(k_x, k_y)|^2 = \text{rect}[-\text{DST}'(k_x, k_y)]$$

其中 rect(x) 是半整流函数：如果 x > 0，则 rect(x) = x，如果 x < 0，则 rect(x) = 0。

这些方程给出了两个滤波器的幅度。由于可以自由选择相位，我们选择了零相位滤波器，因为这使我们能够获得具有空间局部脉冲响应的滤波器。

这两个滤波器的脉冲响应是能够最好地区分两组图像的接受场。Figure 10 分别显示了两个滤波器对人造场景和自然场景的 *自然度* DST 和 *开放度* DST 的脉冲响应。

在傅里叶域中计算的 DST 相当于将图像与两个空间不变滤波器进行卷积，然后计算它们的总输出能量差。两个脉冲响应 $h_+(x, y)$ 和 $h_-(x, y)$ 揭示了在要考虑的两组相反图像之间具有判别性的空

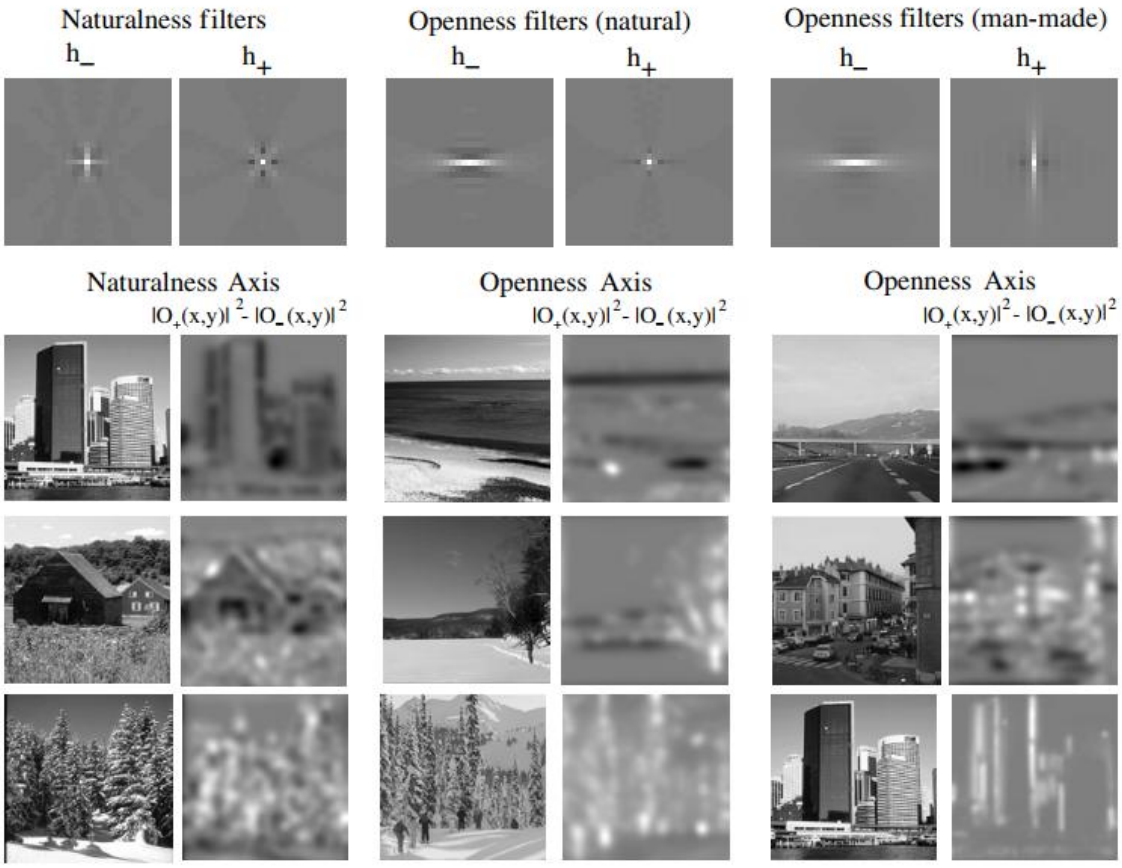

图 10: 自然度和开放度维度的对手过滤器 h- 和 h+ 图示。对于每个维度，我们都展示了对手能量通道图像的示例。

间特征。对于人造场景与自然场景，我们看到交叉脉冲响应与各向同性（略微倾斜）脉冲响应。对于开放与封闭的自然场景，我们发现水平边缘检测器与各向同性脉冲响应。对于人造场景，用于测量空间开放度的脉冲响应类似于水平边缘检测器与垂直边缘检测器。

如果我们通过将图像与相应的脉冲响应进行卷积来计算两个滤波器的输出，$o_+(x,y) = i(x,y) * h_+(x,y)$ 和 $o_-(x,y) = i(x,y) * h_-(x,y)$，那么结构特征也可以通过以下公式获得：

$$w = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (|o_+(x,y)|^2 - |o_-(x,y)|^2).$$

我们将经济能量通道定义为通过差值 $d(x,y) = |o_+(x,y)|^2 - |o_-(x,y)|^2$ 获得的图像。该信号给出了对判别器特征 w 有贡献的空间位置。Figure 10 显示了不同图像的对手能量通道。城市图片以黑色显示人造环境典型的垂直和水平分量。自然图像成分以白色表示。农场场景显示草和树的自然纹理成分（白色）和人造成分（黑色）。与 *oppness* 相对应的组件能量通道将全景场景的地平线（黑色负值）与其他光谱成分（白色）对立。

5. 语义分类

在开创性的视觉方法中，视觉处理被描述为复杂程度不断增加的模块的层次化组织，其中最后一个模块衍生出场景的类别（Barrow 和 Tenenbaum 1978、Marr 1982、Biederman 1987）。最近的计算方法将场景识别呈现为一个过程，该过程结合了一组基于图像的特征（例如颜色、方向、纹理）以形成更高级的表示，例如区域（Carson *et al* 2002 中的 *blobword*）、简单块（Biederman 1987 中的 *geons*）或对象（Barnard 和 Forsyth 2001）。这种方法建立在生物学证据的基础上，表明视觉编码机制基于多尺度和多方向的表示（Hubel 和 Wiesel 1968）。在视觉处理的最早阶段（从视网膜到 V1），图像由编码线条和边缘的局部特征表示（例如，Atick 和 Redlich 1992）。独立成分模型：Bell 和 Sejnowski（1997）、van Hateren 和 van der Schaaf（1998）。稀疏编码模型：Field（1994）、Olshausen 和 Field（1996）、Olshausen *et al*（2001）、Vinje 和 Gallant（2000）、Simoncelli 和 Olshausen（2001）。在下一个处理阶段，更复杂的特征家族在视觉皮层 V4 和 TEO 中编码。例如，这些区域中的细胞已被证明对曲线轮廓有选择性（例如 Gallant *et al* 1993）或调整到 3D 方向（Hinkle 和 Connor 2002）。对复杂物体斑块作出反应的细胞（Tanaka 1993、Fujita *et al* 1992、Logothetis *et al* 1995）后来被发现于颞下皮质的前部区域（IT，请参阅 Gallant 2000 的综述和 Ullman *et al* 2002）。最后，有人提出现实世界场景布局的皮质表征位于 PPA（海马旁皮质的一个区域，Epstein 和 Kanwisher (1998)）。

渐进重建方案的另一种方法是直接从低级特征池构建语义信息。这种方法利用了前面几节中描述的场景类别特征统计分布的规律性。使用这种观点，高级机制可以对现实世界场景图像进行分类，而无需完全基于区域和对象分割阶段（Schiele and Crowley 2000、Vailaya *et al* 1998、Oliva and Torralba 2001、2002、Torralba and Oliva 2002、Torralba 2002）。这种方法与实验研究一致，实验研究表明，人类观察者能够在一次注视中识别复杂场景的图片（Potter 1975）。当物体质量下降到无法单独识别时，观察者也可以恢复场景图像的身份（Oliva and Schyns 1997、2000、Torralba 2003）。与这种直接场景分类方案相一致的计算方法已经提出了简单的方法，可以成功地对不同抽象层次的现实世界场景进行分类。可以根据多尺度特征的统计数据来区分现实世界图像的上级类别（Gorkani 和 Picard

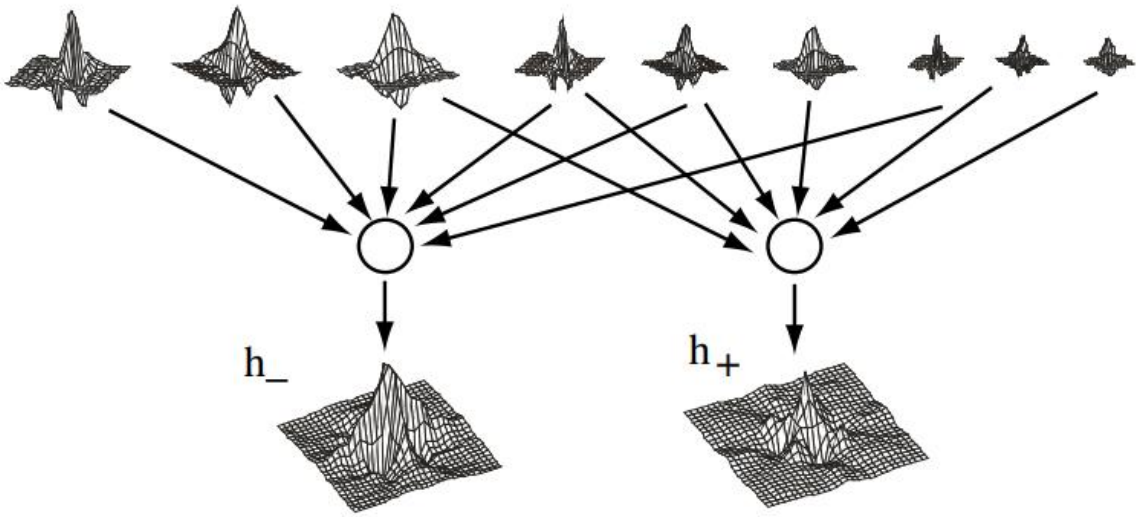

图 11: 图中展示了根据多方位和多尺度调整的简单滤波器的线性组合如何组成代表自然度检测器的复杂滤波器。

1994、Szummer 和 Picard 1998、Guerin-Dugue 和 Oliva 2000、Oliva 和 Torralba 2001，Vailaya *et al* 1998，1999）。场景分类在基本级别类别中（精确到街道、建筑物、高速公路）也可以通过选择相关空间频率和方向调整滤波器直接执行（Oliva 和 Torralba 2001、2002、Torralba 和 Oliva 2002）。从第一个光谱主成分（Figure 9 ）简单地出现有意义的图像组织以及编码复杂场景属性（例如自然性和开放性）的可行性表明，场景分类可能来自一组与生物系统中视觉世界的统计属性相匹配的维度。在 Oliva 和 Torralba (2001) 中，展示了场景所涵盖的空间固有的 3D 属性（例如深度范围、开放性、扩展性、粗糙度和崎岖性等）如何可以通过低级特征的线性组合来估计。Figure 11 显示了用于区分自然场景和人造场景的两个潜在接受场，它们是 Gabor 型细胞的线性组合。调整到用于沿 自然性和 开放性维度对图像进行分类的复杂特征的细胞可能位于视觉处理架构的不同级别，尽管它们的形状可能与 V4 中的接受场非常相似（Gallant *et al* 1993，Hinkle 和 Connor 2002）。

直接从低级特征基础计算“高级场景”基元的系统的结果对识别的理论和计算问题很有吸引力。如果可以以这种前馈方式计算出稳健的场景分类（Van Rullen 和 Thorpe 2001），它将在视觉处理链的早期提供自上而下的上下文信息，这可能对物体定位、分割和识别过程有益。事实证明，人类观察者在寻找物体时使用自上而下的机制来找到感兴趣的区域，而不管物体的物理特征是否存在（Henderson *et al* 1999）。下一节专门讨论这种上下文效应对物体处理的影响。

6. 图像统计和上下文对象处理

正如介绍中所提到的，全局图像统计数据也与场景中存在的物体相关。这不仅是因为物体形状对全局图像统计数据有影响（当物体较小时，这种影响可以忽略不计），还因为存在的物体与其环境之间存在很强的相关性。

Figure 12 显示了场景图片的叠加，唯一的约束是平均场景应包含场景中的特定对象。平均图片

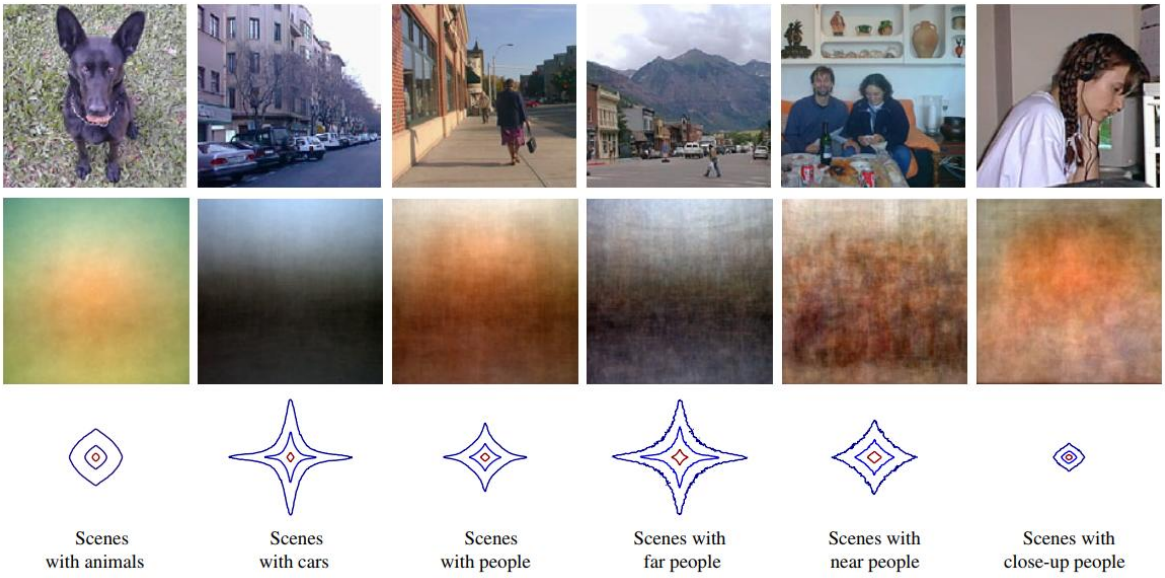

图 12: 包含特定物体的图像集的平均强度和光谱特征。图像统计数据可以预测场景中是否存在特定物体。

显示包含每个对象的场景图片的平均强度值和光谱特征。虽然没有施加其他约束，但不同的平均图片对应于所选对象的典型环境特征。例如，动物应该在自然环境中。汽车在道路和城市环境中。人们可以在许多不同的室内和室外地方找到。当将对象限制为具有特定图像大小时，光谱特征也会发生变化（这相当于固定观察者和物体之间的距离，因此与场景尺度有关）。产生人物仅占据几个像素的图像的场景是室外场景（大街、道路）。具有面部特写视图的图像可以是室内的，也可以是室外的，这将产生不同的光谱特征。同样，全局图像属性和场景中物体的位置之间存在相关性。全局图像统计数据可以预测物体的存在、它们在图像中的尺度和位置。即使感兴趣的物体太小以至于对图像的统计数据没有直接贡献，情况也是如此。例如，小型汽车是大型街道场景的典型特征，这些场景以特定的全局图像统计数据为特征。

使用图像统计数据提供了一种在不需要检测其他对象的对象检测方法中引入上下文信息的方法。为了预测感兴趣的对象的存在/不存在，我们使用了比第 4 节中执行的线性判别更可靠的统计框架。函数 $P(O|\vec{v}_C)$ 给出了给定一组表示上下文信息的全局图像统计数据 $\vec{v}_C$ 时对象类 O 存在的概率。

用于学习概率 $P(O|\vec{v}_C)$ 的训练集是一组带注释的图片。我们使用贝叶斯规则来写出 $P(O|\vec{v}_C) = P(\vec{v}_C|O)P(O) + P(\vec{v}_C|\overline{O})P(\overline{O})$。我们学习 PDF $P(\vec{v}_C|O)$，从一组存在物体的图像中获取。我们使用混合高斯函数来近似 PDF，并使用 EM 算法（Torralba 和 Sinha 2001）进行学习。以同样的方式，我们使用一组不存在物体的图像来学习 PDF $P(\vec{v}_C|O)$。我们假设 $P(\overline{O}) = P(O) = 1/2$。上下文特征 $\vec{v}_C$ 是通过将图像功率谱投影到前 16 个 SPC 中获得的特征。一旦完成学习，函数 $P(\vec{v}_C|O)$ 就会评估物体 O 与上下文 $\vec{v}_C$ 的一致性，因此，在扫描图片寻找物体之前提供有关一个物体类别可能存在/不存在的信息。

Figure 13 显示了利用图像的二阶统计预测真实世界图像中动物、人和车辆存在/不存在的性能。对于每个对象，我们使用了约 1000 张图像进行训练，并使用了 1000 张新图像进行测试。对于所有

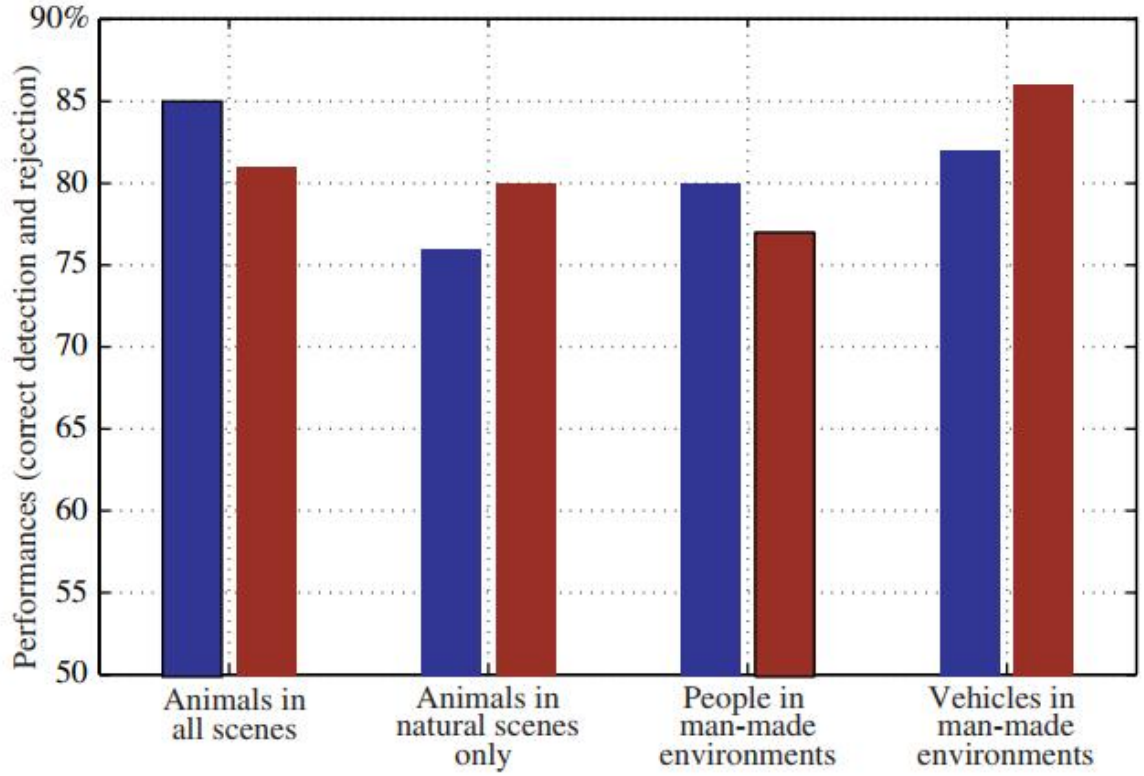

图 13: 物体预测性能。对于每个物体类别，我们显示了预测物体存在（左侧条形图）和预测物体不存在（右侧条形图）的性能。

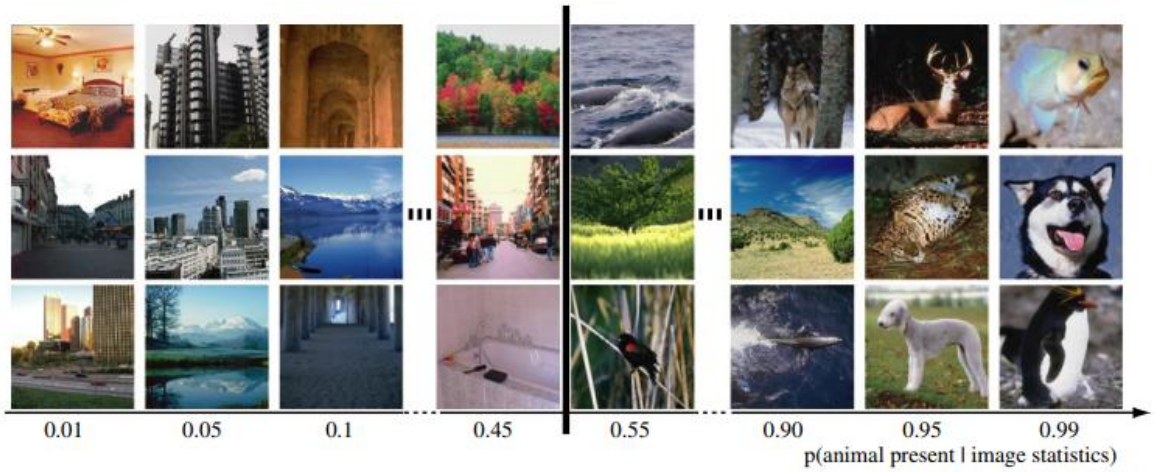

图 14: 根据图像统计预测的动物出现可能性排列的图像示意图。位于中间的图像在图像统计方面是模糊的，不能做出可靠的预测。位于两端的图像可以可靠地预测没有动物或有动物。

对象，其性能都明显高于偶然性（平均 80%），这表明全局图像统计信息为对象检测提供了相关信息。Figure 14 展示了当任务是预测场景中是否有动物时，PDF $P(O|\vec{v}_C)$ 的值。当 PDF $P(\ |\vec{v}_C) \simeq 0$ 时，系统甚至在扫描场景之前就能可靠地判断场景中没有动物。当 PDF $P(animal|\vec{v}_C) \simeq 1$ 时，系统可以预测动物的存在，并使用更专业的机制来检测动物。当使用这些简单特征时，位于中心的图像（$P(animal|\vec{v}_C) \simeq 0.5$）不能提供可靠的预测。这些结果证实了索普及其合作者的研究（索普等人，1992 年；卢塞莱等人，2002 年），这些研究表明，像动物与非动物分类这样的认知任务可以通过前馈方式完成，而不需要连续的注意力集中或分割阶段。

图像统计数据还可以预测其他对象属性，如尺度（由于场景尺度和图像统计数据之间的相关性）或其在场景中的位置。例如，我们可以了解图像中人脸位置 $\vec{x}$ ）和全局图像统计数据 $(\vec{v}_C)$ 之间的统计关系：$P(\vec{x}|\vec{v}_C)$。为了能够获取空间信息，我们在图像特 $(\vec{v}_C)$ 中包含了在图像中多个位置计算出的光谱特征。具体来说，我们将图像分成 4×4 个块，然后计算每个块的功率谱。这也使我们能够 $\vec{x}$ 中编码有关图像空间组织的信息（Oliva 和 Torralba 2001）。

概率密度函数 $P(\vec{x}|\vec{v}_C)$ 的学习提供了上下文与图像中感兴趣对象的更典型位置之间的关系。为了建模这个 PDF，我们使用了高斯混合（Gershnfeld 1999）：

$$P(\vec{x}, \vec{v}_C) = \sum_{i=1}^{M} b_i G(\vec{x}; \vec{x}_i, X_i) G(\vec{v}_C; \vec{v}_i, V_i).$$

图 15: 根据上下文特征预计包含人脸的图像和选定区域示例。这些区域是根据全局图像统计数据而非感兴趣对象的实际存在情况来选择的。

我们使用 EM 算法和带注释图像的训练数据库（Torralba and Sinha 2001，Torralba 2002）来学习该模型的参数。该 PDF 形式化了注意力焦点的上下文控制的一个方面。在寻找人脸（或任何其他感兴趣的物体）时，根据系统的过去经验，注意力将集中在具有最高可能性 $P(\vec{x}|\vec{v}_C)$ 包含目标物体的候选区域。请注意，尽管上下文特征使用 4 × 4 块对图像进行编码，但变量 x 是一个连续变量。给定图像中的 $\vec{x}$，PDF $P(\vec{x}, \vec{v}_C))$ 作为位置 $\vec{x}$ 的函数，将是高斯斑点的混合（Figure 15）。

Figure 15 显示了两个图像示例和根据上下文特征预计包含面部的选定区域。

根据图像统计数据 $(\vec{v}_C)$ 选择区域。请注意，在左侧示例中，行人体型较小，对全局图像统计数据没有贡献。在考虑完整测试数据库时，90% 的人脸位于由函数 $P(\vec{x}|\vec{v}_C)$. 的最大值定义的图像大小的 35% 的区域内。

7. 结论

自然图像的统计数据随着观察者与世界之间的互动而发生很大变化。本文展示了图像的二阶统计量如何与场景规模和场景类别相关联，并提供信息以执行快速可靠的场景和对象分类。统计规律可能是自上而下和上下文启动的相关来源，在视觉处理链的早期阶段。结果表明，直接基于低级特征的视觉分类（无需分组或分割阶段）可以有利于对象定位和识别，并可用于在探索图像之前预测场景中是否存在对象。

8. 致谢

作者特别感谢 David Field 和 Bill Freeman 的富有成效的讨论和建议，以及两位匿名评论者。还要感谢 Monica Castelhano、Dan Gajewski、Aaron Pearson 和 Michael Mike 对最终稿件的有益评论。两位作者对这项研究的贡献相同，作者身份是任意确定的。通讯可以发送给任何一位作者。

9. 参考文献

Atick J J and Redlich N A 1992 What does the retina know about natural scenes? *Neural Comput.* 4 196–210

Baddeley R 1996 Searching for filters with 'interesting' output distributions: an uninteresting direction to explore? Network 7 409–21

Baddeley R 1997 The correlational structure of natural images and the calibration of spatial representations Cogn. Sci. 21 351–72

Barnard K and Forsyth D A 2001 Learning the semantics of words and pictures Proc. Int. Conf. on Computer Vision (Vancouver) (Los Alamitos, CA: IEEE Computer Society Press) pp 408–15

Barrow H G and Tenenbaum J M 1978 Recovering intrinsic scene characteristics from images Computer Vision Systems ed A Hanson and E Riseman (New York: Academic) pp 3–26

Bell A and Sejnowski T J 1997 The 'independent components' of natural scenes are edges filters *Vis. Res.* 37 3327–38

Biederman I 1987 Recognition-by-components: a theory of human image interpretation *Psychol. Rev.* 94 115–48

Burton G J and Moorhead I R 1987 Color and spatial structure in natural scenes *Appl. Opt.* 26 157–70

CarsonC, BelongieS, GreenspanH andMalikJ 2002Blobworld: imagesegmentationusingexpectation-maximization and its application to image querying *IEEE Trans. Pattern Anal. Mach. Intell.* 24 1026–38

Craw I and Cameron P 1991 Parameterising images for recognition and reconstruction British Machine Vision Conf.ed P Mowforth (London: Springer) pp 367–70

DeValois R L and DeValois K K 1988 *Spatial Vision* (New York: Oxford)

Epstein R and Kanwisher N 1998 A cortical representation of the local visual environment *Nature* 4 598–601

Field D J 1987 Relations between the statistics of natural images and the response properties of cortical cells J. Opt. Soc. Am. 4 2379–94

Field D J 1994 What is the goal of sensory coding? *Neural Comput.* 6 559–601

Field D J 1999 Wavelets, vision and the statistics of natural scenes *Phil. Trans. R. Soc.* A **357** 2527–42

Fujita I, Tanaka K, Ito M and Cheng K 1992 Columns for visual features of objects in monkey inferotemporal cortex Nature **360** 343–6

Gallant J L 2000 The neural representation of shape *Seeing* ed K K DeValois and R L DeValois (San Diego, CA: Academic)

Gallant J L, Braun J and Van Essen D C 1993 Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex *Science* **259** 100–3

Gershnfeld N 1999 *The Nature of Mathematical Modeling* (Cambridge: Cambridge University Press)

Gorkani M M and Picard R W 1994 Texture orientation for sorting photos 'at a glance' Proc. Int. Conf. on Pattern Recognition (Jerusalem) vol 1 (New York: IEEE) pp 459–64

Guerin-DugueA andOliva A 2000 Classificationofscenephotographs from localorientations features Pattern Recogn. Lett. 21 1135–40

Hancock P J, Baddeley R J and Smith L S 1992 The principal components of natural images *Network* 3 61–70

Henderson J M, Weeks P A and Hollingworth A 1999 Effects of semantic consistency on eye movements during scene viewing *J. Exp. Psychol. Hum. Percept. Perform.* 25 210–28

Hinkle D A and Connor C E 2002 Three-dimensional orientation tuning in macaque area V4 *Nat. Neurosci.* 5 665–81

Hubel D H and Wiesel T N 1968 Receptive fields and functional architecture of monkey striate cortex J. Physiol. (Lond.) **195** 215–43

Jepson A, Richards W and Knill D 1996 Modal structures and reliable inference *Perception as Bayesian Inference* ed D Knill and W Richards (Cambridge: Cambridge University Press) pp 63–92

Liu Y and Shouval H 1994 Localized principal components of natural images—an analytic solution Network: Comput. Neural Syst. 5 317–25

Logothetis N K, Pauls J, Bulthoff H H and Poggio T 1995 Shape representation in the inferior temporal cortex of macaque *Curr. Biol.* 5 552–63

Marr D 1982 *Vision* (San Francisco, CA: Freeman) Oliva A and Schyns P G 1997 Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli *Cogn. Psychol.* 34 72–107

Oliva A and Schyns P G 2000 Diagnostic color blobs mediate scene recognition *Cogn. Psychol.* 41 176–210

Oliva A and Torralba A 2001 Modeling the shape of the scene: a holistic representation of the spatial envelope Int. J. Comput. Vis. 42 145–75

Oliva A and Torralba A 2002 Scene-centered representation from spatial envelope descriptors Proc. Biologically Motivated Computer Vision (Springer Lecture Notes in Computer Science vol 2525) ed H H Bulthoff *et al* (Berlin: Springer) pp 263–72

Oliva A, Torralba A, Guerin-Dugue A and Herault J 1999 Global semantic classification using power spectrum templates *Proc. Challenge of Image Retrieval (Electronic Workshops in Computing Series)* (Newcastle: Springer)

Olshausen B A and Field D J 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9

Olshausen B A, Sallee P and Lewicki M S 2001 Learning sparse image codes using a wavelet pyramid architecture Adv. Neural Inform. Process. Syst. 12 887–93

Potter M C 1975 Meaning in visual search *J. Exp. Psychol.* 2 509–22

Ripley B D 1996 *Pattern Recognition and Neural Networks* (Cambridge: Cambridge University Press)

Rosch E, Mervis C B, Gray W D, Johnson D M and Boyes-Braem P 1976 Basic objects in natural categories Cogn. Psychol. 8 382–439

Rousselet G A, Fabre-Thorpe M and Thorpe S J 2002 Parallel processing in high-level categorization of natural images Nat. Neurosci. 5 629–30

Ruderman D L 1994 The statistics of natural images *Network* 5 517–48

Ruderman D L 1997 Origins of scaling in natural images *Vis. Res.* 37 3385–98

Schiele B and Crowley J L 2000 Recognition without correspondence using multidimensional receptive field histograms *Int. J. Comput. Vis.* 36 31–50

Simoncelli E P and Olshausen B A 2001 Natural image statistics and neural representation *Annu. Rev. Neurosci.* 24 1193–216

Sirovich L and Kirby M 1987 Low-dimensional procedure for the characterization of human faces *J. Opt. Soc. Am.* 4 519–24

Swets D L and Weng J J 1996 Using discriminant eigenfeatures for image retrieval IEEE Trans. Pattern Anal. Mach. Intell. 18 831–6

Switkes E, Mayer M J and Sloan J A 1978 Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis *Vis. Res.* 18 1393–9

Szummer M and Picard R W 1998 Indoor–outdoor image classification 1998 IEEE Int. Workshop on Content-Based Access of Image and Video Databases (New York: IEEE) pp 42–51

Tanaka K 1993 Neuronal mechanisms of object recognition *Science* **262** 685–8

Thorpe S, Fize D and Marlot C 1992 *Nature* **381** 520–2

Tolhurst D J, Tadmor Y and Tang C 1992 The amplitude spectra of natural images *Ophthalmic Physiol. Opt.* 12 229–32

Torralba A 2002 Contextual modulation of target saliency *Advances in Neural Information Processing Systems* vol 14, ed T G Dietterich, S Becker and Z Ghahramani (Cambridge, MA: MIT Press)

Torralba A 2003 Contextual priming for object detection *Int. J. Comput. Vis.* 53 169–91

Torralba A and Oliva A 2002 Depth estimation from image structure *IEEE Trans. Pattern Anal. Mach. Intell.* 24 1226–38

Torralba A and Sinha P 2001 Statistical context priming for object detection Proc. Int. Conf. on Computer Vision (Vancouver) (Los Alamitos, CA: IEEE Computer Society Press) pp 763–70

Turk M and Pentland A 1991 Eigenfaces for recognition *J. Cogn. Neurosci.* 3 71–86 Tversky B and Hemenway K 1983 Categories of environmental scenes *Cogn. Psychol.* 15 121–49

Ullman S, Vidal-Naquet M and Sali E 2002 Visual features of intermediate complexity and their use in classification Nat. Neurosci. 5 682–7

VailayaA, FigueiredoM, JainA andZhangH-J 1999Content-basedhierarchical classification ofvacationimages Proc. IEEE Multimedia Systems'99 (ICMCS'99, Proc. Int. Conf. on Multimedia, Computing and Systems, Florence, June 1999)

Vailaya A, Jain A and Zhang H-J 1998 On image classification: city images versus landscapes *Pattern Recognit.* 31 1921–36

van der Schaaf A and van Hateren J H 1996 Modeling of the power spectra of natural images: statistics and information Vis. Res. 36 2759–70

van Hateren J H and van der Schaaf A 1998 Independent components filters of natural images compared with simple cells in the primary visual cortex *Proc. R. Soc.* B **265** 359–66

Van Rullen R and Thorpe S J 2001 Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex *Neural Comput.* 13 1255–83

Vinje W E and Gallant J L 2000 Sparse coding and decorrelation in primary visual cortex during natural vision Science 297 1273–6

Affiliation:

Antonio Torralba[3]
E-mail: torralba@ai.mit.edu

Aude Oliva[4]
E-mail: aoliva@msu.edu

翻译：林绪虹[5]
E-mail: linxuhong@yahoo.com

Silkman Statistical Journal http://cookwhy.com/
published by the Funny Project of Silkman Press http://cookwhy.com
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

[3]Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA (email: torralba@ai.mit.edu).

[4]Department of Psychology and Cognitive Science Program, Michigan State University, East Lansing, MI 48824, USA (email: aoliva@msu.edu).

[5]软件工程师，数学史爱好者，本文基于原作者 2003 年发表于 NETWORK: COMPUTATION IN NEURAL SYSTEMS 的原文翻译，翻译时间 2024 年 4 月）