DOI:10.13878/j.cnki.jnuist.2017.06.007

梅舒欢[1,2]　闵巍庆[2]　刘林虎[2,3]　段华[1]　蒋树强[2]

基于 Faster R-CNN 的食品图像检索和分类

摘要

面向食品领域的图像检索和分类等方面的研究成为多媒体分析和应用领域越来越受关注的研究课题之一.当前的主要研究方法基于全图提取视觉特征,但由于食品图像背景噪音的存在使得提取的视觉特征不够鲁棒,进而影响食品图像检索和分类的性能.为此,本文提出了一种基于 Faster R-CNN 网络的食品图像检索和分类方法.首先通过 Faster R-CNN 检测图像中的候选食品区域,然后通过卷积神经网络(CNN)方法提取候选区域的视觉特征,避免了噪音的干扰使得提取的视觉特征更具有判别力.此外,选取来自视觉基因库中标注好的食品图像集微调 Faster R-CNN 网络,以保证 Faster R-CNN 食品区域检测的准确度.在包括 233 类菜品和 49 168 张食品图像的 Dish-233 数据集上进行实验.全面的实验评估表明:基于 Faster R-CNN 食品区域检测的视觉特征提取方法可以有效地提高食品图像检索和分类的性能.

关键词

食品图像;图像检索;图像分类;深度学习;Faster R-CNN;卷积神经网络

中图分类号 TP391.41

文献标志码 A

收稿日期 2017-07-28

资助项目 国家自然科学基金(61532018,61602437,61672497,61472229,61202152);北京市科技计划(D161100001816001);山东省自然科学基金(ZR2017MF02);山东省科技发展计划(2016ZDJS02A11,2014GGX101035,2014BSB01020)

作者简介

梅舒欢,男,硕士,研究方向为多媒体检索及其应用.shuhuan.mei@vipl.ict.ac.cn

段华(通信作者),女,博士,副教授,主要从事支持向量机、机器学习、图论等方向的研究工作.huaduan59@163.com

1 山东科技大学 数学与系统科学学院,青岛,266590

2 中国科学院计算技术研究所 智能信息处理重点实验室,北京,100190

3 中国科学院大学 人工智能技术学院,北京,100049

0 引言

Web2.0 的迅速发展使得食品图像分享网站得到迅速的发展,例如国内的大众点评网和国外的 Yummly 网站等.食品图像检索和识别可以对网络中的食品图像实现有效的组织、总结和检索.自动的食品图像检索和识别也是食品和健康等许多领域中最有前途的应用之一,可以进一步帮助估计食品的热量和分析人的饮食习惯,实现个性化的服务.因此本文主要解决面向食品领域的图像检索和分类问题.

由于食品图像检索和识别广泛的应用价值,近年来,越来越多的研究者开始研究面向食品图像的分析、检索和分类等问题.例如文献[1]用一种基于统计的方法计算食品图像的特征,实现食品图像的识别.但是该方法仅适用于在标准的食品图像数据集上,不具有泛化性.文献[2]提出用随机森林的方法提取图像中的局部视觉特征实现食品图像的分类.随着深度学习技术的发展,基于 CNN 的方法[3]已经成为提取图像视觉特征的主流方法.例如文献[4]采用 AlexNet 网络提取图像的视觉特征实现食品图像的检测和分类.文献[5]采用 Google 的 Inception 网络提取图像的视觉特征实现食品图像的分类.文献[6]采用 GoogLeNet 网络提取图像视觉特征实现食品图像和非食品图像的分类.这些方法主要针对整张食品图像进行视觉特征提取,没有考虑食品图像背景信息对食品图像分类的影响.在现实世界中,拍摄的食品图片不仅包含食品本身的视觉信息,还包含各种各样的背景信息.图 1 展示了来自 Dish 数据集[7]一些菜品图像的例子,比如 CBD 寿司图片中包含人的背景信息;三文鱼图片包含调料、虾等其他物的背景信息.现有的方法由于针对整张图像信息提取视觉特征,会不加区分地把不相关的背景信息也作为菜品视觉表示的一部分,从而影响食品图像检索和分类的性能.

图 1　来自 Dish-233 数据集的一些菜品图片

Fig. 1　Some food images from Dish-233 dataset

为了解决该问题,本文提出了一种基于 Faster R-CNN 的食品图像检索和分类方法,如图 2 所示.该方法主要包括以下 2 个步骤:首先微调(fine-tune)Faster R-CNN,使用训练的 Faster R-CNN 网络检测食品图像的食品区域;然后基于检测的食品区域,利用 CNN 深度神经网络提取该候选区域的视觉特征;最后将提取的食物图像的视觉特征应用到食品图像检索和分类任务中.本文在 Dish-233 食品数据集上进行实验,该数据集为菜品数据集[7]的一个子集,包括 233 类菜品和 49 168 张图像.全面的实验评估表明本文提出的方法相较于其他方法,提取的视觉特征更具有判别性,在食品图像检索和分类任务中,性能均得到了改善.

本文的主要贡献包括以下 2 个方面:1)提出了一种基于 Faster R-CNN 的食品图像检索和分类方法,由于提取的视觉特征仅针对图片的食物区域因而更为鲁棒和更具有判别力;2)将提出的方法应用到 Dish-233 数据集上,全面的实验评估验证了本文所提方法的有效性.

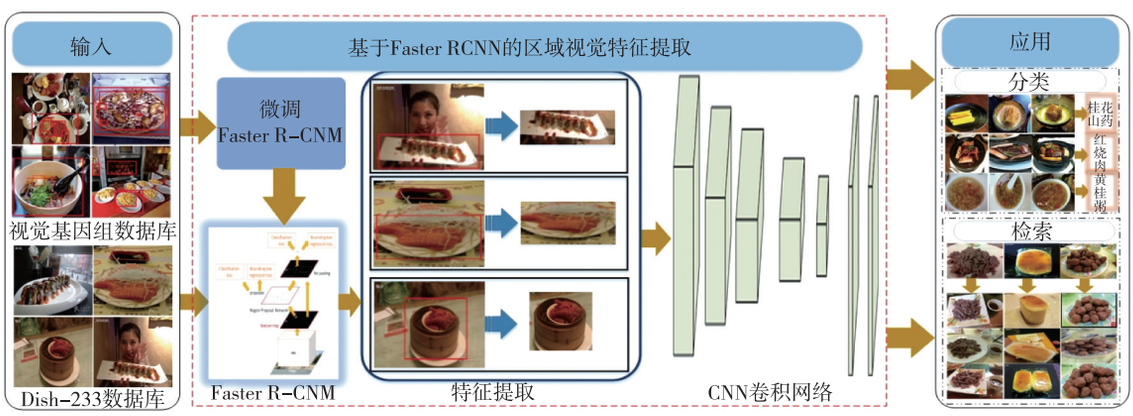

图 2　基于 Faster R-CNN 的食品图像检索和分类方法

Fig. 2　Faster R-CNN based food image retrieval and classification method

1　相关工作

本节将对近年来食品图像检索和分类相关技术和方法进行介绍,主要包括 2 部分,第 1 部分主要介绍食品图像检索和分类的相关技术方法;第 2 部分主要介绍基于物体检测的卷积神经网络的相关方法.

1.1　食品图像检索和分类

近年来,食品图像检索和分类受到了越来越多研究者的关注.例如 Farinella 等[8]提出了一种基于纹理 Anti-Textons 特征用于食品图像的检索与分类的方法.Yang 等[1]提出了一种基于统计的食品图像的特征表示方法用于食品图像的分类,但该方法仅局限于标准的食品.Xu 等[7]通过随机森林的方法学习食品图像中判别性的特征表示实现菜品识别.相比于以上浅层模型的方法,Kagaya 等[4]利用当前的深度卷神经网络(CNN)提取食品图像的特征用于食品图像的检测和识别.Hassannejad 等[5]进一步采用更深层的网络实现食品图像的识别.还有一些工作集中在面向餐馆上下文信息的菜品识别[10]和面向移动端的菜品识别[11-12].此外,近来的一些工作例如文献[10,13-14]则进一步考虑图像的原料信息,以一种多任务的方法建模原料信息、视觉信息和类别信息之间的关联.Salvador 等[15]学习食品图像和原料信息等不同模态信息的嵌入,实现跨模态的检索.

本文也研究基于面向食品领域的图像检索和分类.不同以上基于整张图片提取视觉特征的方法,考虑食品图像中包含许多和食品无关的背景信息,因此提出了首先用 Faster R-CNN 方法检测食品的图像区域,然后利用 CNN 方法提取目标物体区域的视觉特征实现面向食品图像的检索与分类.

1.2　基于物体检测的卷积神经网络(CNNs)

卷积神经网络(CNNs)已作为一种主流的特征提取方法成功应用到许多任务中,例如图像的分类[16]和检索[17-18].相比于传统的视觉方法,CNNs 能够提取更为丰富的语义信息.为了将基于 CNN 的方法应用到物体检测任务中,许多基于检测的 CNN 方法被相继提出.例如 Girshick 等[19]提出了 R-CNN 深

度框架,该框架首先利用 Object Proposal 算法提取图像的候选区域,将这些候选区域作为输入进行模型训练.为了改进 R-CNN 的速度和准确度,有学者提出了 SPP-Net[20] 和 Fast-RCNN[21].Ren 等[22] 进一步引入了 Faster R-CNN 网络,提出了一个 Region Proposal Network(RPN)方法用于克服 object proposal 的依赖.近来一些学者为了能够更快地检测物体,提出了 YOLO9000[23].考虑到资源的消耗和准确度性能等的因素,本工作充分利用 Faster R-CNN 算法提取食品图像的区域,提取食品区域的特征并将其应用到食品图像的检索和分类任务中.

2 基于 Faster R-CNN 的区域视觉特征提取

为了有效提取食品图像的视觉特征,采用了以下 2 个步骤:1)微调 Faster R-CNN;2)基于食品检测区域的 CNN 特征提取.

如上分析,大多数食品图像都含有不相关的背景信息,如果只针对食品区域进行特征提取,将会降低食品图像背景信息带来的影响.为了解决该问题,需要对图像的食品区域进行检测.Faster R-CNN[9] 已经成为物体区域检测的有效方法,因而也可采用 Faster R-CNN 检测图像的食品区域.但是现有的 Faster R-CNN 模型主要基于 VOC2007 中 20 类常见的物体预训练得到的模型,并没有涉及到与食品相关的类别,为此首先从视觉基因库中选择和食品类别相关的标定好的图像,然后利用选择的食品图像数据集微调 Faster R-CNN,最终得到每张食品图像的候选区域.

$$L(p_i,t_i)=\frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(p_i,p_i^*)+\lambda\frac{1}{N_{\text{reg}}}\sum_i p_i^* L_{\text{reg}}(t_i,t_i^*), \tag{1}$$

这里 i 是一个 mini-batch 中 anchor 的索引,p_i 表示的是 anchor i 的预测概率,如果 anchor 正向,p_i^* 为 1,反之 p_i^* 为 0.t_i 表示预测边界框的 4 个参数坐标,t_i^* 是与正 anchor 相对应的 ground-truth box 的坐标向量,分类损失 L_{cls} 是 2 个类别(目标与非目标)的对数损失.

微调 Faster R-CNN 后,利用微调的 Faster R-CNN 网络模型获得食品图像的候选框和对应的每个候选框的得分.

CNNs 网络已经成为视觉特征提取的有效方法.对于每张图像得分较高的候选区域,根据候选框的坐标,用 AlexNet 网络提取 FC7 层的特征,然后将得分较高的区域的特征进行串联得到最终图像的特征表示.

对于图像检索任务,给定一张查询图像,基于微调的 Faster R-CNN 和 CNN 提取食品图像的视觉特征,通过和查询数据库进行相似度计算返回检索的结果.对于食品图像分类任务,对所有的训练集和测试集的食品图像通过上述方法提取视觉特征,然后通过训练集训练分类器,基于训练的分类模型得到分类的结果.在提取图像食品区域的视觉特征时,通常选取得分最高的图像候选区域.

3 实验评估

在本节中,首先描述实验设置,包括数据集和实现细节,然后在 Dish-233 食品数据集上验证所提方法在食品图像检索和分类任务中的有效性.

3.1 数据库

利用 Dish-233 数据集来验证本文所提方法的有效性.原始的菜品数据集[7] 包含 117 504 张图像和 11 611 种菜品类别.从中挑选出图像数量大于或者等于 15 的菜品类别,最终获得 233 种菜品类别和 49 168 张图像,称这个数据集为 Dish-233.图 3 展示了 Dish-233 数据集中的一些食品类别的例子.

图 3 Dish-233 数据集的一些食品类别的图像

Fig. 3 Some food images in Dish-233 dataset

3.2 实现细节

为了利用 Faster R-CNN[9] 对食品图像区域进行检测,需要微调(fine-tune)Faster R-CNN,为此需要带有区域边框标注的食品图像数据集.视觉基因组数据库(visual genome)[24] 包含了标定的 108 077 张图像,其中包括大量的食品图像.因此利用 Dish-233 数据集的类别名,将其翻译成成英文,作为查询词,构建查询词典列表,利用关键词匹配从视觉基因数

据集中选取食品图像.为了得到更多的食品图片,进一步选用其他食品数据库的类别信息(比如 Food-101),选取食品图片,然后经过人工进一步的筛选,去掉非食品图像,最终得到 10 641 张食品图像及对应标定的区域,称之为 VisGenome-11K.图 4 展示了来自视觉基因带有标注框的菜品图像.

图 4　来自视觉基因库带有标定区域的一些菜品例子

Fig. 4　Some annotated food images from the visual genome dataset

对于模型的参数设置,在 Faster R-CNN 训练过程中,最小批尺寸(mini-batch)为从一张图像中提取的 256 个 anchor,迭代次数为 80 000.其中在前 60 000 迭代,学习率设为 0.001,在后 20 000 迭代,学习率设为 0.000 1.momentum 参数设为 0.9,权重衰减参数设为 0.000 5.在微调 AlexNet 模型时,将初始学习率设为 0.001,每 20 个时期(epoch)之后,将学习率调整为之前的 0.1.最大迭代次数设为 60 个时期(epoch).

将 VisGenome-11K 图像集划分成 2 部分,80%用于训练集,20%用于验证集.利用 VisGenome-11K 微调 Faster R-CNN,然后通过微调的 Faster R-CNN 模型对 Dish 数据集的食品图像进行区域检测得到图像中的食品区域,再利用在 ImageNet 上预训练的 AlexNet 模型从检测的食品区域提取 4096-D 的视觉特征.图 5 是经过微调 Faster R-CNN 网络得到的一些食品图片的检测结果,从中可以看到经过食品区域检测,可以排除背景噪音的干扰,使得提取的视觉特征更具有判别性.

对所有的图像进行区域检测和视觉特征提取之后,将其应用到食品图像检索和分类任务中.

3.3　检索任务

对于检索任务,从 Dish-233 数据集的每一类菜品图像集中随机选取 25%的图像作为查询图像,然后将数据集的全集作为查询数据库进行检索.

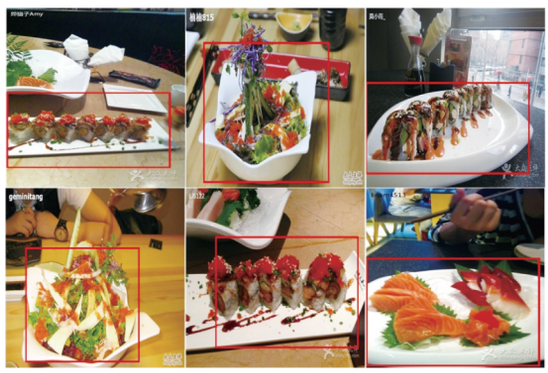

图 5　基于 Faster R-CNN 食品区域检测的一些食品例子

Fig. 5　Some detected results by Faster R-CNN

3.3.1　评价指标和比较方法

采用 Precision 和 MAP 2 个评价指标,这 2 个指标均为信息检索中常用的指标.为了验证本文方法的有效性,和以下方法进行了比较:1) CNN-G[25].该方法主要是利用 7 层的 AlexNet 直接提取全局图像的视觉特征;2) CNN-G-F.相比 CNN-G,CNN-G-F 首先利用训练集对 AlexNet 网络进行微调,利用微调的网络提取整张图像的视觉特征;3) Faster R-CNN-G.该方法直接用 Faster R-CNN 网络检测图像的候选食品区域,然后对得分最高的候选区域用微调的 AlexNet 网络提取视觉特征.方法 1)和 2)均未用 Faster R-CNN 进行区域检测,方法 3)是为了进一步说明通过 VisGenome-11K 微调 Faster-RCNN 产生的影响.

3.3.2　检索结果及分析

分别用上述 4 种方法在 Dish-233 数据集上进行检索实验.具体来说,采用 Precision@K 和 MAP@K (K 表示在检索过程中返回候选图像的数量),K=$\{1,20,40,60,80,100\}$.图 6 展示了 4 种不同方法检索在这 2 个指标的检索结果.从中可以得出以下结论:1) CNN-G-F 要比 CNN-G 的检索性能好,说明通过微调 AlexNet 网络可以得到更适合 Dish 数据集的视觉特征;2) Faster R-CNN-G 和本文方法要超过 CNN-G-F,说明经过 Faster R-CNN 对食品图像区域进行检测后可以有效减少食品图像背景信息产生的干扰,进而提高了检索性能;3)本文方法比 Faster R-CNN-G 的性能有适度的提升,说明利用 VisGenome-11K 微调可以改进食品图像检测的准确度.

图 7 展示了一些例子的检索结果.从中可以看到本文方法在所有方法中检索结果是最好的,这进

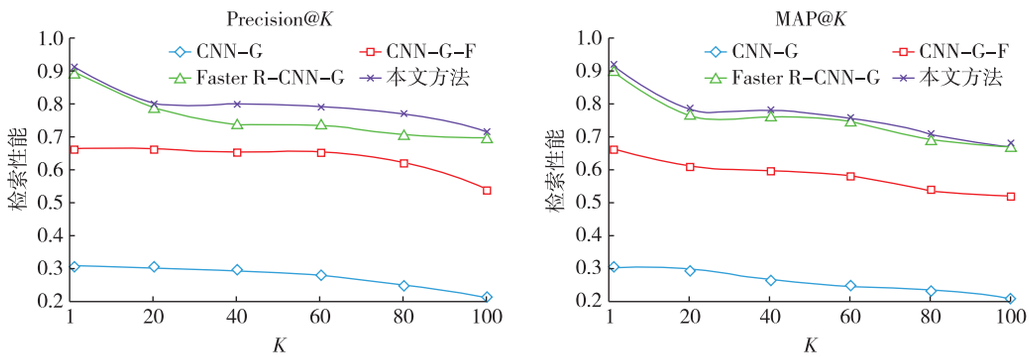

图 6　不同方法的检索性能比较

Fig. 6　Retrieval performance comparison between the proposed and other methods

图 7　不同方法检索结果的例子

Fig. 7　Some retrieval results from different methods

一步验证了其有效性.

3.4　分类任务

对于分类任务,将每一类 75%的数据集作为训练集,25%的数据集作为测试集.由于分类任务为单标签的,所以采用准确率(accuracy)作为评价指标.为了验证本文方法的有效性,采用和检索任务相同的比较算法进行比较

图 8 展示了不同方法的分类性能.从中可以看到本文方法的性能最好,相比于 CNN-G-F,性能提升了 5 个百分点.

图 9 展示了随机选取的 20 类菜品在不同方法的分类结果.可以看到在大多数例子中,本文方法的分类性能是最好的,进一步验证了其有效性.

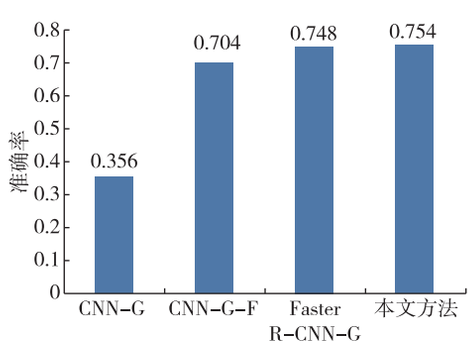

图 8　不同方法的分类准确率比较

Fig. 8　Classification accuracy performance of different methods

4　总结与展望

本文提出了一种基于 Faster R-CNN 的食品图像

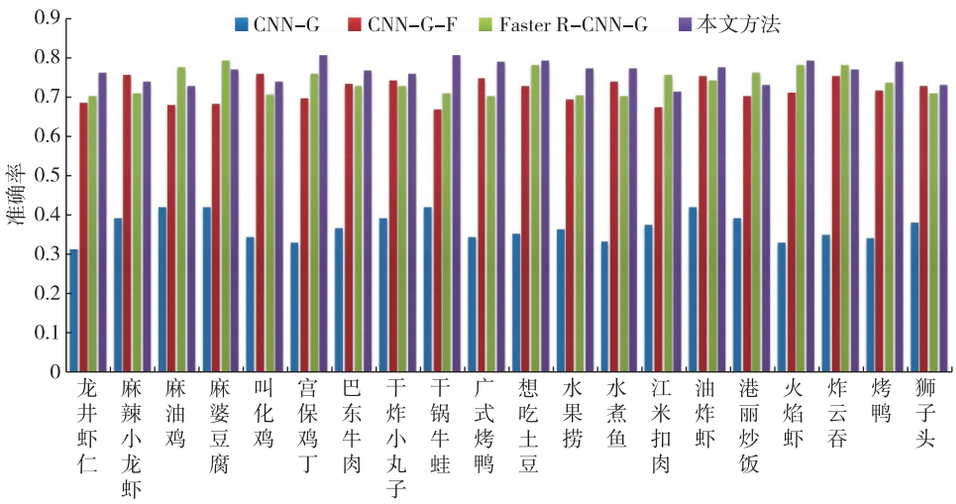

图 9　随机选取的 20 类菜品在不同方法的分类结果

Fig. 9　Classification performance of different methods on randomly selected 20 food categories

检索和分类方法.该方法主要利用 Faster R-CNN 网络检测图像的食品区域,对检测的食品图像区域通过 CNN 网络提取视觉特征.相比于传统的基于 SIFT 和 CNN 的全局视觉特征提取方法,本文方法所提取的视觉特征更为鲁棒.将本文方法应用到食品图像检索和分类任务中,并在 Dish-233 数据集上进行实验,实验结果验证了其有效性.在未来的研究中将考虑以下研究方向:1)在更多更大规模的食品数据集上进行实验,比如 Food-101[2],以验证本文方法的可扩展性;2)考虑更多食品图像的上下文信息,比如地理位置信息等实现基于上下文的食品图像检索和分类.另外,将其应用到移动设备中实现移动食品图像的检索和分类,以及针对食品区域的热量估计[25]等也将作为后续工作探索的研究方向.

参考文献

References

[1] Yang S L,Chen M,Pomerleau D,et al.Food recognition using statistics of pairwise local features[C]//IEEE Conference on Computer Vision and Pattern Recognition,2010:2249-2256

[2] Bossard L,Guillaumin M,Van Gool L.Food-101-mining discriminative components with random forests[C]//European Conference on Computer Vision,2014:446-461

[3] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems,2012:1097-1105

[4] Kagaya H,Aizawa K,Ogawa M.Food detection and recognition using convolutional neural network[C]//ACM International Conference on Multimedia,2014:1085-1088

[5] Hassannejad H,Matrella G,Ciampolini P,et al.Food image recognition using very deep convolutional networks[C]//International Workshop on Multimedia Assisted Dietary Management,2016:41-49

[6] Singla A,Yuan L,Ebrahimi T.Food/non-food image classification and food categorization using pre-trained GoogLeNet model[C]//International Workshop on Multimedia Assisted Dietary Management,2016:3-11

[7] Xu R,Herranz L,Jiang S Q,et al.Geolocalized modeling for dish recognition[J].IEEE Transactions on Multimedia,2015,17(8):1187-1199

[8] Farinella G M,Allegra D,Moltisanti M,et al.Retrieval and classification of food images[J].Computers in Biology & Medicine,2016,77:23-39

[9] Krishna R,Zhu Y K,Groth O,et al.Visual genome:Connecting language and vision using crowdsourced dense image annotations[J].International Journal of Computer Vision,2017,123(1):32-73

[10] Min W Q,Jiang S Q,Wang S H,et al.A delicious recipe analysis framework for exploring multi-modal recipes with various attributes[C]//ACM International Conference on Multimedia,2017(in press)

[11] Dehais J,Anthimopoulos M,Mougiakakou S.Dish detection and segmentation for dietary assessment on smartphones[C]//International Conference on Image Analysis and Processing,2015:433-440

[12] Tanno R,Okamoto K,Yanai K.Deepfoodcam:A DCNN-based real-time mobile food recognition system[C]//International Workshop on Multimedia Assisted Dietary Management,2016:89

[13] Chen J J,Ngo C-W.Deep-based ingredient recognition for cooking recipe retrieval[C]//ACM on Multimedia Conference,2016:32-41

[14] Min W Q,Jiang W Q,Sang J T,et al.Being a super cook:Joint food attributes and multi-modal content modeling for recipe retrieval and exploration[J].IEEE Transactions on Multimedia,2017,19(5):1100 - 1113

[15] Salvador A,Hynes N,Aytar Y,et al.Learning cross-modal embeddings for cooking recipes and food images[C]// IEEE Conference on Computer Vision and Pattern Recognition,2017:3020-3028
[16] Szegedy C,Liu W,Jia Y Q,et al.Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition,2015:1-9
[17] Tolias G,Sicre R,Jégou H.Particular object retrieval with integral max-pooling of CNN activations[J].arXiv e-print,2015,arXiv:1511.05879
[18] Radenovic F,Tolias G,Chum O.CNN image retrieval learns from DoW:Unsupervised fine-tuning with hard examples[C]//European Conference on Computer Vision,2016:3-20
[19] Girshick R,Donahue J,Darrell T,et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition,2014:580-587
[20] He K M,Zhang X Y,Ren S Q,et al.Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European Conference on Computer Vision,2014:346-361
[21] Girshick R.Fast R-CNN[C]// IEEE International Conference on Computer Vision,2015:1440-1448
[22] Ren S Q,He K M,Girshick R,et al.Faster R-CNN:Towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149
[23] Redmon J,Farhadi A.YOLO9000:Better,faster,stronger[J].arXiv e-print,2016,arXiv:1612.08242
[24] Min W Q,Bao B K,Mei S H,et al.You are what you eat:Exploring rich recipe information for cross-region food analysis[C]// IEEE Transactions on Multimedia,2017(In public)
[25] Meyers A,Johnston N,Rathod V,et al.Im2Calories:Towards an automated mobile vision food diary[C]// IEEE International Conference on Computer Vision,2015:1233-1241

Faster R-CNN based food image retrieval and classification

MEI Shuhuan[1,2] MIN Weiqing[2] LIU Linhu[2,3] DUAN Hua[1] JIANG Shuqiang[2]
1 College of Mathematics and Systems Science,Shandong University of Science and Technology,Qingdao 266590
2 Key Lab of Intelligent Information Processing,Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190
3 School of Artificial Intelligence,University of Chinese Academy of Sciences,Beijing 100049

Abstract Automatic understanding of food images has various applications in different fields,such as food intake monitor and food calorie estimation.Thus,the research on food related tasks,such as food image retrieval and classification has been one of the hot research topics in the field of multimedia analysis and applications recently.Existing methods mainly extract the visual features from the whole food image for further food analysis.The extracted features are lacking in robustness because of the background interference from the images.In order to solve this problem,we propose a Faster R-CNN (Region-based Convolutional Neural Network) based food retrieval and classification method.For the solution,we first detect the food candidate regions using Faster R-CNN,and then adopt the CNN network to extract the visual features from the detected food regions.Such extracted features are more discriminative for reducing the background interference.Furthermore,we select the annotated food images from the Visual Genome dataset to fine-tune the Faster R-CNN to guarantee its performance.We conduct the experiment on two datasets:Food-101 with 101 classes and 10 641 food images,and Dish-233 with 233 dishes and 49 168 images.The extensive evaluation demonstrates the effectiveness of the proposed Faster R-CNN based food visual feature extraction method in food image retrieval and classification.

Key words food image;image retrieval;image classification;deep learning;Faster R-CNN;convolutional neural network