### Google视频：一种用于视频中对象匹配的文本检索方法

**（对象和场景检测方法，并与数据库对接）**

**概要**

【可以和论文3整合】 **我们描述了一种对象和场景检索的方法**，该方法可搜索并定位视频中用户概述的对象的所有出现情况。该对象由一组视点不变区域描述符表示，因此尽管视点，照明和部分遮挡发生了变化，也可以成功进行识别。镜头中视频的时间连续性用于跟踪区域，以拒绝不稳定的区域并减少描述符中的噪声影响。 与文本检索的类比是在实现中，其中预先计算了描述符的匹配（使用矢量量化），并反转了文件系统和文档使用排名。 结果是检索是即时的，并以Google的方式返回了关键帧/镜头的排名列表。现将该方法在两部完整的film上进行匹配说明。

1. **引言**

这项工作的目的是以Google检索包含特定单词的文本文档（网页）的便捷性、速度和准确性来检索视频中那些包含特定对象的关键帧和镜头。**本文研究是否可以成功地采用文本检索方法进行物体识别。在图像数据库中识别出一个（相同的）对象**。

现在正达到一定的成熟度。这仍然是一个具有挑战性的问题，因为物体的视觉外观可能会因为视点和光线的不同而有很大的不同，而且可能会被部分遮挡，但现在已经有了成功的方法。通常情况下，一个对象由一组重叠区域表示，每个区域由从区域外观计算的向量表示。区域分割和描述符是建立在一个可控程度的视点不变性和图解的基础上的采集条件。

**我们探讨这种类型的识别方法**（对数据库中的所有图像都计算出类似的描述符。识别一个特定的对象是通过描述符向量的最近邻匹配进行的，然后利用局部空间相干性（如邻域、排序或空间相干性）进行模糊识别。布局)，或全局关系(如表极几何)。例子包括[5，6，8，11，13，12，14，16，17]。）**是否可以重构为文本检索。在本质上，这需要一个词的可视化模拟，这里我们通过对描述符向量进行向量量化来提供**。然而，我们会看到，追求与文本检索的类比，不仅仅是对不同向量量化的简单优化。

在文本检索文献中，有许多经验教训和经验法则已经被学习和发展，值得确定的是，这些是否也可以运用到视觉检索中。

这种方法的好处是，匹配是有效的预先计算，因此在运行时，包含任何特定对象的帧和镜头可以用nodelay检索。这意味着，即使在为视频建立描述符时，对这些对象没有明确的兴趣，也可以检索视频中出现的任何对象（以及对象的联合）。然而，我们还必须确定这种向量量化检索是否会错过任何如果使用前一种最近邻匹配方法就会得到的匹配。

回顾文本检索：文本检索系统一般采用一些标准步骤[1]。首先将文档解析成单词。第二，用词干来表示这些词，例如'walk'、'walking'和'walks'就用词干'walk'来表示。第三，使用停止列表来拒绝非常常见的词，如'the'和'an'，这些词在大多数文档中出现，因此对某一特定文档没有辨别能力。其余的词被分配一个唯一的标识符，每个文档用一个向量来表示，其分量由文档所包含的词的出现频率给出。此外，这些分量以不同的方式进行加权（详见第4节），在Google的情况下，一个网页的权重取决于链接到该特定网页的网页数量[3]。以上所有步骤都是在实际检索前进行的，并且将代表语料库中所有文档的向量集组织成一个倒置文件[18]，以方便高效检索。一个倒置文件的结构就像一个理想的图书索引。它为语料库中的每一个词设置了一个条目，后面是该词出现的所有文档（以及该文档中的位置）的列表。

检索文本的方法是计算其词频向量，并返回与该词最接近的文档。(用角度测量)向量。此外，匹配上的 语序分明 返回的文件。

**本论文大纲**：在这里，我们探索每个步骤的视觉类比。 第2节描述了使用的视觉描述符。 然后，第3节将它们的向量量化描述为可视化的“单词”，第4节则对向量模型进行加权和索引。 然后，在第5节中基于一组框架的地面真实性对这些想法进行评估。最后，在第6节中介绍了停止列表和排名（通过空间布局的匹配），并用于评估整个过程中的对象检索。

两部长片：“ Run Lola Run”（劳拉·雷恩特）（Tykwer，1999年）和“ Groundhog Day”（Ramis，1993年）。 尽管以前的工作已经从文本检索文献中借用了从数据库中检索图像的想法（例如[15]使用加权和倒排文件方案），但据我们所知，这是将这些想法首次应用于视频中的对象检索。

1. **Viewpoint invariant description？**
2. **构建视觉词汇表**
3. **实验**
4. **总结**

与文本检索的类比确实证明了其价值：尽管许多帧中的视点发生了重大变化，我们仍可以在整个电影数据库中立即进行运行时对象检索。该对象被指定为图像的子部分，事实证明这足以用于准平面刚性对象。

当然，可以进行一些改进，主要是为了克服视觉处理中的问题。当前排名较低是由于缺少某些场景类型的视觉描述符。但是，框架允许添加其他现有的仿射协变区域（它们将定义扩展的视觉词汇），例如[17]的那些。

另一个改进将是在一个以上的框架上定义感兴趣的对象，以允许在其所有视觉方面进行搜索。

文本检索类比也为将来的工作提出了有趣的问题。在文本检索系统中，文本词汇不是一成不变的，随着新文档被添加到集合中而不断增长。同样，我们并不主张矢量量化对所有图像都是通用的。到目前为止，我们已经了解了足以用于两部电影的矢量量化，

但是将需要找到升级视觉词汇的方法。可以想到学习不同场景类型（例如城市景观与森林）的视觉词汇。

最后，我们现在有一个有趣的可能性，可以追踪文本检索社区的其他成功，例如潜在的语义索引来查找内容，以及自动聚类以查找整个电影中出现的主要对象。

**Abstract：**

我们描述了一种对象和场景检索的方法，该方法可搜索并定位视频中用户概述的对象的所有出现情况。该对象由一组视点不变区域描述符表示，因此尽管视点，照明和部分遮挡发生了变化，也可以成功进行识别。镜头中视频的时间连续性用于跟踪区域，以拒绝不稳定区域并减少描述符中噪声的影响。与文本检索类似的是在实现中，其中预先计算了描述符上的匹配（使用矢量量化），并使用了倒排文件系统和文档排名。结果是检索是即时的，并以Google的方式返回关键帧/镜头的排名列表。说明了该方法用于在两个全长特征膜上进行匹配。

**Introduction：**

这项工作的目的是通过Google轻松，快速和准确地检索包含特定单词的文本文档（网页），来检索包含特定对象的视频的关键帧和镜头。本文研究了文本检索方法是否可以成功地用于对象识别。识别图像数据库中的（相同）对象现在已经成熟。这仍然是一个具有挑战性的问题，因为对象的视觉外观可能由于视点和光照而有很大不同，并且可能会部分被遮挡，但现在存在成功的方法。通常，一个对象由一组重叠的区域表示，每个重叠的区域都由从该区域的外观计算出的向量表示。区域分割和描述符的构建具有对视点和照明条件的可控制不变性。将为数据库中的所有图像计算相似的描述符。特定对象的识别通过描述符向量的最近邻匹配进行，然后使用局部空间一致性（例如邻域，有序或空间布局）或全局关系（例如对极几何）进行歧义消除。例子包括[5、6、8、11、13、12、14、16、17]。

我们探讨了这种类型的识别方法是否可以作为文本检索重铸。本质上，这需要一个单词的视觉类比，这里我们通过向量量化描述符向量来提供它。但是，可以看出，对文本检索进行类比不仅仅是对不同矢量量化的简单优化。在文本检索文献中已经学习和发展了许多课程和经验法则，值得确定的是，这些知识也可以用于视觉检索中。

这种方法的好处是有效地预先计算了匹配，因此在运行时可以使用nodelay检索包含任何特定对象的帧和镜头。这意味着，即使在为视频构建描述符时对这些对象没有明确的兴趣，也可以检索视频中出现的任何对象（以及对象的结合）。但是，我们还必须确定，如果使用以前的最接近邻居匹配方法，则该矢量量化检索是否会错过任何匹配项。

**文本检索的回顾：**文本检索系统通常采用许多标准步骤[1]。首先将文档解析为单词。其次，单词用词干表示，例如“ walk”，“ walking”和“ walks”将用词干“ walk”表示。第三，停止清单用于拒绝非常常见的单词，例如“多数”和“一个”，它们出现在大多数文档中，因此不会区分特定文档。然后，其余单词被分配一个唯一的标识符，并且每个文档都由一个向量表示，该矢量的分量由该文档包含的单词出现的频率给出。另外，以各种方式对组件进行加权（在第4节中有更详细的说明），在Google的情况下，网页的加权取决于链接到该特定页面的网页的数量[3]。上述所有步骤都是在实际检索之前执行的，代表语料库中所有文档的向量集被组织为倒排文件[18]，以促进有效检索。反向文件的结构类似于理想的书本索引。它为语料库中的每个单词都有一个条目，后跟出现该单词的所有文档的列表（以及该文档中的位置）。

通过计算文本的单词频率矢量并以最接近的矢量（以角度测量）返回文档，即可检索文本。另外，单词排序和分隔上的匹配可用于对返回的文档进行排名。

**论文概述：**在这里，我们探索每个步骤的视觉类比。第2节描述了使用的视觉描述符。然后，第3节将它们的向量量化描述为可视化的“单词”，第4节则对向量模型进行加权和索引。然后，在第5节中基于框架的真实性对这些想法进行评估。最后，在第6节中介绍了停止列表和排名（通过空间布局的匹配），并用于评估两个特征膜中的对象检索： Lola Run”（“ Lola Rennt”）[Tykwer，1999]和“ Groundhog Day” [Ramis，1993]。

尽管以前的工作已经从文本检索文献中借用了从数据库中检索图像的想法（例如[15]使用加权和倒置文件方案），但据我们所知，这是将这些想法首次系统地应用于视频中的对象检索。

2.观点不变描述

为每帧计算两种类型的视点协变区域。第一个是通过围绕兴趣点的椭圆形状自适应来构造的。该方法涉及迭代确定椭圆中心，比例和形状。比例尺由拉普拉斯算子的局部极值（横向比例尺）确定，形状由最大化椭圆区域上的强度梯度各向同性确定[2，4]。实现细节在[8，13]中给出。此区域类型称为“适应形状（SA）”。

通过从强度分水岭图像分割中选择区域来构造第二类型的区域。这些区域是随着强度阈值的变化而面积大致固定的区域。实现细节在[7]中给出。此区域类型称为最大稳定（MS）。

使用两种类型的区域，因为它们检测不同的图像区域，从而提供帧的互补表示。 SA区域倾向于以类似拐角的特征为中心，而MS区域则相对于周围环境（例如，灰墙上的暗窗口）对应高对比度的斑点。两种类型的区域都用椭圆表示。以原始检测到的区域大小的两倍计算这些值，以使图像外观更具区分性。对于720576像素的视频帧，计算出的区域数通常为1600。图1显示了一个示例。

每个椭圆形不变区域都使用由Lowe [5]开发的SIFT描述符以128维向量表示。在[9]中，该描述符被证明优于文献中使用的其他描述符，例如一组可控滤波器[8]或正交滤波器[13]的响应，并且我们还发现SIFT更好（通过比较如第5.1节所述，针对地面真相的场景检索结果。之所以具有如此优异的性能，是因为SIFT与其他描述符不同，其设计是不变的，使得该区域位置上几个像素的位移不变，并且这种定位误差是经常发生的误差。将SIFT描述符与适当的协变区域结合起来，可以得到区域描述向量，该向量对于图像的适当变换是不变的。注意，区域检测和描述都是基于帧的单色版本计算的，当前在此工作中未使用颜色信息。

为了减少噪声并拒绝不稳定区域，信息在一系列帧上进行聚合。使用简单的恒定速度动力学模型和相关性跟踪在视频的每个帧中检测到的区域。存活时间超过三帧的任何区域都将被拒绝。轨道的每个区域都可以视为对公共场景区域（检测到的区域的前图像）的独立测量，并且通过对整个轨道上的描述符求平均，可以计算出该场景区域的描述符的估计值。这在描述符的信噪比方面提供了可测量的改进（再次使用第5.1节的地面真实性测试证明了这一点）。

3.建立视觉词汇

此处的目标是将描述符向量量化成簇，这些簇将成为用于文本检索的可视化“单词”。然后，当观察到电影的新帧时，该帧的每个描述符将分配给最近的群集，这将立即为整个电影中的所有帧生成匹配项。该词汇表是由电影的一个子部分构成的，并且它的匹配精度和表达能力将在电影的其余部分上进行评估，如以下各节所述。

矢量量化是通过K-means聚类进行的，尽管其他方法（K-medoids，直方图合并等）当然是可能的。

4.使用文本检索方法进行可视索引

在文本检索中，每个文档都由单词频率的向量表示。然而，通常是对向量[1]的分量进行加权，而不是直接使用频率向量进行索引。在这里，我们描述了采用的标准加权，然后将文档检索与框架检索进行视觉类比。标准加权被称为“术语频率与文档频率成反比” tf-idf，其计算方法如下。假设有一个k个单词的词汇表，则每个文档都由一个k-向量Vd t 1 t i t k加权单词频率表示，分量为t i¤log n d ni

其中n id是文档d中单词i的出现次数，nd是文档d中单词i的总数，ni是整个数据库中单词i的出现次数，N是整个文档中的单词数数据库。权重是两个项的乘积：单词频率n id n d和反向文档频率logN n i。直觉是单词频率对在特定文档中经常出现的单词进行加权，从而对其进行了很好的描述，而反向文档频率对在数据库中经常出现的单词进行了加权。

在检索阶段，通过查询矢量V q和数据库中所有文档矢量V d之间的归一化标量积（角度余弦）对文档进行排序。

在我们的情况下，查询向量是由用户指定的一帧子部分中包含的视觉词给出的，而其他帧则根据其加权向量与该查询向量的相似性进行排序。下一节将评估各种加权模型。

总结：

文本检索的类比确实证明了其价值：尽管在许多帧中视点发生了显着变化，我们仍可以在整个电影数据库中立即进行运行时对象检索。该对象被指定为图像的子部分，这已证明对于准平面刚性对象是足够的。

当然，可以进行一些改进，主要是为了克服视觉处理中的问题。当前排名较低的原因是某些场景类型缺少视觉描述符。但是，该框架允许添加其他现有的亲信协变区域（它们将定义扩展的视觉词汇），例如[17]的那些。另一个改进将是在一个以上的框架上定义感兴趣的对象，以便在其所有视觉方面进行搜索。

文本检索类比也为将来的工作提出了有趣的问题。在文本检索系统中，文本词汇不是一成不变的，随着新文档被添加到集合中而不断增长。同样，我们并不主张矢量量化对所有图像都是通用的。到目前为止，我们已经学习了足以用于两部电影的矢量量化，但是将需要找到提升视觉词汇量的方法。可以想到学习不同场景类型（例如城市景观与森林）的视觉词汇。

最后，我们现在有一个有趣的可能性，可以追随文本检索社区的其他成功，例如对内容的潜在语义索引，以及自动聚类以发现整个电影中出现的主要对象。