计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150689
53(1):113-122,2016

面向智能交互的图像识别技术综述与展望

蒋树强　闵巍庆　王树徽
(中国科学院智能信息处理重点实验室(中国科学院计算技术研究所)　北京　100190)
(sqjiang@ict.ac.cn)

Survey and Prospect of Intelligent Interaction-Oriented Image Recognition Techniques

Jiang Shuqiang, Min Weiqing, and Wang Shuhui
(*Key Laboratory of Intelligent Information Processing (Institute of Computing Technology, Chinese Academy of Sciences), Chinese Academy of Sciences, Beijing* 100190)

Abstract　Vision plays an important role in both the human interaction and human-nature interaction. Furthermore, equipping the terminals with the intelligent visual recognition and interaction is one of the core challenges in artificial intelligence and computer technology, and also one of lofty goals. With the rapid development of visual recognition techniques, in recent years the emerging new techniques and problems have been produced. Correspondingly, the applications with the intelligent interaction also present a few new characteristics, which are changing our original understanding of the visual recognition and interaction. We give a survey on image recognition techniques, covering recent advances in regarding to visual recognition, visual description, visual question and answering (VQA). Specifically, we first focus on the deep learning approaches for image recognition and scene classification. Next, the latest techniques in visual description and VQA are analyzed and discussed. Then we introduce visual recognition and interaction applications in mobile devices and robots. Finally, we discuss future research directions in this field.

Key words　image recognition; intelligent visual recognition; intelligent interaction; visual description; visual question and answering (VQA); deep learning

摘　要　视觉在人与人交互以及人与自然界的交互过程中起到非常重要的作用,让终端设备具有智能的视觉识别和交互能力是人工智能和计算机技术的核心挑战和远大目标之一.可以看到,近年来视觉识别技术发展飞速,新的创新技术不断涌现,新的研究问题不断被提出,面向智能交互的应用呈现出一些新的动态,正在不断刷新人们对此领域的原有认识.从视觉识别、视觉描述和视觉问答3个角度对图像识别技术进行综述,对基于深度学习的图像识别以及场景分类技术进行了具体介绍,对视觉描述和问答技术的最新技术进行了分析和讨论,同时对面向移动终端和机器人的视觉识别和交互应用进行了介绍,最后对该领域的未来研究趋势进行了分析.

收稿日期:2015-07-26;修回日期:2015-10-20
基金项目:国家自然科学基金重点项目(61532018);国家自然科学基金优秀青年科学基金项目(61322212);国家自然科学基金青年科学基金项目(61303160);国家"九七三"重点基础研究发展计划基金项目(2012CB316400)
This work was supported by the National Key Natural Science Foundation of China (61532018), the National Natural Science Foundation for Excellent Young Scholars of China (61322212), the National Natural Science Foundation of China Young Scientists Fund (61303160), and the National Basic Research Program of China (973 Program) (2012CB316400).

关键词 图像识别;智能的视觉识别;智能交互;视觉描述;视觉问答;深度学习

中图法分类号 TP391

人类得以在自然界中长期生存,一个重要的原因就是拥有迅速认识并理解其所处环境的能力,而这其中的关键环节是利用人类视觉系统完成对目标的定位与识别,同时实现视觉场景的理解与描述.如果计算机能够实现自动的图像识别,必将进一步丰富与方便人类生活,这促使图像识别技术成为当前人工智能领域内重要的研究方向之一.图像识别是指利用计算机视觉、模式识别、机器学习等技术方法,自动识别图像中存在的一个或多个语义概念,广义的图像识别还包括对识别的概念进行图像区域定位等.图像识别技术可以满足用户在不同场景下的视觉应用需求,主要包括面向互联网的图像检索与挖掘、面向移动设备和机器人等智能终端的人机对话与信息服务等.

最早的图像识别技术可以追溯到20世纪60年代[1],自20世纪90年代以来,随着计算机的处理能力越来越强,图像识别技术得到了很大的进步与发展.从最早的数字识别、手写文字识别逐渐发展到人脸识别、物体识别、场景识别、属性识别、精细目标识别等,所采用的技术也从最早的模板匹配、线性分类到现在所广泛使用的深层神经网络与支持向量机分类的方法.特别是进入21世纪10年代以来,随着计算能力的大幅度提升、新的计算方法的不断提出、可利用的数据资源的大规模增长、新型应用模式不断涌现,图像识别及其应用技术无论在研究的广度和深度上、在识别效果的性能上、在技术及应用的扩展上,都呈现出新的趋势.其中有4个特点比较突出:1)图像的特征表示已经从传统的手工设定演变为如今的自动学习方法,这主要得益于深度神经网络技术的广泛应用;2)图像识别的概念已由早期个别概念(如特定概念、十几个概念的识别)转变为成百上千的概念,这主要是由于大规模图像数据集的发展所推动的,如ImageNet[2],Places[3],SUN397[4]等;3)图像识别技术正在和自然语言理解技术进行融合,形成了图像描述技术,有别于图像识别只是对图像进行个别概念的标注,图像描述可以自动对一副图像进行一句话或一小段话的描述,从而可以更全面地描述图像内容;4)在应用模式上,传统的图像识别技术或者是为了服务于监控、检索等特定的应用场景,或只是为了突破计算机视觉的挑战性问题,在技术研究时并未过多考虑全面图像识别技术的应用场景.随着技术发展,一些面向智能交互与服务的应用模式也逐渐引起了研究者的关注,这也进一步促进了图像识别技术的发展.

本文将对图像识别与应用技术的最新进展进行介绍.在方法上,将首先对基于深度学习的图像识别技术进展进行讨论,主要从物体识别和场景识别2个角度探讨相关技术的特点.ImageNet是最新的常用数据集,主要是物体概念的图像,也包括少量场景概念的图像,该数据集是当前不同深度学习模型的训练数据来源,也是算法性能的主要测试场地;而随着SUN397,Places等大规模场景数据集的出现和普及,场景分类技术成为当前图像识别的重要研究问题,在分类方法和模型训练上都有新的推进,本文也将进行介绍.此外,我们对近一两年来研究颇多的图像描述与问答技术也进行介绍,这是最新研究方向.在面向视觉交互的图像识别应用上,将主要对面向移动终端与面向机器人的视觉识别技术进行讨论,同时对基于图像理解的智能交互的不同应用模式进行分析.在本文的最后,将对未来的研究趋势进行展望和讨论.

1 基于深度学习的图像识别技术

自从Krizhevsky等人[5]在ImageNet上训练一个8层的深度模型并在ImageNet竞赛上取得非常好的效果后,卷积神经网络(convolutional neural network, CNN)在图像分类与识别领域受到了广泛关注,取得了巨大成功.之后,在很多图像识别的应用场景中,卷积神经网络也都取得了很大的性能改进.卷积神经网络能够逐层学习图像的特征,其中低层是具有普遍性的(general)特征,如图像的边缘、角点、纹理等;高层特征是低层特征的组合,是针对特定任务的有针对性的(specific)特征[6-7].逐层特征学习模拟了人脑分层处理信息机制,能够直接从原始像素得到图像特征.将卷积神经网络用于图像识别与分类,可以归纳为3种途径:

1) 直接在待分类的数据集上训练一个深层的网络.随着CNN深度和宽度的增加,CNN的分类性能有着明显的提升.Simonyan等人[8]提出了一个19层的CNN模型(VGG-19),该模型在原来Krizhevsky[5]提出的模型的基础上通过增加卷积层来增加该模型

的深度,由于在所有的层上采用比较小的卷积滤波核(3×3),因而可在实践中实现.相比之下,Szegedy等人[9]基于Hebbian原理和多尺度处理的启发提出了一个22层的深度学习模型GoogLeNet[9],它是由多个Inception Model堆叠而成.该模块中,利用不同带宽的卷积核对前一层的输出做卷积,最后合并形成后一层的输入.不同尺寸大小的卷积核能够捕获多尺度的视觉特征,这些特征的融合能够使整个网络更好地适应图像物体的表观多尺度特性.另外针对不同的分类任务,如场景分类和物体分类等,不同数据集上训练的模型也有不同的特性,例如Zhou等人在Places[3]上训练的深度模型,对于场景的分类有非常好的效果.

2) 在训练好的网络上直接提取特征.训练好的CNN模型可以直接用来当特征提取器,提取的特征可以用做其它的后续操作.Donahue等人[10]利用Krizhevsky提出的模型将CNN的全连接层的特征与SVM分类器结合,在多个数据集上取得了很好的分类效果,这表明CNN的高层全连接层的特征可以作为通用的视觉特征.相比之下,Liu等人[11]采用跨卷积层池化技术将卷积层的特征作为通用特征在MIT-67等数据库上取得了更好的分类效果.Gong等人[12]在多个尺度下基于图像块提取CNN特征,然后通过主成分分析(principal component analysis, PCA)降维以及局部聚合的描述子向量(vector of locally aggregated descriptors, VLAD)[13]编码等形成图像的特征.相比于直接从整幅图片上提取CNN特征,该方法提取的特征具有几何不变性.Li等人[14]更进一步在提取图像的多个块级特征的基础上,通过关联规则来发现隐藏在这些特征之间的模式,从而实现图像的分类和识别.

3) 在目标数据集上对现有深度模型进行"精细化"调整(fine-tuning).在特定数据集上训练好的模型有很强的泛化性能,但是fine-tuning能够进一步提升分类性能[15].fine-tuning是在目标数据集上重新调整网络参数,从而使深度模型能够捕获针对目标任务更具有区分性的特征[16-17].

表1给出了基于CNN的分类方法在不同的数据集上最好的分类准确率.从Krizhevsky等人的8层的CNN-S[5]网络到Simonyan等人的22层网络GoogLeNet[9],随着网络层次的增加,CNN的性能有很大提升.表2给出了2014年ImageNet大规模视觉识别挑战(ILSVRC 2014)[2,18]的排名前7的结果,这些团队均是采用深度学习模型得到测试结果.如表2所示,GoogLeNet由于采用最多的22层网络而达到最好的测试性能;VGG采用19层网络紧随其后;相比于增加深度学习模型的层数,SPPNet[18]网络通过将空间金字塔模型引入到深度学习模型中,消除了输入图像尺寸的限制,在网络层数最多只有7层的条件下组合多个深度学习模型,达到了第3名的测试结果.从表2我们可以看出,这些深度学习模型的架构基本没有什么变化,可以通过1)增加网络层数学习更为抽象的表示;2)消除深度学习中的某些限制或者瓶颈,比如输入图像尺寸的限制等途径继续通过深度学习模型提高识别性能.

Table 1 Object Classification Accuracy on Different Datasets

表1 不同数据集的物体分类准确率

Datasets	Classes	Total Samples	Best Methods	Accuracy /%
Caltech101	101	9 144	SPPNet	93.42
Caltech256	256	30 607	CNN-S	77.61
VOC2007	20	9 963	HCP[19]	85.20
ImageNet	10 000	1 281 167	GoogLeNet	93.33

Table 2 Results of ILSVRC 2014 Classification[18]

表2 不同深度学习模型在ILSVRC 2014的物体分类结果[18]

Rank	Methods	Top-5Test
1	GoogLeNet	6.66
2	VGG	7.32
3	SPPNet	8.06
4	Howard	8.11
5	DeeperVision	9.50
6	NUS-BST	9.79
7	TTIC ECP	10.22

2 场景分类技术

场景分类技术一般分为2步:1)提取图像的中层特征描述;2)基于中层特征描述训练分类器,并进行场景分类.近10年来,场景分类技术的发展主要体现在中层特征描述能力的不断增强.典型的中层描述特征为词袋(bag-of-word)[20],该方法利用聚类得到视觉特征码书,根据码书进行编码,得到词袋特征,进而用SVM进行分类.Li等人[21]提出了一种基于物体描述的中层特征,预先学习物体检测器,检测器的响应即为其物体描述特征.Rasiwasia等人[22]利用场景类别概率分布作为中层描述,对每一场景类别学习狄利克雷混合模型,以预测未知图像属于

该场景类别的概率,所有场景类别概率的分布即为该图像的中层特征描述.具体来说,对于每一个在语义空间中的每个场景类别通过如下狄利克雷混合分布表示:

$$P_{\Pi|Y}(\pi \mid y;\Lambda^y)=\sum_k \beta_k^y Dir(\pi;\alpha_k^y). \quad (1)$$

这里模型参数为 $\Lambda^y=\{\beta_k^y,\alpha_k^y\}$,$Dir(\pi;\boldsymbol{\alpha})$ 参数为 $\boldsymbol{\alpha}=\{\alpha_1,\alpha_2,\cdots,\alpha_L\}$.该工作仅仅考虑全局的共生模式,为了改进图像的特征表示能力,Song 等人[23]利用局部空间和多特征上下文信息优化了文献[22]的中层描述,增强了特征描述能力.相比于以上方法,当前最有效的场景分类方法是深度学习方法,即训练 CNN,利用末层决策层分类.CNN 虽不同于传统 2 步分类框架,并没有明显的中层特征描述,但网络的中间层结果也可被认为是一种中层特征描述,且也可结合 SVM 分类器用以分类.近两年 CNN 的发展主要体现在 2 方面:1)更深层的网络,如 VGG-NET[8]和 GoogLeNet;2)更丰富的训练图像数据,如 Places.随着网络深度的增加,识别正确率也大幅度提升;同时由于数据集的丰富,训练集能涵盖更多场景类别,场景分类技术已能看到实际应用的曙光.例如,麻省理工学院目前发布的关于场景识别的演示[24]已能达到正确识别大部分室内外和自然场景的效果.表 3 给出了当前最好的不同场景数据集下的分类性能.

Table 3 Scene Classification Accuracy on Different Datasets
表 3 不同数据集的场景分类性能

Datasets	Classes	Total Samples	Best Methods	Accuracy /%
Scene15	15	4 485	CNN-Places[3]	90.2
MIT67	67	1 520	CNN+Fisher[25]	79.2
SUN397	397	108 762	CNN+Fisher[25]	61.7
Places	205	2 448 873	CNN-Places[3]	66.2

3 图像描述技术

通过目标检测和分类技术,可以将图片中用户感兴趣的部分从复杂的背景中分离出来并对其进行分类.在此基础上,通过目标描述技术,我们可以使用更加丰富的信息来产生更进一步的结果:自动产生自然语言来对视觉目标进行描述.

随着计算机视觉和自然语言理解领域相关技术的突破,图片描述[26-33]技术是在 2014—2015 年获得了突飞猛进的发展.在 2015 年微软 COCO 图片标注竞赛中,来自微软[26-27]、谷歌[28]、蒙特利尔大学、多伦多大学[29]和加州大学伯克利分校[30-31]等研究机构的最新工作在人工测评和图灵测试方面都取得了令人惊叹的成绩.谷歌(基于 CNN 视觉特征和 RNN(recurrent neural network)语言模型)和微软(基于区域的单词检测和最大熵语言模型)目前在技术和性能方面处于领先地位.

目前,在目标描述这一方面的解决方案主要都是根据通过编码-解码(encoder-decoder)的想法而来,最有代表性的方法有 2 种:

1) 类似于 Fang 等人[26]使用的流程化方法:根据图片得到单词,再将单词组合为句子,最后对句子进行打分.Fang 等人[26]首先利用多示例学习(MIL)方法,根据图片的各个部分产生相对应的名词、动词和形容词;接下来,使用最大熵语言模型(MELM)产生包含提取词的句子;最后,使用最小错误率训练(MERT)对所产生的所有句子进行打分并排序.

2) 类似于 Vinyals 等人[28]和 Karpathy 等人[30]使用的端到端(end-to-end)方法:受机器翻译技术的启发,将图片整体转化为特征,再将特征转化为一个完整的句子.Karpathy 等人[30]利用 CNN 模型将图片整体转化为一个特征,再利用 RNN 模型根据已产生的单词预测句子中的下一个单词,最终生成一个完整的描述.

对于整体流程中各个步骤的研究也有许多进展,比如对于流程化方法:Kiros 等人[34]提出的 SC-NLM(structure-content neural language model),它与其他模型的不同之处在于它根据已生成的单词预测的并不是下一个单词而是接下来的句子结构.对于端到端方法,Mao 等人[35]提出的 m-RNN(multimodal recurrent neural network)模型,它通过一个 multimodal 的部分将 CNN 和 LM 联系起来.Donahue 等人[31]提出的 LRCNs(long-term recurrent convolutional networks)模型可以在可变长度的输入和可变长度的输出之间直接建立映射关系.这与 Chen 等人[36]在图片和描述映射关系方面提出的方法有类似之处,该方法并未将图片和描述映射到同一空间,而是在图片和描述之间直接建立双向映射关系.最近,Jia 等人[37]则是采用 gLSTM(guiding long-short term memory)模型,如图 1 所示,在 LSTM 模型[28]的基础上引入外部的语义信息生成图像标题.具体来说,gLSTM 块的内存细胞和门定义为

$$i'_l = \sigma(W_{ix}x_l + W_{im}m_{l-1} + W_{ig}g), \quad (2)$$

$$f'_l = \sigma(W_{fx}x_l + W_{fm}m_{l-1} + W_{fg}g), \quad (3)$$

$$o'_l = \sigma(W_{ox}x_l + W_{om}m_{l-1} + W_{og}g), \quad (4)$$

$$c'_l = f'_l \odot c'_{l-1} + i'_l \odot h(W_{cx}x_l + W_{cm}m_{l-1} + W_{cg}g), \quad (5)$$

$$m_l = o'_l c'_l, \quad (6)$$

其中,$\odot$表示逐项相乘;$\sigma(\cdot)$表示 S 形函数;$h(\cdot)$表示双曲正切函数;i'_l,f'_l,o'_l,c'_l 和 m'_l 分别表示输入门、遗忘门、LSTM 细胞的输出门、内存单元细胞的状态门和隐状态;x_l 表示在时间 l 的序列元素;$W_{[\cdot][\cdot]}$代表模型参数;g 为引入的语义信息.相比于标准的 LSTM 架构,gLSTM 引入了新的语义项,该项成为连接视觉和文本域的桥梁.

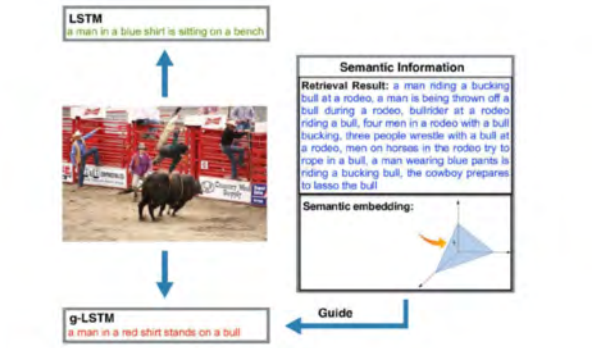

Fig. 1 Image caption generation using LSTM and the proposed gLSTM[37].

图 1 用 LSTM 和 gLSTM 生成图像标题[37]

表 4 给出了不同方法在生成图像标题性能的结果,评价指标采用了 BLEU 量度[38].从表 4 中我们看到最新的方法 Hard-Attention 和 gLSTM 达到最好的性能.

Table 4 Comparison of Different Methods on MS COCO

表 4 不同图像标题生成模型在 MS COCO 的性能比较

Methods	B@1	B@2	B@3	B@4
Multimodal RNN[30]	62.5	45.0	32.1	23.0
Google NIC[28]	66.6	46.1	32.9	24.6
LRCN-CaffeNet[31]	62.8	44.2	30.4	
m_RNN[38]	67.0	49.0	35.0	25.0
Soft-Attention[29]	70.7	49.2	34.4	24.3
Hard-Attention[29]	**71.8**	**50.4**	35.7	25.0
gLSTM	67.0	49.1	**35.8**	**26.4**

4 视觉问答技术

基于图像内容识别与分类的另一个新的应用场景是视觉问答,这也是近期受研究者关注的一个新方向.该技术将自然语言理解与视觉内容描述相结合,可以根据当前图像内容与用户问题产生出相应的回答.针对当前的视觉问答主要有推理和端到端的深度学习 2 种方法.

推理方法比较有代表性的是 Malinowski 等人[39]提出的使用基于不确定输入的多世界(multi-world)方法实现对于真实世界的场景问答:该方法使用带有深度信息的数据集 NVU-Depth V2dataset,对于场景使用语义分割算法[40]构建世界并且收集关于物体的识别信息,例如物体类别、3D 位置和颜色;然后利用对于一个场景的多种 world 解释,这里的 world 解释是由语义分割产生;最后通过概率模型来得到最大后验概率的答案.

端到端的深度学习方法主要输入为自由形式的问题文本.答案的输出主要分为:1)Malinowski 等人[41]和 Gao 等人[42]基于 RNN 框架,可以产生自由形式答案;2)Geman 等人[43]和 Ma 等人[44]提出的基于分类方式产生答案框架.Gao 等人[42]采用 long-short term memory (LSTM)抽取输入问题的表示,同时利用 CNN 抽取视觉图像表示,再利用一个 LSTM 存储答案中的语言环境信息,最后利用一个融合组件将 3 种成分进行融合产生答案.Ma 等人[44]对输入问题使用 CNN 生成输入问题表示,同时利用 CNN 生成图像的视觉表示并使用映射矩阵将其映射到与问题表示相同的向量长度,最后将 2 个表示向量进行混合后再次使用卷积与 softmax 进行分类输出对应的答案,如图 2 所示:

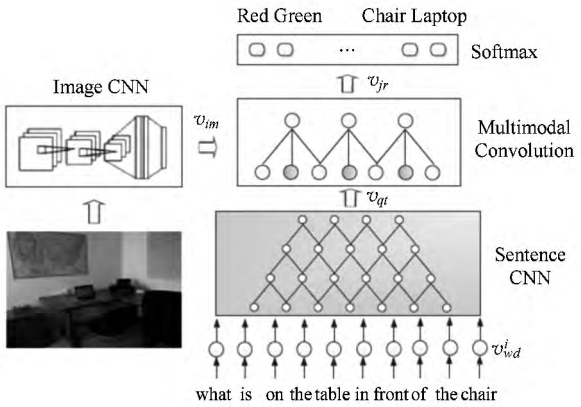

Fig. 2 The proposed CNN model for image QA[44].

图 2 提出的图像问答的 CNN 模型[44]

目前针对视觉问答的工作还不多,但是已经可以看到深度学习在这个领域中已经有了比较好的表现.这主要得益于目前深度学习在视觉表示和自然语言理解等领域都有了长足的发展.

5 面向移动终端的视觉识别技术

近些年来移动设备(如手机、平板)越来越普及,这些设备大多装配有摄像头和图形芯片,此外还有GPS和无线联网等功能.这些都促使移动端的视觉识别应用越来越多,常见的包括地标建筑物识别[45-46]、商品识别[47-48]、食品识别[49-50]、艺术品识别[51]等,上线的APP如Goggles[52]等.

由于面向移动端,一些方法关注移动设备资源的合理利用,如提高传输速度、减小内存开销等.Tsai等人[47]提取低码率的CHoG特征[53],并利用了位置直方图编码对特征描述子的位置进行压缩,最后用几何验证的方法对检索结果进行重排序.He等人[48]将图像的局部特征编码到位数较少的哈希码,而非对视觉单词(VW)进行量化,从而将图像表示成词袋型哈希码,然后采用边界特征对检索结果进行重排序.

移动设备带有丰富的传感器,可以为图像提供拍照时的上下文信息,如GPS获取的地理位置信息、拍摄时间、相机参数等,所以有些工作利用这些信息对图像中的目标进行识别.Runge等人[54]将图像的地理标签、时间、图像主颜色、天气等各种信息与图像的视觉特征组合成一个特征向量,然后利用分类器预测图像的概念标签.Chen等人[45]基于SIFT描述子训练得到的词汇树,计算数据库中的图像与查询图像的相似度,排除地理相距非常远的地标建筑,然后在特征空间使用近似近邻(ANN)的方法对查询图像进行识别.Dhiraj和Luo[55]对视觉和地理检测器分别训练并使用相同的权重在预测阶段进行融合.进一步地,Li等人[56]对不同概念分别学习了不同检测器的权重.Xu等人[49]研究了利用地理信息辅助视觉识别菜品类别的问题.为了对分类模型进行地理约束,该文提出地理局部化模型,将地理上下文信息用于分类模型的训练过程,使得模型从根本上对地理信息更有针对性,最后再根据查询图像的地理坐标对这些分类模型进行自适应组合,从而实现菜品类别的预测.该方法用到的图像特征就是训练好的深度特征.

近年来,由于深度学习很强的特征学习能力已应用到各种移动视觉识别任务中.例如,Teradeep[57]公司已经针对移动和嵌入式设备开发了一套基于深度学习的算法实现移动端的场景理解、物体检测和识别等.百度等搜索公司[58]也将深度学习技术比如DNN[5]等应用到基于移动端的人脸识别、鞋识别和检索等视觉任务中.

6 面向机器人的视觉识别技术

视觉识别技术在机器人的领域也扮演着举足轻重的角色.作为机器人感知外界环境信息的一个重要输入渠道,其对于机器人理解周围场景和辅助完成特定任务具有至关重要的作用.目前视觉识别技术在机器人领域的应用主要有环境理解[59-62]、自学习物体识别[63-64]和智能交互[63]、导航与避障[65]等.

面向机器人的视觉识别技术不同于其他单纯的视觉识别方法,其具有一定的交互能力(语言、动作等)和多感知能力(深度信息感器、定位装置等),对于机器人的视觉能力可以具有一定的辅助作用.从机器人视觉感知方式上可以分为2种:基于2D图像的识别和基于3D视觉信息的识别.

1) 2D图像识别中主要是对获取到的图像进行物体检测和整体场景识别.基于2D图像的识别可以直接对图像进行特征提取或者对图像进行区域特征提取然后使用模型进行标签预测.Rouanet等人[63]的方法在交互过程中利用用户指定区域,从而缩小图像区域,然后对该区域提取特征并进行物体识别,这里为了进行增量式学习,采用了产生式模型进行物体识别.Wang等人[61]给出了一种实例级物体识别方法,利用图像检索方式匹配输入图像与数据库中的图像,再经过空间一致性验证和投票机制实现物体的识别,这种方法识别精度比较高,但是缺点是对于识别的物体不具有很好的泛化能力.

2) 3D图像识别主要是借助可以获取深度信息的传感器例如Kinect或者激光测距实现对于环境内的物体深度感知.额外的深度信息可以帮助机器人感知物体位置及大小.Lv等人[62]利用Kinect采集的深度信息和人体骨骼信息进行手持物体分割,同时提取多种模态特征训练分类模型,从而实现对人手上物体的理解.Filliat等人[59]主要针对室内的物体进行识别.采用PCL库[66]将获取到的3D数据映射到点云空间中,通过检测去除地板和墙壁等噪音同时进行物体分割,然后使用多种特征结合作为前馈神经网络输入学习到综合特征表示.

视觉识别技术是机器人感知外界信息的重要渠道,因此未来在交互过程中利用视觉识别技术以增强机器人理解能力和提升与用户交互体验也具有很重要的研究价值,是一个具有挑战性的方向.例如利

用图像识别技术同时识别人脸和物体,可以帮助关联理解用户意图和兴趣爱好.目前受到大家广泛研究关注的图像描述和问答技术也会很快和机器人的视觉交互应用相结合,产生新的研究内容和应用场景,从而进一步促进视觉识别技术的发展和进步.

7 总结和展望

由于相关理论和技术的长足发展,在过去20年中,视觉识别和智能交互技术发生了日新月异的变化.从小数据到大数据,从手工设计特征到以深度学习为代表的视觉特征学习,从简单内容到自然场景,从简单模型到复杂模型,从单一输出到复杂输出,从视觉识别到视觉理解、进一步到视觉描述和问答,视觉识别和智能交互技术已经逐渐从实验室走向现实的应用场景,相关方法尤其在深度学习方法、视觉和自然语言处理等技术深度结合的方面发展速度快,技术更新多.视觉交互的主要形式从普通设备逐渐迁移到智能终端和机器人,视觉信息处理能力越来越强,人机交互的体验也越来越真实.

通过以上分析和讨论,视觉识别和智能交互技术呈现4个发展趋势:1)深度学习方法由于其突出的泛化能力和视觉特征捕捉能力,将被应用在更深层次、多角度的视觉识别和理解的各项技术当中;2)视觉识别和理解将与语言和认知技术进行更深入全面的结合,使得更加高级的视觉理解和描述性语义输出取代简单的物体、场景识别而成为下一个10年的研究热点;3)视觉识别和理解将会在具体的应用中进行更深层次的融合和适配,如特定内容的图像和视频识别等;4)随着视觉描述和视觉问答的兴起,智能终端和机器人的视觉能力将在人机智能交互中起到越来越重要的作用,并将逐渐从较为局限的人机对话模式,进化为基于多通道智能信息处理的自然交互.

与此同时,在视觉识别和智能交互技术发展的过程中也面临着许多挑战.主要包括3个方面:1)通过深度学习技术提高性能的一种主流方法是通过增加网络层数来增加识别的准确度.但是更深的网络需要更多训练的参数,这就意味着需要更多的训练样本和训练时间.因此,怎样设计网络模型如网络深度、卷积核的个数、卷积核的大小等以及如何快速地训练得到高性能模型将是深度学习技术面临的一个重要挑战.2)尽管现有的视觉识别和理解技术取得了巨大的进展,但是现有的视觉识别技术仍然只能理解简单的场景,设计理解复杂场景的视觉技术也是未来视觉技术发展的一个难点问题.3)现有的视觉识别技术依然以视觉信息为主,但是随着各种传感器的迅速发展,我们可以得到各种各样的上下文信息,如果将视觉信息和这些上下文信息高效有机结合将对提高视觉识别的性能有很大的改进,尤其是在面向基于机器人的视觉识别应用中.如果未来能够比较好地解决这些技术问题,视觉识别和智能交互技术有望在未来越来越多的领域中造福人类社会,更加深入地为人类的生产、生活、消费和娱乐等方面提供智能化、个性化和全面化的服务.

参 考 文 献

[1] Andreopoulos A, Tsotsos J K. 50 years of object recognition: Directions forward [J]. Computer Vision and Image Understanding, 2013, 117(8): 827-891

[2] Russakovsky O, Deng Jia, Su Hao, et al. ImageNet: Large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252

[3] Zhou Bolei, Lapedriza A, Xiao Jianxiong, et al. Learning deep features for scene recognition using Places database [C] //Proc of the 28th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 487-495

[4] Xiao Jianxiong, Hays J, Ehinger K, et al. Sun database: Large-scale scene recognition from abbey to zoo [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3485-3492

[5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] //Proc of the 26th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105

[6] Yosinski J, Clune J, Bengio Y, et al. How transferable features in deep neural networks [C] //Proc of the 28th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 3320-3328

[7] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2014: 297-312

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. CoRR abs/1409.1556, 2014

[9] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9

[10] Donahue J, Jia Yangqing, Vinyals O, et al. DeCAF: A deep convolutional activation feature for generic visual recognition [C] //Proc of the 31st Int Conf on Machine Learning. New York: ACM, 2014: 647-655

[11] Liu Lingqiao, Shen Chunhua, Hengel A. The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4749-4757

[12] Gong Yunchao, Wang Liwei, Guo Ruiqi, et al. Multi-scale orderless pooling of deep convolutional activation feature [C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2014: 392-407

[13] Jegou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3304-3311

[14] Li Yao, Liu Lingqiao, Shen Chunhua. Mid-level deep pattern mining [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 971-980

[15] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets [C] //Proc of the British Machine Vision Conf. Nottingham, UK: British Machine Vision Association, 2014

[16] Agrawal P, Girshick R, Malik J. Analyzing the performance of multilayer neural networks for object recognition [C] // Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2014: 329-344

[17] Azizpour H, Razavian A S, Sullivan J, et al. From Generic to specific deep representation for visual recognition [C] // Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 36-45

[18] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[19] Wei Yunchao, Xia Wei, Huang Junshi, et al. CNN: Single-label to multi-label [J]. CoRR abs/1406.5726, 2014

[20] Dixit M, Chen Si, Gao Dashan et al. Scene classification with semantic Fisher Vectors [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3485-3492

[21] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2006: 2169-2178

[22] Li Lijia, Su Hao, Xing E, et al. Object bank: A high-level image representation for scene classification and semantic feature sparsification [C] //Proc of the 24th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2010: 1378-1386

[23] Rasiwasia N, Vasconcelos N. Holistic context models for visual recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(5): 902-917

[24] Song Xinhang, Jiang Shuqiang, Herranz L. Joint multi-feature spatial context for scene recognition in the semantic manifold [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1312-1320

[25] MIT. Places [EB/OL]. [2015-07-10]. http://places.csail.mit.edu/demo.html

[26] Fang Hao, Gupta S, Iandola F, et al. From captions to visual concepts and back [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1473-1482

[27] Devlin J, Cheng Hao, Fang Hao, et al. Language models for image captioning: The quirks and what works [C] //Proc of the 2015 Conf of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2015: 100-105

[28] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156-3164

[29] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [J]. CoRR abs/1502.03044, 2015

[30] Karpathy A, Li F. Deep visual-semantic alignments for generating image descriptions [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3128-3137

[31] Donahue J, Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2625-2634

[32] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4566-4575

[33] Chen Xinlei, Zitnick C L. Mind's eye: A recurrent visual representation for image caption generation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2422-2431

[34] Kiros R, Salakhutdinov R, Zemel R. Unifying visual-semantic embeddings with multimodal neural language models [J]. CoRR abs/1411.2539, 2014

[35] Mao Junhua, Xu Wei, Yang Yi, et al. Explain images with multimodal recurrent neural networks [J]. CoRR abs/1410.1090, 2014

[36] Chen Xinlei, Zitnick C L. Mind's eye: A recurrent visual representation for image caption generation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2422-2431

[37] Jia Xu, Gavves E, Fernando B, et al. Guiding long-short term memory for image caption generation [J]. CoRR, abs/1509.04942, 2015

[38] Mao Junhua, Xu Wei, Yang Yi, et al. Deep captioning with multimodal recurrent neural networks (m-RNN) [J]. CoRR, abs/1412.6632, 2014

[39] Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input [J]. CoRR, abs/1410.0210, 2014

[40] Gupta S, Arbelaez P, Malik J. Perceptual organization and recognition of indoor scenes from RGB-D images [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 564-571

[41] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images [J]. CoRR, abs/1505.01121, 2015

[42] Gao Haoyuan, Mao Junhua, Zhou Jie, et al. Are you talking to a machine? Dataset and methods for multilingual image question answering [J]. CoRR, abs/1505.05612, 2015

[43] Geman D, Geman S, Hallonquist N, et al. Visual turing test for computer vision systems [J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(12): 3618-3623

[44] Ma Lin, Lu Zhengdong, Li Hang. Learning to answer questions from image using convolutional neural network [J]. CoRR, abs/1506.00333, 2015

[45] Chen D, Baatz G, Koser K, et al. City-scale landmark identification on mobile devices [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 737-744

[46] Lim J H, Li Yiqun, You Yilun, et al. Scene recognition with camera phones for tourist information access [C] //Proc of the IEEE Int Conf on Multimedia & Expo. Piscataway, NJ: IEEE, 2007: 100-103

[47] Tsai S S, Chen D, Chandrasekhar V, et al. Mobile product recognition [C] //Proc of the Int Conf on Multimedia. New York: ACM, 2010: 1587-1590

[48] He Junfeng, Feng Jinyuan, Liu Xianglong, et al. Mobile product search with Bag of Hash Bits and boundary reranking [C] //Proc the IEEE Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 16-21

[49] Xu Ruihan, Herranz L, Jiang Shuqiang, et al. Geolocalized modeling for dish recognition [J]. IEEE Trans on Multimedia, 2015,17(8): 1187-1199

[50] Kawano Y, Yanai K. Foodcam: A real-time food recognition system on a smartphone [J]. Multimedia Tools and Applications, 2015, 74(14): 5263-5287

[51] Kurz D, Himane S B. Inertial sensor-aligned visual feature descriptors [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 161-166

[52] Google. Google Goggles [EB/OL]. [2015-07-05]. http://www.google.com/mobile/goggles

[53] Chandrasekhar V, Takacs G, Chen D, et al. CHoG: Compressed histogram of gradients [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 2504-2511

[54] Runge N, Wenig D, Malaka R. Keep an eye on your photos: Automatic image tagging on mobile devices [C] //Proc of the Int Conf on Human-Computer Interaction with Mobile Devices & Services. New York: ACM, 2014: 513-518

[55] Dhiraj J, Luo Jiebo. Inferring generic activities and events from image content and bags of geo-tags [C] //Proc of the Int Conf on Content-Based Image and Video Retrieval. New York: ACM, 2008: 37-46

[56] Li Xirong, Snoek C G M, Worring M, et al. Fusing concept detection and geo context for visual search [C] //Proc of the Int Conf on Multimedia Retrieval. New York: ACM, 2012: 1-8

[57] TeraDeep Inc. Teradeep [EB/OL]. [2015-07-05]. http://www.teradeep.com

[58] LLRXcom. Chips [EB/OL]. [2015-06-06]. http://www.llrx.com/features/new-chips-are-using-deep-learning-to-enhance-mobile-camera-and-auto-image-processing-capabilities.htm

[59] Filliat D, Battesti E, Bazeille S, et al. Rgbd object recognition and visual texture classification for indoor semantic mapping [C] //Proc of the IEEE Int Conf on Technologies for Practical Robot Applications (TePRA). Piscataway, NJ: IEEE, 2012: 127-132

[60] Lai K, Bo Liefeng, Ren Xiaofeng, et al. RGB-D Object Recognition: Features, Algorithms, and a Large Scale Benchmark in Consumer Depth Cameras for Computer Vision [M]. Berlin: Springer, 2013: 167-192

[61] Wang Shuang, Jiang Shuqiang. INSTRE: A new benchmark for instance-level object retrieval and recognition [J]. ACM Trans on Multimedia Computing, Communications, and Applications, 2015,11(3): 37:1-37:20

[62] Lv Xiong, Jiang Shuqiang, Herranz L, et al. RGB-D hand-held object recognition based on heterogeneous feature fusion [J]. Journal of Computer Science and Technology, 2015, 30(2): 340-352

[63] Rouanet P, Oudeyer P, Danieau Y, et al. The impact of human-robot interfaces on the learning of visual objects [J]. IEEE Trans on Robotics, 2013, 29(2): 525-541

[64] Matuszek C, Bo Liefeng, Zettlemoyer L, et al. Learning from unscripted deictic gesture and language for human-robot interactions [C] //Proc of the 28th Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2014: 2556-2563

[65] Moubarak P M, Ben-Tzvi P. Adaptive manipulation of a hybrid mechanism mobile robot [C] //Proc of the IEEE Int Symp on Robotic and Sensors Environments. Piscataway, NJ: IEEE, 2011: 113-118

[66] Rusu R B, Cousins S. 3D is here: Point cloud library (PCL) [C] //Proc of the IEEE Int Conf on Robotics and Automation (ICRA). Piscataway, NJ: IEEE, 2011: 9-13

Jiang Shuqiang, born in 1977. PhD. Professor in the Institute of Computing Technology, Chinese Academy of Sciences. Member of China Computer Federation. His current research interests include multimedia analysis and multi-modal intelligent technology.

Min Weiqing, born in 1985. PhD. Postdoctor in the Institute of Computing Technology, Chinese Academy of Sciences. Member of China Computer Federation. His current research interests include multimedia analysis and context based visual recognition (minweiqing@ict.ac.cn).

Wang Shuhui, born in 1983. PhD. Associate professor in the Institute of Computing Technology, Chinese Academy of Sciences. Member of China Computer Federation. His current research interests include social media mining, multimedia analysis and machine learning (wangshuhui@ ict.ac.cn).