链滴

# 视频语义分割基准数据集与评估方法

作者：vcjmhg

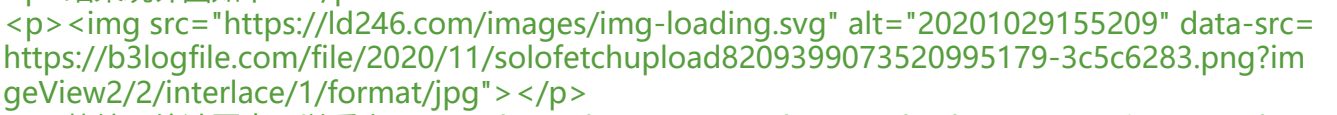
原文链接：https://ld246.com/article/1604236495193

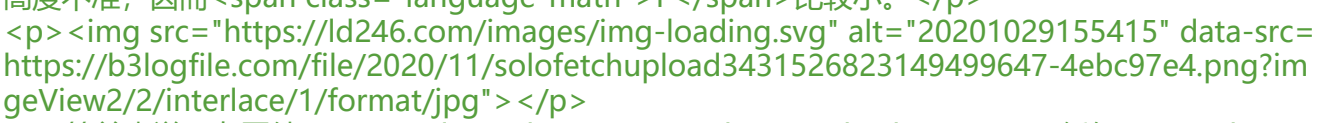
来源网站：链滴

<h2 id="概述">概述</h2>
本文来源于《A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation》，论文主要介绍了一种作者团队提供的针对视频语义分割算法进行评估的基准数据并提供了种指标用于评估算法效果的优劣。本文主要是个人在阅读该论文的的一些所得，但由于论文内容所致本文阅读起来更像一篇说明文档，提供与此，仅供参考。
DAVIS 由 50 个高质量，全高清的视频序列组，包含有多个视频目标分割挑战，如遮挡，运动模和外观变化。每一个视频都是稠密标注，像素级别的精度和逐帧的真值分割（前景对象与背景区域精确像素分离）。同时提出了三种互补的度量标准（区域相似性、轮廓准确性以时间连贯性）来对当时几种最新的的分割方法进行综合分析。
<h2 id="数据集说明">数据集说明</h2>
根据以往的数据集经验，整个数据集重点关注四个关键方面，来创建一个平衡全面的数据集。</>
<h2 id="1--数据的数量和质量">1. 数据的数量和质量</h2>
首先，一个好的数据集要有足量的数据，这是确保内容的多样性并提供一组均匀分布挑战的前提。并且拥有足量的数据对于避免过度拟合和延迟性能起到至关重要的作用，同这在一定程度上也保证数据集具有更长的使用寿命。
另一方面，数据的质量也非常重要， 数据集的质量需能反映现有技术的水平。
基于上边两个因素，构建了 <code>DAVIS数据集</code>，DAVIS 构成包括 50 个列总共 3455 标注帧，视频帧率为 24fps，分率为 1080p。
同时由于当前计算复杂度是视频处理的一个重要瓶颈。因此，为了减少计的复杂程度，DAVIS 中的视频序列采用较短的时间长度（2-4 秒），但是它涵盖了大部分在长视频序中找到的挑战。
<h2 id="2--实验验证">2. 实验验证</h2>
对于视频中的每一帧，提供了像素级别的精度，以二进制掩码的方式手工创建分割结。
每个 DAVIS 数据集细分成分成训练集和测试集两个部分。但在评估的时候，不使用分区,而是考整个数据集，因为大多数的评估方法不需要训练，并且由于计算复杂度，网格搜索最优参数的估计是可行的。
<h2 id="3--对象存在">3. 对象存在</h2>
每个序列都应包含至少一个要与背景区域分开的目标前景对象。
选择不使用多个具有明显运动的不同对象，以便能够公平地将对单个对象进行操作的段方法与对多个对象进行联合分段的方法进行比较。<br>
此外，每个序列只有一个对象，这将消除通过全自动方法执行的检测的歧义（因为检测的目标只有一）。
<h2 id="4--无约束视频挑战">4. 无约束视频挑战</h2>
为了能够对算法的性能进行更深入的分析和理解，识别可能影响算法性能的关键因素和环境是至重要的。因此定义一个扩展的视频属性集合，用于代表特殊情况，如速运动，遮挡和杂乱背景这些典型的视频分割挑战。
具体属性及其含义如下表所示：
<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
<th>描述</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>

<td>Background Clutter. The back- and foreground regions around the object boundaries ha e similar colors (<img src="https://ld246.com/images/img-loading.svg" alt="\chi ^2" data-sr ="https://b3logfile.com/file/2020/11/solofetchupload8189323692753835142-a211a280.gif"> ver histograms).</td>
<td>复杂的背景。在目标边界附近的背景前景区域有相似的颜色。 </td>
</tr>
<tr>
<td>DEF</td>
<td>Deformation. Object undergoes complex, non-rigid deformations.</td>
<td>形变。目标存在复杂的非刚性的形变。 </td>
</tr>
<tr>
<td>MB</td>
<td>Motion Blur. Object has fuzzy boundaries due to fast motion.</td>
<td>运动模糊。由于快速的运动，目标边界模糊。 </td>
</tr>
<tr>
<td>FM</td>
<td>Fast-Motion. The average, per-frame object motion, computed as centroids Euclidean di tance, is larger than<img src="https://ld246.com/images/img-loading.svg" alt="\tau_{fm}" da a-src="https://b3logfile.com/file/2020/11/solofetchupload8189323692753835142-a211a280.g f"> = 20 pixels.</td>
<td>快速运动。目标平均的帧间运动距离大于 20 像素，距离定义为质心的欧几里得距离。 </td>
</tr>
<tr>
<td>LR</td>
<td>Low Resolution. The ratio between the average object bounding-box area and the image area is smaller than tlr = 0.1.</td>
<td>低分辨率（小目标）。平均目标边框区域与图像区域的比值小于 tlr = 0.1。 </td>
</tr>
<tr>
<td>OCC</td>
<td>Occlusion. Object becomes partially or fully occluded.</td>
<td>遮挡。目标部分或全部被遮挡。 </td>
</tr>
<tr>
<td>OV</td>
<td>Out-of-view. Object is partially clipped by the image boundaries.</td>
<td>视野之外。目标被图像边框裁剪了部分，即只有部分目标处于视野中。 </td>
</tr>
<tr>
<td>SV</td>
<td>Scale-Variation. The area ratio among any pair of boundingboxes enclosing the target o ject is smaller than<img src="https://ld246.com/images/img-loading.svg" alt="\tau_{sv}" data src="https://b3logfile.com/file/2020/11/solofetchupload8189323692753835142-a211a280.gif > = 0.5.</td>
<td>尺度变化。存在一对包围目标对象的边界框（两帧），他们的面积比小于 0.5。 </td>
</tr>
<tr>
<td>AC</td>
<td>Appearance Change. Noticeable appearance variation, due to illumination changes and elative camera-object rotation.</td>
<td>外观变化。由光照变化和相对的相机-目标旋转导致的显著外观变化。 </td>
</tr>

<tr>
<td>EA</td>
<td>Edge Ambiguity. Unreliable edge detection. The average groundtruth edge probability ( sing [11]) is smaller than<img src="https://ld246.com/images/img-loading.svg" alt="\tau _{e}" data-src="https://b3logfile.com/file/2020/11/solofetchupload8189323692753835142-a211a20.gif">= 0.5.</td>
<td>边沿模糊。不可靠的边沿检测。平均真值边界的概率小于 0.5.</td>
</tr>
<tr>
<td>CS</td>
<td>Camera-Shake. Footage displays non-negligible vibrations.</td>
<td>相机抖动。画面显示不可忽略的振动。 </td>
</tr>
<tr>
<td>HO</td>
<td>Heterogeneus Object. Object regions have distinct colors.</td>
<td>颜色不均匀的目标。目标区域有不同的颜色。 </td>
</tr>
<tr>
<td>IO</td>
<td>Interacting Objects. The target object is an ensemble of multiple, spatially-connected objects (e.g. mother with stroller).</td>
<td>交互的对象。目标对象是多个空间连接的对象(例如母亲和婴儿车)的集合。 </td>
</tr>
<tr>
<td>DB</td>
<td>Dynamic Background. Background regions move or deform.</td>
<td>动态背景。背景区域移动或者形变。 </td>
</tr>
<tr>
<td>SC</td>
<td>Shape Complexity. The object has complex boundaries such as thin parts and holes.</t>
<td>复杂形状。目标有复杂的边界，比如很细的部分或者洞。 </td>
</tr>
</tbody>
</table>
这些属性并不具备排他性，因此一个视频序列可以被标注多个属性。他们数据集中的分布展示在下图左中，图右显示他们两两之间的依赖关系。
<img src="https://ld246.com/images/img-loading.svg" alt="image-20201029151003915" data-src="https://b3logfile.com/file/2020/11/solofetchupload7226046149100489190-f129e27.png?imageView2/2/interlace/1/format/jpg">
<h2 id="实验验证">实验验证</h2>
在有监督的评估框架中，给定一个特定帧上的标记数据 G 和一个输出的分割结果 M，所有的评指标都是主要为了解决一个问题：即 G 和 M 之间的拟合程度或者说是相似的程度。
因此论文中给了三种评价指标，区域相似性、轮廓准确性以及时间连贯性
<h2 id="1--区域相似度-Region-Similarity--">1. 区域相似度（Region Similarity \mathcal{J}）</h2>
为了测量基于区域的分割相似度，即识别错误像素的数量，此处使用使用 Jaccard 索引!\mathcal{J}。
Jaccard 索引定义如下：
\mathcal{J} = \frac{M\cap G}{M\cup G}
其中M为输出的分割结果，G为真值掩膜（也就是图像的标记结果

。
<h2 id="2--轮廓准确性-Contour-Accuracy---">2. 轮廓准确性（Contour Accuracy <span class="
anguage-math"> \mathcal{F} )）</h2>
从基于轮廓的角度来看，可以将 M 解释为一组限定掩模空间范围的闭合轮廓 c(M)。因此可说出过一个二分匹配来比较 c(M)和 c(G)边缘点的精确度 Pc 和召回率 Rc。进而定了一个F-score来衡量轮廓的整体准确性，其具体定义如：
\mathcal{F} = \frac{2P_{c}R_{e}}{P_{c}+R_e}
<h2 id="3--时间稳定性-Temporal-stability--">3. 时间稳定性（Temporal stability <span class="
anguage-math"> \mathcal{T} ）</h2>
结果的时域稳定性是视频对象分割中的一个相关重要的方面，由于对象形状的演化是识别和抖动一个重要线索，不稳定的边界在视频编辑应用中是不可接受的。
因此，论文引入了一种时间稳定性测量方法来惩罚这种不期望的效果。关键的问题是区分物体的接受的运动和不需要的不稳定性和抖动。
因此估计了在一帧掩码转换到下一帧所需的变形。简单来说，如果转换是平滑和精确，结果可以认为是稳定的。
在形式上，我们将帧 t 的掩膜\mathcal{T}转换为代其轮廓的多边形。 \mathcal{T} 然后，我们使用形状上文描(SCD)[3]述符描述每个点\mathcal{T}。接下来，我将匹配设置为动态时间扭曲(DTW)[39]问题，是我们寻找和 \mathcl{T} \mathcal{T}之间的匹配，它最小化了匹配之间的 SCD 距离，同时保持了点在形状中出现的顺序。
每匹配点的平均成本作为时间稳定性\mathcal{T}的量。直观上，匹配将补偿运动和小的变形，但它不会补偿曲线的振荡和误差，这是我们想要测量的。挡和非常强的变形会被误解为轮廓不稳定，因此在没有这种影响的情况下计算序列子集的测量值。</
>
<h2 id="指标相关性">指标相关性</h2>
结果统计图如下：


从结果统计图中可以看出\mathcal{T}和<span class
"language-math">\mathcal{J} 之间有明显的线性相关。<span class="language-math"
\mathcal{F} 和\mathcal{T} 之间则没有。
<h2 id="指标差异性">指标差异性</h2>
在左边，结果受到 J 的惩罚，因为就像素数量而言，未成功识别的区域头和脚很大，而对于边界量 F，漏掉的百分比更低。在右侧，整个车身都被识别出来了，因此 IoU 是比较大的，但是对应的边高度不准，因而F比较小。


简单来说，左图结果\mathcal{T}低但<span class="
anguage-math">\mathcal{J} 高，右图\mathcal{J} </
an>高但低\mathcal{F}。
<h2 id="结论">结论</h2>
运行时间效率和内存要求是几种视频分割算法的可用性的主要瓶颈。在我们的实验中，我们观察花费大量时间对图像进行预处理以提取边界保留区域，对象建议和运动估计。鼓励未来的研究仔细考那些可能会损害其工作实用性的组件。高效的算法将能够利用此数据集提供的全高清视频和精确的分蒙版。利用高分辨率可能无法在区域相似性方面产生更好的结果，但是改善复杂物体轮廓和微小物体域的分割至关重要。