基于传统特征和卷积神经网络（CNN）特征融合的视觉实例检索算法研究

摘要

关键字

一、背景和意义

随着互联网和电子智能设备的快速普及和发展，企业甚至个人每天需要浏览、存储和处理大量的声音、视频和图像等数据，比如Flickr、Facebook社交网站或者电子商务系统淘宝网等。如何有效地使用和管理这些数据，成为多媒体领域的热门话题。其中，图像数据研究的核心问题之一是图像检索[1]。

一般来说，图像检索要解决的问题是，给定查询图像，在图像库中找到与查询图像最相似的图像，同时满足一定的计算精度和效率。因此如何选择图像的特征表示，对图像检索的准确度至关重要。当前主流的图像检索分为相同类别图像检索和相同物体图像检索：

**相同类别图像检索**

1. 用户感兴趣的是物体、场景的类别，即用户想要获取的是那些具有相同类别属性的物体或场景的图片；
2. 目前已广泛应用于图像搜索引擎，医学影像检索等领域；
3. 面临的主要问题是属于同一类别的图像类内变化巨大，而不同类的图像类间差异小；
4. 如果采用低层的特征去描述，比如颜色、纹理以及形状等特征，其类间差异非常小，直接采用这些特征是很难将这两者分开的。

**相同物体图像检索**

1. 用户感兴趣的是图像中包含的特定物体或目标，并且检索到的图片应该是包含有该物体的那些图片；
2. 不论是在研究还是在商业图像搜索产业中都具有重大的价值，比如购物应用中搜索衣服鞋子、人脸检索等；
3. 在检索相同的物体或目标时，易受拍摄环境的影响，比如光照变化、尺度变化、视角变化、遮挡以及背景的杂乱等都会对检索结果造成较大的影响；
4. 由于受环境干扰比较大，因而对于相同物体图像检索，在选取特征的时候，往往会选择那些抗干扰性比较好的不变性局部特征，比如局部特征尺度不变变换特征（Scale-invariant feature transform，sift）[28]，加速稳健特征（ Speeded Up Robust Features，surf）[29],（Oriented FAST and Rotated BRIEF，ORB）[30]等。

实际上，相同物体图像检索，又叫图像实例检索或者视觉实例检索，与一般的图像搜索相比，实例搜索的搜索条件更为苛刻——是否包含相同的物体，如同一款衣服，同一辆车等等。该问题具有非常广泛的应用前景，如商品搜索[11]，车辆搜索[9]和基于图像的地理位置识别[10]等。举个例子，移动商品图像搜索就是通过分析利用手机相机拍摄的商品照片，从商品库中找到相同或相似的商品。然而在实际场景中，由于姿态，光照和背景等干扰因素，所以两张包含相同物体的图像往往在外观上差异很大。从这个角度来看，视觉实例检索的本质问题就是应该学习什么样的图像特征，从而使得包含相同物体的图像在特征空间上是相似的。

近年来机器学习和深度学习不断发展，卷积神经网络在各种计算机视觉任务上都取得了重大的突破，一个重要因素就是其强大的非线性表示能力，能够理解图像更深层次的信息。其中就包括视觉实例搜索任务。虽然深度学习已经成为解决许多计算机视觉任务的核心技术和方法之一，但是它的应用在视觉实例检索并没有显著的性能提升[8]。

随着各种技术的不断发展，计算机获得的底层图像像素特征和人眼高层语义特征之间存在的语义鸿沟问题仍然存在，因此如何如何消除这种语义鸿沟，实现更好的图像检索性能尤为重要。基于局部特征尺度不变变换特征sift的图像检索发展快速，但是sift特征属于底层特征，存在缺乏集合信息，颜色信息和空间信息，高层语义的表达不够等问题，在处理语义鸿沟和大视角变化有很低的性能，而CNN特征往往缺乏足够的底层信息，为了丰富描述符的信息。与sift相比，CNN全局特征的性能较好，但是其通用性及特征集合不变性都不够强，且缺乏对局部细节的描述，所以卷积层的局部CNN特征逐渐变成了研究热点。为了丰富描述符的信息，通常将sift和CNN融合，有多种融合方式，串联、索引融合和相似度融合等，融合可以有效地利用不同的特征的互补性，提高检索的准确性。实际上，全局的CNN特征适合表示场景，局部的CNN特征适合表示目标层，而sift特征主要适合表示点层。对于预训练的模型而言，如何有效的融合这三种特征，进而改善视觉实例检索的精度，会是一个值得研究的课题。

二、国内外研究现状

图像检索最早开始于基于文本的图像检索(TBIR)和基于内容的图像检索（CBIR：content-based image retrieval），其中，CBIR一直是国内外计算机视觉领域的一个研究热点。CBIR通过分析图像内容，如颜色、纹理和形状等，从图像中提取对应特征，然后采用合适的距离度量方式计算图片的相似度，从而排序得到和查询图像相似的结果。CBIR在电子商务、皮革布料、版权保护、医疗诊断、公共安全、街景地图等工业领域具有广阔的应用前景，通过搜索相似的商品图像、皮革布料图像、版权商标或者病理图像等，为人们的生活带来了极大的便利。表示图像的最直接手段是提取全局的图像特征，CBIR将图像内容的表达和相似性度量交给计算机进行自动的处理，克服了采用文本进行图像检索所面临的缺陷，并且充分发挥了计算机长于计算的优势，大大提高了检索的效率，从而为海量图像库的检索开启了新的大门。但是全局特征没有考虑到光照、角度和变化等因素，会丢失图像的特征不变性,并大大降低图像检索的精度[18]。因此出现了基于局部特征的图像检索。

本文主要集中于视觉实例检索，对于相同物体的图像检索都统称为视觉实例检索。传统的CBIR方法一直应用到2000年，2003年研究者们开始提出局部特征尺度不变变换特征（Scale-invariant feature transform，sift）[28]，加速稳健特征（ Speeded Up Robust Features，surf）[29],（Oriented FAST and Rotated BRIEF，ORB）[30]等等，然后以此为基础通过不同的编码方式构建图像的全局描述,如词袋模型 (Bag of Words，BoW)[31]、 局部特征聚合描述符 (VLAD, Vector of Locally Aggregated Descriptors)[32]以及Fisher向量 (FV, Fisher Vector)[33]也长期被广泛用于图像检索。ｓｉｆｔ一直是ＣＢＩＲ中最受欢迎的手工特征，从本质上看，它具有一定的几何不变性，并且对于视觉实例检索和去重图像检索［３５］［３６］表现出较好的性能。上述一类以SIFT为基础的视觉检索方法，由于结合了SIFT不变性的特性，并且采用了中间层由局部到全局的特征表达方式，从整体上来说能够获得比较好的检索效果，但这一类方法通常其特征维度往往非常高。为了获得较高的检索精度，在聚类大规模图像数据集时聚类数目一般都设置到了几十万，这造成其最终表示的特征其维度高达几十万维，因此为它们设计高效的索引方式显得十分必要。尽管研究者们做了很多必要的工作来提升sift的性能，但是只使用sift特征仍然很难缩小特征表示本身和人类的图像视觉感知之间的语义鸿沟。

2012年研究者们开始转移关注基于深度学习的方法卷积神经网络CNN（convolutional neural networks）进行图像检索。近年来CNN凭借强大的特征学习能力，在计算机视觉领域应用广泛，比如图像分类、图像识别，都获得了较大的性能提升。已有工作[3][4][5]表明，在大规模数据库上预训练的ＣＮＮ特征应用于图像识别等视觉任务，远远胜于手工提取特征的方式。与手工特征相比，CNN特征通过一个深度学习框架从大量的标注数据，例如ImageNet [39]学习而来，同时携带了丰富的高层的语义信息。一些工作基于CBIR[40][41][42]展示了CNN特征作为场景层的表示对于图像检索起到了很好的效果。Gong等人[43]提出了一个方法MOP (Multi-scale Orderless Pooling),通过聚合三种规模的CNN特征表示局部信息。Konda等人[44]和Xie等人[45]检测了目标候选区域，并在目标层提取了每个区域的CNN特征。除此之外，还有一些研究者们更加关注深度卷积层来获取特征[46][47][48][49]完成图像检索。虽然CNN特征实现了比较好的性能，但是我们不能说CNN一定能超越Sift[50]。Vijay等人[51]表明没有哪一种特征会比另一种好，结合两种特征会获得更好的检索性能。

已经有工作表明[50]CNN和Sift是高度互补的，并提出了一种CSS特征表示。zheng[38]也做了相似的工作，在多层上融合了CNN和SIFT，但是他们利用图像区域和块提取CNN特征，并构建bow模型改善局部特征的检索精度。[50]从场景层和目标层提取CNN特征，从点层提取sift特征，然后融合这些特征构建新的图像表示。其中，顶层的CNN特征可以表示场景层，通过目标检测的候选区域得到的CNN特征可以表示局部特征，并采用了VLAD来对这些局部特征进行编码，当然Vlad池化方法会影响局部特征特征表示的性能。接着基于sift特征构建Vlad表示，最后串联三种特征并采用主成分分析方法进行降维处理，从而获得最终比较紧致的特征。然而，Dan等人[52]从多个尺度的局部块提取CNN局部特征，然后直接串联所有层的特征构建最终的表示，他们发现由于由于能够同时提取图像的局部和全局信息，这种方法能够改善检索性能，因此当全局的CNN特征拥有更好的性能时，采用Vlad池化方式是没有必要的。

受这些工作的激励，本文主要研究并实现了一种特征融合检索算法。