**一．引入**

**对象识别**是核心问题之一，在计算机视觉，它是一个非常广泛的研究对象。

题。由于外观上的差异造成的例如，通过非刚性、背景杂乱、差异性、视角、方位、比例或光照条件，它是一个难的问题。

**重要的挑战之一是构建方法规模与数据库的大小相适应的方法，并且可以在可接受的时间内从大量物体中选出一个。**

**本文介绍了一种处理大量对象的方法**。该方法属于目前非常流行的一类算法，这些算法处理局部图像区域，并用从这些局部区域提取的描述符来表示对象[1, 2, 8, 11, 15, 16, 18]。这类算法的优势在于对遮挡的天然鲁棒性和背景杂乱。本文最重要的贡献是建立了一种能够极高效检索的索引机制。在目前的实施中，拟议的方案，在640×480视频帧上进行特征提取。

耗时约0.2秒，而数据库查询则需要25ms，在一个有50000张图片的数据库上。

**二．创新**

**论文提出的方法基于其他人：**

我们的工作很大程度上受西维克（Sivic）和齐瑟曼（Zisserman）的启发[17]。他们使用以下命令从电影中检索镜头文本检索方法。

从局部仿射不变区域提取的描述符被量化为视觉词，这些视觉词由对来自多个训练帧的描述符向量执行的k均值定义。

视觉单词的集合用于图像与以下内容的相关性的TermFrequencyInverseDocument Frequency（TF-IDF）评分中查询。评分是使用反向文件来完成的。

**本文创新是创建了词汇树：**我们建议使用以下方法进行分层TF-IDF评分形成词汇的分层定义的视觉单词树。这样可以更有效地查找视觉对象单词，从而可以使用更大的词汇量，被证明可以显着改善检索质量。

**在本文方法展现了很好的效果**：我们显示了高质量的检索结果，而无需考虑几何布局框内的视觉单词的数量，我们将集中精力展示几何检索前阶段的质量，我们认为这对于扩大大型数据库很重要。

识别质量是通过检索来评估的具有基本事实的数据库，由多个已知组组成相同对象或位置但在不同位置下的图像视点，旋转，比例和照明条件。这个评估表明，词汇树使我们能够无论是在质量和效率方面都有很好的提升。

**词汇树表明使用大量词汇效果会更好：**使用较大的词汇表还可以减少数据库中必须明确考虑的图像比例，从而释放反转文件方法的真正功能。在[17]中，使用的词汇量约为10000个视觉单词。每帧大约1000个视觉单词，这意味着即使在数据库中视觉单词的占用是均匀分布的，在查询过程中也会遍历大约十分之一的数据库。我们表明，使用更大的词汇量（甚至与具有1600万个叶节点的词汇树一样大）可以获得更好的检索质量。在这种词汇量的情况下，必须明确考虑数据库中少几个数量级的特征。因此，可以获得更高的检索质量和效率。特别是，我们获得了一百万个图像的数据库的亚秒级检索时间，而在[17]中，仅尝试了几千帧。我们使用分层评分，这意味着要考虑除叶节点以外的其他节点，但是附加到一个节点的反向文件的图像数量被限制为固定数量，因为较大的反向文件很昂贵，并且提供的图像很少TF-IDF评分中的熵。

词汇树还通过使用分层k均值方法提供了更有效的训练。在[17]中，使用了400个训练框架，而在本文中我们使用了35000个训练框架，这表明使用大词汇量可以提高质量。

**本论文的一创新点——将对象插入数据库存储，同时可以后期检索**：虽然[17]使用离线抓取阶段为视频编制索引，每帧至少需要10秒，但我们可以以与特征提取报告的速率相同的速率将图像插入数据库，即对于640×480分辨率，大约为5Hz。快速将新对象插入数据库的潜力是量化为视觉单词的结果，该单词一劳永逸地定义，同时仍具有较高的检索性能。此功能很重要，但在以前的工作中并不常见。我们计划将其用于基于视觉的同时定位和制图，其中需要即时添加新位置。

**本论文的关键点（沿用，非创新，算是创新点吗）——使用描述符向量与定义词汇树的各种聚类中心的接近度，同时使用离线无监督训练阶段来定义词汇树，实现新图像即时插入数据库：**几位作者表示，树提供了一种索引本地图像区域的有效方法。例如，Lepetit，Lagger和Fua [7]使用图像块的重新渲染来训练用于索引关键点的多个决策树，这在某种程度上让人联想到局部敏感散列[6，3]。它们的树中使用的度量是像素强度的比率。相反，我们使用描述符向量与定义词汇树的各种聚类中心的接近度。他们的方法提供了非常快速的在线操作。它着重于单个对象的检测，但有可能用于更多对象。使用他们的方法，对新对象进行最佳训练需要10到15分钟，而他们最快的方法是对新对象进行训练需要1分钟。我们使用离线无监督训练阶段来定义词汇树，但是一旦确定了词汇树，就可以将新图像即时插入数据库中。

**词汇树方法，使查询效率更高：**在大多数情况下，上述方法会将数据库中的数据量保持在与图像补丁本身或至少区域描述符一样大的数量级上。但是，数据库的紧凑性对于大型数据库中的查询效率非常重要。使用我们的词汇树方法，图像补丁的表示形式只是一个或两个整数，应该与用于描述符向量的数百个字节或浮点数进行对比。

**紧凑性也是我们的方法与Grauman和Darrell [5]使用的分层方法之间最重要的区别。**他们使用直方图的金字塔，在每个级别将沿每个轴的bin数量加倍，而无需考虑数据的分布。通过使用适合于可能的数据分布的词汇表，我们可以使用小得多的树，从而在保持紧凑表示的同时获得更好的分辨率。我们还估计，我们的方法要快1000倍左右。

**方法中涉及——最大稳定极值区域（MSER）和SIFT特征提取**：对于特征提取，我们使用自己的最大稳定极值区域（MSER）实现[10]。已经发现它们在全面的性能评估中表现良好[13，4]。 我们将每个MSER区域周围的椭圆形补丁变形为圆形补丁。然后，Lowe [9]根据SIFT特征提取管道实施特征提取的其余部分。 根据图像梯度上形成的方向直方图找到规范方向。 然后相对于规范方向提取SIFT描述符。 已经发现，SIFT描述符在性能评估中非常有特色[12]。 然后用词汇树对归一化的SIFT描述符进行量化。 最终，应用了一种分层评分方案来从数据库中检索图像。

**三、论文剩下的就是构建和使用词汇树的步骤，以及实验进行评分，然后就是说明一下评分的好坏，最后总结:**

本文已经提出了一种具有索引方案的对象识别方法，该方法比现有技术强大得多。 该方法建立在词汇树上，该词汇树对图像的描述符进行了层次化量化。

结果表明，随着词汇量的增加和图像相似性定义中L 1范数的提高，检索结果得到改善。

同时基于该方法，进行了实时演示，处理了50K CD封面图像，并在1M图像数据库中显示了第二次计时查询。 这些结果使我们寄希望于该方法可能会导致基于互联网规模的基于内容的图像搜索引擎。