学校编码：10384　　分类号_____密级_____
学号：23020081153230　　UDC_____

厦门大学

硕士学位论文

基于网络隐监督的地标图像的搜集与标记研究

Landmark Images Collecting and Labeling Based on Web Implicit Supervision

程燕云

指导教师姓名：曲延云　副教授
专业名称：计算机应用技术
论文提交时间：2011 年 5 月
论文答辩日期：2011 年 月
学位授予日期：2011 年 月

答辩委员会主席：
评阅人：

2011 年 6 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）：程燕云

2011年6月7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）：程燕云

2011年6月7日

摘 要

网络的发展为我们带来了丰富的视觉信息，利用图像搜索引擎可以搜集到成千上万的图像，这其中包含有大量的各地风景名胜图像，它们从各种视角、各种季节、各种时段展现着旅游名胜的美。如何对这些风景图像进行结构化组织是网络视觉研究的一个热点问题。本文针对地标风景图像进行了深入的研究，按视觉一致性对地标图像进行聚类，进行地标图像的摘要，并对地标进行定位。该研究对地标的三维重建、地标图像的浏览具有重要的作用。本文的研究工作和学术贡献如下：

1. 提出一种按空间分布结合语义的地标图像组织和过滤方法。针对网络引擎搜集的噪声图像集，利用 GIST 描述子对图像进行全局特征描述，然后设计分层聚类方法对图像进行聚类。对得到的图像聚类集合，利用局部特征描述子 SIFT 对关键点进行描述，并结合 RANSAC 算法和词袋模型思想，进行图像集的几何一致性验证和共性特征提取，过滤噪声图像，同时为图像集挑选出一张地标图像（Iconic Image）作为该集合的摘要。

2. 提出一种基于视觉词词频挖掘的地标定位方法。在得到的各类地标聚类集合上，利用 SIFT 特征计算具有视觉一致性及空间一致性的兴趣点，并且设计出一种提取含有最关键信息的兴趣点的方法，然后通过这些兴趣点的位置，结合图割技术（GrabCut），预测地标的位置。

3. 在假设地标聚类集合含有地标的前提下，提出了两种基于监督的地标定位算法。第一种方法，将地标定位问题转化为弱监督目标的分类问题：首先采用基于兴趣点的双模板对图像进行 GrabCut 分割，接着利用多示例学习思想对分割结果进行半监督分类，最后从分割结果中筛选出对地标的最优标记。第二种方法，将地标定位问题转化为集合内部元素的近邻匹配问题：首先利用兴趣点匹配技术大致标记目标的位置，接着通过 GrabCut 对标记结果进行优化，最后结合地标面比特征实现对地标的最优标记。

将本文算法应用于从网络检索到的四类地标图像上。在地标图像的组织方面，本文算法取得了较好的效果，能将主观视觉上具有空间及语义一致的地标

类聚合在一起；在图像过滤方面，本文算法对正确地标图像的平均查准率达到89.52%，而利用关键词从网络搜索得到的地标图像集的平均精度为27.97%；在地标定位方面，基于词频挖掘的地标定位方法最高达到95.35%的标记精度，基于弱监督学习的地标定位方法最高达到90.91%的标记精度，基于近邻匹配的地标定位方法最高达到95.74%的标记精度。实验结果证明了本文所提算法的有效性。

关键词：分层 Kmeans 聚类，图像标记，词袋模型，GrabCut，GIST，SIFT，HOG，多示例学习

Abstract

Network development has brought us a wealth of visual information. We can get thousands of landmark images, through the network image search engines. These images embody the beauty of tourist attractions by different perspectives, different seasons, different weather, and different times. How to structure and organize these images is a hot issue in network vision research. In this paper, we study the representation of landmark images, cluster them according to visual consistency, summary an iconic image for each clustered set, and finally label the landmark positions. The study plays an important role in three-dimensional reconstruction, landmark images browsing, and so on. The work and contribution of this paper are as follows:

1. Propose a method for landmark images organization and filtering, which is based on spatial distribution and semantics. We use GIST descriptor to describe the panorama layout feature of each image collected by photo collection engine, and design a hierarchical clustering method to cluster them. For each set of cluster result, we filter the noises by geometric consistency information, using SIFT descriptor combined with RANSAC algorithm and bag-of-word model. And also, we pick out an iconic image for each reservation set.

2. Propose a method for landmark annotation, which is based on the frequency of visual words. We firstly find out all the interest points that are both consistent in visual and spatial, and then design an approach to choose all the key points that contain the most important information in the sets, and finally label landmarks by GrabCut algorithm with the help of the key points.

3. Propose two supervision methods for landmark labeling, on the assumption of each set containing landmarks. In the first method, the labeling issue is transformed into landmark classification problem based on weak supervision; we

firstly segment images by two-template using GrabCut, and then classify the segmentation results by MIL semi-supervised, and finally select the best results. In the second method, the labeling issue is transformed into neighborhood matching problem binding the aspect ratio of targets; we firstly generally mark the targets using points-matching technology, and then optimize the results by GrabCut, and finally achieve the best marks with the help of aspect ratio.

We test our algorithm on four categories landmark images retrieved from network. In the respect of images organization and filtering, it achieves good effectiveness on subjective visual, which can cluster the landmark images having similar space and semantic features together. In the respect of image filtering, the average precision rate of our algorithm is 89.52%, while the precision rate of retrieval is 27.97%. In the respect of landmark labeling, the way based on frequency of visual word up to 95.35% accuracy, the way based on MIL up to 90.91% accuracy, and the way based on neighbor matching up to 95.74% accuracy. Experimental results show the effectiveness of our algorithm.

Keywords: Hierarchical Kmeans Clustering, Image Labeling, Bag-of-Word, GrabCut, GIST, SIFT, HOG, MIL

目　录

Contents

第一章 绪论

1.1 研究背景及课题难点

随着网络的发展以及摄像设备的普及，越来越多人开始习惯以照相机记录自己的生活，同时通过网络与全世界的人们分享自己的乐趣。据不完全统计，Flickr 网站（全球著名的网络相册网站）管理着大约 4 亿张照片，Facebook 网站(开启“个人网页”的先师)拥有 15 亿张以上的照片资源，要是再算上 YouTube 这一类的在线视频网站，我们可以从网络上直接获取的图像资源就超过了 675 亿张！而且随着时间的推移，这个数字还将继续增大。——这意味着：如果我们将 Flickr 网站上的图像全都打印出来，叠放到一起，就可以达到 410 英里的高度，几乎是人造地球卫星绕地飞行高度（220 英里）的 2 倍；如果将 Facebook 网站上的图像全部打印出来，可以达到 1539 英里；如果将 YouTube 网站中的所有视频帧图像也全部打印出来，就可以达到惊人的 69247 英里——只要有 4 个与 YouTube 网站规模类似的视频网站，人类就可以踩着由这几个网站的打印出来的照片登上月球了！

如此多的图像资源，为网络图像资源的管理和检索提出新的挑战，基于关键词检索的图像检索已远远不能满足检索的需要。这其中对风景图像的处理引起了图像视觉研究者们的极大兴趣。风景图像往往记录了旅游者的旅游信息，为旅游者留下美好的回忆，当人们看到这些图像，就能联想到其地理信息，达到一图胜千言的效果；另外，这些风景名胜图像具有地理位置的唯一性，能为观赏者提供明确的地理方位，或者，为哪些询问者提供地理方位信息。从观赏的角度，人们不仅仅满足于对 2D 图像的浏览，特别是对国内外著名的名胜古迹，人们更愿意身临其境，要达到这样的效果需要对这些名胜古迹进行 3D 重建；利用网络上丰富的视觉资源，特别是名胜古迹图像资源进行 3D 重建是当前学术界的研究热点。无论是基于内容的地理信息检索还是利用网络图像进行 3D 重建，一个先决条件必须从网络上搜集足够多足够好的地标图像。本文的研究正是在这样一个需求的驱动下，利用多个搜索引擎基于关键词搜索地标图像，在这样一个网络隐监督条件下研究地标图像的搜集与标记。其研究内容主要包

括：

1）地标图像的特征表示；

2）地标图像集的分层组织；

3）地标图像的视觉一致性和几何一致性验证；

4）地标图像的标记。

本文运用计算机视觉及模式识别的理论与方法，探讨解决网络图像结构化组织、目标定位等问题，其研究具有科学意义，以及广泛的应用价值。地标图像作为一个地方的标志与广告，其检索与 3D 显示对旅游业具有很大的吸引力。本文的研究还可拓展到网络购物等系统中，存在潜在的商业应用前景。

地标图像虽是一特殊目标，但因相机的质量、拍摄者的技术与意愿或者天气等因素，使得同一目标，其视觉外观差异性极大，与一般类目标有着同等处理难度。其困难如下：

1）利用检索得到的地标图像集含有大量噪声图像。通过网络搜索获得的图像集属于非结构性数据集，“噪声”是其中不可避免的，如图 1.1 所示，这是我们根据关键词“自由女神像”利用 Google 图像搜索引擎检索得到的某一页图像显示，其中错误图像超过三分之一，说明当前的关键词检索会引入大量噪声图像。这些噪声图像不仅不包含我们所需要的信息，而且会带来误导；因此，必须尽可能多、并且准确地去除这些噪声图像。

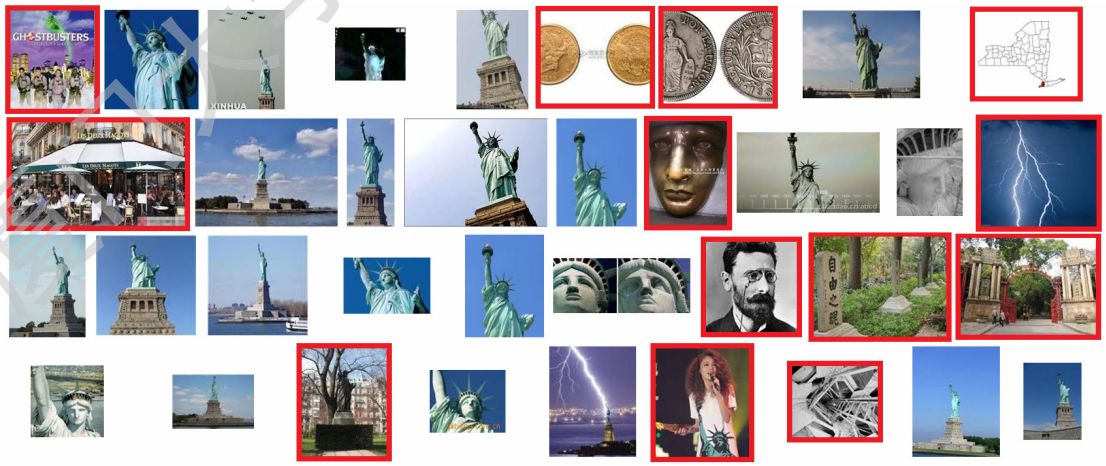

图1.1 “自由女神像”某一搜索页的图像显示，其中红框标注的图像为错误检索图像，共 37 幅图像，其中 13 幅错误

2）地标图像集中包含了多视角、多尺度的地标。因为我们使用的图像库是