基于 SIFT 的图像检索算法

刘嘉伟

Hefei University of Technology

2020 年 11 月 6 日

摘要

图像检索有基于文本的检索和基于内容的检索，如果是基于语义的检索的话，在检索之前需要对海量的图片进行语义属性的标注，这种标注有主观性偏差，时间成本很高，而且语义属性也不能完全表达图像中的包含的丰富的信息，检索效果是有限的。基于内容的检索，“以图搜图”就有他独特的优势。本文介绍的图像检索方式就是一种基于内容的图像检索，通过提取已有图像内容的特征然后对新图像进行预测，是典型的以图搜图的检索方式。

关键字：计算机科学与技术，尺度不变特征变换，机器视觉，图像检索，词袋模型

Abstract

Image retrieval includes text-based retrieval and content-based retrieval. If it is a semantic retrieval, a large number of images need to be labeled with semantic attributes before retrieval. This kind of labeling has subjective bias, high time cost, and semantics. Attributes cannot fully express the rich information contained in the image, and the retrieval effect is limited. Content-based retrieval, "searching for pictures with pictures" has its unique advantages. The image retrieval method introduced in this article is a kind of content-based image retrieval, which extracts the features of the existing image content and then predicts the new image, which is a typical retrieval method of searching for images.

Keywords: CS ，SIFT，Computer Vision，Image Retrieval，BOW

1 图像检索概述

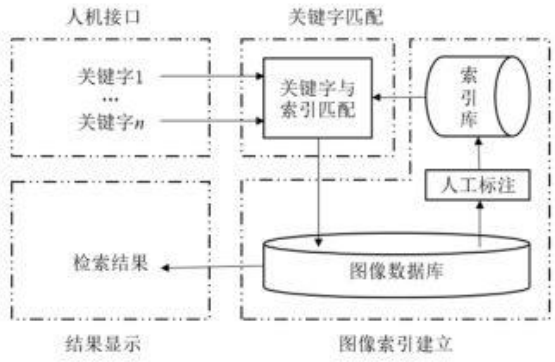
在检索原理上，无论是基于文本的图像检索还是基于内容的图像检索，主要包括三方面：一方面对用户需求的分析和转化，形成可以检索索引数据库的提问；另一方面，收集和加工图像资源，提取特征，分析并进行标引，建立图像的索引数据库；最后一方面是根据相似度算法，计算用户提问与索引数据库中记录的相似度大小，提取出满足阈值的记录作为结果，按照相似度降序的方式输出。

为了进一步提高检索的准确性，许多系统结合相关反馈技术来收集用户对检索结果的反馈信息，这在 CBIR 中显得更为突出，因为 CBIR 实现的是逐步求精的图像检索过程，在同一次检索过程中需要不断地与用户进行交互。

1.1 基于文本的图像检索

基于文本的图像检索(KBIR,Keywords-Based Image Retrieval)研究已经非常成熟而且应用也非常

广泛，对图像数据的管理就是对图像给予一些指定的属性信息，并且在关系数据库中将两者联系起来看作格式化数据，图像的检索只能局限于对这些属性关键字字符串的匹配。这种方法通过利用图像文件名和图像相关的一些关键字属性，为图像建立索引。索引的提取一般为利用人工标注来实现。用户通过输入对应属性关键字来检索图像，系统通过匹配输入的关键字和图像的索引，返回查询结果。如图（1）所示，有人形象的称之为“以字搜图”。此方法的优点是实现过程简单，容易理解，查询速度快，能够反映用户的查询意图，并且检索结果较为精确。但是人工标注需要耗费大量的人力，无法满足大型的多媒体数据库，也难以适应大量新数据的出现，无法解决标注人员在内容感知和描述上的主观性。



图（1）基于文本的图像检索

1.2 基于内容的图像检索

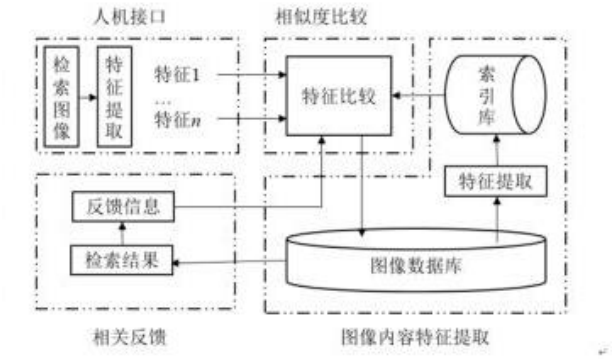
基于内容的图像检索，英文简写为 CBIR。是现在图像检索中的一个重要研究方向。用户通过输入相关图像进行搜索，系统首先提取检索图像的特征，后计算检索图像的特征和特征库中特征之间的相似度，输出检索结果。这种根据特征相似度，给出查询结果的方法，又称之为“以图搜图”。如图（2）所示，为 CBIR 框架的基本组成。系统的运行过程如下：

（1）首先是建立图像特征数据库，具体描述是对图像数据库中的每张图像进行特征提取，用提取的特征来建立图像特征数据库，图像的特征提取也是整个系统的核心部分，所以选取的图像特征对检索结果有着决定性作用；

（2）人机接口通过人机交互提取检索图像特征，并对提取的特征进行匹配，匹配过程其实就是有个相似度计算的过程，计算检索图像的特征和特征数据库中特征的相似度，根据相似性度量结果进行系统输出；

（3）对检索结果进行评价打分，并将评价结果反馈给系统，系统会根据反馈的信息调整相应的比重权重，达到优化系统的目的。

最初 CBIR 研究主要集中在如何用合适的图像特征去描述图像的内容和怎样计算图像特征间的相似度。现在，仍然没有一种有效的特征描述子能够获取图像的高级语义特征，但是，如何提取一些有效的图像描述算子以便于检索和储存却有着实际的研究价值。



图（2）基于内容的图像检索

二 SIFT 算法

SIFT，即尺度不变特征变换（Scale-invariant feature transform，SIFT），是用于图像处理领域的一种描述。这种描述具有尺度不变性，可在图像中检测出关键点，是一种局部特征描述子。

2.1 SIFT 特征介绍

SIFT 特征是基于物体上的一些局部外观的兴趣点而与影像的大小和旋转无关。对于光线、噪声、微视角改变的容忍度也相当高。基于这些特性，它们是高度显著而且相对容易撷取，在母数庞大的特征数据库中，很容易辨识物体而且鲜有误认。使用 SIFT 特征描述对于部分物体遮蔽的侦测率也相当高，甚至只需要 3 个以上的 SIFT 物体特征就足以计算出位置与方位。在现今的电脑硬件速度下和小型的特征数据库条件下，辨识速度可接近即时运算。SIFT 特征的信息量大，适合在海量数据库中快速准确匹配。

2.1 算法特点

SIFT 算法具有如下一些特点：

1）SIFT 特征是图像的局部特征，其对旋转、尺度缩放、亮度变化保持不变性，对视角变化、仿射变换、噪声也保持一定程度的稳定性；

2）区分性好，信息量丰富，适用于在海量特征数据库中进行快速、准确的匹配；

3）多量性，即使少数的几个物体也可以产生大量的 SIFT 特征向量；

4）高速性，经优化的 SIFT 匹配算法甚至可以达到实时的要求；

5）可扩展性，可以很方便的与其他形式的特征向量进行联合。

2.2 特征检测

SIFT 特征检测主要包括以下 4 个基本步骤：

1、尺度空间极值检测

搜索所有尺度上的图像位置。通过高斯微分函数来识别潜在的对于尺度和旋转不变的兴趣点。

2. 关键点定位

在每个候选的位置上，通过一个拟合精细的模型来确定位置和尺度。关键点的选择依据于它们的稳定程度。

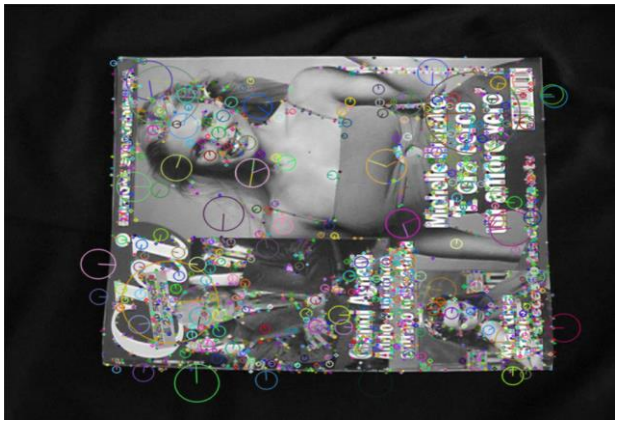
3. 方向确定

基于图像局部的梯度方向，分配给每个关键点位置一个或多个方向。所有后面的对图像数据的操作都相对于关键点的方向、尺度和位置进行变换，从而提供对于这些变换的不变性。

4. 关键点描述

在每个关键点周围的邻域内，在选定的尺度上测量图像局部的梯度。这些梯度被变换成一种表示，这种表示允许比较大的局部形状的变形和光照变化。

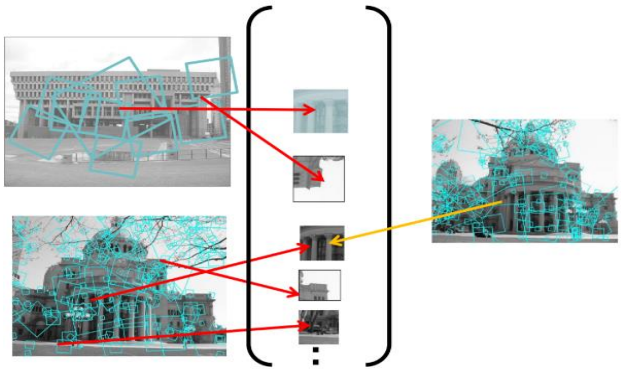
2.3 特征点展示



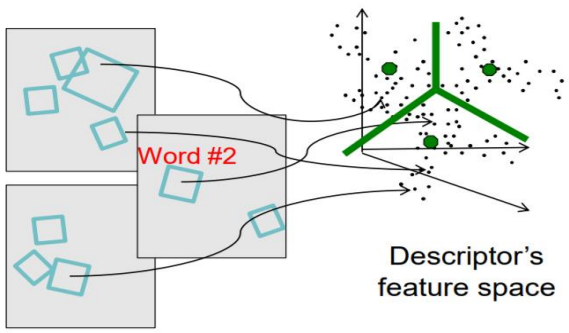
三 词袋模型

最初的 Bag of words，也叫做“词袋”，在信息检索中，Bag of words model 假定对于一个文本，忽略其词序和语法，句法，将其仅仅看做是一个词集合，或者说是词的一个组合，文本中每个词的出现都是独立的，不依赖于其他词是否出现，或者说当这篇文章的作者在任意一个位置选择一个词汇都不受前面句子的影响而独立选择的。

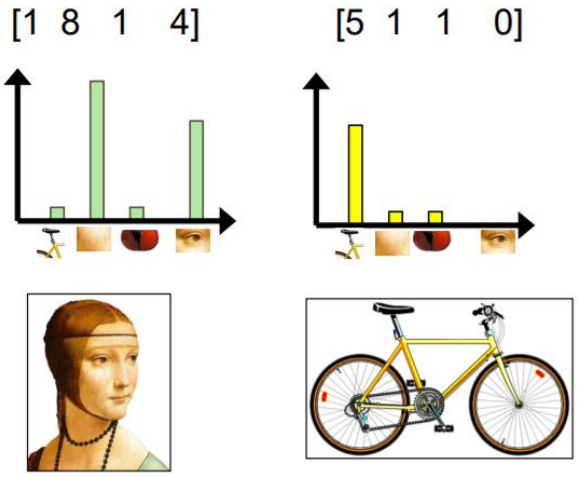
现在 Computer Vision 中的 Bag of words 来表示图像的特征描述也是很流行的。大体思想是这样的，假设有 5 类图像，每一类中有 10 幅图像，这样首先对每一幅图像划分成 patch（可以是刚性分割也可以是像 SIFT 基于关键点检测的），这样，每一个图像就由很多个 patch 表示，每一个 patch 用一个特征向量来表示，假设用 Sift 表示的，一幅图像可能会有成百上千个 patch，每一个 patch 特征向量的维数 128，如下图所示。



接下来就要进行构建 Bag of words 模型了，假设 Dictionary 词典的 Size 为 100，即有 100 个词。那么咱们可以用 K-means 算法对所有的 patch 进行聚类，k=100，我们知道，等 k-means 收敛时，我们也得到了每一个 cluster 最后的质心，那么这 100 个质心（维数 128）就是词典里德 100 个词了，词典构建完毕。



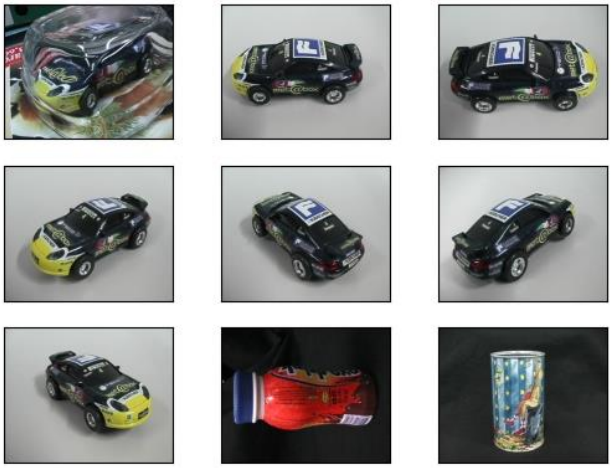
词典构建完了怎么用呢？是这样的，先初始化一个 100 个 bin 的初始值为 0 的直方图 h。每一幅图像不是有很多 patch 么？我们就再次计算这些 patch 和和每一个质心的距离，看看每一个 patch 离哪一个质心最近，那么直方图 h 中相对应的 bin 就加 1，然后计算完这幅图像所有的 patches 之后，就得到了一个 bin=100 的直方图，然后进行归一化，用这个 100 维的向量来表示这幅图像。对所有图像计算完成之后，就可以进行分类聚类训练预测之类的了。



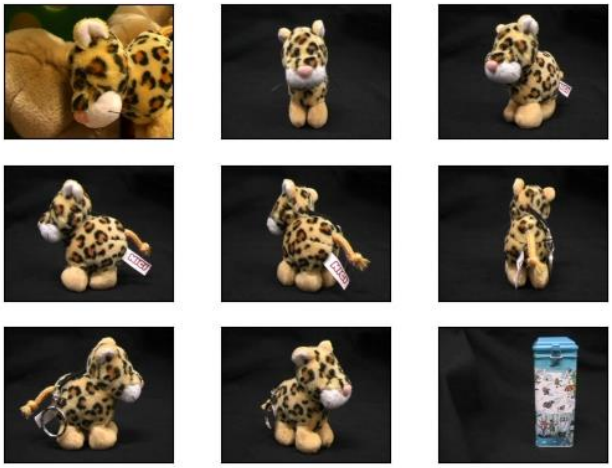
然后我们将每张图片的 100 维向量当作训练集，该向量对应的图片标签为标签，使用 SVM 进行对特征向量进行训练。最后，对于待预测图片，我们按照上述步骤生成该图片的特征向量，然后使用训练好的分类器进行分类预测，由于 SVM 分类结果只会输出一类，所以将最高的那个概率的 80%作为阈值，只要输出概率高于该阈值，则可以将该图片作为匹配的图片输出。

四　结果展示

结果一

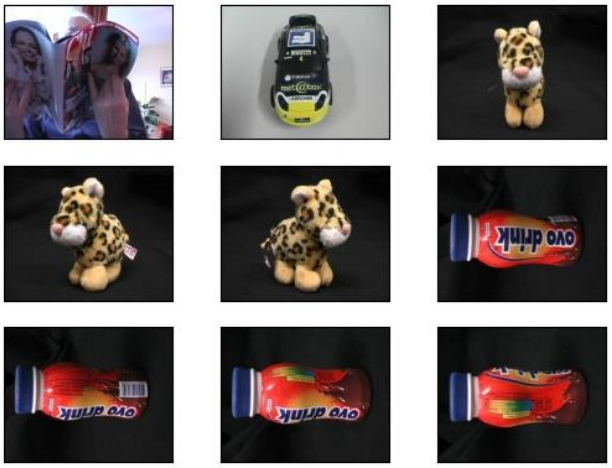


结果二



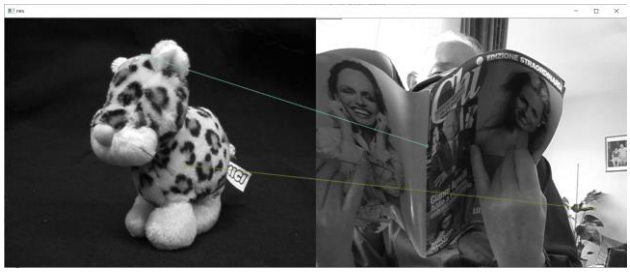
结果三

五　算法评价



从这张图来看，我们发现没有一个匹配的图，将该图的 sift 匹配结果一张一张输出来查看具体发生了什么情况：

匹配点一



匹配点三

可以看到，对于该图的匹配，算法找到了匹配的 sift 点，但是很显然，这几个匹配上的 sift 点都不正确，因为 sift 算法匹配的是特征点，并没有任何结构性的特征，如纹理特征等，所以这种匹配模式完全是基于图片的内容进行匹配，对于不同的图片，上面的点特征很可能会相似，就如上面这几张图片所见，虽然人眼看见觉得毫不相关，但是对于电脑来说，他们就是相似的图片。如果加上一些纹理特征进行匹配，那么匹配的精度肯定会更高，因为纹理特征相当于计算机读出了图片中的物体的形状，可以根据物体的形状来判断是否相似，而不是仅仅单纯的通过图像上的点来判断是否相似。

匹配点二