JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
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

2022
07
VOL.27

ISSN1006-8961
CN11-3758/TB

舰船目标检测 *P2078*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第7期（总第315期）
2022年7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 中国科学院空天信息创新研究院
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 吴一戎
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@aircas.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 Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
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@aircas.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

2022年7月 第27卷 第7期（总第315期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

深度学习背景下视觉显著性物体检测综述(第2112页)

提取全局语义信息的场景图生成算法(第2214页)

融合弱监督目标定位的细粒度小样本学习(第2226页)

Volume 27, Number 7
Published July 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Review of deep learning based salient object detection(P2112)

Global semantic information extraction based scene graph generation algorithm(P2214)

Weakly-supervised object localization based fine-grained few-shot learning(P2226)

中图法分类号:TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)07-2199-15

论文引用格式: Bao W X, Mao L L, Wang N, Tang J, Yang X J and Zhang Y. 2022. Non-local attention dual-branch network based cross-modal barefoot footprint retrieval. Journal of Image and Graphics,27(07):2199-2213(鲍文霞,茅丽丽,王年,唐俊,杨先军,张艳. 2022. 非局部注意力双分支网络的跨模态赤足足迹检索. 中国图象图形学报,27(07):2199-2213)[DOI:10.11834/jig.200806]

非局部注意力双分支网络的跨模态赤足足迹检索

鲍文霞[1],茅丽丽[1],王年[1],唐俊[1],杨先军[2*],张艳[1]

1. 安徽大学电子信息工程学院,合肥　230601; 2. 中国科学院合肥物质科学研究院,合肥　230031

摘　要: **目的**　针对目前足迹检索中存在的采集设备种类多样化、有效的足迹特征难以提取等问题,本文以赤足足迹图像为研究对象,提出一种基于非局部(non-local)注意力双分支网络的跨模态赤足足迹检索算法。**方法**　该网络由特征提取、特征嵌入以及双约束损失模块构成,其中特征提取模块采用双分支结构,各分支均以 ResNet50 作为基础网络分别提取光学和压力赤足图像的有效特征;同时在特征嵌入模块中通过参数共享学习一个多模态的共享空间,并引入非局部注意力机制快速捕获长范围依赖,获得更大感受野,专注足迹图像整体压力分布,在增强每个模态有用特征的同时突出了跨模态之间的共性特征;为了增大赤足足迹图像类间特征差异和减小类内特征差异,利用交叉熵损失 L_{CE}(cross-entropy loss)和三元组损失 L_{TRI}(triplet loss)对整个网络进行约束,以更好地学习跨模态共享特征,减小模态间的差异。**结果**　本文将采集的 138 人的光学赤足图像和压力赤足图像作为实验数据集,并将本文算法与细粒度跨模态检索方法 FGC(fine-grained cross-model)和跨模态行人重识别方法 HC(hetero-center)进行了对比实验,本文算法在光学到压力检索模式下的 mAP(mean average precision)值和 rank1 值分别为 83.63% 和 98.29%,在压力到光学检索模式下的 mAP 值和 rank1 值分别为 84.27% 和 94.71%,两种检索模式下的 mAP 均值和 rank1 均值分别为 83.95% 和 96.5%,相较于 FGC 分别提高了 40.01% 和 36.50%,相较于 HC 分别提高了 26.07% 和 19.32%。同时本文算法在 non-local 注意力机制、损失函数、特征嵌入模块后采用的池化方式等方面进行了对比分析,其结果证实了本文算法的有效性。**结论**　本文提出的跨模态赤足足迹检索算法取得了较高的精度,为现场足迹比对、鉴定等应用提供了研究基础。

关键词:图像检索;跨模态足迹检索;非局部注意力机制;双分支网络;赤足足迹图像

Non-local attention dual-branch network based cross-modal barefoot footprint retrieval

Bao Wenxia[1], Mao Lili[1], Wang Nian[1], Tang Jun[1], Yang Xianjun[2*], Zhang Yan[1]

1. College of Electronic Information Engineering, Anhui University, Hefei 230601, China;

2. Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China

Abstract: **Objective** Footprints are the highest rate of material evidence left and extracted from crime scene in general. Footprint retrieval and comparison plays an important role in criminal investigation. Footprint features are identified via the

收稿日期:2020-12-31;**修回日期**:2021-04-14;**预印本日期**:2021-04-21

*通信作者:杨先军　xjyang@iim.ac.cn

基金项目:国家重点研发计划资助(2020YFF0303803);国家自然科学基金项目(61772032);安徽高校自然科学研究重点项目(KJ2021ZD0004,KJ2019A0027)

Supported by: National Key R&D Program of China (2020YFF0303803); National Natural Science Foundation of China (61772032); Natural Science Research Project of Anhui Provincial Education Department (KJ2021ZD0004, KJ2019A0027)

foot shape and bone structure of the person involved and have its features of specificity and stability. Meanwhile, footprints can reveal their essential behavior in the context of the physiological and behavioral characteristics. It is related to the biological features like height, body shape, gender, age and walking habits. Medical research results illustrates that footprint pressure information of each person is unique. It is challenged to improve the rate of discovery, extraction and utilization of footprints in criminal investigation. The retrieval of footprint image is of great significance, which will provide theoretical basis and technical support for footprint comparison and identification. Footprint images have different modes due to the diverse scenarios and tools of extraction. The global information of cross-modal barefoot images is unique, which can realize retrieval-oriented. The retrieval orientation retrieves the corresponding image of cross-modes. The traditional cross-modal retrieval methods are mainly in the context of subspace method and objective model method. These retrieval methods are difficult to obtain distinguishable features. The deep learning based retrieval methods construct multi-modal public space via convolutional neural network (CNN). The high-level semantic features of image can be captured in terms of iterative optimization of network parameters, to lower the multi-modal heterogeneity. **Method** A cross-modal barefoot footprint retrieval algorithm based on non-local attention two-branch network is demonstrated to resolve the issue of intra-class wide distance and inter-class narrow distance in fine-grained images. The collected barefoot footprint images involve optical mode and pressure mode. The median filter is applied to remove noises for all images, and the data augmentation method is used to expand the footprint images of each mode. In the feature extraction module, the pre-trained ResNet50 is used as basic network to extract the inherent features of each mode. In the feature embedding module, parameter sharing is realized by splicing feature vectors, and a multi-modal sharing space is constructed. All the residual blocks in the Layer2 and Layer3 of the ResNet50 use a non-local attention mechanism to capture long-range dependence, obtain a large receptive field, and highlight common features quickly. Simultaneously, cross-entropy loss and triplet loss are used to better learn multi-modal sharing space in order to reduce intra-class differences and increase inter-class differences of features. Our research tool is equipped with two NVIDIA 2070TI graphics CARDS, and the network is built in PyTorch. The size of the barefoot footprint images is 224 × 224 pixels. The stochastic gradient descent (SGD) optimizer is used for training. The number of iterations is 81, and the initial learning rate is 0.01. The trained network is validated by using the validation set, and the mean average precision (mAP) and rank values are obtained. In addition, the optimal model is saved in accordance with the highest rank1 value. The backup model is based on the test set, and the data of the final experimental results are recorded and saved. **Result** A cross-modal retrieval dataset is collected and constructed through a 138 person sample. Our comparative experiments are carried out to verify the effect of non-local attention mechanism in related to the retrieval efficiency, multiple loss functions and different pooling methods based on feature embedding modules. Our illustrated algorithm is compared to fine-grained cross-modal retrieval derived fine-grained cross-model (FGC) method and the RGB-infrared cross-modal person re-identification based hetero-center (HC) method. The number of people in the training set, verification set and test set is 82, 28 and 28, respectively, including 16 400 images, 5 600 images and 5 600 images each. The ratio of query images and retrieval images in the verification set and test set is 1∶2. The evaluation indexes of the experiment are mAP mean (mAP_Avg) and rank1 mean (rank1_Avg) of two retrieval modes. Our analysis demonstrates that the algorithm illustrated has a higher precision, and the mAP_Avg and rank1_Avg are 83.95% and 96.5%, respectively. Compared with FGC and HC, the evaluation indexes of the proposed algorithm is 40.01% and 36.50% (higher than FGC), and 26.07% and 19.32% (higher than HC). **Conclusion** A cross-modal barefoot footprint retrieval algorithm is facilitated based on a non-local attention dual-branch network through the integration of non-local attention mechanism and double constraint loss. Our algorithm considers the uniqueness and correlation of in-modal and inter-modal features, and improves the performance of cross-modal barefoot footprint retrieval further, which can provide theoretical basis and technical support for footprint comparison and identification.

Key words: image retrieval; cross-modal footprint retrieval; non-local attention mechanism; two-branch network; barefoot footprint image

0 引 言

足迹是指人体穿鞋、穿袜或赤足站立/行走条件下,脚掌通过体重压力作用在承痕体形成的痕迹(Gurney 等,2008)。医学研究表明,足迹数据具有唯一性和独特性(Nirenberg 等,2019),不仅能够反映个体的生理特征,还可以反映个体的行为特征。在侦查犯罪案件中,足迹特征相对于指纹和面部等其他人体特征来说更加不易伪装,刑侦人员可以从现场遗留足迹中挖掘到更多有效的信息和线索。因此,研究人员在足迹方面开展了形式多样的研究。Kulkarni 和 Kulkarni(2015)使用摄像机采集了 30 人的足底图像,提取足弓水平最小截距、足弓内侧最大距离、痕迹指数和痕迹几何指数等 4 种特征进行研究。Osisanwo 等人(2014)使用捺印技术在纸板上印出足迹,比较分割后的前脚、中间和后脚 3 个部分的压力面积和压力值标准偏差并进行图像匹配。Khokher 等人(2015)通过平板扫描仪获取了 21 人的足底图像,分别使用主成分分析(principal component analysis, PCA)和独立成分分析(independent components analysis,ICA)线性投影技术提取足底的纹理和形状特征来进行足迹识别,识别率分别达到 95.24% 和 97.23% 。Heydarzadeh 等人(2017)使用压力传感器平面板采集了 35 人的足迹数据,将每帧图像叠加合成一幅融合图像,使用支持向量机(support vector machine, SVM)对融合图像进行分类识别,识别率达到 97% 。鲍文霞等人(2020a)使用光学足迹采集器采集 134 人的 2 680 枚足迹数据,通过度量学习方法,结合支持向量机分类器实现足迹识别,达到 96.66% 的准确率。朱明等人(2020)使用压力足迹采集器采集了 100 名受试者的 1 000 枚压力赤足足迹,通过多尺度自注意卷积模块自适应地提取可判别足迹特征,以实现足迹图像检索,其中 mAP(mean average precision)值为 81.64% ,rank1 值为 95.63% 。鲍文霞等人(2020b)使用压力足迹采集器构建了 100 人的 2 000 枚足迹图像,通过设计一种空间聚合加权模块,有效实现了足迹识别,其中算法准确率达到 91.20% 。

从上述足迹方面的研究可以看出,目前尚未存在公共的足迹图像数据集,并且在研究过程中的足迹图像采集规范和采集设备多样化,相同采集对象在不同采集设备下获取的足迹图像模态不同,而它们包含的信息既有共性又有各自的独特性。目前公安部门使用的足迹采集设备不统一,主要有光学和压力两种足迹图像采集仪。压力成像的足迹在图像中间或边缘都能比较准确地体现压力值的大小,压力变化层次丰富。光学成像的足迹压力值是通过算法估计得到的相对值,并且图像轮廓的边缘模糊、压力值不准确,但是光学足迹图像的这种特点更接近现场足迹。因此,通过压力和光学两种模态足迹图像的互检索,一方面可以使目前由不同设备采集构建的足迹库中的图像互查,另一方面也为下一步现场足迹比对、鉴定等应用提供了研究基础。

基于图像的跨模态检索方法主要包括子空间方法、主题模型方法、哈希变换方法和深度学习方法。子空间方法的原理是利用跨模态样本对的信息构建共享子空间,通过度量相似性实现跨模态检索(Liang 等,2016),但存在不能有效建模不同模态的高阶相关性的问题。主题模型法是通过对各模态都学习一个潜在主题模型,再建模不同模态间的关系(Zheng 等,2014),但在建立不同模态语义关联的同时未能综合考虑不同模态的结构。哈希变换法是将高维特征转换为便于存储和计算的二进制码,其中相似的数据会得到相似的二进制码,并将不同模态特征映射到一个汉明(Hamming)二值空间再进行学习(Cao 等,2016),可能会造成图像信息上的损失。深度学习方法是利用能够学习到高层语义特征的神经网络为多个模态构造一个公共语义空间,减小跨模态的异构性,然后在公共空间进行模态之间的相似性度量(Wang 等,2015),具有更好的检索效果,因此成为目前跨模态检索的研究热点。例如,在深度学习领域,Wu 等人(2017)提出一个深度零填充网络来自适应地学习模态共享特性,实现了可见光行人图像与红外光行人图像之间的跨模态检索;Ye 等人(2018a)引入了双流网络(two stream network)来建模特定的和可共享的信息,进行模态内和模态间的特征学习,实现了可见光行人图像与红外光行人图像的模态内和模态间的特征学习;Zhu 等人(2019)提出了一个新的异质中心损失(hetero-center loss),以减少类内交叉模态的变化,再结合交叉熵损失和双分支结构,扩大类间差异和尽可能提高类内跨模态相似性,在红外光和可见光的跨模态行人重识别领域取得了较优的检索效果。

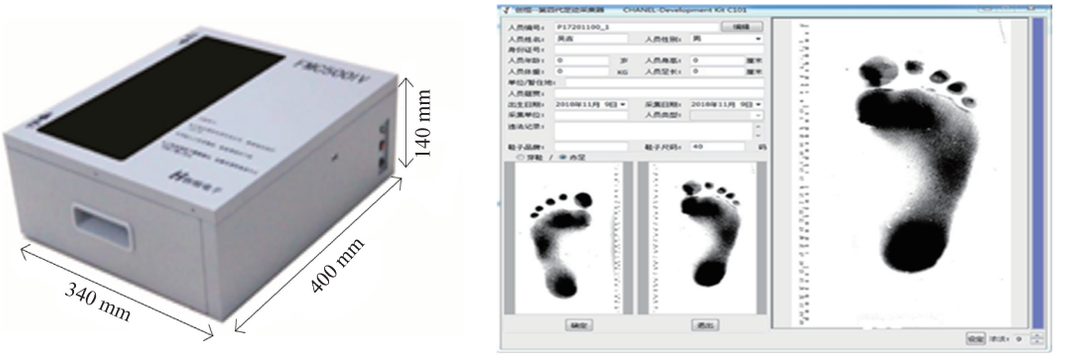
在行人重识别、草图识别以及图像、文本、声音及视频的多模态领域中,跨模态检索取得了一定进展,但关于足迹图像的跨模态检索方面的研究还很少。因此,本文根据赤足足迹图像的特点,结合深度学习方法研究足迹的跨模态检索问题。首先设计了人在自然行走状态下的足迹图像采集规范和流程,利用光学足迹和压力足迹图像采集仪器,采集并建立了一个包含 138 人的 5 520 枚赤足足迹的跨模态检索数据集。针对细粒度赤足足迹图像存在的类内差异大类间差异细微的特点,设计了一种基于非局部(non-local)注意力机制的双分支网络用于赤足足迹的跨模态检索,该网络在特征提取模块采用了双分支结构,各分支均采用 ResNet50 网络作为基础网络,提取不同模态的有效特征,过滤干扰特征;在特征嵌入模块通过参数共享构建一个多模态的共享空间,并在网络的 layer2 层和 layer3 层的每个残差块处引入非局部注意力机制模块,增强每个模态的有用特征,以得到更具辨别性的足迹特征,同时获取跨模态的共享特征空间;为了实现每个模态以及跨模态类间距离的最大化和类内距离的最小化,本文损失函数采用了交叉熵损失和三元组损失,以学习到更有效的跨模态共享特征。

1 足迹图像获取和预处理

本文使用光学足迹图像采集仪器和压力足迹图像采集仪器获取足迹图像。

光学足迹图像采集仪器由正面为脚踏面的等腰三棱镜、均匀直流光源和拍摄装置组成,采集时不需要在足底涂抹油墨,而是利用棱镜全反射原理形成足迹图像并由拍摄装置获取。光学足迹采集图像的分辨率为 1 362 × 2 871 dpi(dots per inch)。光学足迹图像采集器及软件界面如图 1 所示。

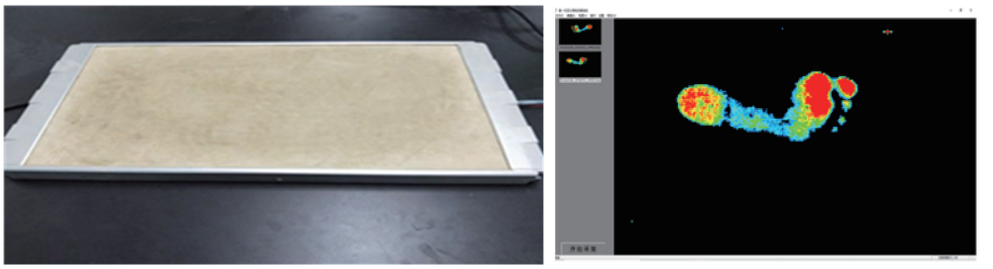
从生物力学角度来看,压力足迹图像采集仪器可以获取采集者自然行走过程中的压力变化以及足迹特征等信息。该采集器的性能稳定,能够较好地保证采集图像的质量。压力足迹图像采集器及软件界面如图 2 所示。表 1 给出了采集仪器的主要技术参数,其中采集频率为 100 Hz,可以满足人体行走时图像采集的要求;传感器密度为 25 个/cm^2;幅面为 50 cm × 30 cm 的有效区域,便于清晰地观察采集的压力图像。压力足迹采集图像的分辨率为 250 × 150 dpi。



(a) 光学足迹图像采集器　　(b) 软件界面

图 1　光学足迹图像采集仪器及软件界面

Fig. 1　Optical footprint image collector and its software interface((a) optical footprint image collector; (b) software interface)



(a) 压力足迹图像采集器　　(b) 软件界面

图 2　压力足迹图像采集仪器及软件界面

Fig. 2　Pressure footprint image collector and its software interface((a) pressure footprint image collector; (b) software interface)

表 1　压力图像采集仪器主要参数

Table 1　Parameters of pressure footprint image collector

主要参数	参数设置
分辨精度	25 点/cm^2
采集区域	50 cm × 30 cm
通讯接口	USB 3.0
采集帧率	50 帧/s
分辨率	250 ×150 dpi
采集时间	<20 s
仪器尺寸	58 cm（长）× 32 cm（宽）× 5 cm（高）
仪器重量	8.5 kg

1.1　足迹数据采集流程

本次实验总共采集了 138 人的光学赤足图像和压力赤足图像，其中男性 90 人，女性 48 人。数据采集前，首先在系统中录入被采集人员的身高、体重等个人基本信息。表 2 给出了被采集人员的基本信息分布。

表 2　被采集人员的信息分布

Table 2　Information distribution of the person to be collected

属性	男	女
人数	90	48
身高/cm	166 ～195	150 ～175
体重/kg	52 ～100	40 ～65
足迹/cm	24.5 ～28	22 ～25
年龄/岁	20 ～26	20 ～26

在图像采集过程中，可能会受到灰尘、仪器噪声以及被采集人员脚部的施力状态和行走姿势等因素的影响。为了提高采集的规范性，采集过程要求采集仪器表面保持一定的清洁。足迹是体现人体心理的重要载体，心理活动与足迹之间有着必然的联系（薛亚龙和岳佳，2012），被采集人员的心理活动也会影响采集数据的质量。因此，采集人员会提前向被采集者介绍采集内容和方式，然后向被采集者示范正确的采集规范。具体的采集流程如下：

1）将采集仪器水平嵌入地板槽内，保持仪器与地面持平。随后打开电源连接设备进行测试，确认设备无误后，根据被采集者信息和采集方式建立文件夹，例如“学号_行走趟数_左脚”。为了达到采集标准，在光学足迹和压力足迹正式采集前预留足够时间，要求被采集人员行走多趟适应地面环境以达到自然行走状态。然后点击“采集”按钮，并示意被采集人员按示范的标准开始行走。

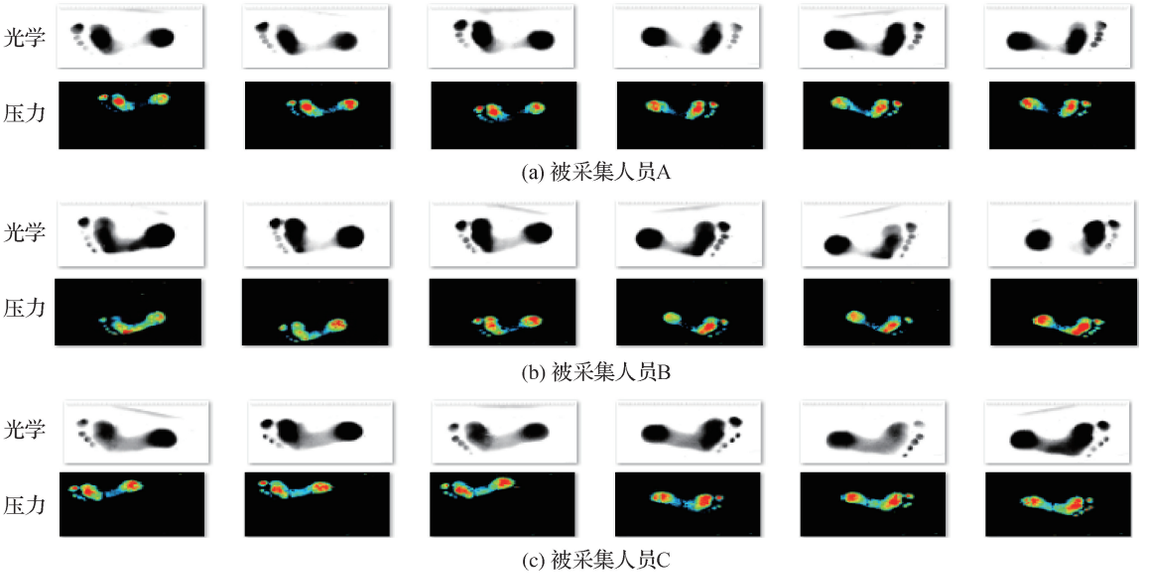
2）在被采集人员平稳自然行走经过采集仪器时，查看生成的图像是否存在缺损严重、偏离采集页面中心以及行走速度过快等问题，若有则重新采集，以提高采集图像的质量；如果被采集者和采集的图像满足采集规范，则将图像数据保存到事先建立好的个人文件夹中。

3）为了维持采集设备的性能稳定性，每天在早中晚 3 个时间段只采集两个人的赤足足迹图像，并在每次采集后进行仪器表面的清理，减少采集图像的背景噪声。

正式采集时，被采集人员在赤足条件下自然地来回走过采集仪器 10 趟，得到 20 幅赤足足迹图像，其中左右脚图像各 10 幅，最终共采集 5 520 幅赤足足迹图像，其中光学赤足图像和压力足迹图像各有 2 760 幅。图 3 展示了 3 个不同被采集人员在每种模态下的 6 幅赤足足迹图像，其中左右脚图像各 3 幅。由于被采集人员的身高、体态以及行走习惯等因素，被采集人员的光学和压力足迹在形态以及图像中的具体分布位置等方面存在很大差异；同时，在单模态内同一个被采集人员的图像存在较大差异，而不同被采集人员之间的图像差异细微，导致跨模态赤足足迹检索难度显著提升。

1.2　足迹数据预处理

采集的光学赤足图像中含有标尺部分，并且在采集的过程中受灰尘等因素的影响，从而产生少量噪声，因此首先对采集的赤足足迹图像进行去标尺和滤波等预处理。在去噪方法中最常用的是中值滤波（赵博文 等，2020），因为中值滤波可以保留图像轮廓信息，更能起到保护图像细节信息的作用，因此本文采用中值滤波方法分别对光学赤足图像和压力赤足图像进行去噪处理，并在足迹图像送入网络训练之前将分辨率统一调成了 224 × 224 dpi。光学赤足图像和压力赤足图像的预处理分别如图 4 和图 5 所示。为了提升网络模型的泛化能力和鲁棒性，对采集图像进行了垂直翻转、水平翻转、逆时针旋转 10°和顺时针旋转 10°的数据增广，光学赤足图像和压力赤足图像的增广操作分别如图 6 和图 7 所示。经扩充后，光学赤足图像和压力赤足图像分别从2 760 幅扩展为 13 800 幅。



(a) 被采集人员A

(b) 被采集人员B

(c) 被采集人员C

图 3　采集的光学赤足图像和压力赤足图像

Fig. 3　Collected optical barefoot images and pressure barefoot images

((a) the collected person A; (b) the collected person B; (c) the collected person C)

(a) 光学原图　(b) 去标尺结果　(c) 滤波结果

图 4　光学赤足图像的预处理效果图

Fig. 4　Preprocessing of optical barefoot images ((a) optimal original image; (b) after removing the scale; (c) after filtering)

(a) 压力原图　(b) 滤波结果

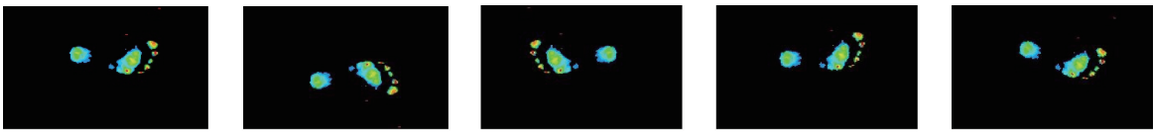
图 5　压力赤足图像的预处理效果图

Fig. 5　Preprocessing of pressure barefoot image ((a) pressure original image; (b) after filtering)

(a) 滤波　(b) 垂直翻转　(c) 水平翻转　(d) 逆时针旋转10°　(e) 顺时针旋转10°

图 6　光学赤足图像的数据增广效果图

Fig. 6　Data augmentation of optical barefoot images ((a) filtering; (b) vertical flipping; (c) horizontal flipping; (d) rotating 10 degrees counterclockwise; (e) rotating 10 degrees clockwise)



(a) 滤波　(b) 垂直翻转　(c) 水平翻转　(d) 逆时针旋转10°　(e) 顺时针旋转10°

图 7　压力赤足图像的数据增广效果图

Fig. 7　Data augmentation of pressure barefoot images ((a) filtering; (b) vertical flipping; (c) horizontal flipping; (d) rotating 10 degrees counterclockwise; (e) rotating 10 degrees clockwise)

2　基于非局部注意力的双分支网络

2.1　网络总体结构

采集的光学赤足图像纹理信息较为完好，但是边缘模糊，压力值不准确；采集的压力赤足图像的轮廓较为清晰，压力变化对比鲜明，重压面区域较清晰。由于不同模态足迹图像的特征之间既有共性又有各自的独特性，且赤足足迹图像属于细粒度图像，不同个体赤足足迹图像差异细微，而同一个体由于受环境、心理以及走路姿态等影响使得足迹图像差异较大，即存在较大的类内差异和较小的类间差异。针对足迹图像的这些特点，本文设计了一个双分支卷积神经网络用于跨模态赤足足迹图像的检索。整个网络包含特征提取(feature extractor)、特征嵌入(feature embedding)和双约束损失(dual-constrained loss)3 个模块，如图 8 所示。其中特征提取模块采用两个分支分别用于提取不同模态赤足图像的特征；在特征嵌入模块中，将特征提取模块输出的各模态特征进行拼接(concat)实现参数共享，以构建一个多模态的共享空间，同时引入非局部注意力机制(Wang 等，2018)，快速捕获长范围依赖，获得更大感受野，专注赤足图像整体压力分布，从而学习到更优的压力分布多模态共享特征；为了增大赤足图像特征的类间差异和减小类内差异(Wang 等，2019)，减小光学模态与压力模态图像的语义鸿沟，在双约束损失模块引入交叉熵损失(cross-entropy loss，CE loss)和三元组损失(triplet loss，TRI loss)进行网络优化。

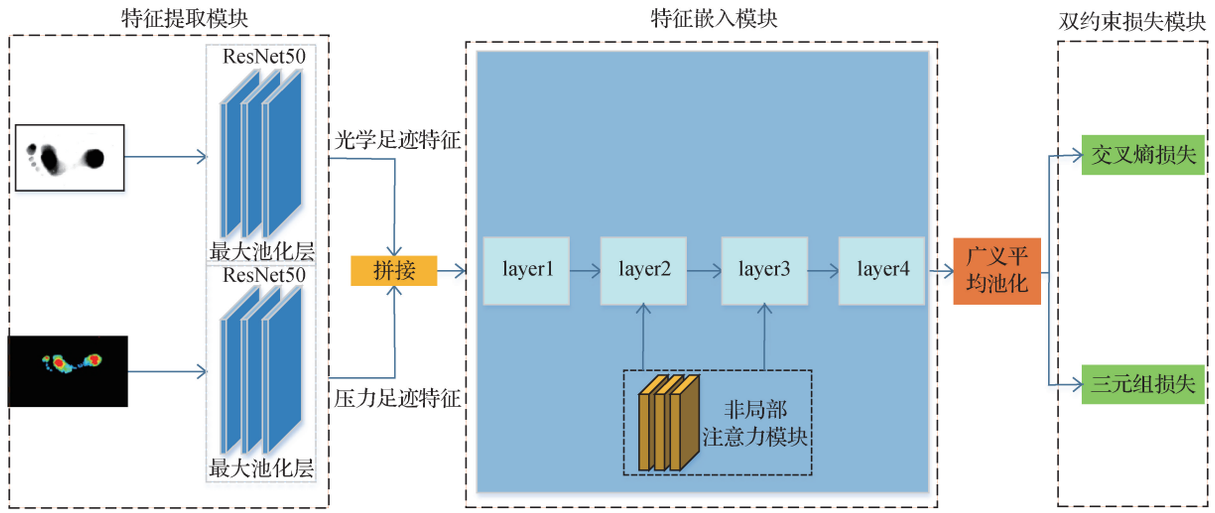


图 8　跨模态足迹检索网络结构

Fig. 8　The network architecture of cross-modal retrieval for footprint images

2.2　特征提取模块

在深度学习领域，随着网络深度的增加，可能会伴随梯度消失或梯度爆炸等问题，不利于网络的收敛优化。常用的 ResNet50 网络(He 等，2016)通过短接操作构建了恒等映射，使原始粗略的特征与训练后得到的精细化特征相互补充融合，可以突出有用特征，过滤掉一些无用特征。同时 ResNet50 网络的学习过程是去拟合残差，不是直接拟合输入和

输出，优化更加简单，不仅能够学习有效的足迹特征，还使得梯度反向传播时，更加不容易出现梯度消失等问题。因此本文采用经过预训练的 ResNet50 网络作为双分支结构中的每个单分支的基础网络。

ResNet50 的网络结构图如图 9 所示，特征提取模块由 ResNet50 网络的 layer1 层之前的卷积层（convolution layer，Conv）、批量正则化（batch-normalization，BN）、ReLU（rectified linear unit）激活函数以及最大池化（max pooling，MaxPool）构成，分别提取光学赤足图像的特征和压力赤足图像的特征。

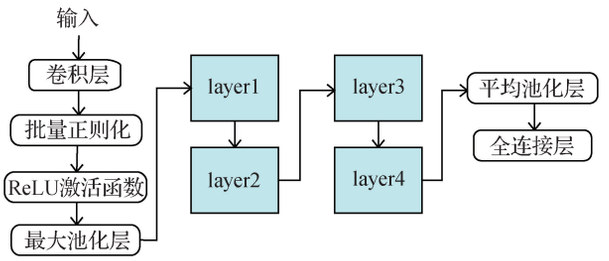


图 9　ResNet50 网络结构图

Fig. 9　The network structure of ResNet50

2.3　特征嵌入模块

将特征提取模块输出的各模态特征拼接（concat）后送入特征嵌入模块，用于构建多模态的共享空间。为了获取更加充分全面的共享特征信息，特征嵌入模块由 ResNet50 中的 layer1—layer4 层构成，同时在 layer2 层和 layer3 层的每个残差块处都引入了非局部注意力模块（Wang 等，2018）。

在常规的卷积操作输出特征图中，每个空间位置的输出值由输入特征图和卷积核在局部范围内进行对应元素相乘再相加获得，而 non-local 注意力机制在计算每个空间位置输出时，不是仅考虑局部元素，而是与图像中所有空间位置计算相关性，并将相关性作为权重以表征当前空间位置与其他空间位置的相似度。通过关注所有空间位置并取其加权平均值作为当前空间位置的响应值，从而增强赤足足迹图像中的有效特征。同时，非局部注意力机制的输入尺度多样化，易与其他基础网络模型相结合。如图 8 所示，本文在 ResNet50 基础网络的 layer2 和 layer3 中每个残差块处都引入了 non-local 注意力机制，该注意力机制的完整表达式为

$$\boldsymbol{z}_i = W_z \boldsymbol{y}_i + \boldsymbol{x}_i \tag{1}$$

式中，i 是输出特征图的空间位置索引，$\boldsymbol{x}_i$ 表示空间位置 i 处的输入特征向量，$\boldsymbol{y}_i$ 是对所有空间位置进行加权学习全局信息的特征向量，形状与 $\boldsymbol{x}_i$ 相同，W_z 表示训练过程中输出特征向量 $\boldsymbol{z}$ 的学习参数。其中 $\boldsymbol{y}_i$ 的表达式为

$$\boldsymbol{y}_i = \frac{1}{C(\boldsymbol{x})} \sum_{\forall j} f(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_j) \tag{2}$$

式中，j 表示所有空间位置的索引，$g(\boldsymbol{x}_j)$ 表示进行信息变换的一元输入函数，$C(\boldsymbol{x})$ 为保证变换前后整体信息不变的归一化函数，$f(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 表示特征图中空间位置 i 与空间位置 j 之间的相似度权重关系函数，$f(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 的表达式为

$$f(\boldsymbol{x}_i, \boldsymbol{x}_j) = \theta(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j) \tag{3}$$

式中，$\theta(\boldsymbol{x}_i)$ 表示用 1 × 1 卷积学习当前位置的信息，$\phi(\boldsymbol{x}_j)$ 表示用 1 × 1 卷积学习全局信息。通过将 $\theta(\boldsymbol{x}_i)$ 的转置与 $\phi(\boldsymbol{x}_j)$ 进行矩阵相乘操作，计算位置 i 与位置 j 的相似度大小。

该注意力机制的实现细节如图 10 所示。其中，$\boldsymbol{x}$ 表示形状为［b，c，h，w］的特征向量，b 表示每个模态的批量处理数目，c 表示特征图的通道数，h 和 w 分别表示特征图的高度和宽度。对 $\boldsymbol{x}$ 经过 $1 \times c \times 1 \times 1$ 卷积（输入通道为 c，输出通道为 1 的 1×1 卷积）操作，得到形状为［b，1，h，w］的特征向量 $\boldsymbol{fea}_1$。$\boldsymbol{fea}_1$ 经过变维操作 1 后得到形状为［b，$h \times w$，1］的特征向量，记为 $g(\boldsymbol{x})$；$\boldsymbol{fea}_1$ 经过变维操作 2 后得到两个形状均为［b，1，$h \times w$］的特征向量，分别记为 $\theta(\boldsymbol{x})$ 和 $\phi(\boldsymbol{x})$，将 $\theta(\boldsymbol{x})$ 的转置与 $\phi(\boldsymbol{x})$ 进行对应元素相乘，得到 $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$；经过 $1/N$（$N = h \times w$）得到归一化的特征向量，与 $g(\boldsymbol{x})$ 进行对应元素相乘的加权操作，并经过变维操作 3 得到形状为［b，1，h，w］的特征向量 $\boldsymbol{fea}_2$；然后通过 $c \times 1 \times 1 \times 1$ 卷积（输入通道为 1，输出通道为 c 的 1×1 卷积）操作和BN正则化操作，输出与 $\boldsymbol{x}$ 形状相同的特征

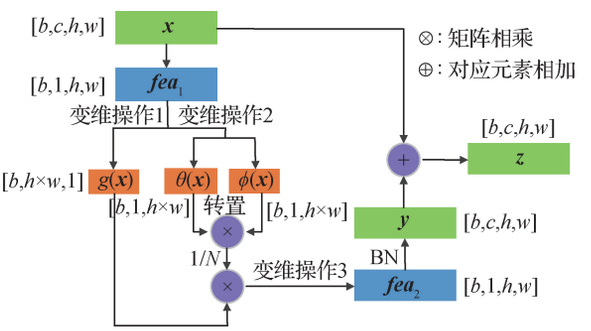


图 10　Non-local 注意力模块原理图

Fig. 10　The schematic diagram of non-local attention module

向量。最后将 $\boldsymbol{y}$ 与原始输入特征向量 $\boldsymbol{x}$ 进行对应元素相加，从而学习到更有关联的共享特征 $\boldsymbol{z}$，减小了模态间的特征差异。

2.4 双约束损失模块

为了能够同时考虑到模态内和模态间的特征差异，使网络学习到更好的跨模态共享特征，本文损失函数采用交叉熵损失和三元组损失双约束方式进行网络优化。损失函数具体结构如图 11 所示，其中光学足迹特征和压力足迹特征为图 8 中特征提取模块输出的光学模态特征向量和压力模态特征向量，通过在第 1 维度上 concat 后得到 $\boldsymbol{fea}_1$ 特征向量，再依次经过 ResNet50 的 layer1、layer2、layer3、layer4，以实现特征嵌入模块中的参数共享。为获得更有辨别性的模态特征，学习到一个更好的模态共享空间，本文在 layer2 和 layer3 中的每个残差块处引入 non-local 注意力机制进行训练，将最后输出的特征向量经过两种不同处理分别得到交叉熵损失（CE loss）和三元组损失（TRI loss）。本文总的损失函数为

$$L_{\text{total}} = L_{\text{CE}} + L_{\text{TRI}} \tag{4}$$

式中，L_{CE} 为交叉熵损失，L_{TRI} 为三元组损失。

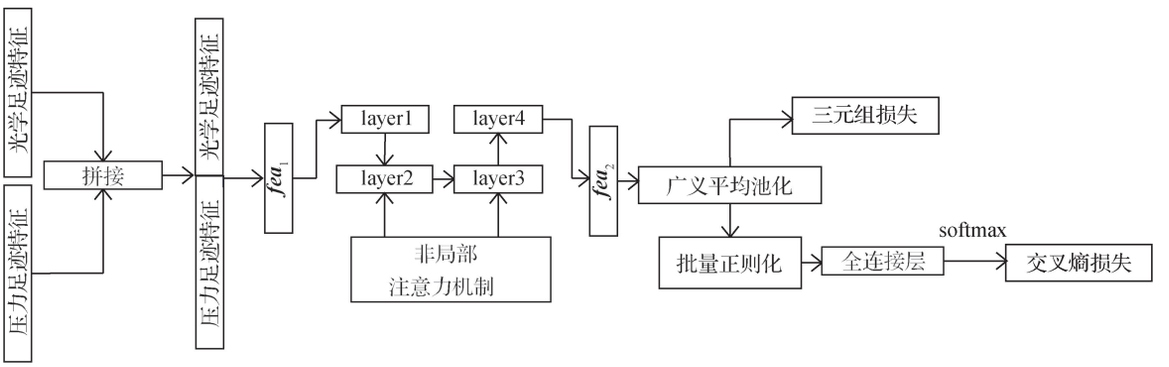


图 11 损失函数结构图

Fig. 11 Structure diagram of loss function

交叉熵损失是利用特定模态信息学习类别特征，并对各模态的类内特征进行约束，使类间差异增大，同时也有助于增加跨模态样本的相关性（Ye 等，2018b）。如图 11 所示，将由广义平均池化（generalized mean pooling，GmPool）操作产生的特征向量，经过正则化（BN）、全连接层（FC1）和 softmax 函数，结合标签计算得到交叉熵损失，其中正则化可以在一定程度上提升网络训练速度，加快收敛过程。

交叉熵损失（CE loss）计算为

$$L_{\text{CE}} = -\frac{1}{K}\sum_{i=1}^{K}\ln\frac{\mathrm{e}^{\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_{y_i}}}{\sum_{j=1}^{n}\mathrm{e}^{\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_j}} \tag{5}$$

式中，i 表示样本索引，K 表示 2 × batch_size，其中 batch_size 为网络训练的超参数，本文设置 batch_size 为 16；j 表示类别索引，n 表示总的类别数，本文设置 n 为 82；$\boldsymbol{x}_i$ 为第 i 个样本的特征向量，y_i 表示样本 i 的真实类别，$\boldsymbol{w}_j$ 表示权重的第 j 列，$\boldsymbol{w}_{y_i}$ 表示权重的第 $\boldsymbol{y}_i$ 列，$\boldsymbol{b}_{y_i}$ 表示偏置的第 y_i 列。该损失值越小，表示预测概率分布与真实概率分布之间的差异越小。

三元组损失（Liu 等，2020）通过在跨模态共享空间中促使每幅足迹图像与相同类别图像之间的距离小于其与不同类别图像之间的距离，以进一步解决图像检索中类内距离大于类间距离的问题。如图 11 所示，将经过池化层 GmPool 的特征向量，结合每个模态的标签计算可得到三元组损失。

三元组损失（TRI loss）计算为

$$L_{\text{TRI}} = \sum_{i=1}^{P}\sum_{a=1}^{2K}\left[\rho + \max_{p=1,\cdots,2K}\|\boldsymbol{x}_a^i - \boldsymbol{x}_p^i\|_2 - \min_{\substack{j=1,\cdots,P(j\neq i)\\ n=1,\cdots,2K}}\|\boldsymbol{x}_a^i - \boldsymbol{x}_n^j\|_2\right] \tag{6}$$

式中，P 和 K 分别表示各模态足迹图像在一个 batch 中的类别数目和每类包含的图像数目，实验中设置 $P = K = 4$。$\boldsymbol{x}_a^i$ 表示第 i 类别对应的两个模态的 $2K$ 幅图像中的第 a 幅图像的特征向量，$\boldsymbol{x}_p^i$ 表示类别 i 对应的两个模态的 $2K$ 幅图像中的第 p 幅图像的特征向量，$\boldsymbol{x}_n^j$ 表示类别 j 对应的两个模态的 $2K$ 幅图像中的第 n 幅图像的特征向量，距离的度量方式采用的是 L2 正则化，max 为 batch 中每幅图像与其他相同类别的图像之间的最大距离，min 为 batch 中每幅图像与不同类别图像之间的最小距离。ρ 表示最大距离与最小距离之间的间隔，该值若设置得较小，损

失函数很容易趋近0,难以很好地区分图像;若设置得较大,损失函数较难趋近0,可能导致网络不收敛。本文首先借鉴检索领域中Liu等人(2020)对参数ρ的设置规则,然后针对赤足足迹图像的特点进行多组实验调参,最终将参数ρ设置为0.3最佳。通过不断迭代训练,使得三元组损失函数值越来越小,即某一样本与最近的类间样本之间的距离比该样本与最远的类内样本的距离远高于参数ρ,实现减小类内差异和增大类间差异的目的。

3 实 验

3.1 实验环境

实验所用计算机带有两块NVIDIA 2070ti,系统环境为Linux操作系统,网络模型均在PyTorch中搭建。本文应用sampler方式选取训练图像,各模态赤足图像在每次迭代时的批处理大小均设置为16,其中随机选择4个不同类别,每个类别对应4幅不同的图像。同时各模态赤足图像在每次迭代时选取的类别须对应。图像尺寸设置为224 × 224像素,并采用翻转、旋转等预处理操作。损失函数使用交叉熵损失(CE loss)和三元组损失(TRI loss)进行双约束。优化器采用随机梯度下降法(stochastic gradient descent,SGD)对网络进行训练,初始学习率设置为0.01,训练轮数在[0, 9]范围内时学习率按0.01进行增加,训练轮数在[10, 20)范围内时学习率为0.1,训练轮数在[20, 50)范围内学习率为0.01,训练轮数在[50, 80]范围内时学习率为0.001。本算法共训练81个epoch。

3.2 评价指标

本文选用Dai等人(2019)和Zhao等人(2019)采用的评价指标评估赤足图像跨模态检索的性能,包括mAP和CMC(cumulative match characteristic curve)中的rank1(R1)、rank5(R5)和rank10(R10)。rank1表示首位检索到的概率,相较于其他rank值使用更为广泛。为进一步验证本文算法对赤足足迹图像跨模态互检索的有效性,实验过程中将两种检索模式下的mAP均值(mAP_Avg)和rank1均值(R1_Avg)也同时作为评价指标。其中mAP_g2y和R1_g2y分别表示光学到压力检索模式下的mAP值和rank1值,mAP_y2g和R1_y2g分别表示压力到光学检索模式下的mAP值和rank1值。

3.3 数据集的划分

实验数据集划分为训练集、验证集和测试集。其中,训练集包含82人,验证集包含28人,测试集包含28人,人数比例为6∶2∶2。为了保证实验的公平性和有效性,须保证训练集、验证集和测试集的类别和图像各不交叉重叠。训练集包括82人的8 200幅光学赤足图像和8 200幅压力赤足图像。验证集包括两种检索模式,一种是根据光学图像检索压力图像,另一种是根据压力图像检索光学图像。同时将查询集和检索集的数据量比例设置为1∶2,即查询集中每人50幅图像,检索集中每人100幅图像。测试集的图像数目分布与验证集相同。在每轮训练结束后都对验证集进行性能评测,得到mAP和rank值,并且按照rank1最高值保存最优模型。为证明模型的有效性,最终运用保存的最优模型对测试集进行性能评测,记录和保存最后的实验结果数据。

3.4 实验结果及分析

本文设计了基于非局部(non-local)注意力双分支网络架构的跨模态赤足足迹检索算法。对比实验包括以下4个方面:1)non-local注意力机制对检索性能的影响;2)采用不同损失函数的对比;3)特征嵌入模块后采用不同池化方法的对比;4)将本文算法与跨模态检索领域的FGC(fine-grained cross-media)方法以及跨模态行人重识别领域的HC(hetero center)方法进行比较。

3.4.1 非局部注意力机制对检索性能的影响

为验证本文算法中non-local注意力机制对足迹图像跨模态检索性能的影响,对网络是否使用non-local注意力机制以及将该注意力机制应用在ResNet50网络的不同位置进行对比分析,实验结果如表3所示。其中Non0、Non23、Non146和Non46分别表示不使用non-local注意力机制和将non-local注意力机制应用在ResNet50网络的不同残差块上的4种实验方法,Non46为本文方法。ResNet50网络在layer1、layer2、layer3和layer4上的残差块数目分别是3、4、6和3。Non23表示注意力机制应用在layer2层中的后2个残差块和layer3层中的前3个残差块;Non46表示注意力机制应用在layer2层中的4个残差块和layer3层中的6个残差块;Non146表示注意力机制应用在layer1层中的最后一个残差块、layer2层中的4个残差块和layer3层中的6个残

差块。实验结果表明，Non46 方法的效果最优，其中 mAP_Avg 和 R1_Avg 分别达到 83.95% 和 96.5%，相较于 Non0，分别高出 0.63% 和 0.78%；相较于 Non23，分别高出 0.72% 和 0.07%；相较于 Non146，分别高出 1.66% 和 0.21%。

表 3 中 Non0 方法的 mAP_Avg 和 R1_Avg 分别为 83.32% 和 95.72%，Non23、Non46 和 Non146 方法的 R1_Avg 均高于 Non0，但 Non23 和 Non146 方法的 mAP_Avg 比 Non0 低 0.09% 和 1.03%，仅本文采用的 Non46 方法的 mAP_Avg 比 Non0 高出 0.63%。上述结果说明，该 non-local 注意力机制通常可以在一定程度上帮助网络更好地进行跨模态检索的特征学习，进一步提高检索精度。同时，non-local 注意力机制应用在 ResNet50 网络的具体位置会影响最终的检索结果。

Non23 和 Non46 方法皆是将注意力机制应用在 ResNet50 的 layer2 层和 layer3 层，区别仅在于具体应用的残差块上。表 3 中的实验结果比较直观地体现了相同层不同残差块上的检索效果差异。其中，Non46 方法的 mAP_Avg 和 R1_Avg 分别比 Non23 高 0.72% 和 0.07%，说明将注意力机制应用在 layer2 层和 layer3 层的所有残差块可以进行远距离深层次的信息交互(Wang 等，2018)，更有利于本文针对跨模态检索的研究。

表 3 中 Non146 方法的 mAP_Avg 和 R1_Avg 分别为 82.29% 和 96.29%，相较于最优实验效果的 Non46 方法降低了 1.66% 和 0.21%。该实验结果表明在 layer1 层引入 non-local 注意力机制不利于跨模态特征的学习，与 layer1 层学习到的浅层足迹信息相比，layer2 层和 layer3 含有更为丰富的细节信息，使得 non-local 注意力机制能够捕获更具辨别性的特征。

表 3　本文算法中 non-local 注意力机制对检索性能的影响

Table 3　The effect of non-local attention mechanism in the algorithm on retrieval performance

/%

方法	mAP_g2y	mAP_y2g	mAP_Avg	R1_g2y	R1_y2g	R1_Avg
Non0	82.88	83.75	83.32	98.00	93.43	95.72
Non23	82.88	83.57	83.23	**98.43**	94.43	96.43
Non146	82.23	**84.34**	82.29	98.21	94.36	96.29
Non46(本文)	**83.63**	84.27	**83.95**	98.29	**94.71**	**96.50**

注：加粗字体表示各列最优结果。

3.4.2　不同损失函数的对比实验

实验对网络采用的损失函数进行对比，结果如图 12 所示。CE 表示仅使用交叉熵损失，TRI 表示仅使用三元组损失，CE_TRI表示交叉熵损失和三

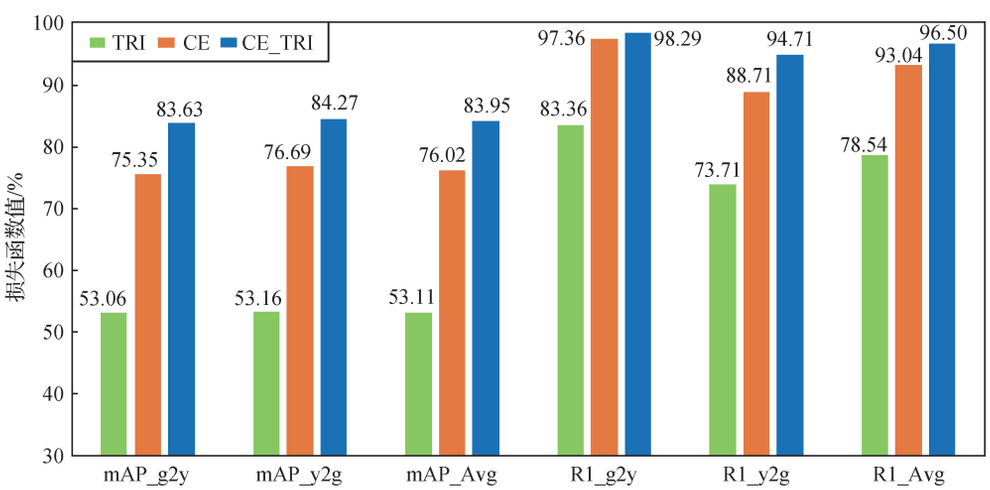


图 12　采用不同损失函数的结果柱状图

Fig. 12　Histogram of results using different loss functions

元组损失的双约束损失。实验结果表明,使用交叉熵损失(CE loss)和三元组损失(TRI loss)的双约束损失下检索效果最优,mAP_Avg 值为 83.95%,R1_Avg 值为 96.5%,相较于仅使用 CE 损失分别提高了 7.93% 和 3.46%;相较于仅使用 TRI 损失分别提高了 30.84% 和 17.96%。

交叉熵损失衡量的是实际输出与期望输出之间的差异大小。差异越大,参数调整得越快,收敛速度也越快。虽然三元组损失可以针对细粒度赤足足迹图像的特点实现增大类间距离且减小类内距离,但仅使用三元组损失而省略交叉熵损失,在一定程度上加大了训练难度。从图 12 中对比数据可知,只采用 TRI 时,mAP_Avg 值为 53.11%,R1_Avg 值为 78.54%,相较于 CE 损失,分别降低了 22.91% 和 14.5%。因此,在本文研究任务上,交叉熵损失的实验效果更优于三元组损失。而本文采用双约束损失 CE_TRI,融合了 CE 损失和 TRI 损失的优点,均优于仅使用 CE 损失或仅使用 TRI 损失的实验效果。

3.4.3 特征嵌入模块后采用不同池化方法的对比

为验证不同池化方法对检索性能的影响,本文分别在特征嵌入模块后采用 3 种池化方法进行对比分析,即最大池化(MaxPool)、平均池化(AvgPool)和广义平均池化(GmPool)。该对比实验结果如图 13 所示,其中横坐标为 3 种不同池化方法的检索精度。从对比数据可以看出,采用广义平均池化效果最好,mAP_Avg 和 R1_Avg 分别为 83.95% 和 96.5%,相较于最大池化分别提高了 4.79% 和 2.18%;相较于平均池化分别提高了 3.07% 和 1.46%。

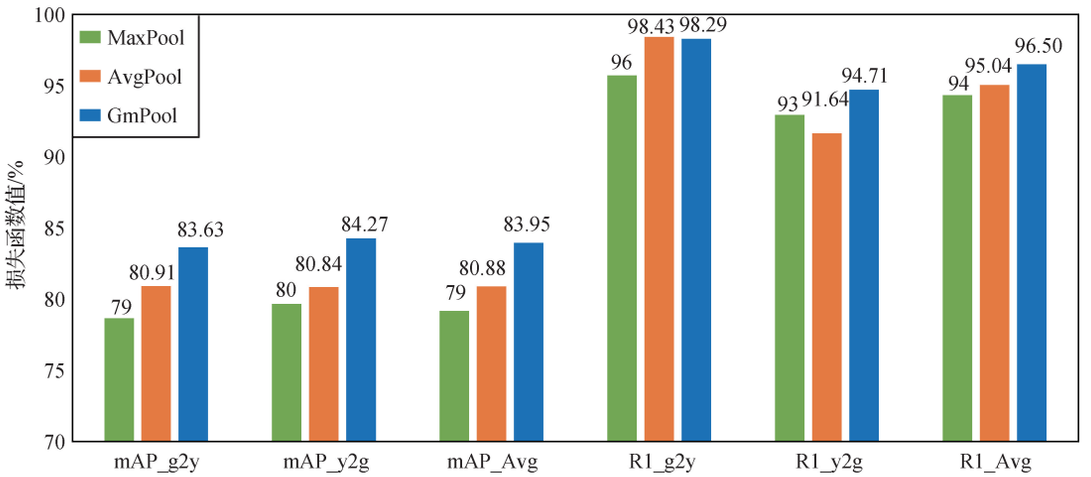


图 13 特征嵌入模块后采用不同池化方法的结果柱状图

Fig. 13 Histogram of results adopting different pooling methods after the feature embedding module

最大池化是对邻域内特征点取最大,能够更多地保留图像的纹理特征;均值池化是对邻域内特征点求平均,能够更多地保留图像的背景信息;而广义平均池化是介于这两种池化方式之间的,在特定的情况下可以转换为最大池化和平均池化,其池化层中的超参数可以在网络的反向传播过程中不断学习(Radenović等,2019)。相较于最大池化和平均池化,本文将广义平均池化应用到卷积神经网络中,可使网络具备更高的灵活性和实用性,以更好地适应跨模态检索任务。

3.4.4 与其他跨模态检索方法的比较

为了验证本文算法的有效性,在本文采用的赤足足迹数据集上,分别与跨模态检索领域提出的具有代表性的 FGC 方法(He 等,2019)和 HC 方法(Zhu 等,2019)进行对比分析。FGC 的研究对象是细粒度图像及其对应的文本、声音和视频的多模态数据,取得了一定的跨模态检索效果;HC 的研究对象分别是三通道彩色图的可见光行人图像和单通道灰度图的红外光行人图像,也取得了不错的效果。实验结果如表 4 所示。结果表明,本文算法效果最优,mAP_Avg、R1_Avg、R5_Avg 和 R10_Avg 值分别为 83.95%、96.50%、98.13% 和 98.74%,相较于 FGC 分别高出 40.01%、36.50%、20.42% 和 16.1%;相较于 HC 分别高出 26.07%、19.32%、20.95% 和 6.81%。

FGC 是将每个模态原始输入样本在第 1 维度上进行拼接后再经过一个单分支 ResNet50 网络进行多模态的学习,损失函数采用交叉熵损失和中心

表 4　本文算法与 FGC 和 HC 方法的结果对比

Table 4　Comparison of results among FGC, HC and ours

/%

方法	mAP_Avg	R1_Avg	R5_Avg	R10_Avg
FGC	43.94	60.00	77.71	82.64
HC	57.88	77.18	77.18	91.93
本文	**83.95**	**96.50**	**98.13**	**98.74**

注：加粗字体表示各列最优结果。

损失，以实现各模态类别的分类以及增加各模态中同类别特征的紧致性，但是该中心损失仅有益于拉近同模态的类内特征距离，忽略了跨模态的类内之间的距离。

HC 方法采用的也是双分支网络架构，损失函数包括交叉熵损失和异质中心（hetero-center）损失。其中，异质中心损失通过拉近跨模态同类别样本的中心距离减少跨模态类内差异，但是由于双分支网络结构的复杂性以及嵌入模块位于网络的后端，难以有效地学习跨模态赤足足迹特征。

本文设计的跨模态赤足足迹检索算法采用双分支网络，结构简单，有助于网络学习到不同模态的足迹特征。同时，运用 non-local 注意力机制获取更有辨别性的特征，采用交叉熵损失和三元组损失的双约束方式，考虑到了模态内和模态间的特征差异问题，对整个网络进行有效优化，在赤足足迹图像的跨模态检索任务上，相较于 FGC 方法和 HC 方法，具有更高效的跨模态检索性能。

4　结　论

足迹检索在犯罪案件侦查、身份识别等领域发挥着重要作用。针对目前足迹检索中采集设备种类多样化的问题，本文以光学和压力赤足足迹为研究对象，构建了 138 人的光学和压力赤足足迹图像数据库，提出一种基于细粒度足迹图像的非局部注意力双分支网络模型，对赤足足迹图像的跨模态互检索问题进行了研究。该模型由特征提取、特征嵌入和双约束损失 3 个模块组成。在特征提取模块的各个分支上均采用 ResNet50 作为基础网络进行足迹有效特征的学习。为了减小跨模态的异构性，在特征嵌入模块通过参数共享以学习到一个多模态的共享空间。同时，为了获得更有辨别性的多模态足迹特征，该模型在网络的 layer2 层和 layer3 层的所有残差块上采用了 non-local 注意力机制，用于学习特征图中每个空间位置与所有空间位置之间的相似度权重关系，从而增强足迹有用特征，同时突出跨模态的共享特征。为了增大不同类别之间差异和减小相同类别之间的差异，本文算法还采用了交叉熵损失和三元组损失的双约束损失进行网络优化。

在跨模态赤足足迹数据集上的实验结果验证了本文采用的 non-local 注意力机制、不同损失函数和在特征嵌入模块后采用不同池化方法对跨模态赤足足迹检索效果较好。同时，与细粒度图像的跨模态检索方法 FGC 和跨模态行人重识别方法 HC 在足迹数据集上的实验效果进行对比分析，结果表明，本文方法的 mAP 和 rank1 指标均取得最优效果。

由于本文主要采用注意力机制构造一个跨模态共享空间，且针对的研究对象是基于光学和压力两个模态的赤足足迹图像，所以未来考虑使用对抗网络学习更有效的足迹特征。下一步将考虑在穿鞋、穿袜和捺印等其他模态足迹图像中，利用对抗网络中的生成器和判别器，更好地解决模态内和模态间的特征差异问题，进一步提升跨模态检索精度。

参考文献（References）

Bao W X, Wang Y F, Wang N and Tang J. 2020a. Identification algorithm of optical footprint image based on metric learning kernel function. Journal of Huazhong University of Science and Technology (Natural Science Edition), 48(11): 11-16 (鲍文霞, 王云飞, 王年, 唐俊. 2020a. 基于度量学习核函数的光学足迹图像识别算法. 华中科技大学学报(自然科学版), 48(11): 11-16) [DOI: 10.13245/j.hust.201103]

Bao W X, Qu J J, Wang N, Tang J and Lu X L. 2020b. Force-tactile footprint recognition based on spatial aggregation weighted convolutional neural network. Journal of Southeast University (Natural Science Edition), 50(5): 959-964 (鲍文霞, 瞿金杰, 王年, 唐俊, 鲁玺龙. 2020b. 基于空间聚合加权卷积神经网络的力触觉足迹识别. 东南大学学报(自然科学版), 50(5): 959-964) [DOI: 10.3969/j.issn.1001-0505.2020.05.023]

Cao Y, Long M S, Wang J M and Zhu H. 2016. Correlation autoencoder hashing for supervised cross-modal search//Proceedings of 2016 ACM on International Conference on Multimedia Retrieval. New York, USA: ACM: 197-204 [DOI: 10.1145/2911996.2912000]

Dai Z Z, Chen M Q, Gu X D, Zhu S Y and Tan P. 2019. Batch DropBlock network for person re-identification and beyond//Proceedings

of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3690-3700 [DOI: 10.1109/ICCV.2019.00379]

Gurney J K, Kersting U G and Rosenbaum D. 2008. Between-day reliability of repeated plantar pressure distribution measurements in a normal population. Gait and Posture, 27(4): 706-709 [DOI: 10.1016/j.gaitpost.2007.07.002]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He X T, Peng Y X and Xie L. 2019. A new benchmark and approach for fine-grained cross-media retrieval//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 1740-1748 [DOI: 10.1145/3343031.3350974]

Heydarzadeh M, Birjandtalab J, Pouyan M B, Nourani M and Ostadabbas S. 2017. Gaits analysis using pressure image for subject identification//Proceedings of 2017 IEEE EMBS International Conference on Biomedical and Health Informatics. Orland, USA: IEEE: 333-336 [DOI: 10.1109/BHI.2017.7897273]

Khokher R, Singh R C and Kumar R. 2015. Footprint recognition with principal component analysis and independent component analysis. Macromolecular Symposia, 347(1): 16-26 [DOI: 10.1002/masy.201400045]

Kulkarni P S and Kulkarni V B. 2015. Human footprint classification using image parameters//Proceedings of 2015 International Conference on Pervasive Computing. Pune, India: IEEE: 1-5 [DOI: 10.1109/PERVASIVE.2015.7087011]

Liang J, He R, Sun Z N and Tan T N. 2016. Group-invariant cross-modal subspace learning//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: AAAI Press: 1739-1745

Liu H J, Cheng J, Wang W, Su Y Z and Bai H W. 2020. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. Neurocomputing, 398: 11-19 [DOI: 10.1016/j.neucom.2020.01.089]

Nirenberg M S, Ansert E, Krishan K and Kanchan T. 2019. Two-dimensional metric comparison between dynamic bare and sock-clad footprints for its forensic implications——a pilot study. Science and Justice, 59(1): 46-51 [DOI: 10.1016/j.scijus.2018.09.001]

Osisanwo F Y, Adetunmbi A O and Álese B K. 2014. Barefoot morphology: a person unique feature for forensic identification//Proceedings of the 9th International Conference for Internet Technology and Secured Transactions. London, UK: IEEE: 356-359 [DOI: 10.1109/ICITST.2014.7038837]

Radenović F, Tolias G and Chum O. 2019. Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7): 1655-1668 [DOI: 10.1109/TPAMI.2018.2846566]

Wang C, Yang H J and Meinel C. 2015. Deep semantic mapping for cross-modal retrieval//Proceedings of the 27th IEEE International Conference on Tools with Artificial Intelligence. Vietri sul Mare, Italy: IEEE: 234-241 [DOI: 10.1109/ICTAI.2015.45]

Wang X L, Girshick R, Gupta A and He K M. 2018. Non-local neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Wang Z X, Wang Z, Zheng Y Q, Chuang Y Y and Satoh S. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 618-626 [DOI: 10.1109/CVPR.2019.00071]

Wu A C, Zheng W S, Yu H X, Gong S G and Lai J H. 2017. RGB-Infrared cross-modality person re-identification//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5390-5399 [DOI: 10.1109/ICCV.2017.575]

Xue Y L and Yue J. 2012. Effect of perpetrators' mentality on footprints. Journal of Fujian Police College, 26(4): 55-60 (薛亚龙, 岳佳. 2012. 论案犯心态对足迹反映的影响. 福建警察学院学报, 26(4): 55-60)

Ye M, Wang Z, Lan X Y and Yuen P C. 2018a. Visible thermal person re-identification via dual-constrained top-ranking//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI: 1092-1099 [DOI: 10.24963/ijcai.2018/152]

Ye M, Lan X Y, Li J W and Yuen P C. 2018b. Hierarchical discriminative learning for visible thermal person re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 7501-7508

Zhao B W, Zhang L F and Pan Z F. 2020. Comparison of image filtering methods based on OpenCV. China Computer & Communication, 32(15): 78-80 (赵博文,张力夫,潘在峰. 2020. 基于 OpenCV 的图像滤波方法比较. 信息与电脑,32(15):78-80)

Zhao Y B, Lin J W, Xuan Q and Xi X G. 2019. HPILN: a feature learning framework for cross-modality person re-identification. IET Image Processing, 13(14): 2897-2904 [DOI: 10.1049/iet-ipr.2019.0699]

Zheng Y, Zhang Y J and Larochelle H. 2014. Topic modeling of multimodal data: an autoregressive approach//Procedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1370-1377 [DOI: 10.1109/CVPR.2014.178]

Zhu M, Wang T S, Wang N, Tang J and Lu X L. 2020. Footprint pressure image retrieval algorithm based on multi-scale self-attention convolution. Pattern Recognition and Artificial Intelligence, 32(12): 1097-1103 (朱明, 汪桐生, 王年, 唐俊, 鲁玺龙. 2020. 基于多尺度自注意卷积的足迹压力图像检索算法. 模式识别与人工智能, 33(12): 1097-1103) [DOI: 10.16451/j.

cnki. issn1003-6059. 202012004］

Zhu Y X, Yang Z, Wang L, Zhao S, Hu X and Tao D P. 2019. Hetero-center loss for cross-modality person re-identification. Neurocomputing, 386: 97-109［DOI:10. 1016/j. neucom. 2019. 12. 100］

作者简介

鲍文霞,1980 年生,女,副教授,主要研究方向为图像处理、模式识别和机器学习。E-mail: bwxia@ ahu. edu. cn

杨先军,通信作者,男,研究员,主要研究方向为智能信号处理、检测技术和智能系统。E-mail: xjyang@ iim. ac. cn

茅丽丽,女,硕士研究生,主要研究方向为基于足迹图像的跨模态检索研究。E-mail: 2961192214@ qq. com

王年,男,教授,主要研究方向为计算机视觉、模式识别和智能系统。E-mail: wn_xlb@ ahu. edu. cn

唐俊,男,教授,主要研究方向为计算机视觉、机器学习和模式识别。E-mail: 151746212@ qq. com

张艳,女,副教授,主要研究方向为传统机器学习、深度学习。E-mail: zhangyan@ ahu. edu. cn