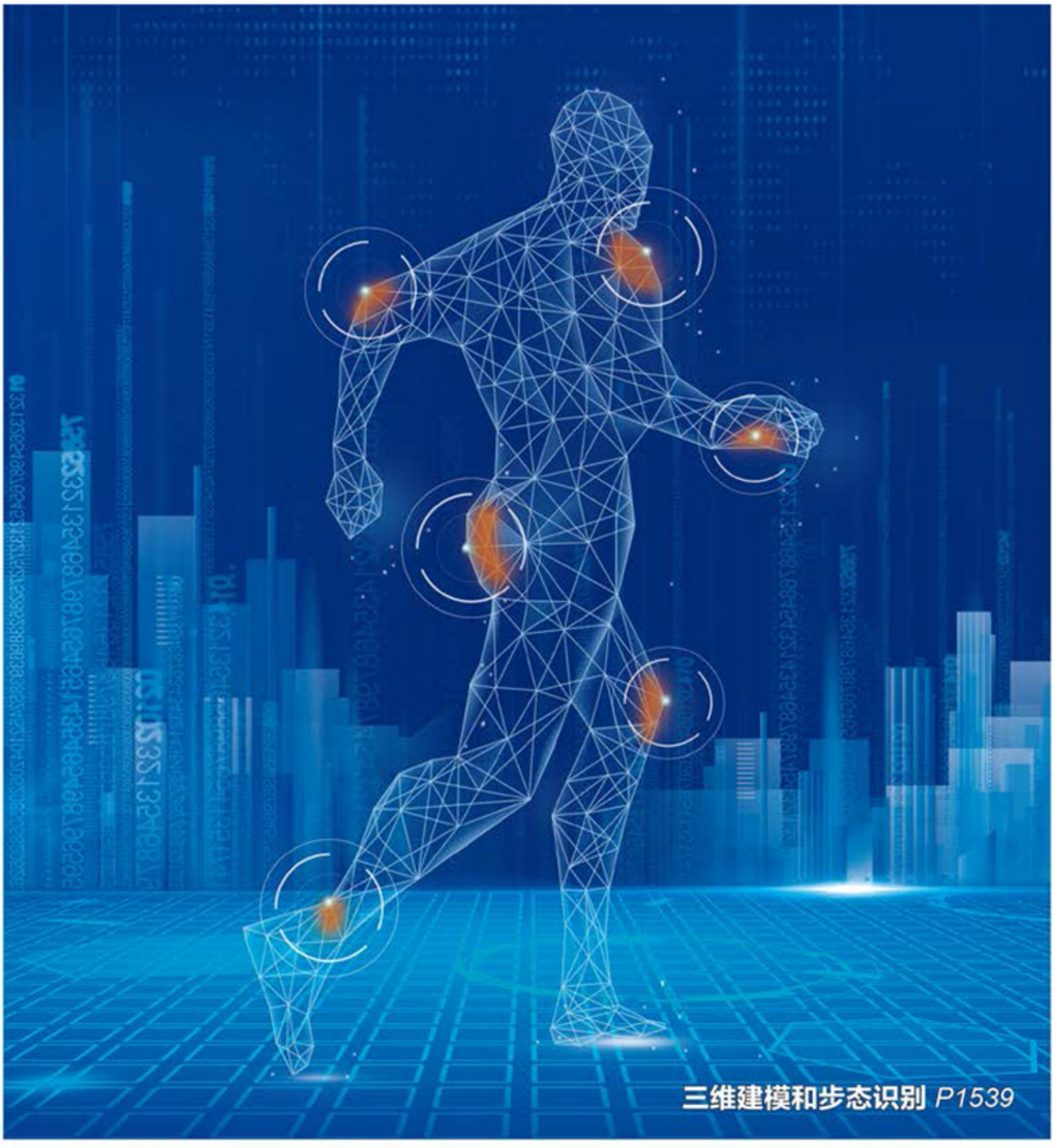
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
08
VOL.25

ISSN1006-8961
CN11-3758/TB



三维建模和步态识别 *P1539*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第8期（总第292期）
2020年8月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

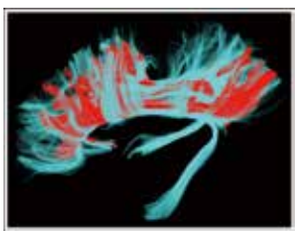
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

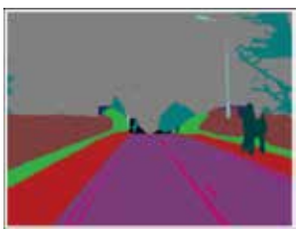
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年8月 第25卷 第8期（总第292期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

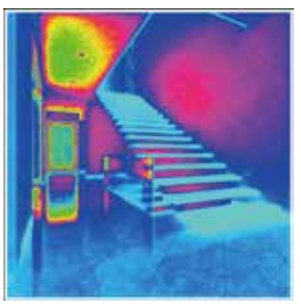
目次



神经纤维跟踪算法研究进展(第1513页)



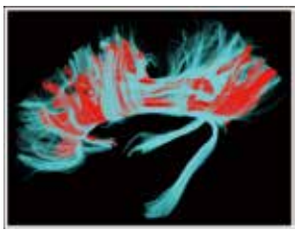
结合注意力机制的双路径语义分割(第1627页)



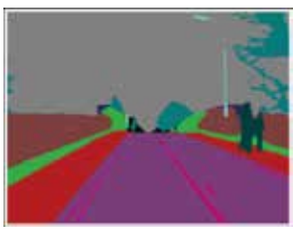
突变策略下多通路Metropolis光照传播(第1658页)

Volume 25, Number 8
Published August 16, 2020

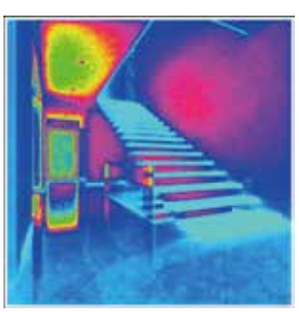
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Research progress of neural fiber tracking(P1513)



Two-path semantic segmentation algorithm combining attention mechanism(P1627)



Improved multiplexed Metropolis light transport based on fusion mutation strategy(P1658)

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1006-8961(2020)08-1578-13

论文引用格式:Chen Q, Liu L, Fu X D, Liu L J and Huang Q S. 2020. Fine-grained shoe image retrieval by part detection and semantic network. Journal of Image and Graphics,25(08):1578-1590(陈前,刘骊,付晓东,刘利军,黄青松. 2020. 部件检测和语义网络的细粒度鞋类图像检索. 中国图象图形学报,25(08):1578-1590)[DOI:10.11834/jig.190467]

部件检测和语义网络的细粒度鞋类图像检索

陈前[1],刘骊[1,2],付晓东[1,2],刘利军[1,2],黄青松[1,2]

1. 昆明理工大学信息工程与自动化学院,昆明　650500;　2. 云南省计算机技术应用重点实验室,昆明　650500

摘　要:**目的**　细粒度图像检索是当前细粒度图像分析和视觉领域的热点问题。以鞋类图像为例,传统方法仅提取其粗粒度特征且缺少关键的语义属性,难以区分部件间的细微差异,不能有效用于细粒度检索。针对鞋类图像检索大多基于简单款式导致检索效率不高的问题,提出一种结合部件检测和语义网络的细粒度鞋类图像检索方法。**方法**　结合标注后的鞋类图像训练集对输入的待检鞋类图像进行部件检测;基于部件检测后的鞋类图像和定义的语义属性训练语义网络,以提取待检图像和训练图像的特征向量,并采用主成分分析进行降维;通过对鞋类图像训练集中每个候选图像与待检图像间的特征向量进行度量学习,按其匹配度高低顺序输出检索结果。**结果**　实验在UT-Zap50K数据集上与目前检索效果较好的4种方法进行比较,检索精度提高近6%。同时,与同任务的SHOE-CNN(semantic hierarchy of attribute convolutional neural network)检索方法比较,本文具有更高的检索准确率。**结论**　针对传统图像特征缺少细微的视觉描述导致鞋类图像检索准确率低的问题,提出一种细粒度鞋类图像检索方法,既提高了鞋类图像检索的精度和准确率,又能较好地满足实际应用需求。

关键词:细粒度图像检索;鞋类图像;部件检测;语义网络;特征向量;度量学习

Fine-grained shoe image retrieval by part detection and semantic network

Chen Qian[1], Liu Li[1,2], Fu Xiaodong[1,2], Liu Lijun[1,2], Huang Qingsong[1,2]

1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; 2. Computer Technology Application Key Lab of Yunnan Province, Kunming 650500, China

Abstract: **Objective** Fine-grained image retrieval is a major issue in current fine-grained image analysis and computer vision. Traditional methods typically retrieve similar replicated images, which are primarily based on large-scale coarse-grained retrieval but with low precision. Fine-grained image retrieval belongs to fine-grained image identification and retrieval subclasses. The traditional image retrieval task extracts only the coarse-grained features of images and cannot be effectively used for fine-grained retrieval. It also lacks key semantic attributes, and this deficiency brings difficulty in distinguishing the nuances among parts. The difficulty in fine-grained image retrieval is that the traditional coarse-grained feature extraction cannot represent images effectively. Fine-grained images of the same subclasses also cause a significant difference due to such factors as shape, posture, and color; consequently, search results cannot be effectively applied to actual needs. Compared with conventional image analysis problems, fine-grained image retrieval is more challenging due to the inter-level subcategories of its smaller class differences and the class differences within the larger ones. A fine-grained image retrieval method by part detection and semantic network for various shoe images is therefore proposed to solve the

收稿日期:2019-09-09;**修回日期**:2020-01-07;**预印本日期**:2020-01-14

基金项目:国家自然科学基金项目(61862036, 61962030, 81860318);云南省中青年学术和技术带头人后备人才培养计划项目(201905C160046);云南省应用研究基础计划面上项目(2017FB097)

Supported by: National Natural Science Foundation of China (61862036, 61962030, 81860318)

above-mentioned problems. **Method** First, part-based detection is conducted to detect undetected shoe images through an annotated training dataset of shoe images. Second, the semantic network is trained based on the semantic attributes of the detected shoe and training images, and feature vectors are extracted. Third, principal component analysis is used for dimensionality reduction. Finally, the results are implemented and output by metric learning to calculate the similarity among images, and fine-grained image retrieval is implemented. On the UT-Zap50K dataset, fine-grained shoe attributes are defined for the shoe images in combination with the component area of shoes. The toe area defines a shape attribute that contains five attribute values. Two attributes of shape and height are defined for the heel area, which contains 13 attribute values. A height attribute is defined for the upper area, which contains four attribute values. A closed-mode attribute is defined for the upper area, which contains nine attribute values. Footwear global properties are defined to include colors and styles, which contain 20 attribute values. **Result** The experiment is compared with four methods with good retrieval performance on the UT-Zap50K dataset. The retrieval accuracy is improved by nearly 6%. Compared with the semantic hierarchy of attribute convolutional neural network (SHOE-CNN) retrieval method of the same task, the proposed method has higher retrieval accuracy. The proposed semantic network is compared with traditional GIST (generalized search trees) features, the linear support vector machine (LSVM) method, and the deep learning method to illustrate the effectiveness of the proposed retrieval method. The performance is evaluated in terms of the accuracy of top-K retrieval. Results show that the method based on deep learning is much better than the traditional GIST features and LSVM method. The retrieval accuracy of this method is better than that of the metric network and SHOE-CNN by combining a metric learning algorithm. **Conclusion** A fine-grained shoe image retrieval method is proposed to address the low accuracy of shoe image retrieval caused by the lack of fine visual description of traditional image features. The method can accurately detect different parts of a shoe image and define the detailed semantic attributes of the shoe image. The visual attribute features of the shoe image are obtained by training the semantic network. The problem of unsatisfied accuracy of shoe image retrieval caused by using only coarse-grained features to represent images is solved. The experimental results show that the proposed method can retrieve the same image as the image to be detected on the UT-Zap50K dataset. The accuracy can reach 80% and 86% while ensuring the running efficiency. However, this method exhibits shortcomings. On the one hand, the accuracy of partial image detection is low because of the many styles and complexity of shoes. On the other hand, the prediction of some semantic attributes is inaccurate, and the fine-grained semantic attributes of shoe images are imperfect. The follow-up work will focus on these issues to improve the search accuracy, and the application issues will be extended to different scenarios.

Key words: fine-grained image retrieval; shoe image; part detection; semantic network; feature vector; metric learning

0 引 言

图像检索(Zhou 等,2010)以检索信息的形式,分为“以文搜图”和“以图搜图”。线上购物的便利性满足了用户采购的需求,但仅文字介绍和图片展示不能为用户提供良好的购物体验,“以图搜图”逐渐成为网络购物方式的主流。传统图像检索任务一般是检索类似复制的图像,且大多基于大类的粗粒度检索,从而导致准确率较低。而细粒度图像检索(Xie 等,2015)是要对细粒度级别图像中的子类进行定位、识别及检索,广泛应用于图像检索和细粒度图像分析领域。然而,目前细粒度图像检索的难点在于:1)由于图像的粒度划分十分细微,导致提取的粗粒度特征不能有效表示图像;2)属于相同子类的细粒度图像本身也会因形态、姿势、颜色等因素形成巨大差异,从而使得检索结果不能有效适用于实际需求。因此,相较于传统图像分析问题,细粒度图像检索由于其子类别间较小的类间差异和较大的类内差异,更具挑战。

由于鞋类产品本身款式、材质、颜色、形状复杂多样,以鞋头形状为例,可细分为鱼嘴型、空腹型、圆型、尖角型和方型等。区别于通用图像,鞋类图像所属类别的粒度更为精细,且其语义属性多且复杂,难以区分部件间细微的属性差异,属于不同类别的鞋类图像在视觉上彼此相似。仅选取粗粒度的特征不能有效地进行细粒度检索,需要为其标注更多的语义属性,才能对复杂的鞋款式、材质、颜色和纹理等类别实现有效的细粒度检索。已有的鞋类图像检索方法主要有词包法(Li 等,2013)、基于视觉特征

(Huang 等,2014)以及深度学习(Zhan 等,2017)等方法,已有方法存在以下问题:1)检索对象局限于粗粒度,主要集中于检索同一类别的图像,忽略了同一类别下较细粒度级别的不同子类类别间的细微差异,导致检索效果不够准确;2)选取粗粒度的特征不能对图像不同子类类别间的细微差异进行有效表示。综上,仅依靠已有的鞋类图像检索方法无法有效实现鞋类图像的细粒度检索。

目前,鞋类图像细粒度检索尚有 3 大问题亟需解决:1)由于鞋类图像款式种类繁多,类别间的属性差异细微,导致难以利用区分性高的局部区域进行精确的细粒度检索。2)缺少关键的语义属性或属性划分不精确等,此外,针对鞋类图像的这些属性,还需添加特定的语义标签以提高后续的检索准确率。3)已有的检索方法主要针对鞋类图像的粗粒度检索及选取的特征不能有效表示不同子类类别间的细微差异,导致难以有效描述细粒度特征。因此,针对以上问题,本文提出结合部件检测和语义网络的鞋类图像细粒度的检索方法。首先采用 DPM(deformable parts model)部件检测方法对待检图像进行检测,再结合部件区域定义的细粒度语义属性训练语义网络得到待检图像的特征向量,最终通过度量学习实现细粒度图像检索。

本文方法流程如图 1 所示,首先采用 DPM 方法实现部件检测,然后结合语义属性训练语义网络得到待检图像的特征向量,最后通过度量学习进行相似度匹配并输出检索结果。本文的主要贡献有:

1)通过部件检测有效定位鞋部件区域并找出具有区分性的关键信息,这些关键信息能够有效区分鞋类图像同一类别下不同子类别之间的细微差异,提高了后续检索的准确率。

2)定义能描述不同鞋图像子类的语义属性,并结合检测后的鞋部件区域训练语义网络,以获得细粒度特征,既减少和降低了数据存储和计算成本,又提高了后续检索的精度。

3)采用主成分分析(principal components analysis,PCA)降低所选特征的维数,通过对鞋类图像训练集中每个候选图像与待检图像间的特征向量进行度量学习,在保证效率的同时,实现了鞋类图像细粒度检索。

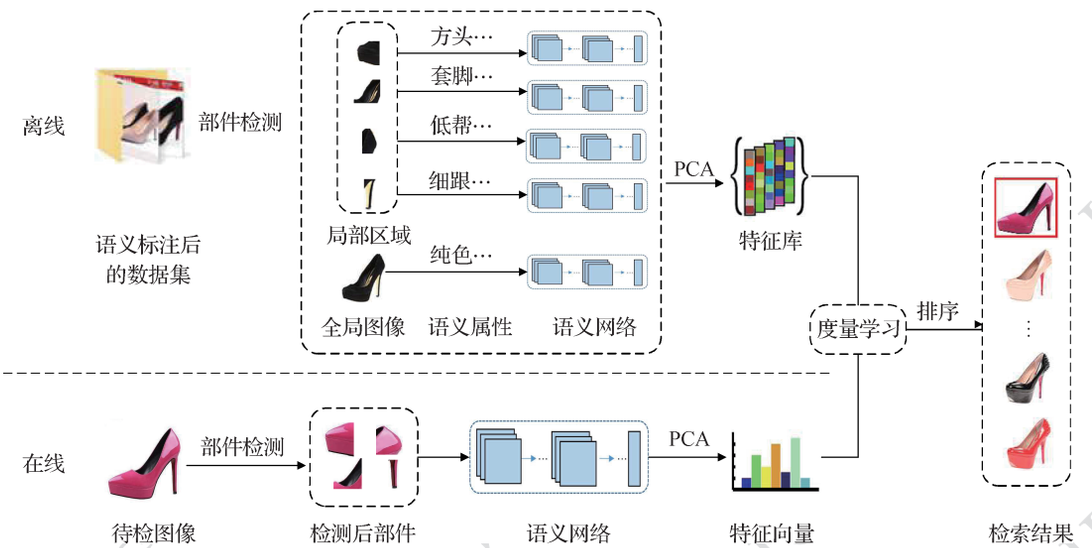


图 1 鞋类图像细粒度检索方法流程图

Fig. 1 Framework of fine-grained shoe image retrieval

1 相关工作

传统的粗粒度图像检索方法(Bourdev 等,2011;Guillaumin 等,2009)主要对类别级的图像检索进行研究,尤其在电子商务领域的产品图像检索中较注重如颜色、纹理和形状等底层特征。通常先提取 SIFT(scale invariant feature transform)和 HOG(histogram of oriented gradient)等特征,再通过特征学习获得相似度模型以实现检索。然而,上述方法在很大程度上受到人造图像特征表现力的限制,不能有效区分同一类别中图像间的细微差异。Li 等

人(2013)提出了一种结合全局特征和局部特征的匹配方法,以交互方式精确检索项目图像,一定程度上提高了检索的性能,但仍局限于粗粒度的图像检索。Xie 等人(2015)提出细粒度图像“搜索”的概念,通过构造一个层次数据库将多种现有的细粒度图像数据集和传统图像检索融合。该方法虽然提高了细粒度图像检索精度,但提取的底层特征不能较好地对图像的局部细节进行描述。区别于传统图像检索的深度学习方法,Wei 等人(2017)首次提出卷积特征优于全连接层特征,同时选择卷积描述子进行细粒度图像检索,但由于未考虑图像的语义信息,导致细粒度检索准确率不高。He 等人(2014)利用基于区域的卷积神经网络(region-based convolutional neural networks,R-CNN)算法对细粒度图像进行物体级别与其局部区域的检测,再分别训练 CNN(convolutional neural networks),学习该物体/部位的特征,最终将全连接层特征级联作为整幅细粒度图像的特征表示,该方法在识别细粒度图像类别上取得了较好的效果,但在一定程度上忽略了区域检测和细粒度特征学习的关联性。李振东等人(2018)提出一种基于深度卷积特征的快速人脸图像检索算法,先构建高效的卷积特征向量进行人脸初步检索,然后对待检人脸图像的卷积特征向量均值融合后再次执行检索,该算法显著地提高了检索结果的准确率,但未考虑语义属性对检索结果的影响。针对粗粒度的特征不能较好表示细粒度图像的问题,受 Wei 等人(2017)启发,本文提出面向鞋类图像检索的语义网络。与该方法不同,本文将鞋类图像中的各部件语义属性与提取到的细粒度特征相结合,更好地区分图像子类别间细微的局部差异。

由于细粒度图像的细微差异仅体现在局部的部件区域,如何有效地对细粒度图像进行检测,并找出具有区分性的部件区域成为提高后续检索准确率的关键。针对目标检测,吴圣美等人(2019)采用 k-poselet($k>1$)可变形部件模型分别对少数民族服装人体的各个独立的 poselet 进行检测,一定程度上提高了检测的准确率。Felzenszwalb 等人(2008)采用多尺度可变形部件混合模型(deformable part model,DPM)对人体部位进行检测,能有效定位不同的人体部位并获取相关信息。Cao 等人(2019)提出采用多尺度卷积特征融合的(faster region-based CNN,Faster R-CNN)算法,在小物体上获得了较好的检测效果。

Kovashka 等人(2012)开发了一种利用相对属性成对比较的鞋推荐系统。Huang 等人(2014)基于手工部件标注,通过在部件检测和检索过程中集成属性,实现鞋类图像检索。Kiapour 等人(2015)采用类别独立度量网络来评估图像对的相似性,从而实现了精确的街拍产品检索。Zhan 等人(2017)提出了语义层次的属性卷积神经网络(semantic hierarchy of attribute convolutional neural network,SHOE-CNN),该模型具有 3 层特征表示,用于区分鞋子特征表达和高效检索,并通过计算损失函数减少外观属性间的预测误差。上述方法对粗粒度鞋类图像检索均具有较高的准确率,但存在视觉属性特征利用较少和检索精度较低的问题。主要原因在于:1)未考虑部件检测获得的部件区域的合理性,是否都对检索有效、不可缺少;2)仅通过部件检测和属性标注很难获得细粒度图像检索中所需的细节视觉描述。针对现有的细粒度图像检索较少利用视觉属性特征的问题,受 Zhang 等人(2014)方法的启发,本文在 AlexNet(Krizhevsky 等,2012)基础上,在 FC1 层之后添加树状结构的全连接层捕获属性信息,同时定义鞋头、鞋跟、鞋面和鞋帮 4 个部件的语义属性以有效区分鞋类图像子类别间的细微差异。区别于该方法,本文方法侧重鞋类图像本身的细粒度属性。

2 部件检测

DPM 部件检测对于检测形态、视点变化较大的物体具有精度高、效率快的优势,本文将其应用到鞋类图像检测上,有效定位鞋部件区域。

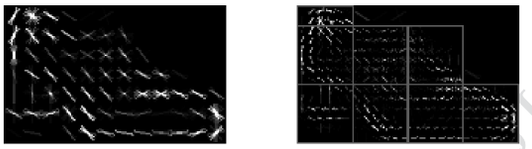
提出的鞋部件检测的具体步骤如下:

1)提取 DPM 特征。对输入的待检鞋类图像 $\boldsymbol{I}$ 以及从鞋类图像库中输入标注后的训练集 $\boldsymbol{S}$,分别计算梯度方向直方图(histogram of oriented gradient,HOG)特征并进行归一化、截断及降维,得到 DPM 特征。

2)鞋部件检测模型。给定训练集 $\boldsymbol{S}$,正样本为 $\boldsymbol{M}=\{(\boldsymbol{I}_1,B_1),\cdots,(\boldsymbol{I}_n,B_n)\}$,$\boldsymbol{I}$ 是图像,B 是目标,从正样本 $\boldsymbol{I}$ 中对目标 B 进行标注得到检测框,并将覆盖整个目标分辨率为 8×8 的根滤波器设为根模型,随机从框内选取与根模型相同大小的区域作为训练集,同时选择较小样本的平均尺寸 130×90 作

为每一类正样本的尺寸。负样本是 $\boldsymbol{N}=\{(\boldsymbol{J}_1,\cdots,\boldsymbol{J}_m)\}$，$\boldsymbol{J}$ 为图像，采用 SVM(support vector machine)训练得到根模型，并用 6 个矩形框尽可能覆盖到根模型。通过双线性插值将这些矩形框内的参数的超分辨率增加到原来的两倍，得到初始化后的各部件模型，并固定其大小和相对部件模型的位置。

鞋部件检测模型包含一个根模型和若干个部件模型，如图 2 所示。含 n 个部件的模型可以通过根滤波器 F_0 和一系列部件模型 $(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n)$ 来定义，其中 $\boldsymbol{P}_i=(F_i,\boldsymbol{v}_i,\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{b}_i)$。$F_i$ 是第 i 个部件的滤波器；$\boldsymbol{v}_i$ 和 $\boldsymbol{s}_i$ 都是 2 维向量，$\boldsymbol{v}_i$ 指明第 i 个部件位置的矩形中心点相对于根位置的坐标，$\boldsymbol{s}_i$ 是此矩形的大小；$\boldsymbol{a}_i$ 和 $\boldsymbol{b}_i$ 也都是 2 维向量，指明一个二次函数的参数，此二次函数用来对第 i 个部件的每个可能位置进行评分。



(a) 根模型　　(b) 部件模型

图 2　鞋部件检测模型

Fig. 2　The part detection model for shoes

((a) root model; (b) part model)

3)响应得分计算。模型在 HOG 金字塔中的位置可以用 $\boldsymbol{z}=(\boldsymbol{p}_0,\cdots,\boldsymbol{p}_n)$ 来表示，当 $i=0$ 时，$\boldsymbol{p}_i=(x_i,y_i,l_i)$ 表示根滤波器的位置；$i>0$ 时，$\boldsymbol{p}_i=(x_i,y_i,l_i)$ 表示第 i 个部件滤波器的位置。空间位置的得分等于每个部件滤波器的得分加上每个部件的位置相对于根的得分。其响应得分计算为

$$\sum_{i=0}^{n}F_i\cdot\phi(H,\boldsymbol{p}_i)+\sum_{i=1}^{n}(\boldsymbol{a}_i\cdot(\tilde{x}_i,\tilde{y}_i)+\boldsymbol{b}_i\cdot(\tilde{x}_i^2,\tilde{y}_i^2)) \tag{1}$$

式中

$$(\tilde{x}_i,\tilde{y}_i)=\frac{(x_i,y_i)-[2(x_0,y_0)+\boldsymbol{v}_i]}{\boldsymbol{s}_i} \tag{2}$$

式中，$\phi(H,\boldsymbol{p}_i)$ 为滤波器的空间位置得分，(x_0,y_0) 是根滤波器在其所在层的坐标，为了统一到部件滤波器所在层需将其乘以 2。$\boldsymbol{v}_i$ 是部件 i 相对于根的坐标偏移，$2(x_0,y_0)+\boldsymbol{v}_i$ 表示未发生形变时部件 i 的坐标，$(\tilde{x}_i,\tilde{y}_i)$ 表示部件 i 的变形程度，$\tilde{x}_i^2,\tilde{y}_i^2$ 为变形特征，$\boldsymbol{a}_i$ 和 $\boldsymbol{b}_i$ 为 2 维向量，指明一个二次函数的参数，此二次函数用来对第 i 个部件的每个可能位置进行评分。

空间位置 $\boldsymbol{z}$ 的得分 $f_\beta(z)$ 表示为 $\boldsymbol{\beta}\cdot\boldsymbol{\psi}(H,z)$，即

$$\begin{aligned}\boldsymbol{\beta}&=(F_0,\cdots,F_n,\boldsymbol{a}_1,\boldsymbol{b}_1,\cdots,\boldsymbol{a}_n,\boldsymbol{b}_n)\\\boldsymbol{\psi}(H,z)&=(\phi(H,p_0),\phi(H,p_1),\cdots,\phi(H,p_n)\\&\tilde{x}_1,\tilde{y}_1,\tilde{x}_1^2,\tilde{y}_1^2,\cdots,\tilde{x}_n,\tilde{y}_n,\tilde{x}_n^2,\tilde{y}_n^2)f_\beta(z)=\\&\max_z\boldsymbol{\beta}\cdot\boldsymbol{\psi}(H,z)\end{aligned} \tag{3}$$

式中，$\boldsymbol{\psi}(H,z)$ 为 HOG 金字塔 H 中模板 Z 对应的特征向量，$\boldsymbol{\beta}$ 为模型参数向量，$\tilde{x}_n,\tilde{y}_n,\tilde{x}_n^2,\tilde{y}_n^2$ 为变形特征，$f_\beta(z)$ 通过对每个空间位置的综合响应得分计算，最终实现鞋类部件检测，得到检测后的待检索图像 $\boldsymbol{I}'$ 和数据集 $\boldsymbol{S}'$。

3　语义网络

区别于 AlexNet 网络，本文增加了全连接层树型结构，并结合定义的语义属性以获得不同鞋类部件对应的高维特征。

3.1　语义属性

Kovashka 等人(2012)针对鞋的款式定义了“高跟”、“运动型”和“舒适型”等 10 个属性，其属性在局部性和粒度方面不够精细。为了更好地描述鞋类图像，结合部件检测后的鞋部件区域，进一步定义了鞋类图像的细粒度语义属性。鞋头区域定义了形状属性，含 5 个属性值；鞋跟区域定义了形状和高度 2 个属性，含 13 个属性值；鞋帮区域定义了高度属性，含 4 个属性值；鞋面区域定义了闭合方式属性，含 9 个属性值。此外，还定义了鞋类全局属性(包括颜色、样式两个属性)，含 20 个属性值，见表 1。

以检测后的待检索图像 $\boldsymbol{I}'$ 为例，$\{l_i=(x_i,y_i)\}_{i=1}^{N}$ 是 $\boldsymbol{I}'$ 中鞋部件 N 的位置，$\{\boldsymbol{\alpha}_i\}_{i=1}^{M}$ 对应图像 $\boldsymbol{I}'$ 中 M 个属性。$\boldsymbol{\alpha}_i=[\alpha_i^1,\cdots,\alpha_i^K]$ 为每个属性与之对应的属性值，其中 K 是属性 $\boldsymbol{\alpha}_i$ 的值的数量。通过定义函数 $i_\alpha=f_\alpha(i_p)$ 获得鞋部件和属性之间的映射关系，其中 i_p 表示部件的索引。给定部件的索引 i_p，可通过 $f_\alpha(i_p)$ 找出它的相应属性。

3.2　网络结构

提出的语义网络结构如图 3 所示，网络包含 5 个卷积层和两个全连接层 FC1，FC2 及 FC2 分支出的树型结构层 FC3。由于在鞋类图像中，属性通常是指某些部件(如鞋头形状、鞋跟高度)或鞋(如鞋

表 1　鞋类语义属性

Table 1　The semantic attributes of shoes

区域	属性	属性值
全局	颜色	红　橘　黄　绿　蓝　紫　黑　白　灰　棕　混色
	样式	网状　条纹　斑点　手绘　纯色　拼色　涂鸦　花卉　卡通
鞋头	形状	鱼嘴型　空腹型　圆型　尖角型　方型
鞋跟	形状	平底　厚平底　坡跟　猫跟　细跟　锥形跟　粗跟　马蹄跟　方跟
	高度	恨天高　高　中　低
鞋帮	高度	超长筒　长筒　中筒　低筒　超低筒
鞋面	闭合方式	套脚　系带　搭扣　魔术贴　拉链　套筒　一字扣带　交叉绑带　T 字绑带

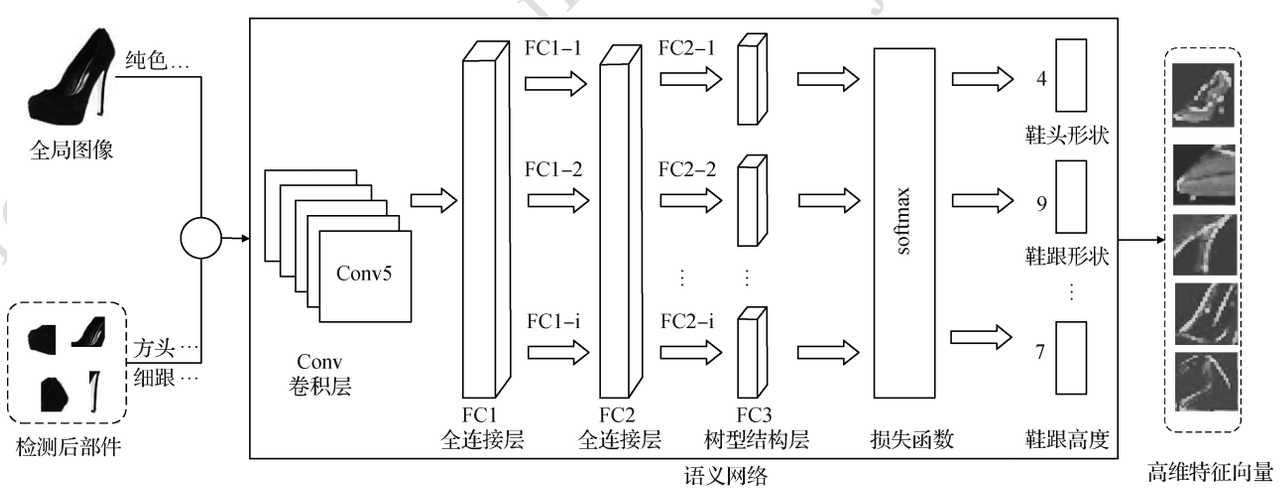


图 3　语义网络结构图

Fig. 3　Structure of semantic network

颜色、鞋样式）的具体描述，作为视觉外观的补充，该信息可用于实现鞋类检索的强大语义表示。为了在语义层次上表示鞋，本文在 FC2 层后增加了一个全连接层的树型结构 FC3 来实现语义属性的嵌入，

以捕获属性信息。具体地,对于每个属性,全连接层的树型结构 FC3 根据表 1 中与之对应的属性值的数量分支出几个单元。以表 1 中鞋跟区域的属性为例,形状属性含 9 个值,则 FC3 分支出 9 个单元,表示具有相应属性值的概率,高度属性含 4 个值,则 FC3 分支出 4 个单元,表示具有相应属性值的概率。

从鞋类图像数据集中定义的语义属性大多属于多类属性类型(即取多个实数值),因此本文使用多分类归一化损失函数(softmax),该函数定义为

$$P(i) = \frac{\exp(\boldsymbol{\theta}_i^{\mathrm{T}}\boldsymbol{x})}{\sum_{k=1}^{k}\exp(\boldsymbol{\theta}_k^{\mathrm{T}}\boldsymbol{x})} \tag{4}$$

式中,$P(i)$ 有多个值,所有值加起来刚好等于 1,每个输出都映射到了[0,1]区间,可将其视做概率问题,$\boldsymbol{\theta}_k^{\mathrm{T}}\boldsymbol{x}$ 对应多个输入。

以表 1 中的鞋头形状属性 a_i 为例,其余属性遵循相同的过程。训练集 $\boldsymbol{S}$ 包含 N 个图像 $\boldsymbol{x}_i$,$i \in \{1,2,\cdots,N\}$,相应的属性标签表示为 $y_i \in \{1,2,\cdots,K\}$,总共有 K 个分类,其中 K 是指定属性 a_i 的值的数量。设 $h_j^{(i)}(j = 1,2,\cdots,K)$ 表示来自图 3 中最后一个完全连接层(FC3)的节点 j 的激活值,则给定训练样本 $\boldsymbol{x}_i$ 属于 j 类的概率可表示为 $P_j^{(i)}$,计算为

$$P_j^{(i)} = \frac{\exp(h_j^{(i)})}{\sum_{g=1}^{K}\exp(h_g^{(i)})} \tag{5}$$

给定 softmax 损失层 $P_j^{(i)}$ 的概率,softmax 的代价函数定为

$$J(h) = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{K}1(y^{(i)} = j)\log P_j^{(i)}\right] \tag{6}$$

式中,示性函数 $1(j^{(i)} = j)$ 表示如果第 i 个样本的类别为 j,则 $y_{ij} = 1$。代价函数可看做最大化似然函数,即最小化负对数似然函数。一般使用梯度降优化算法来最小化代价函数,而其中会涉及偏导数,即 $h''_j = h_j - \alpha\delta_{h_j}J(h)$,则 $J(h)$ 对 h_j 求偏导,得

$$\frac{\partial J_0}{\partial h_j^{(i)}} = -\frac{1}{N}\sum_{i=1}^{N}x_i\left[1\{y^{(i)} = j\} - p(y^{(i)} = j \mid x_i;h)\right] \tag{7}$$

式中,δ 是指标函数,α 是正则化参数。通过对全连接层的树型结构 FC3 训练 softmax 损失函数,得到图像 $\boldsymbol{x}_i$ 属于完全连接层 FC3 的节点 j 类的概率 $P_j^{(i)}$,根据概率 $P_j^{(i)}$ 完成对属性 a_i 的多分类。

3.3 特征向量

仅获取整个图像的全局信息不能有效区分不同部件区域间的差异,通过输入两个不同层次的鞋类图像训练语义网络来获得全局和局部的图像视觉属性特征,见图 3。

给定待检鞋类图像,结合全局图像的样式、颜色属性作为语义网络第 1 层次的输入,通过训练语义网络获取整个图像的特征向量。同时为了有效地区分局部部件区域间的差异,将检测后的待检图像 $\boldsymbol{I}'$ 以及对应各个部件的语义属性作为语义网络第 2 层次的输入,获取局部部件区域的特征向量。具体地,输入两个不同层次的鞋类图像 $\boldsymbol{I}$ 和 $\boldsymbol{I}'$,首先将图像调整为 136 × 102 像素并将其输入至前馈网络并从最后一个卷积层(Conv5)中提取特征。然后,将提取得到的特征输入至两个全连接层,结合树形结构层 FC3 多个分支对应输入的语义属性,对 FC3 层训练 softmax 损失函数,得到部件间具有区分性的视觉属性特征。本文在 Conv5 特征图上应用了 2 级金字塔均值池(2 × 2,1 × 1),大小为(6 × 6)生成两个特征输出,其大小分别等于 2 × 2 × 256 = 1 024 组和 1 × 1 × 256 = 256 组。然后,在空间金字塔池(spatial pyramid pooling,SPP)之后获得 1 024 + 256 = 1 280 维的特征向量。本文通过组合 Conv5 和 FC3 层获得图像的特征表示:1 280 + 4 096 = 5 376 维的特征向量。待检鞋类图像 $\boldsymbol{I}$ 最终的特征表示为 5 376 + 1 024 = 6 400 维的特征向量。

4 细粒度图像检索

通过训练语义网络得到的待检图像的高维特征向量含有一定的噪声且会造成较大的算法计算开销。因此,本文采用主成分分析(PCA)降低所选特征的维数以获得更高的检索准确率。

4.1 PCA 降维

如图 4 所示,首先输入训练得到的鞋类图像的特征向量集 $\boldsymbol{D} = \{\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_n\}$,其中,训练得到的鞋类图像表示为 6 400 维的特征向量,对其进行 L2 归一化。然后,采用主成分分析将图像的 6 400 维特征向量降至 1 024 维。因此,每个鞋类图像由 L2 归一化后的 1 024 维特征向量描述,在减少数据存储和降低数据存储计算成本的同时也提高了

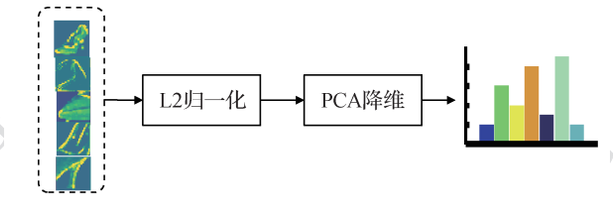


图 4　PCA 降维

Fig. 4　Principe component analysis for dimensionality reduction

后续检索的精度。

4.2　度量学习

给定一组表示为 $\boldsymbol{X}$ 的鞋类图像，每个图像可以由向量 $\boldsymbol{x}_i \in \mathbf{R}^d$ 表示，其中 $d=1\,024$。假设对于每个图像 $\boldsymbol{x}_i$，其相关图像和不相关图像分别表示为 $\boldsymbol{x}_i^+$ 和 $\boldsymbol{x}_i^-$。利用图像的三元组 $(\boldsymbol{x}_i,\boldsymbol{x}_i^+,\boldsymbol{x}_i^-)$，相似性度量任务的核心是学习映射函数 S，用于比较相关对 $(\boldsymbol{x}_i,\boldsymbol{x}_i^+)$ 与不相关对 $(\boldsymbol{x}_i,\boldsymbol{x}_i^-)$ 之间的相似性。由 W 参数化的相似度函数 S 表示为

$$S_W(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^{\mathrm{T}}W\boldsymbol{x}_j \tag{8}$$

式中，$\boldsymbol{x}_i \in \mathbf{R}^d$ 和 $\boldsymbol{x}_j \in \mathbf{R}^d$ 可以是集合 $\boldsymbol{X}$ 中图像的任意两个特征向量，W 为参数。

图像的三元组是由训练数据集中随机选一个样本 x_i，然后再随机选取一个属于同一类 $\boldsymbol{x}_i$ 的样本 $\boldsymbol{x}_i^+$ 和不同类的样本 $\boldsymbol{x}_i^-$ 构成，为了让 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_i^+$ 特征向量之间的距离尽可能小，将每个三元组的损失函数定义为

$$l_W^t(\boldsymbol{x}_i,\boldsymbol{x}_i^+,\boldsymbol{x}_i^-)=\max\{0,\gamma-S_W(\boldsymbol{x}_i,\boldsymbol{x}_i^+)+S_W(\boldsymbol{x}_i,\boldsymbol{x}_i^-)\} \tag{9}$$

式中，γ 是边缘常数，损失函数对相距较远的正匹配对和相距较近的不匹配对进行惩罚，最大限度减少训练阶段的全局损失 L_w 为

$$L_w=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_i^+,\boldsymbol{x}_i^-)\in\boldsymbol{X}} l_W^t(\boldsymbol{x}_i,\boldsymbol{x}_i^+,\boldsymbol{x}_i^-) \tag{10}$$

对于每个查询图像 $\boldsymbol{s}_i \in \boldsymbol{S}$，结合标注后的鞋类图像训练集计算其相似性得分，并根据相似性得分对图像进行排序，最后按其高低顺序输出检索结果。

5　实验结果与分析

5.1　实验数据集

实验选用 CPU Intel i9-9900k、GPU 2080ti 11 GB、内存 32 GB 的 PC 机的硬件平台以及 MATLAB 软件平台。

实验数据 UT-Zap50K（UT Zappos50K）包含全面注释的大型鞋类图像数据集（Yu 和 Grauman，2014）。该数据集由 50 025 个目录图像组成，多为白色背景，并以相同的方向描绘。为了考虑不同鞋类以及同种鞋类间的细微变化，使用了其中 40 000 幅图像作为训练集，10 000 幅图像作为待检索图像数据集，图像尺寸大小统一为 136 × 102 像素。针对 UT-Zap50K 数据集缺乏细粒度的标签，本文从 40 000 幅训练集图像中选取了 5 000 幅图像分不同的部件区域进行标注；为了便于后期实验，对不同类别的鞋进行了归纳和编号，并将图像统一设置成 BMP（Bitmap）格式。

数据集中包含各类鞋图像，款式有靴子、高跟鞋、休闲鞋、单鞋、凉鞋、帆布鞋、运动鞋、拖鞋、篮球鞋、雨鞋、工装鞋等，鞋属性包括样式、鞋跟高度、闭合方式、形状等。

数据集排名前 16 的不同类别鞋子数量分布如图 5 所示，可以看出，样式、属性较丰富的鞋类图像数量相对较多，如高跟鞋、休闲鞋和运动鞋；一些类别的鞋款式、属性较单一，如雨鞋、拖鞋和工装鞋，这些类别的鞋子数量相对较少。

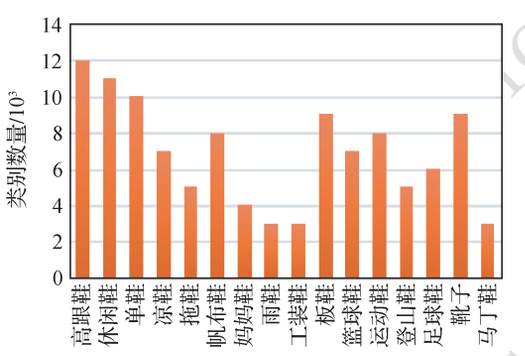


图 5　前 16 种鞋类别数量分布图

Fig. 5　Number distribution of the top 16 shoe categories

数据集中不同属性类别排名前 5 的数量分布如图 6 所示，样式属性数量位列前 5 的属性分别是网状、条纹、拼色、纯色和手绘；鞋头形状属性数量位列前 5 的属性分别是鱼嘴型、圆型、尖角型、方型和空腹型；鞋跟形状属性数量位列前 5 的属性分别是平底、细跟、坡跟、粗跟和方跟。本文数据集包含了各个类别的鞋类图像以及对应鞋不同部件的语义属性，便于后续进行检索实验。

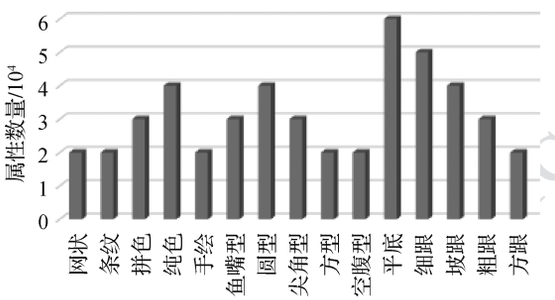


图 6　不同鞋类别中前 5 名属性数量分布图

Fig. 6　The quantity distribution of the top 5 attributes in different shoe categories

5.2　实验结果与性能分析

5.2.1　检测实验对比分析

以待检测的鞋类图像及训练集为输入，对检测结果进行统计分析。表 2 给出了待检测的鞋类图像及训练集部件检测结果。

表 2　鞋类图像检测结果实例

Table 2　Example of shoe image detection result

待检图像	检测结果	待检图像	检测结果
			
			
			

从准确率 P 和召回率 R 两个指标分析部件检测方法性能，计算为

$$P = \frac{h_i \cap q_i}{h_i \cup q_i} \quad R = \frac{f_i}{q_i} \tag{11}$$

式中，h_i 是待检测图像 $\boldsymbol{x}_i$ 中鞋所占的区域；q_i 是利用部件检测方法检测出的待检测图像 $\boldsymbol{x}_i$ 中鞋所占的区域；f_i 是待检测图像 $\boldsymbol{x}_i$ 被正确提取的鞋部件区域。本文的准确率和召回率均在像素级别上进行计算。准确率是指检测出的真实鞋区域的像素占检测出的鞋像素与真实鞋区域像素并集的比重，召回率是指检测出的正确鞋像素占算法检测出的总的鞋像素的比重。

为了验证本文方法的有效性，与目前传统方法中 Hosang 等人(2017)的方法以及基于深度学习的 Redmon 等人(2016)，Zhu 等人(2019)的方法进行比较和检索实验。结果如表 3 所示。

表 3　4 种检测方法的比较

Table 3　Comparison of four detection methods

检测方法	准确率 P	召回率 R
Hosang 等人(2017)	0.745 6	0.632 5
Redmon 等人(2016)	0.762 6	0.839 2
Zhu 等人(2019)	0.810 2	0.763 5
本文(OPM)	0.792 5	0.800 5

Hosang 等人(2017)采用非极大值抑制(Non-maximum suppression，NMS)算法对所有的检测框按得分进行排序，选出得分最大的检测框，其准确率 P 较高，召回率 R 较低。Redmon 等人(2016)提出的 YOLO(you only look once)算法直接对输入图像应用算法并输出类别和相应的定位，其 P 较 NMS 有一定提高。Zhu 等人(2019)提出的 ZSD(zero-shot object detection)算法将语义属性预测与视觉特征融合在一起，获取可见类和不可见类的对象边界框，在保证高准确率的同时其 R 有很大改进。鞋类图像是以不同的部件区域进行划分，而 DPM 是一种基于部件的检测算法，能有效地定位部件区域，便于寻找具有区分性的关键信息。结合 P 和 R 指标，本文采用的 DPM 方法的 P 相对较高，R 较好，检测效果更为理想。

5.2.2　属性预测对比分析

属性预测的目的是为了更好地区分部件间的差异，不同部件间属性的预测对应学习不同部件间的特征。本文基于平均精度(mean average precision，MAP)评估属性预测性能，表 4 是几种属性预测方法的结果。

Chen 等人(2012)利用传统特征(如 SIFT、LAB 空间中的颜色等)和 CRF(conditional random field)模型进行属性预测，在鞋头类型和鞋跟类型属性上表现出不错的性能。Ozeki 和 Okatani(2014)、Zhang 等人(2019)的方法是基于深度学习的方法，其平均精度优于传统方法，但没有在属性预测中结合具有区分性的部件区域信息。本文语义网络方法的大多数属性的预测精度优于对比方法。平均而言，使用语义网络的属性预测的 MAP 比传统 CNN 高 1.2%，这表明了语义网络的有效性。

表 4　4 种属性预测方法平均精度比较

Table 4　Comparison of four attribute prediction methods for average accuracy

/%

属性名	Chen 和 Okatani(2012)	Ozeki 和 Okatani(2014)	Zhang 等人(2019)	本文(DPM)
鞋头类型	71.00	80.37	81.12	**83.14**
鞋跟类型	70.12	72.96	74.23	**76.56**
鞋跟高度	62.52	64.57	65.40	**67.83**
鞋帮高度	68.24	71.18	72.65	**74.56**
闭合方式	64.32	71.45	73.02	**73.90**
样式	35.08	35.67	38.02	**39.09**

注:加粗字体表示最优结果。

5.2.3　检索对比实验

1)检索精度。为了评估所提方法的有效性,将提出的语义网络(semantic network)与传统的最先进的管道方法:具有 1 024 维的 GIST(generalized search trees)特征(Li 等,2013)、基于深度学习的 Chen 和 Deng(2019)、Yu 等人(2018)的方法以及同任务检索的 Huang 等人(2014)和 Zhan 等人(2017)的方法进行比较,以验证本文方法的有效性。根据 top-K 检索精度评估性能,如果 top-K 返回的结果包含与查询项完全相同的鞋,则认为它是成功的。

如图 7 所示,图中的曲线分别表示 GIST 特征、DeML(decoupled metric learning)、S^2GA(semantics-guided attention)以及用于同任务检索的 LSVM(linear support vector machine)、SHOE-CNN 等 6 种方法对应的前 36 个图像检索精度。结果表明,基于深度学习的方法相对于传统的GIST特征(Li等,2013)、LSVM 方法(Huang 等,(2014))有很大的改进,本文方法通过结合度量学习算法较于深度学习的 SHOE-CNN(Zhan 等,2017)、DeML 方法(Chen 和 Deng,2019)在检索精度上有了一定的提高,S^2GA 方法(Yu 等,2018)的检索精度虽稍高于本文方法,但检索效率较低,见表 5。通过进一步度量学习,使用具有相同特征的余弦相似性,本文方法的检索精度相对优于其他几种方法,表 6 中显示了示例检索结果。结果表明:1)输入待检图像为高跟鞋,鞋头区域对应鱼嘴型属性,鞋跟区域对应细跟属性,全局属性为白色,根据不同部件间视觉属性特征的相似度,检索得到与待检图像完全相同的鞋类图像;2)输入待检图像为运动鞋,主要语义属性为鞋面区域的系带、拼色属性,其子类别间的细微差异仅体现在拼色属性,结合其视觉属性特征的相似度,本文方法能有效地区分部件间的细微差异;3)输入待检图像为靴子,对应属性依次为高帮、长筒、超长筒,本文方法可有效地区分同一类别下不同子类之间的细微差异,排序最高的是与待检图像完全相同的鞋类图像。

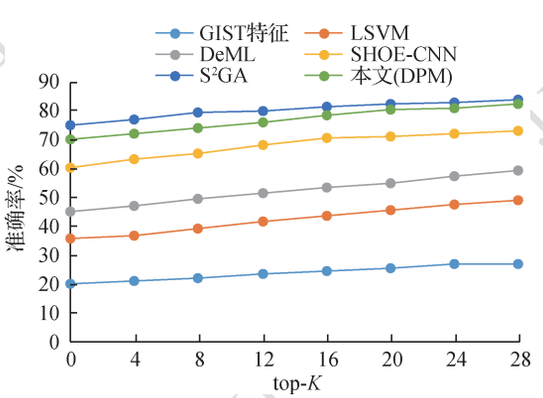


图 7　6 种检索方法检索精度对比

Fig. 7　Comparison of retrieval accuracy of six retrieval methods

表 5　6 种检索方法运行时间比较

Table 5　Running time comparison of six retrieval methods

检索方法	运行时间/s
Li 等人(2013)	0.123 8
Huang 等人(2014)	0.152 6
Zhang 等人(2014)	0.362 1
Zhan 等人(2017)	0.261 6
Yu 等人(2018)	0.302 5
本文(DPM)	0.221 8

2)检索效率。为了进一步验证本文方法的可应用性,针对本文方法以及对比方法进行运行效率分析,如表 7 所示,本文检索方法采用 DPM 部件检测相对于 NMS、YOLO 和 ZSD 算法在有效区分鞋部件区域的同时保证了检索效率。从表 5 可以看出,本文方法与较为传统的 GIST 特征(Li 等,2013)、LSVM 方法(Huang 等,2014)相比,虽然运行速度稍慢,但是其准确率有明显提升;相较于 Metric network(Zhang 等,2014)、SHOE-CNN(Zhan 等,2017)以及 S^2GA 方法(Yu 等,2018)运行速度较快,且准确率有一定的提高,由此可得本文方法具有一定的适用性。

表 6　检索结果实例

Table 6　The table of retrieval result

待检图像	检索结果	待检图像	检索结果
鱼嘴型		鱼嘴型	
系带、拼色		系带、拼色	
低靴		低靴	
长筒		超长筒	

表 7　4 种检测方法的检索效率对比

Table 7　Comparison of four detection methods on retrieval efficiency

检测方法	运行时间/s
Hosang 等人(2017)	0.253 8
Redmon 等人(2016)	0.376 3
Zhu 等人(2019)	0.462 1
本文(DPM)	0.221 8

3）为了更好地说明本文方法的性能，从准确率 P，召回率 R 两个方面对其他同类任务的方法 GIST 特征(Li 等，2013)、LSVM(Huang 等，2014)、SHOE-CNN(Zhan 等，2017)进行比较。召回率指的是所有“正确被检索的图像”占“所有应该检索到的图像”的比例，召回率是在像素级别上计算的，具体公式为 $R = r_i / s_i$，r_i 为正确检测到的服装像素，s_i 为利用算法检测到的服装总像素。

比较结果如表 8 所示，GIST 特征方法、LSVM 方法采用传统的粗粒度特征，其 P、R 偏低，SHOE-CNN 较传统的 GIST 特征方法、LSVM 方法，其 P、R 有一定程度的提高，本文方法相较于这 3 种方法，具有较高的 P 和 R，图 8 为准确率和召回率曲线图。

表 8　4 种检索方法比较

Table 8　Comparison of recall rates of four retrieval methods

检索方法	P	R
Li 等人(2013)	0.352 4	0.542 7
Huang 等人(2014)	0.462 9	0.652 6
Zhan 等人(2017)	0.733 3	0.812 5
本文(DPM)	0.860 5	0.837 2

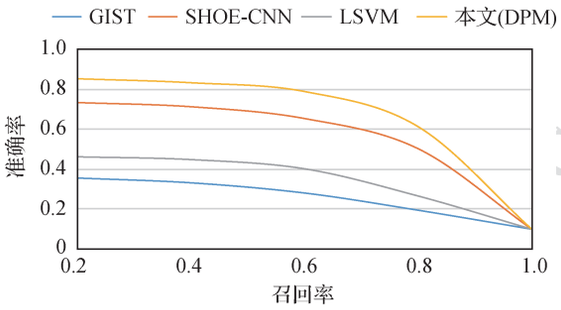


图 8　P-R 曲线

Fig. 8　Curve of P-R

6　结　论

针对细粒度图像的检索问题，以鞋类图像为例，提出一种结合部件检测和语义网络的细粒度图像检索方法。本文方法能够较准确地检测出鞋类图像不同的部件区域，并定义了细致的鞋类图像语义属性，通过训练语义网络得到鞋类图像的视觉属性特征，解决了仅使用粗粒度的特征表示图像而导致鞋类图像检索准确率和精度低的问题。实验结果表明，本文在 UT-Zap50K 数据集上，能够检索到与待检图像完全相同的图像，在保证运行效率的同时，准确率达到 86% 。

然而，本文方法还存在一些不足：一方面由于鞋的款式多、复杂导致部分图像检测的准确率较低。另一方面，存在部分语义属性预测不准确以及鞋类图像的语义属性定义还不够完善等问题。因此，本文的后续工作将着重围绕这些问题展开，以进一步提高检索的准确率和精度，并将应用问题扩展到不同场景上。

参考文献(References)

Bourdev L, Maji S and Malik J. 2011. Describing people: a poselet-based approach to attribute classification//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain, IEEE: 1543-1550 [DOI: 10.1109/ICCV.2011.6126413]

Cao C Q, Wang B, Zhang W R, Zeng X D, Yan X, Feng Z J, Liu Y T and Wu Z Y. 2019. An improved faster R-CNN for small object detection. IEEE Access, 7: 106838-106846 [DOI: 10.1109/ACCESS.2019.2932731]

Chen B H and Deng W H. 2019. Hybrid-attention based decoupled metric learning for zero-shot image retrieval//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2745-2754 [DOI: 10.1109/CVPR.2019.00286]

Chen H Z, Gallagher A and Girod B. 2012. Describing clothing by semantic attributes//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer: 609-623 [DOI: 10.1007/978-3-642-33712-3_44]

Felzenszwalb P, McAllester D and Ramanan D. 2008. A discriminatively trained, multiscale, deformable part model//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2008.4587597]

Guillaumin M, Mensink T, Verbeek J and Schmid C. 2009. Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE: 309-316 [DOI: 10.1109/ICCV.2009.5459266]

He K M, Zhang X Y, Ren S Q and Sun J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition//Proceedings of European Conference on Computer Vision, 346-361 [DOI: 10.1109/TPAMI.2 10.1109/TPAMI.2015.2389824]

Hosang J, Benenson R and Schiele B. 2017. Learning non-maximum suppression//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6469-6477 [DOI: 10.1109/CVPR.2017.685]

Huang J S, Liu S, Xing J L, Mei T and Yan S C. 2014. Circle and search: attribute-aware shoe retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications, 11(1): 1-21 [DOI: 10.1145/2632165]

Kiapour M H, Han X F, Lazebnik S, Berg A C and Berg T L. 2015. Where to buy it: matching street clothing photos in online shops//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE: 3343-3351 [DOI: 10.1109/ICCV.2015.382]

Kovashka A, Parikh D and Grauman K. 2012. WhittleSearch: image search with relative attribute feedback//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 2973-2980 [DOI: 10.1109/CVPR.2012.6248026]

Krizhevsky A, Sutskever I and Hinton G E. 2012. Imagenet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems Advances in neural information processing systems. NV: NIPS: 1097-1105 [DOI: 10.1145/3065386]

Li H J, Wang X H, Tang J H and Zhao C X. 2013. Combining global and local matching of multiple features for precise item image retrieval. Multimedia Systems, 19(1): 37-49 [DOI: 10.1007/s00530-012-0265-1]

Li Z D, Zhong Y and Cao D P. 2018. Deep convolution feature vector for

fast face image retrieval. Journal of Computer-Aided Design and Computer Graphics, 30(12): 2311-2317 (李振东, 钟勇, 曹冬平. 2018. 深度卷积特征向量用于快速人脸图像检索. 计算机辅助设计与图形学学报, 30(12): 2311-2317) [DOI: 10.3724/SP.J.1089.2018.17119]

Ozeki M and Okatani T. 2014. Understanding convolutional neural networks in terms of category-level attributes//Proceedings of the 12th Asian Conference on Computer Vision. Singapore: Springer: 362-375 [DOI: 10.1007/978-3-319-16808-1_25]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Wei X S, Luo J H, Wu J X and Zhou Z H. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing, 26(6): 2868-2881 [DOI: 10.1109/TIP.2017.2688133]

Wu S M, Liu L, Fu X D, Liu L J and Huang Q S. 2019. Human detection and multi-task learning for minority clothing recognition. Journal of Image and Graphics, 24(4): 562-572 (吴圣美, 刘骊, 付晓东, 刘利军, 黄青松. 2019. 结合人体检测和多任务学习的少数民族服装识别. 中国图象图形学报, 24(4): 562-572) [DOI: 10.11834/jig.180500]

Xie L X, Wang J D, Zhang B and Tian Q. 2015. Fine-grained image search. IEEE Transactions on Multimedia, 17(5): 636-647 [DOI: 10.1109/TMM.2015.2408566]

Yu A and Grauman K. 2014. Fine-grained visual comparisons with local learning//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio: IEEE: 192-199 [DOI: 10.1109/CVPR.2014.32]

Yu Y L, Ji Z, Fu Y W, Guo J, Pang Y and Zhang Z. 2018. Stacked semantics-guided attention model for fine-grained zero-shot learning//Proceedings of the 32nd Conference on Neural Information Processing Systems. Canada: MIT Press: 5995-6004

Zhan H J, Shi B X and Kot A C. 2017. Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network. IEEE Transactions on Image Processing, 26(12): 5867-5881 [DOI: 10.1109/TIP.2017.2736346]

Zhang N, Paluri M, Ranzato M A, Darrell T and Bourdev L. 2014. Panda: pose aligned networks for deep attribute modeling//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 1637-1644 [DOI: 10.1109/CVPR.2014.212]

Zhang S Y, Song Z J, Cao X C, Zhang H and Zhou J. 2019. Task-aware attention model for clothing attribute prediction. IEEE Transactions on Circuits and Systems for Video Technology: 1051-1064 [DOI: 10.1109/TCSVT.2019.2902268]

Zhou W G, Lu Y J, Li H Q, Song Y B and Tain Q. 2010. Spatial coding for large scale partial-duplicate web image search// Proceedings of the 18th International Conference on Multimedea, Firenze, Italy, ACM: 511-520 [DOI: 10.1145/1873951.1874019]

Zhu P K, Wang H X and Saligrama V. 2019. Zero shot detection. IEEE Transactions on Circuits and Systems for Video Technology, 30(14): 998-0101 [DOI: 10.1109/TCSVT.2019.2899569]

作者简介

陈前,1995 年生,男,硕士研究生,主要研究方向为图形图像处理。
E-mail: 542231969@qq.com

刘骊,通信作者,女,教授,主要研究方向为计算机辅助设计与图形学、图像处理、计算机视觉。
E-mail: kmust_mary@163.com

付晓东,男,教授,主要研究方向为服务计算、决策理论与方法。E-mail: xiaodong_fu@hotmail.com
刘利军,男,副教授,主要研究方向为图像处理、云计算、信息检索。E-mail: cloneiq@126.com
黄青松,男,教授,主要研究方向为机器学习、数据挖掘、智能信息系统。E-mail: 1912443688@qq.com