显著性检测

朱亚菲

2015 年 2 月

目录

1. 引言

由论文“A closer look at context: from coxes to the contextual emergence of object saliency”知道

2. 多尺度的概念

多尺度

3. Hierarchical Saliency Detection

这篇论文主要解决的是当图像中显著前景或背景中存在小尺度大对比度 patterns，而在生成的显著图中并不突出这些 patterns 的情况。论文框架如图 1。主要步骤是三步：首先从原图像中提取 layers，然后从每个 layer 中计算 saliency cues，最后把它们融入一个分层模型以得到最终的结果。

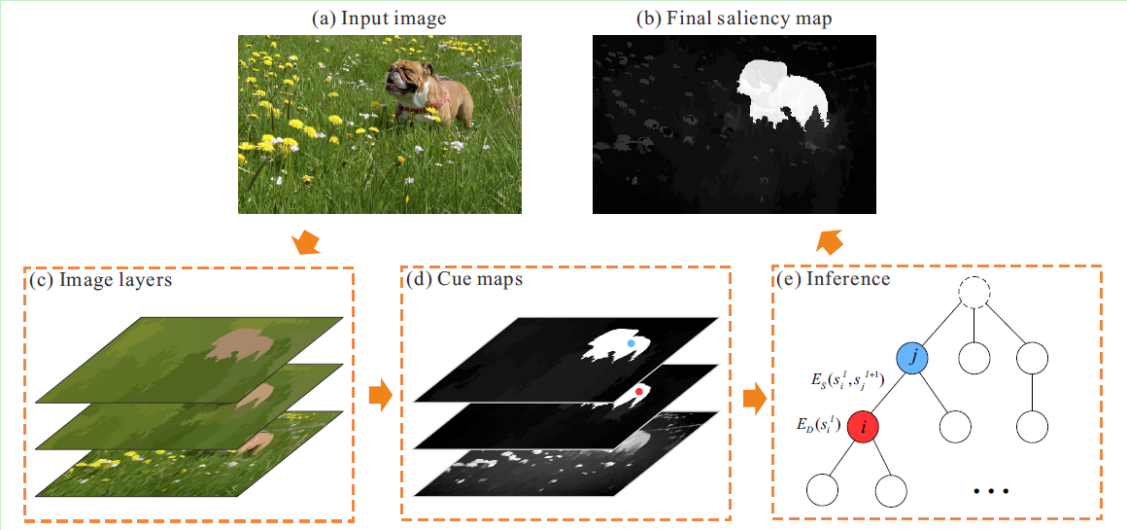

图 1: 框架

分层与多尺度、多分辨率的区别?

1. 如何提取这三个 layers? 如图 2

图 2: 不同尺度下的区域合并结果

先对原图像 (400×300) 用 watershed-like 方法 [1] 进行初始的过分割，对每个分割区域计算一个 scale 值，然后对所有区域的 scale 值按从小到大排序，如果一个区域的 scale 值小于 3，就将它和最近的区域合并 (通过判断两个区域内 CIELUV 颜色均值的距离)，然后更新它的 scale，并更新合并区域的颜色均值，等对所有区域都处理后，得到的结果就是 L^1 层。L^2 层是通过对 L^1 层采取同样的步骤，只不过用一个更大的阈值 17。L^3 层也是如此，阈值取 33。

2. 如何求每个区域的 scale?

通常在 Mean shift、graph-based segmentation 等超像素分割方法中，区域的 size 是指该区域内所有像素的个数。本论文指出了这样的不合理性，就人类视觉感知而言，较多的像素个数和大尺度的区

域并不完全符合。如图 3，尽管弯曲的区域 a 包含了很多像素，但对于我们的视觉感受却并不觉得它很大，而 b 看着会更大一些，尽管它的像素个数并不是很多。根据这样的现象，作者基于 shape uniformities 定义了一个新的 encompassment scale measure，以用来在合并阶段获取区域的 size。

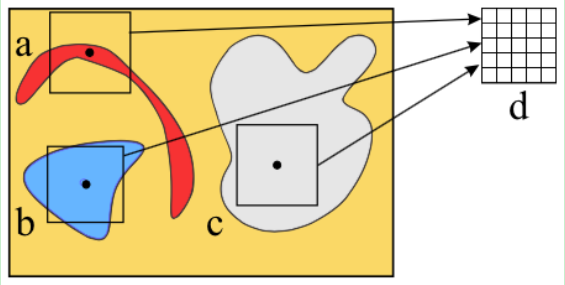

图 3: scale

关于 scale 的定义如下：

$$scale(R) == arg\max_{i} R_{t\times t} | R_{t\times t} \subseteq R \tag{1}$$

其中，$R_{t\times t}$ 是一个 $t \times t$ 的正方形区域。也就是说，一个区域的 scale 是指该区域内所能包含的最大方形区域的边长。这里并不需要通过复杂的计算来算出每个区域的 scale 是多大，只要判断其相对于阈值是大还是小，这样就简化了，可以对每个区域用一个 $t \times t$ 的模板进行滤波，如果滤完后该区域内所有像素值都被更新了，说明该区域的 scale 小于 t，反之说明大于 t，如图 4所示。

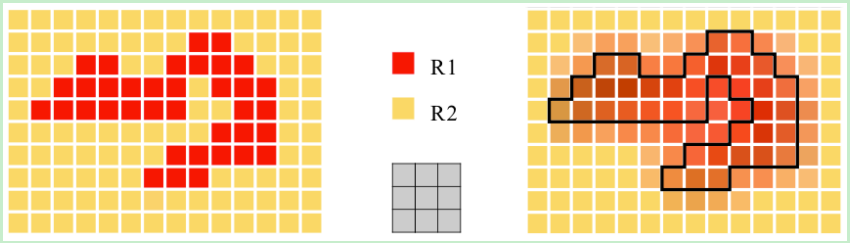

图 4: scale

3. 如何计算每一层的 saliency cues?

主要从颜色、位置、大小三个方面提取 saliency cues，以找到该层比较重要的 pixels，作者用了两种 cues：

1）local contrast

$$C_i = \sum_{j=1}^{n} w(R_i)\Phi(i,j)||c_i - c_j||_2 \tag{2}$$

其中 c_i 和 c_j 分别表示区域 R_i 和 R_j 中的颜色,$w(R_j)$ 指 R_j 中像素的个数。$\Phi(i,j) = exp\{-D(R_i,R_j)/\sigma^2\}$，控制了区域 R_i 和 R_j 之间的空间距离，其中 $D(R_i,R_j)$ 是区域 R_i 和 R_j 的中心的欧几里得距离的平方。

2）location heuristic

心理物理学方面的研究表示人类视觉注意偏好图像的中央区域，所以在通常情况下越靠近图像中央的像素越显著。

$$H_i = \frac{1}{w(R_i)} \sum_{x_i \in R_i} exp\{-\lambda ||x_i - x_c||^2\} \tag{3}$$

其中 $\{x_0, x_1 \ldots\}$ 是区域 R_i 中的像素坐标的集合，x_c 是图像中心坐标。

然后将 C_i 与 H_i 组合起来，得到

$$\bar{s}_i = C_i \cdot H_i \tag{4}$$

由于 local contrast 和 location cues 都被归一化到 $[0,1)$，它们各自的重要性由 λ 来控制。当对三个 layers 均计算完 $\bar{s}_i$ 后，就可以得到每一层的初始显著图，最后通过一种 hierarchical inference procedure 来对多尺度显著性检测结果进行融合。

4. 最后一步 Hierarchical Inference 是怎么进行的?

Cue maps 显示了不同尺度下的显著性，效果很不一样。在底层，会产生很多小区域，而在高层会包含大尺度的结构。由于图像的多样性，单独的一层并不能保证效果是完美的，也很难判别哪一层是效果最好的。

由于背景或前景的复杂性，单纯通过求这三层产生的显著图的平均值来融合并不是一个好的选择。作者构造了一个基于树结构的图，见图 1中的 (e)，其中的节点代表相应层中的区域。节点 j 在下一层中包含两个分割区域，因而有两个子节点。其中父节点代表整幅图像的最粗糙表示。

将图中对应于第 L_l 层中第 i 个区域的节点上的显著性定义为变量 s_i^l，设 S 是包含图中所有节点的集合。最小化如下的能量方程:

$$E(S) = \sum_l \sum_i E_D(s_i^l) + \sum_l \sum_{i, R_i^l \subseteq R_j^{l+1}} E_S(s_i^l, s_j^{l+1}) \tag{5}$$

4. 2014CVPR-LuSong-Learning optimal seeds for diffusion-based salient objectt detection

5. 2013CVPR-MaiLong-Saliency Aggregation: A Data-driven Approach

方法动机：目前在视觉显著性分析方面有很多的方法，每种方法可能对某一些图像或图像中的某些部分适用，能产生较好的结果，却没有哪一种方法能对所有的图像都适用。并且不同的显著性计算方法之间通常是互补的。因而可以尝试对这些显著性计算方法进行组合以获得比使用单一的一种方法好的效果。一般的融合方法都是预先定义好一个组合方程，并且其中每种方法的重要程度都是一样的。而论文中采用的是 CRF aggregation 框架来进行显著性聚合，不仅考虑了各个显著图的贡献，还考虑了相邻像素间的相互作用。由于每种方法在不同图像上的表现也不同，所以对每幅图像使用的方法组合也是不同的。考虑到每幅图像上聚合的依赖性，该方法首先从训练数据集中选择与输入图像近似的图像子集，以在这个子集上而不是在整个训练集上训练这个 CRF aggregation model。

该方法有以下两点优势：

1. 考虑了不同显著性方法的性能 gaps，可以更好地决定每种方法在融合时的贡献大小。

2. 考虑到各个显著性方法在处理不同图像时效果有所不同，该方法能对不同图像自适应选择合适的融合 model。

步骤如下：

对给定的图像 I，首先运行 m 种显著性方法 $\{M_i||1 \le i \le m\}$，以产生 m 幅显著图 $\{S_i||1 \le i \le m\}$。显著图中的每个元素 $S_i(p)$ 代表了像素点 p 处的显著值。显著图中的显著值都被归一化到 $[0, 1]$。输入这 m 幅显著图，输出最终的显著图 S。

6. Saliency Detection via Graph-Based Manifold Ranking

方法动机：以往的显著性检测方法大多是通过计算某像素或区域在局部的上下文中或整幅图像中的对比度来得到该像素或区域上的显著性值，这篇论文则是通过计算图像元素 (像素或区域) 和 foreground cues 以及 background cues 的相似度来得到该图像元素上的显著度，相似度是通过 graph-based manifold ranking 求得的。

background cue：背景通常与图像的四个边界呈现局部或整体上的外观关联性

foreground cue：foreground presents appearance coherence and consistency

算法步骤如图 5，先对图像进行超像素分割，然后将分割后的图像映射为图，每个超像素为图中的节点。第一个阶段是将图像的每一条边界上 (共 4 条) 的节点看作是 labelled background queries，依次算图中的每个节点与这些 queries 的相关性来得到 4 幅 labelled maps，然后对其进行融合得到显著图。在第二阶段，对第一阶段得到的显著图二值化，然后将得到的 labelled foreground 节点看作是 salient queries，最终每个节点的显著度就是通过计算其与 foreground queries 的相关性得到。

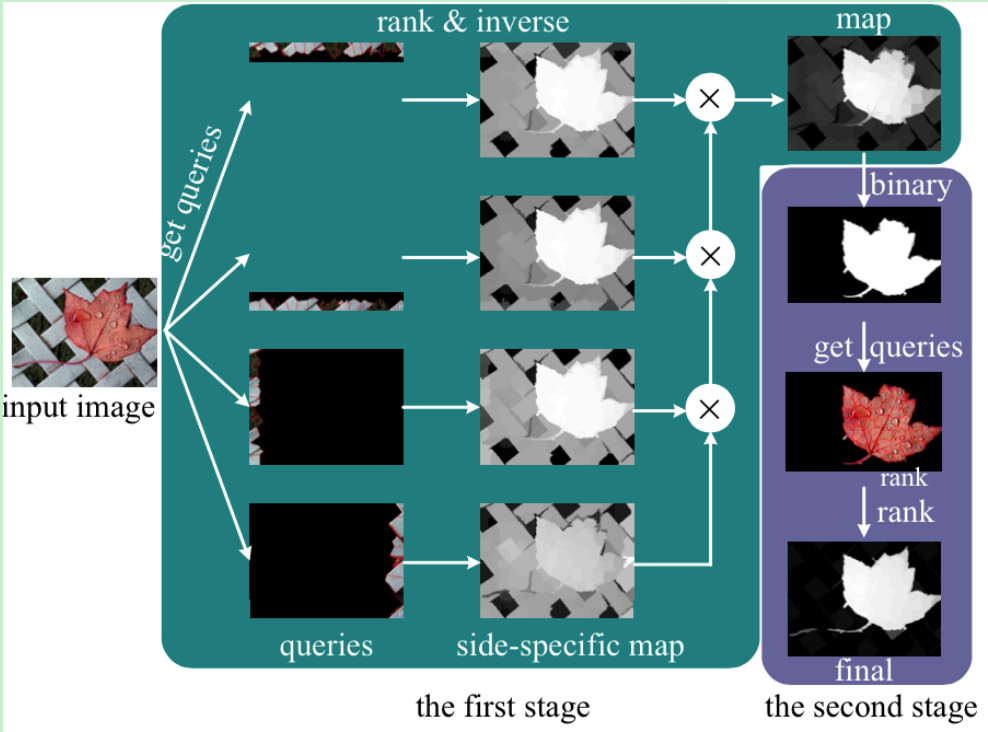

图 5: 算法步骤

第一阶段与第二阶段中的 ranking function 是什么呢?

7. Saliency Optimization from Robust Background Detection

这个论文是基于假设：an image patch is background only when the region it belongs to is heavily connected to the image boundary.

动机：目前虽然已有方法利用了 boundary prior，但一般是基于假设：将图像边缘区域看作是背景。这些方法有两方面的缺点：一是当目标物体稍微接触图像边缘时，效果就会变差。二是它们只是探索性的，并没有告诉我们怎样将其跟其他 saliency cues 相融合。

这篇论文主要有两个方面的贡献：

1. 提出了一种非常鲁棒的 boundary connectivity 方法
2. 提出了融合多种 low level cues 的优化框架

关于 boundary connectivity 的一个示例：

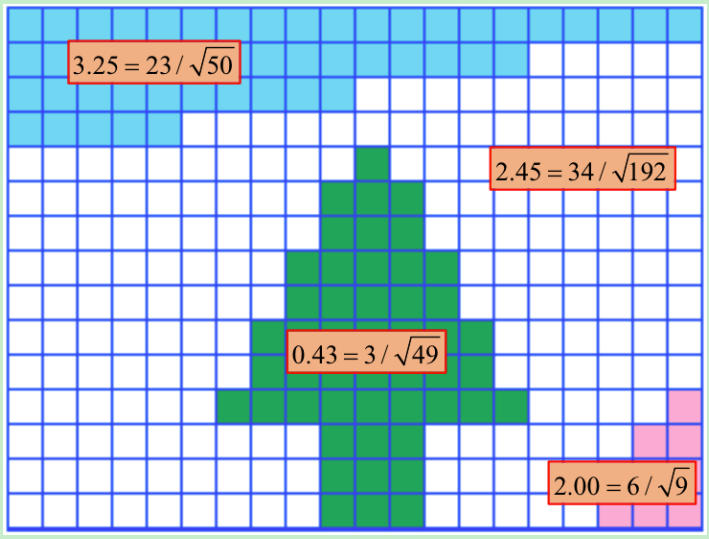

图 6: 关于 boundary connectivity 的一个图例

图 6是一幅人造图，共有四个区域。对于人类的视觉感知来说，图中绿色的区域很显然是显著区域，因为它满足以下条件：大、紧凑、仅仅轻微地接触图像边缘。粉色区域与图像边缘接触也很少，但它的 size 太小了，使得它更像是一个被部分切断了的物体，因而不是显著的。基于此，论文中提出了一种方法来将区域 R 与图像边缘的连接程度量化，叫做 boundary connectivity，定义如下：

$$BndCon(R) = \frac{|\{p|p \in R, p \in Bnd\}|}{\sqrt{|\{p|p \in R\}|}} \tag{6}$$

，其中 Bnd 是处于图像边缘的图像块的集合，p 是图像块。该公式在几何上的解释是：它是一个区域在边缘上的 perimeter 与该区域的整个 perimeter 的比值。这里用的是区域面积的平方根，目的是保持尺度不变性：在不同 resolution 的图像块上的效果保持稳定。如图 6所示，对于背景区域，boundary connectivity 值通常较大，而对目标区域，boundary connectivity 值通常较小。

有了 boundary connectivity 的定义，要计算它的值仍然是比较困难的，因为图像分割本身就是一个复杂的问题。用像图 6中的 hard segmentation 不仅会涉及分割方法选择的难题，而且会在区域边界产生不想要的间断性。

论文中指出精确的 hard image segmentation 是没有必要的，只需要用一种“soft”的分割方法。

论文的基本步骤如下：

1. 用 SLIC 方法将图像先分割成超像素，对一幅 300×400 的图像，取超像素个数为 200 个足够了。

2. 构造无向图，超像素作为节点，相邻的超像素之间连成边，边的权值用两个超像素区域在 CIE-Lab 颜色空间上的颜色均值的欧几里得距离 $d_{app}(p, q)$ 表示。

两个超像素之间的 deodesic distance 被定义为它们在图上最短路径的累计权值和：

$$d_{geo}(p,q) = \min_{p_1=p,p_2,\ldots,p_n=q} \sum_{i=1}^{n-1} d_{app}(p_i, p_{i+1}) \tag{7}$$

为方便体现，定义 $d_{geo}(p,p) = 0$。每个超像素 p 的 spanning area 定义为：

$$Area(p) = \sum_{i=1}^{N} exp(-\frac{d_{geo}^2(p,p_i)}{2\sigma_{clr}^2}) = \sum_{i=1}^{N} S(p,p_i) \tag{8}$$

其中，N 是超像素的个数。可以看到，$S(p,p_i)$ 的范围一直在 $(0,1]$ 之间，并且表征了超像素 p_i 对 p 的贡献大小。当 p_i 和 p 在一个较平坦的区域内，则有 $d_{geo}(p,p_i) = 0$，并且 $S(p,p_i) = 1$，保证 p_i 对区域 p 增加了一个单位区域。而当 p_i 和 p 在不同的区域中时，它们的最短路径中至少有一个边满足 $d_{app}(*,*) \geq 3\sigma_{clr}$)，使得 $S(p,p_i) \approx 0$，保证 p_i 不对 p 的 area 产生影响。通过实验发现参数 σ_{clr} 在 $[5,15]$ 之间时效果是稳定的，在作者试验中取 $\sigma_{clr} = 10$。

同样地，将 length along the boundary 定义为：

$$Len_{bnd}(p) = \sum_{i=1}^{N} S(p,p_i) \cdot \delta(p_i \in Bnd) \tag{9}$$

最后，计算 boundary connectivity

$$BndCon(p) = \frac{Len_{bnd}(p)}{\sqrt{Area(p)}} \tag{10}$$

到这里，只是计算出了某区域的 boundary connectivity 值，还没有求出最终的显著图。以前的显著性方法通常将某区域和周围区域之间的 region contrast 看成一种 saliency cue，计算如下：

$$Ctr(p) = \sum_{i=1}^{N} d_{app}(p,p_i) w_{spa}(p,p_i) \tag{11}$$

其中，$d_{app}(p,p_i)$ 是两个区域在 CIELab 空间上的颜色均值的欧几里得距离，$w_{spa}(p,p_i)$ 是区域之间的空间距离，$w_{spa}(p,p_i) = exp(-\frac{d_{spa}^2(p,p_i)}{2\sigma_{spa}^2})$，$d_{spa}(p,p_i)$ 是超像素 p 和 p_i 的中心间的距离，σ_{spa} 取 0.25。

论文中对上述方程进行了扩展，加入了一项新的权重项背景概率 w_i^{bg}。w_i^{bg} 是从超像素的 boundary connectivity 值映射来的。当 boundary connectivity 值较大时，w_i^{bg} 接近 1，当 boundary connectivity

值较小时，w_i^{bg} 接近。定义如下：

$$w_i^{bg} = 1 - exp(-\frac{BndCon^2(p_i)}{2\sigma_{bndCon}^2}) \tag{12}$$

这里令 $\sigma_{bndCon} = 1$。

于是扩展后的 contrast 公式如下：

$$wCtr(p) = \sum_{i=1}^{N} d_{app}(p, p_I) w_{spa}(p, p_i) w_i^{bg} \tag{13}$$

当 p 是背景区域时，$wCtr(p)$ 是与其相邻的目标区域的背景关联度的加权，值较小；当 p 是目标区域时，$wCtr(p)$ 是与其相邻的背景区域的背景关联度的加权，值较大。满足对显著性的定义。

全局优化：定义 cost function 如下

$$\sum_{i=1}^{N} w_i^{bg} s_i^2 + \sum_{i=1}^{N} s_i^{fg} (s_i - 1)^2 + \sum_{i,j} w_{ij} (s_i - s_j)^2 \tag{14}$$

其中第一项是背景约束，当某区域的 w_i^{bg} 较大时，是背景的概率较大，方程的第一项占的比重较多，为使整个方程值最小，需使 s_i 近似为 0，也就是使该区域的显著性近似为 0。第二项是前景约束，当某区域的 w_i^{fg} 较大时，第二项占的比重较大，需使 s_i 近似为 1。w_i^{fg} 可以通过目前已有的一些显著性方法或它们的组合来计算。第三项是平滑约束，保证显著值的连续性，w_{ij} 定义如下：

$$w_{ij} = exp(-\frac{d_{app}^2(p_i, p_j)}{2\sigma_{clr}^2}) + \mu \tag{15}$$

超像素 p_i 与 p_j 越相似，$d_{app}(p_i, p_j)$ 越小，w_{ij} 越大，越需要 s_i，s_j 近似相等。

8. Graph-Regularized Saliency Detection With Convex-Hull-Based Center Prior

简称 PBS(the proposed prior based saliency) [2]。

文章动机：Center prior 虽然被广泛用到了显著性检测方法中，但是它还有一定的缺陷。比如在很多图像中，显著物体可能会出现在偏离图像中心的位置，这就使得在用 center prior 时会错误地抑制掉离图像中心很远的目标区域而将靠近图像图像中心的背景区域高亮显示。另外之前的一些方法在计算每个像素点或区域的显著值时是各自独立的，而忽略了相似像素点上显著性值应该近似的现象。基

于这两个方面，文章中引入了 convex-hull-based center prior 和 smoothness prior。

算法步骤：

1. 先将给定图像用 SLIC 方法分割成超像素。

2. 计算初始显著图。先计算了 spatially weighted contrast

$$S_{co}(i) = \sum_{j \neq i} ||c_i - c_j|| \cdot exp\left(-\frac{||p_i - p_j||^2}{2\sigma_p^2}\right) \tag{16}$$

然后又计算了 convex-hull-based center prior map，即首先估计显著目标的中心位置 (x_0, y_0)，再计算每个超像素 i 的显著性：

$$S_{ce}(i) = exp\left(-\frac{||x_i - x_0||^2}{2\sigma_x^2} - \frac{||y_i - y_0||^2}{2\sigma_y^2}\right) \tag{17}$$

最后将两者相乘得到初始显著图

$$S_{in}(i) = S_{co}(i) \times S_{ce}(i) \tag{18}$$

使用 convex-hull-based center prior 后虽然会改善当显著目标偏离图像中心的情况，但仍然存在以下局限性：1. convex-hull-based center prior 的效果取决于所求 convex hull 的精确度。2. 由于取 σ_x 和 σ_y 相等并且固定，所以当显著目标不是对称的并且大小是随图像不同而变化的时候，会导致初始显著图的不准确性。

引入 smoothness prior 可以解决这一问题，得到更精细的显著图。平滑约束通常是用在基于图的目标分割中，目的是使图像中相邻的像素拥有同样的 label 值。论文中先将图像分割成超像素，然后将其映射为图，每个超像素对应图中的节点，有共同的边界的超像素之间有一条边，边上的权值为 $w_{ij} \in W$：

$$w_{ij} = exp\left(-\frac{||c_i - c_j||}{2\sigma_w^2}\right) \tag{19}$$

其中 c_i 和 c_j 是 CIELab 空间超像素区域内像素的颜色均值。可以看到关系矩阵 W 是一个稀疏矩阵。定义如下 saliency cost function 来表示这种 smoothness prior：

$$E(S) = \sum_i (S(i) - S_{in}(i))^2 + \lambda \sum_{i,j} w_{ij}(S(i) - S(j))^2 \tag{20}$$

$S(i)$ 和 $S(i)$ 分别表示节点 i 和 j 的所要求的显著值，$S_{in}(i)$ 是节点 i 的初始显著值，λ 是规范化系数。其中等式右边第一项是 fitting constraint，表示一幅好的显著图与初始显著图之间不会变化太多。第二项是 smoothness constraint，一幅好的显著图上相邻超像素的显著值不会相差太多。超像素上的最优显著值是通过最小化该 cost function 来计算。令该方程关于 S 的导数为 0 可得

$$S^* = \mu(D - W + \mu I)^{-1} S_{in} \tag{21}$$

其中 D 是三角矩阵，并且有 $d_{ii} = \sum_j (w_{ij})$，$\mu = 1/(2\lambda)$。

9. Saliency Detection via Absorbing Markov Chain

马尔可夫链，因安德烈 · 马尔可夫 (A.A.Markov, 1856-1922) 得名，是数学中具有马尔可夫性质的离散事件随机过程。该过程中，在给定当前知识或信息的情况下，只有当前的状态用来预测将来，过去 (即当前以前的历史状态) 对于预测将来 (即当前以后的未来状态) 是无关的。

$X_1, X_2, X_3 \ldots$ 马尔可夫链 (Markov Chain)，描述了一种状态序列，其每个状态值取决于前面有限个状态。马尔可夫链是具有马尔可夫性质的随机变量的一个数列。这些变量的范围，即它们所有可能取值的集合，被称为“状态空间”，而 X_n 的值则是在时间 n 的状态。如果 X_{n+1} 对于过去状态的条件概率分布仅是 X_n 的一个函数，则

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n) \tag{22}$$

这里 x 为过程中的某个状态。上面这个恒等式可以被看作是马尔可夫性质。

在马尔可夫链的每一步，系统根据概率分布，可以从一个状态变到另一个状态, 也可以保持当前状态。状态的改变叫做过渡，与不同的状态改变相关的概率叫做过渡概率。随机游走就是马尔可夫链的例子。随机游走中每一步的状态是在图形中的点，每一步可以移动到任何一个相邻的点，在这里移动到每一个点的概率都是相同的 (无论之前漫步路径是如何的)。

在马尔可夫链中，称 $P_{ij} = 1$ 的状态为吸收状态。如果一个马尔科夫链中至少包含一个吸收状态，并且从每一个非吸收状态出发，都可以到达某个吸收状态，那么这个马尔可夫链称为吸收马尔可夫链 (Absorbing Markov Chains)。如图 7，这是一个醉汉游走模型，当醉汉处于位置 1、2 或者 3 时，他将会以等概率 (1/2) 向左或者向右走，他一直走，直到他到达位置 0(他的家) 或者位置 4(酒吧) 才停止游走。

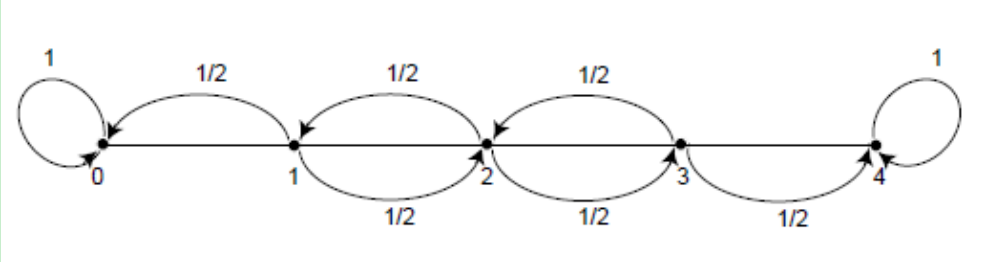

图 7: Drunkard's Walk

参考文献

[1] Rafael C Gonzalez. *Digital image processing*. Pearson Education India, 2009.

[2] Chuan Yang, Lihe Zhang, and Huchuan Lu. Graph-regularized saliency detection with convex-hull-based center prior. *Signal Processing Letters, IEEE*, 20(7):637–640, 2013.