# 网页美感，眼动一致性和眼动实验数据的新可视化指标

>>>>>>>>>>颜色标注说明

红色标出的是有待补充或考证的

图片image 外部文件为jpg格式

图表graph 外部文件为fig格式

表格table 外部文件为excel

公式formula 暂无

<<<<<<<<<<<<<<

## - introduction

眼动检测技术在今天已经不算什么新鲜的话题了，在各类平面、网页乃至产品的设计中，它已经有了相当长一段时间的应用。很大一部分的眼动研究旨在评估图像上的目标区域对用户的注意力的吸引能力，从而确定一个设计的有效性，商业价值等。也有一部分的研究聚焦在行为学的研究上。

现今用于分析或是可视化眼动实验数据结果的指标有如下：

heatmap：一种评估画面上不同区域所受到眼动聚焦强度的二维图像；

【1-1热图例图】

bee swarm，一段沿时间播放的表现每个时刻点每个用户所注视的位置的视频；【1-2 bee swarm例图】

gaze plot，一种把所有用户的眼动转移路径通过圆点和直线标注出来的复杂图像；

【1-3 gazeplot例图】

AOI cluster，一种对用户的眼动数据进行空间上的聚类，从而在二维图像上标注出若干个视觉重点区块的技术。

【1-4 AOI热图】

针对上述的可视化，尤其是AOI，又有相当多的指标来评估眼动的结果。以一个聚类得到的AOI为例，常见的指标有用户平均首次聚焦到AOI的时刻，在AOI上的平均停留时长，最长单次停留时长，首次停留时长，注视次数等，在此就不一一展开了。

不难发现，这些常见的眼动指标大都旨在通过不同的维度表现眼动的一种“强度”属性，可能是沿时间来展开，如首达时间、持续时间，可能是沿空间来展开如热图，AOI。我们的实验旨在研究一群用户的眼动行为，尤其是这些行为的一致性，与美感之间的联系。现有的可视化和指标在这方面，尤其是表现一致性方面，可能略显力不从心。因而本文中，我们将一定程度上基于现有的眼动指标，探讨优化和改进他们的方法，并提出一些新的指标，来考察他们与网页美感的联系，从而试图探寻人类眼动行为和美感之间的关联性。

## - related works

## - hypothesis：

一，猜想：一个公认好看的网页应该具有对视觉的较强的引导作用，引导被试以较小的眼动路径选择的代价来获取较多的页面信息量，从而在实验数据上，使得浏览者各自的浏览轨迹之间表现出更强的一致性（相对于较差的页面）。

\*需要注意的是，我们这里所讨论的一致性指的是个体之间的一致性而非单个个体多次浏览间的一致性。 后者由于存在印象残留等干扰因素，在本实验中不予探讨。

为了评估这种一致性，我们提出如下的眼动熵的概念。

二，概念： 眼动熵。现有的对于熵的定义包括热力学和信息论，概念上，熵表达一个系统的有序度、混乱度、或者反过来，纯度，一致度。

与此对应的，当我们把熵的概念应用在一个群体的被试者对同一个网页对象的观察所产生的一系列眼动行为上时，眼动熵的概念表达为这些被试者的眼动行为在时空上的一致性。越大的熵代表越弱的一致性，越小的熵代表越强的一致性。

在此，我们试图以基于概率散布的香侬信息熵为基础，来计算这种眼动的一致性。下文中我们将通过充分的实验数据分析，尝试定义能反映上述视觉引导作用的眼动熵的合理表达，分析其与美感的关系和其美学理论的合理性。

## - experiment：

一，网页的收集：由于眼动实验的数据规模难以做到很大，我们挑选已经具有先验评价的网页以增强代表性。我们对一个评选优秀网页设计的网站（\*）和一个评选丑陋网页设计的网站（www.websitesfromhell.com），从第一页开始，挑选基数序号的网页。在此过程中，我们滤除了包含具有非拉丁字母字符和具有高识别度的商标、符号或人脸的网站以减少非形式层面的干扰因素。如此直到获得来自上述两个网站的各20张网页。对这些网页我们统一在1280\*800的分辨率（\*一方面我们使用的眼动仪Tobii T50是1280\*1024的，一方面为了照顾到被试常用的横屏， 我们选择了这个分辨率作为实验分辨率）下截屏为图片（包含了使用的浏览器窗口），用作实验用的刺激物。这些刺激物（stimuli）如下。

【image 4 - 1】好的所有网页

【image 4 - 2】差的所有网页

二，被试的邀请：一共有30名被试（13男，17女）参与了眼动的实验，这些人全部都是来自上海交通大学的学生或老师。他们大都来自上海交通大学媒体与设计学院。他们来自中国的各地以及韩国。

三，实验环境：实验在上海交通大学的IXD交互设计实验室内进行。实验过程中室内没有除了实验操作者和被试者之外的人员以及噪声的干扰，房间的窗帘被拉上以避免不可控的光源和反光因素，实验用的眼动仪被布置在一面纯净的白墙前，以减少背景对用户注意力的分散。

四，实验设备：我们采用的实验设备为一台Tobii T50眼动仪，配合Tobii Studio 9软件使用。该眼动仪的分辨率为1280\*1024，如之前所描述，我们的实验页面采用1280\*800的尺寸。

五，眼动实验：对于每个被试，先进行的是眼动实验。所有被试都被保证是第一次浏览实验的网页，并被告知以平时放松地在网上闲逛的心态去浏览这些页面。为了减少疲劳的干扰，40张实验页面被随机（对于每个被试，这样的分组都是随机产生的）分为等量的两组，分前后两次进行实验，期间有半分钟的休息时间。每次，眼动仪会先自动播放3张Dummy页面以使被试进入实验状态，其后是以随机顺序出现的20张的实验页面，每张页面显示3秒钟，之后会有1秒钟的黑屏休息。这样每次实验大概进行1分半。

整个实验过程中，被试对鼠标是具有控制的，但由于所浏览的网页本质上是一张截图，滚动和交互是不被允许的。

较短的曝光时间（3秒）足够使我们考察早期的下意识的眼动行为同时避免了过长的实验时间带来的疲劳。

六，评分实验：为了得到较为准确的评分同时减少评分过程对眼动的干扰，评分实验被安排在眼动实验之后分开进行。在评分实验中，被试得以再次浏览一遍刚才的40张实验网站，这次没有时长限制，被试需要对每张页面给出“好看”或是“不好看”的二值化评分。

七，奖励：每个被试在完成实验后会得到20元RMB的奖励。

## - raw data collection & format：

1，评分数据的产生：作为之后评估指标与美感关联性的一个基准，每个网页被赋予class和score两个值来评价其美感的得分：

评分(score)：令“不好看”为得0分，“好看”为得1分，得到如下的每个网页的评分。代表着每个网页被被试评为好看的比例。

【表5-1\_score\_category\_sorted】见excel

【图5-1】排序网页图片

类别class：对score值>=0.5的页面，我们将它归入“好看”网页类别，剩余的归入“不好看”网页的类别。归类的结果如下，比较实验开始前的先验分类结果，可以看到是完全一致的。也就是说，参与实验的40张网页是具有公认的代表性的“好看的”或“不好看”的网页。

2，眼动数据格式：眼动数据的收集原理在眼动仪和相关软件内部发生：眼动仪以一定的频率（对于我们的设备T50而言，这个频率是50Hz）采集用户聚焦的平面坐标，并通过插值得到每个时刻的眼动速率的值，并通过一定的阈值得到基于注视的眼动数据。

基于注视的眼动数据是一般能通过眼动仪直接获得的最原始的数据。它是由一系列的注视构成的，每个注视包含了四个参数：注视开始时间，注视持续时间，注视横坐标，注视纵坐标。通俗而言，记录了一个注视发生的时间和地点。下文中我们的一切分析都仅来自于这些基于注视的原始眼动数据。

【表5-2一个被试对一个页面的眼动注视数据举例】

|  |  |  |  |
| --- | --- | --- | --- |
| start\_time | duration | x | y |
| 24224 | 700 | 483 | 341 |
| 24924 | 400 | 173 | 356 |
| 25324 | 350 | 58 | 351 |
| 25673 | 200 | 696 | 462 |
| 25873 | 150 | 738 | 483 |
| 26023 | 217 | 928 | 442 |
| 26240 | 300 | 561 | 239 |
| 26539 | 683 | 638 | 163 |

## - analysis

一，指标的评价手段

在开始前，我们首先需要约定初步评估一个计算指标与美感的相关性强弱的手段。这里我们采用分别基于score和class的两种方式来评估：

1. pearson系数：通过计算代表每个网页美感评分的score与指标的线性相关系数，可以快速获得一个指标是否与美感存在潜在的联系。

计算公式如下：

【公式】

pearson系数的绝对值越接近1代表指标与美感的关联越显著，反之越接近0则越不显著。

1. 分类准确率：通过指标对网页样本进行分类，可以统计一系列反映分类优度的指标，本实验中由于好坏网页的个数相等，不存在数据扭曲的情况，故直接统计分类的准确率即可。

二，眼动熵的定义——信息熵及分布估计算法

我们首先给出如下的基于信息熵的眼动熵的定义：

定义：对于一个固定分辨率的画面（本实验中1280\*800），以画面中的全体像素点构成的集合作为概率空间（即每个像素点都作为眼动注视的一种可能的选择），则在此概率空间上的眼动熵定义为，entropy = sum(-p(x, y)\*log(p(x,y))) (for all (x, y) s.t. p(x, y) != 0)

理论上，眼动熵反映被试之间的眼动行为在空间分布上的一致性或是集聚性， 当所有注视集中在一个像素上的时候它取得最小的指，当注视分散在不同空间位置的时候它取得最大的值。需要注意的是，该值不受具体的空间分布状况无关而只在乎被试的注视之间的聚拢程度。

该指标中的p是眼动注视的二维离散空间的概率分布, 满足sum(p) = 1，p(x, y)是眼动注视像素点(x, y)的可能性，是待估计确定的。下面我们讨论几种获得p的估计的方法，并逐一评估基于他们的熵与美感的关联性。为了有一个直观的大体的认识，在之后的指标解释中我们用如下的实验网页页面来帮助解释：

【image 6-1】举例的页面

一）传统热图：一个能够直接想到的，也是现有的能够直观反映注视在画面上的空间分布的可视化方式是热图：

热图是将整个眼动记录时间内的所有被试的眼动注视（fixation）以其注视时长为强度，以其注视坐标为位置叠加的一系列高斯核组成的图像。具体地，对于一个选定的标准差r，一个持续时长为t，坐标为(x0, y0)的注视会叠加的高斯核的表达式为-1/2/r/r\*e^(-((x-x0)^2+(y-y0)^2)/2/r/r)。 如此叠加所有被试的每一个的注视之后对得到的热图进行归一化处理，就得到了可以用于计算眼动熵的概率分布p。

【image6-2】传统热图举例，取不同r（r = 20， 60， 100）

r值（即高斯核的标准差）的选取应该是具有一定的理论合理性的的。它应该一定程度上反映人类眼动的聚焦范围（黄斑角度）或是反映眼动仪的记录误差。不过为了讨论的周密性，同时也是实验性，我们对从1到30px的所有r值下得到的基于热图的眼动熵都做了计算。评估他们的pearson系数和分类准确率，得到如下的图像。

【figure 6-1】传统热图的半径迭代

可以发现在整个r值的区间里，基于传统热图的眼动熵都没有表现出任何与美感的关联性。这是令人有点吃惊的。

二）去权热图：由于空间分布熵只包含空间维度的信息，我们尝试对传统的热图做一些改进，在计算热图的阶段只考量每个注视的坐标位置，并给每予个注视相同的权重而不再根据他们的持续时长来决定权重。 这样就得到了我们称之为去权热图的概率分布估计。

【image6-3】去权热图举例，取不同半径（r = 20, 60, 100）

类似的，对于基于去权热图的眼动熵进行关于高斯核标准差r的遍历计算，并评估他们的pearson系数和分类准确率，得到如下的图像。

【figure 6-2】去权热图的半径迭代

虽然只是改变了注视的权重比例，但与美感的结果却产生了很大的差别。在从13px-60px的广泛标准差区间里，眼动熵与美感的相关系数的绝对值都高于0.5，表现除了与美感的稳定而显著的负相关性。

其中相关系数表现最优的标准差等于20px的去权热图的眼动熵数据如下

【table 6-1】标准差等于20px的去权热图的眼动熵数据

其盒图和二维散点图如下：

【figure 6-3】r=20的去权热图的盒图

【figure 6-4】r=20的去权热图的散点图

三）对分布估计算法的小结

上述的计算似乎表明与美感相关的眼动空间分布的一致性是一个较为简洁的概念，其依赖的概率分布与单个眼动注视的时长（duration）没有直接关系。在计算一致性的过程中加入duration造成的权重反而会使得眼动熵几乎毫无效果。

三、眼动熵的有效性——浏览者的人数

为了进一步验证眼动熵的与美感之间的关联，同时探究眼动熵的稳定性与被试人数的关系，我们从对被试的人数进行限制。从随机抽取2人、随机抽取3人一直到抽取全部30个被试，并分别计算基于抽出的被试的数据的眼动熵。这里的眼动熵的半径我们仍然取20px。

【figure 6-5】传统热图和去权热图的眼动熵随人数的变化

四、眼动熵的归一化——先验条件的讨论

通过上述基于眼动熵的讨论和实验，在一定程度上可以得出“较好看的网页的用户的眼动注视在画面上的分布更为一致和集中”这样的推论。这看似和我们的猜想是相符的，却同时也会引出一些例如“只有单一极强的视觉重点的页面会是最好看的页面“这样不合理的结论。 显然，仅仅一个纯白色的画面配上中央的一个黑点并不会是最好看的页面，但同时他的眼动熵确实可能会非常低，这是不合理的。

【6-4不合理的网页举例】

会造成上述的不合理推论的一个重要的原因是没有纳入考量一些先验的与信息量相关的因素。不同的网页拥有不同的信息量，把建立在这些的不同基础上的分部一致性直接进行比较是不恰当的。而在我们的实验中，由于样本量较小，各网页的信息量间没有出现极端差异的情况。 下面我们通过纳入眼动注视个数作为先验条件，对眼动熵进行归一化处理，并计算这种归一化的眼动熵与美感的关联性和理论上的合理性。

对于一个页面上的注视总数而言，如果所有的注视都是随机落到画面的随机位置的，那么显然越多的注视个数会导致越高的信息熵。 是否实际上是因为好看的网页相比难看的网页在同样的实验时间内获得了更少的注视从而使得其眼动熵较小呢？实验数据的分析结果表明，恰恰相反！

统计得到的注视个数与网页美感表现出微弱的正相关性，代表着好看的网页在总体上反而更有可能获得更多的注视。

【table 6-1】每个页面的fixation总数的表格

【figure 6-6】总数的盒图

【figure 6-7】总数的散点图

为了考量注视个数的大前提下，眼动熵的值，我们用眼动熵/log(注视个数)，得到一种“归一化的眼动熵“（这里取log是为了与眼动熵的非线性计算方式匹配）。理论上，由于这种归一化的眼动熵排除了眼动注视个数的前提影响，他反映的是纯粹的在同样多的眼动注视下，被试之间注视位置的一致性。如下是该指标在半径值的遍历下得到的与美感的相关系数的变化图像。

【figure 6-8】三种分布算法的归一化眼动熵的半径枚举

可以发现，相比没有归一化的眼动熵，它的指标效果明显更好一些。同时，通过对先验眼动个数的考察，使其在理论上更为完备，更有可能规避前面讨论到的极端情况造成的不合理。

事实上，通过考察一些其他的具有复杂度度量能力的先验条件，也能对眼动熵起到归一化的作用，并使其在理论上自洽，例如以所有眼动注视的在画面上的外界矩形的面积作为先验条件，图像如下。在此就不过多展开了。

【figure6-9】以外接矩形面积归一的半径枚举

五、眼动熵随时间的发展

在一个时间维度上，眼动熵的值是如何发展变化的，其与美感的关联是怎样随时间变化的？计算多长时间的眼动熵能取得与美感的更好的关联性。这部分，我们讨论沿时间发展的眼动熵的一些属性。

给定一个时刻t，归一化的眼动熵

【image6-5时间切面上的眼动分布r=20的去权重热图】

下面从眼动熵自身随时间的变化规律和不同时刻的眼动熵与美感的相关性两个方面展开讨论：

1. 眼动熵与曝光时间的关系

对每一张页面，可以计算得到一条E(A(t))的曲线，把所有页面曲线在一起展示，如下图：

【figure6-10】沿时间切片的所有眼动熵值以及他们在每个时间点的均值。包括好的坏的分别得均值，还有眼动熵理论的最大值和最小值

可以看到，整体上，网页无论好看与否，其眼动熵的是随时间不断增大的。这一定程度上表明，早期的眼动行为是偏向生理性的，本能性的，从而在个体间表现出一定的一致性。而随着时间的发展，个体间的认知、阅读习惯等差异逐渐表现出来，从而使得眼动熵愈来愈高。

二）时间切片的眼动熵与网页美感的关联性

下图给出了不同时刻的世界切片的眼动熵与网页美感的关联性。

【figure 6-11】沿时间的时间切片的眼动熵与美感的关联性

可以初步得出以下的观察结果：

1. 在最早的500ms左右的时间内，眼动行为的一致性与美感几乎没有关联性。
2. 最早的相关性高峰出现在800ms-1000ms左右，大概是人的眼动足以表现出与美感的关联性的时间。
3. 这种关联性在时间范围内并不,平稳，存在着一定幅度的波动。并在一段时间后逐步减小。

上述的观察结果在一定程度表明了人类的眼动行为的一些特征：

1. 在刚被曝光到一张页面时，眼动行为似乎倾向于表现出的一种无差别的快速扫略，
2. 在800ms-1000ms左右，好的网页的视觉引导作用开始表现，使得被试开始聚焦。
3. 进一步的浏览过程中，这种最初的快速扫略和聚焦行为可能会不停地交替进行，从而导致关联性的波动。

在之后的discussion中，将进一步讨论这些观察背后可能的成因以及对人类眼动行为的一些推测。

## - discussion

一、半径意味着什么？与黄斑有没有关系

二、美感与眼动的关联与信息论的关系

三、基于沿时间的眼动熵对人类眼动行为的一些推测。

## - further discussion

眼动实验的新可视化指标

基于眼动熵的对现有眼动数据可视化的探索，如去权热图，布尔饼图和相关眼动熵数据用来配合传统热图表现空间眼动的一致性；

【image 8-1】去权热图的可视化界面

通过时间切片上的眼动熵配合Bee swarm进行眼动一致性沿时间的表现。

【image 8-2】bee swarm配合时间切片的眼动熵可视化界面

## - conclusion