学校编码：10384　　分类号_____密级 _____
学　　号：23020111153085　　UDC_____

厦门大学

硕 士 学 位 论 文

基于判别性特征表示的图像检索算法研究

Research onImage Discriminative Representation for Image Retrieval

宋书阳

指导教师姓名：　曲延云　副教授
专 业 名 称：　计算机应用技术
论文提交日期：　2014 年　5 月
论文答辩时间：　2014 年　5 月
学位授予日期：　2014 年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2014 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：宋书阳

2014 年 05 月 14 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：[illegible]

2014年05月14日

摘要

随着互联网的发展、社交媒体的兴起以及图像采集设备的普及，大量图像数据涌现在互联网上。图像数量的爆炸式增长，给图像检索带来了巨大的挑战。在图像检索中，通常使用词袋模型(Bag Of Words,简称 BOW)对图像进行描述，得到检索结果之后使用 RANSAC(RANdom SAmple Consensus，简称 RANSAC)进行几何验证或者进行匹配验证实现重排序。这一检索框架存在三方面的不足:1) 词袋模型完全忽略了图像中的空间结构信息，在图像的特征表示上没有充分利用空间信息增强判别性；2) 面向规模较大的图像检索问题，需要相应的大规模的视觉词典，直接针对视觉词的度量方法，其计算复杂度高；3) 基于 RANSAC 的几何验证或者匹配验证计算复杂度高。后两条导致检索效率不高。

针对以上三点不足，本文主要研究如何利用空间信息提高图像的判别性表示，如何利用哈希算法加快图像的检索速度，如何利用空间位置的粗匹配，加快图像验证。本文进一步研究了如何利用哈希算法解决自然场景中中文字符识别问题。本文的工作主要集中在以下两个方面：

(一)设计一个融合空间判别性信息的图像检索框架：在第一层使用粗粒度的几何信息，设计了空间最小哈希方法。哈希表示是词袋模型的零阶逼近，它随机的抽取了词袋表示的部分视觉词进行比较，提高了计算速度，然而丧失了部分判别信息。为了增加哈希表示的判别性，本文将图像先进行空间金字塔表示，然后在各个局部空间进行最小哈希算法，改善了检索的性能。在第二层图像验证层，使用细粒度的空间信息--局部空间金字塔表示，进行图像之间的配准验证。利用最大极值稳定区域(Maximally Stable Extremal Regions，简称 MSER)和角点之间的空间位置关系，进行配准验证。该验证避免了图像之间所有点的完全匹配验证，通过特征分层验证，降低了计算量，加快了验证速度。

(二)针对于自然场景中文字符识别存在的字体不一致、数据集不平衡、常用中文字符类别多、类内样本少等问题，本文将图像检索的技术应用在自然场景汉字识别中。利用迭代量化算法用于中文字符识别，并结合编辑距离对识别的结果进行纠

正。

关键词：最小哈希 空间最小哈希 局部空间金字塔表示 迭代量化

Abstract

With the development of the Internet, the rise of social media and the popularization of image acquisition devices, a large number of images have emerged on the Internet .The explosive growth of images bring great challenge for image retrieval. In image retrieval, the Bag-of-words model is usually first used for image description ,then the returned images are proposed by RANSAC(RANdom SAmple Consensus) for geometry verification to RE-RANK, that usually give a better result. This image retrieval framework has 3 shortcomings: 1) Spatial information of images is ignored completely, thus the description is less discriminative; 2) For large scale image retrieval, a great amount of visual vocabularies are needed for image description. Therefore the computation cost is high if the Bag-of -words model is used directly. 3) RANSAC-based geometry validation or image verification is time-consuming, due to the computation complexity. Item 2 and item 3 lead to inefficient image retrieval.

To overcome the shortcomings mentioned above, this paper is focused on improve the performance of image retrieval, by using spatial information. This paper also further the study on natural scene Chinese character recognition using image retrieval method. Works in this paper is mainly covers 2 aspects:

1) This paper designs a retrieval framework with identifying spatial information: on the first layer, spatial min-Hash algorithm is proposed using coarse-grained geometry information. Hash representation is zero-order approximation to Bag-of-words model. It randomly selects and compares a subset of visual words represented by Bag-of-words. The computation is speed up while scarifying some discrimination. In order to increase the recognition performance of Hash representation, this paper will first use spatial pyramid to present images and then process each part of the image by the min-Hash method to improve retrieval performance. On the second image verification layer, fine-grained geometry information, presented by detailed spatial pyramid, is used for verification between images. MSER regions and spatial relationship between angles and points are used for matching and verification. The verification avoids exact matching between all points of images. It reduces computation amount and accelerates verification by layered verifications of features.

2) This paper adopts image retrieval techniques in Chinese character recognition in natural scenes to address problems of font inconsistency, dataset imbalance, variety of Chinese character categories and shortage of samples within a category etc. Iterative quantization is used in Chinese character recognition and the recognized result will be rectified with the help of editing distance.

Keywords: min-Hash; spatial min-Hash; local spatial pyramid; iterative quantization

目录

Contents

第一章 绪论

1.1 研究背景与意义

如今，WEB2.0 已经渗透到了人们的生活，各种社交网站和分享平台充斥在我们的周围。各种图像采集设备，如数码相机、具有摄像功能的手机，也随着硬件成本的下降开始普及。人们开始使用图像来记录自己的生活，并利用人人、微博、朋友圈等社交工具分享自己的生活。据德国著名统计公司 Statista 的研究数据[1]，如图 1-1 所示。按新增用户数计算,2013 年 Twitter 的 Vine 是全球增长最快的应用程序，它的用户群在今年第一季度和第三季度之间增长了 403%。而雅虎旗下的照片分享服务 Flickr，其应用程序的用户数量增加了 146%，列第二位。列第三位的是 Instagram，增长了 130%。前 3 名的应用程序都与图像分享有关。

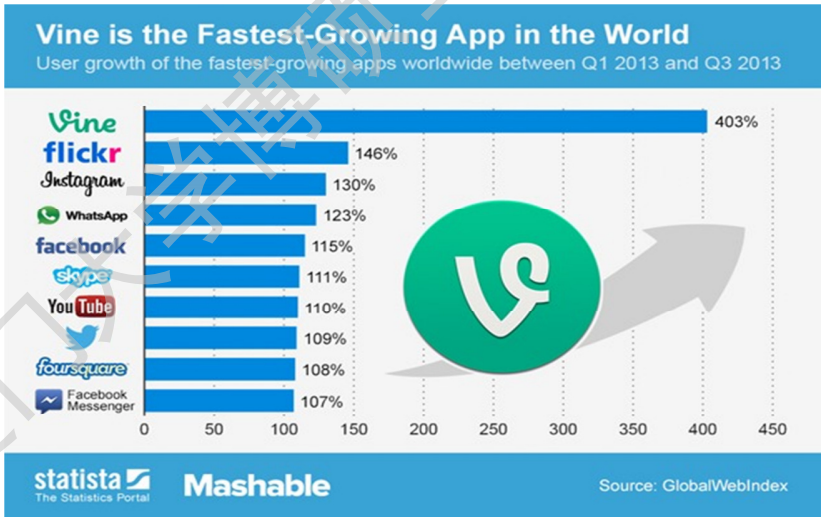

图 1-1 Statista 的研究数据

各种社交网络和分享平台的出现导致海量图像数据涌现在互联网上。据社交搜索引擎 Topsy 提供的数据显示，自 2005 年 10 月以来，Facebook 上分享的图像超过 5000 亿张；截止 2010 年，flickr 网站上的图像突破 50 亿张，并以每年 10 亿的数

量增长(如图 1-2 所示)；图像存储网站 Photobucker 有 80 亿张，Picasa 上已有 70 亿张。在 2011 年，Twitter 上分享的图像数量增长了 421%，仅在 12 月份用户分享的图像数量就达到了 5840 万张。

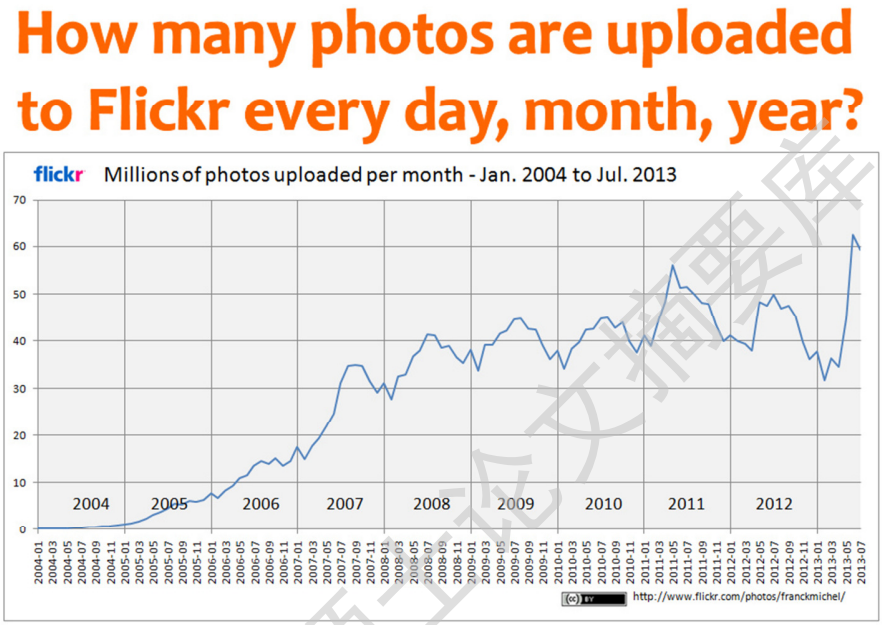

图 1-2 flickr 网站每月上传图像数量统计

相对于文字等传统的信息载体，图像信息简单、直观、信息量大、易于理解和接受。图像在给人们记录信息方面带来便利性的同时，也带来了一个问题：想要在互联网上找到需要的图像变得越来越困难。因此，如何在浩瀚的图像数据中快速、精准的找到用户所需要的图像成为当前一个具有巨大挑战和意义的一个课题。目前，针对于海量图像检索的研究已经成为了计算机视觉和多媒体领域的一个热点。

本课题以统计学、计算机视觉和信息检索等知识为理论依据，以当前最新的图像处理和检索技术为实践指导，主要针对大规模图像数据的结合空间信息的判别性表示以及快速搜索算法的设计开展研究，以期达到快速、精准的检索效果。

1.2 图像检索概述

图像检索就是给定一张查询图像，找到与该图像相似或者相关的其他图像。从

上个世纪70年代开始，人们就开始思考如何才能从海量的图像中检索到所需要的图像[2]。到目前为止，图像检索的技术大概分为两个阶段。

1.2.1 基于文本的图像检索

基于文本的图像检索(Text-Based Image Retrieval，简称TBIR)回避对图像可视化元素的分析，先人工对图像进行关键词标记。把这些关键词作为图像的描述存入数据库中。当查询时，使用关键字通过数据库技术实现对关键字的检索[3]。最终，将关键字对应的图像作为查询结果返回。该方法难度不大，可以利用比较成熟的文本检索技术。但是，由于图像需要人工标注，工作量大，当图像数量比较大时难以实现。并且由于人类认知的主观性，难以解决不同人对同一图像标注不同的问题。

1.2.2 基于内容的图像检索

为了克服TBIR的缺点，在上个世纪90年代LewM等[4]提出了基于内容的图像检索(Content-Based Image Retrieval，简称CBIR)的概念。CBIR的主要思想是：对于图像，从图像的内容出发摒弃人工标注的步骤，使用图像中的视觉特征，如颜色、纹理、形状等来表示图像。然后通过在这些特征空间中的相似度匹配找到相似的图像。一个开创性的工作是在1984年Chang[5]等提出了对图像进行抽象表示来对一个画报图像库进行检索，在文章中作者使用目标类别之间的集合关系对图像进行表示。这种方法结合了概率论、人工智能、机器学习、模式识别等多个学科的相关理论。相对于TBIR方法，这种方法具有以下优点：1，不需要人工标注，节省了时间和成本，可以扩展到大的图像数据库。2，避免了人工标注的主观性，保证了对同一图像描述的一致性。

基于内容的图像检索经过学者们几十年的研究，已经取得了长足的进展。已经出现了Photobook ,TinyEye等基于内容的图像搜索引擎。然而现有的算法，多使用图像的底层特征来表示图像，而图像的底层特征难以对图像的高层语义进行有效描述，得到的结果与人类对图像的视觉感知和心理感知有较大的差异。这种方法面临着语义鸿沟的问题，得到的结果不够理想。

针对于图像检索中存在的语义鸿沟问题，一些学者提出了基于语义的图像检索技术(Semantic-Based Image Retrieval，简称SBIR)。SBIR方法，主要通过构建

图像中的底层特征与高层语义之间的映射关系来解决语义鸿沟的问题，从而提高图像检索的准确率。现在 SBIR 中常用的技术有语义模型、相关反馈等[6]。相关反馈技术使用了人机交互的方式来进行图像检索，在检索的过程中需要跟用户进行多次交互。系统根据用户的反馈，修改其中的参数，通过多轮反馈可以实现较高的检索精度。

本文主要研究关注于 CBIR，CBIR 是 SBIR 的基础，好的 CBIR 算法能极大地提升 SBIR 的效果。

1.2.3 图像检索相关技术介绍

“一图胜千言”，这句谚语说明了数字图像作为一种重要的信息载体与表现形式，在信息传播中的巨大推动作用。图像数目的爆炸式增长，进一步推动了图像检索的发展，同时也给图像检索带来了巨大的挑战。对于海量图像检索系统主要分为两大部分：图像检索中的图像表示和索引方法。

(一) 图像检索中的图像表示

图像数量庞大，类别众多，类间差别大。同一类的目标也会因为扭曲，遮挡，光照，尺度变化等，在图像上表现出较大差异。如何对图像进行特征表示是基于内容的图像检索面临的巨大挑战。目前的图像表示算法按照特征的形成方式可分为全局特征表示和局部特征表示两种。

1) 全局特征表示

图像的全局特征表示将整张图像作为整体进行描述，常用的整体特征描述有颜色特征、纹理特征、GIST[7]、VLAD(Vector of Locally Aggregated Descriptors，简称 VLAD)[8]、HOG(Histogram of Oriented Gradients，简称 HOG)[9]等。

颜色特征是最早被用在图像检索中的特征，颜色特征有多种表达方式，如颜色统计直方图、累积直方图、颜色距、颜色聚合向量等。在实际应用中颜色特征通常和其他特征组合使用。纹理特征被用来描述图像区域所对应物体表面的性质。但是纹理特征只描述了物体的表面特性，因此纹理特征不能够有效的表示物体的本质属性，所以只使用纹理特征不能够有效的获得图像的高层语义。当图像的分辨率变化

较大时，纹理特征的描述的一致性大大降低。纹理特征对于疏密或者粗细变化较大的图像库中进行检索时准确率较高。对于纹理之间疏密、粗细变化较小的图像，纹理特征通常很难准确反映出人类视觉的感受。

GIST 描述子以离散傅里叶变换为基础，先计算图像的能量光谱图，而后采用多种滤波方法以获得图像的各种属性。而 VLAD 描述子先提取 SIFT((Scale-Invariant Feature Transform,简称 SIFT))描述子并进行量化，并将量化结果进行局部累积，累积结果串联起来作 PCA 降维得到图像全局描述符。GIST 描述子对于图像整体结构比较相似的检索效率比较好，但对旋转、遮挡比较敏感，VLAD 描述子不受旋转、遮挡的干扰，但需要提取大量的 SIFT 描述子。

图像的全局特征信息将图像视为一个整体，其优点是提取特征信息的复杂度低，特征表示形式更加紧凑，主要关注的是图像的整体信息。对于图像的变换和遮挡，全局特征的鲁棒性不高。

2） 局部特征表示

20 世纪 70 年代末，局部特征的研究开始兴起[10]。Moravec[64]在 1980 年提出了角点特征。Harris[11]在 1988 年提出了角点特征检测算法，使用微分算子和矩阵值来定位角点，所以 Harris 角点具有更高的检测率和重复性。同时 Harris 角点对于灰度变化和旋转都保持不变性。在 20 世纪 90 年代，Lindebrg[12]系统的提出了基于信号的尺度空间理论。借助尺度空间的概念,Mikolajezyk 提出了 Harris-Laplace[13]和 Harris-affine[13]检测子。2000 年，Lowe[14]提出了 SIFT,通过使用金字塔和高斯滤波差分来快速求解高斯拉普拉斯空间中的极值点。结果大量的实验证明 SIFT 描述子的性能几乎是局部描述子里面最好的。2006 年 Bay[15]沿着 Lowe 的思路，提出了 SURF(Speeded Up Robust Features，简称 SURF)。SIFT 描述子的变化版本还有 PCA-SIFT[16]，GLOH(Gradient Location-Orientation Histogram ，简称 GLOH)等。在 2004 年,Matas 等[17]提出了 MSER(Maximally Stable Extremal Regions,简称 MSER)特征区域检测方法。该方法对于仿射变化具有不变性。

（二） 索引方法