

文章编号:1001-9081(2016)11-3217-05

DOI:10.11772/j.issn.1001-9081.2016.11.3217

对象级特征引导的显著性视觉注意方法

杨凡^{1,2}, 蔡超^{1,2*}

(1. 华中科技大学 自动化学院, 武汉 430074; 2. 多谱信息处理技术国家重点实验室(华中科技大学), 武汉 430074)

(*通信作者电子邮箱 caichao@hust.edu.cn)

摘要:针对已有视觉注意模型在整合对象特征方面的不足,提出一种新的结合高层对象特征和低层像素特征的视觉注意方法。首先,利用已训练的卷积神经网(CNN)对多类目标的强大理解能力,获取待处理图像中对象的高层次特征图;然后结合实际的眼动跟踪数据,训练多个对象特征图的加权系数,给出对象级突出图;紧接着提取像素级突出图,并和对象级突出图融合获得显著图;最后,在OSIE和MIT数据集上验证了该方法,并与国际上流行的视觉注意方法进行对比,结果显示该算法在OSIE数据集上获得的AUC值相对更高。实验结果表明,所提方法能够更加充分地利用图像中对象信息,提高显著性预测的准确率。

关键词:视觉注意;自顶向下;显著性;对象信息;卷积神经网

中图分类号:TP391.41 **文献标志码:**A

Significant visual attention method guided by object-level features

YANG Fan^{1,2}, CAI Chao^{1,2*}

(1. School of Automation, Huazhong University of Science and Technology, Wuhan Hubei 430074, China;

2. National Key Laboratory of Science and Technology on Multi-spectral Information Processing,
(Huazhong University of Science and Technology), Wuhan Hubei 430074, China)

Abstract: Concerning the defects of fusing object information by existing visual attention models, a new visual attention method combining high-level object features and low-level pixel features was proposed. Firstly, high-level feature maps were obtained by using Convolutional Neural Network (CNN) which has strong understanding of multi-class targets. Then all object feature maps were combined by training the weights with eye fixation data. Then the saliency map was obtained by fusing pixel-level conspicuity map and object-level conspicuity map. Finally, the proposed method was compared with many popular visual attention methods on OSIE and MIT datasets. Compared with the contrast methods, the Area Under Curve (AUC) result of the proposed method is increased. Experimental results show that the proposed method can make full use of the object information in the image, and increases the saliency prediction accuracy.

Key words: visual attention; top-down; saliency; object information; Convolutional Neural Network (CNN)

0 引言

视觉注意机制的研究是探索人眼视觉感知的重要一环。在过去几十年中,如何用计算模型模拟人眼视觉注意过程一直是核心问题。尽管取得了很大的进步,但是快速准确地在自然场景中预测人眼视觉注意区域仍然具有很高的挑战性。显著性是视觉注意的一项重要研究内容,它反映了区域受关注的程度。本文的研究着眼于显著性计算模型,更多模型对比和模型分类可以参考Borji等^[1]的文章。视觉注意存在两种机制:自底向上(Bottom-up)和自顶向下(Top-down)。过去的研究中,大多数的计算模型是基于自底向上的信息,即直接从图像像素获取特征。

自底向上显著性计算模型开创性工作源自于文献[2]的Itti模型,该模型是很多其他模型的基础和对照基准,它通过整合多种低层次特征,如颜色、亮度、方向等,给出一个显著度的概率分布图。Harel等^[3]在Itti模型的基础上引入图算法,通过计算节点间特征值相似性和空间位置距离进行差异性度量获取显著图。近年来随着深度学习技术在目标识别领域的

成功应用^[4],研究者们对特征学习产生了更多的兴趣。Borji等^[5]通过稀疏编码方法获取特征,使用图像块的稀疏表示结合局部和全局统计特性计算图像块的稀有性(rarity),稀有反映了当前图像块中心位置的显著性。Vig等^[6]通过训练多个神经网络获取层次特征,然后自动优化特征组合。特征提取的过程可以看作是一种隐式空间映射,在映射空间中使用简单的线性模型进行显著或非显著的分类。以上学习方法获得的特征都是一些低层次特征,对图像中的边缘和特定纹理结构敏感。此外,部分研究人员希望从数学统计和信号处理的角度来度量显著性。Bruce等^[7]根据最大化信息采样的原则构建显著性模型。Li等^[8]总结了多种基于频域的视觉注意研究工作,提出了一种基于超复数傅里叶变换(Hypercomplex Fourier Transform)的视觉注意模型,并展示了其他多种基于频域的模型在某种程度上都是此模型的特例。

以上模型均为数据驱动的显著性模型,模拟人眼视觉注意过程中自底向上的机制。由于人眼视觉注意过程中不可避免地受到知识、任务、经验、情感等因素的影响,因而整合自底向上和自顶向下信息的视觉注意研究受到更多的关注。现有

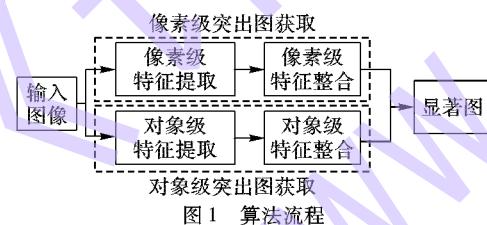
收稿日期:2016-03-18;修回日期:2016-06-26。 基金项目:华为创新基金资助项目(YJCB2010022IN)。

作者简介:杨凡(1990—),男,山东淄博人,硕士研究生,主要研究方向:视觉注意、显著性目标检测、深度学习;蔡超(1971—),男,山东东明人,副教授,博士,主要研究方向:计算机视觉、目标识别、医学图像处理、任务规划。

模型整合的自顶向下信息可以分为三类:任务需求、场景上下文和对象特征。

Borji等^[9]提出了一种构建任务驱动的视觉注意模型的联合贝叶斯方法。Zhang等^[10]提出了一种使用贝叶斯框架整合自底向上和自顶向下显著性信息的方法。Siagian等^[11]利用多种低层次特征对场景主旨进行建模,使用场景主旨引导视觉注意的转移。考虑到任务需求和场景上下文建模的复杂性,研究人员将对象特征视为一种高层次的知识表示形式引入视觉注意模型中。Judd等^[12]和Zhao等^[13]通过将低层次特征和对象特征整合在一个学习框架下来获得特征整合过程中每张特征图的叠加权重,但是模型使用的对象特征只有人脸、行人、车辆等有限的几种。Borji等^[14]遵循了同样的方法,但是在整合过程中添加了更多特征并且结合了其他显著性模型的结果,最后用回归、支撑向量机(Support Vector Machine, SVM)、AdaBoost等多种机器学习算法结合眼动跟踪数据进行训练。实验结果表明对象特征引入较大地提高了模型性能。Xu等^[15]将特征划分为像素级、对象级和语义级三个层次,并重点探索对象信息和语义属性对视觉注意的作用;然而,模型中的对象级和语义级特征是手工标定的,因而不是一种完全意义上的计算模型。

总的来看,虽然部分模型已经使用对象特征作为自顶向下的引导信息,但是在对象特征的获取和整合上仍有很大的局限性。首先,对不包含特定对象的场景适应性较差;其次,对象特征描述困难,通常是通过特定目标检测方法获取对象特征,计算效率低下;此外,对象特征的简单整合方式不符合人眼的视觉感知机制。本文提出了一种结合深度学习获取对象特征的视觉注意计算模型,重点研究了对象级特征的获取和整合方法。算法结构如1所示,其中像素级突出图获取采用现有视觉注意模型的方法,对象级突出图获取采用本文提出的基于卷积神经网(Convolutional Neural Network, CNN)的特征学习和基于线性回归的特征整合方法。实验结果表明,对象级特征的引入可以明显提高显著性预测精度,预测结果更符合人类视觉注意效果。



1 对象信息获取

1.1 对象特征

大量实验证据表明对象特征引导视觉注意的转移。视觉注意中引入对象特征是为了获得图像中对象位置等信息,目的与计算机视觉中的目标检测类似。因而,已有的视觉注意计算模型的对象特征通常是通过特定目标检测方法获得。其中,Viola&Jones人脸检测和Felzenszwalb车辆行人检测是最常用的方法。文献[12~14]均使用此类方法引入对象特征。由于这一类特征针对特定对象样本进行设计和训练,因而推广能力不强。

通过大量样本训练卷积神经网,可以实现多类对象特征的提取。图2显示了AlexNet网络基元可视化结果,训练数据

为ImageNet2012。图2(a)~(d)为采用文献[16]提供的可视化方法获得的四个卷积层的部分基元表示,其中,图2(a)的基元与视觉注意中常用的Gabor特征类似,图2(b)的基元图像包括了拐角和特定纹理等简单结构。更深层次节点对应的复杂特征由之前层的简单特征组合获得。图2(e)~(h)显示了针对网络最后一个卷积层中四个节点采用正则化梯度上升法获得的合成基元图像。图2(i)~(l)为训练样本集中选取的9张对第二行四个节点有最大响应的图像块,可以看出9个图像块具有相似的结构。深层网络节点通过训练学到了一类对象的抽象概念。当场景中出现此类对象时,该节点响应增大。因而,相对于传统的特定目标检测方法,通过训练卷积神经网获取的对象特征种类更丰富、描述更抽象。

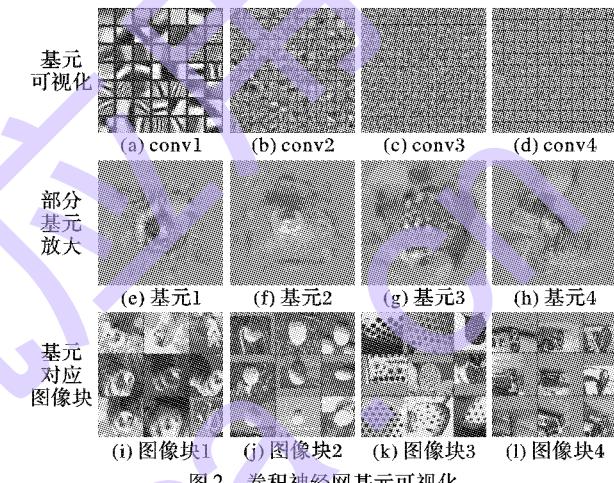


图2 卷积神经网基元可视化

本文利用RCNN(Regions with CNN features)算法^[17]提取图像中的对象信息,该算法利用卷积神经网中的对象特征,对Object Proposal方法^[18]获取的对象可能出现的区域进行处理,判断当前区域是否为对象区域。如果卷积神经网中包含此类对象特征,该算法会给出当前对象区域的多种信息。图3(a)、(b)为该算法对简单和复杂目标场景处理后的结果,用矩形限定框(Bounding Box)给出对象的位置;此外,针对每个对象还给出对象类属、对象类属的置信度。与图3(c)、(d)中用眼动跟踪数据获得的显著图对比,利用卷积神经网深层对象特征给出的对象位置信息,覆盖了人眼视觉注意的大部分区域。

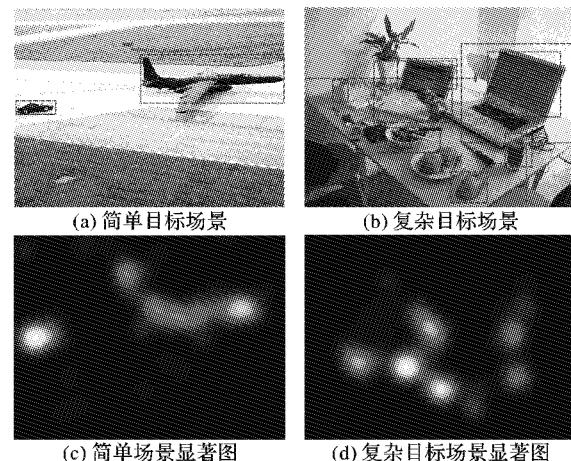


图3 对象位置信息和真实眼动跟踪数据

为了分析不同卷积神经网络结构获取的对象特征对显著

性预测的影响,本文同时也使用 faster RCNN 算法进行对象信息提取。faster RCNN 算法^[19]作为 RCNN 的改进算法,使用了更深的网络结构和更多对象种类的训练样本。本文实验部分对比了这两种方法。由于这两种方法使用训练数据中并不包含人脸这一类对象。而人脸在视觉注意过程中是具有强语义的一类特征。Judd^[12]和 Zhao^[13]模型中均包含人脸特征。本文中使用 face++ 算法进行人脸定位。后序实验分析可以看出,添加人脸特征对预测准确率的提升有一定帮助。

1.2 对象特征图生成

针对从图像中获取的每个对象,利用对象位置信息生成一幅对象特征图。根据人眼视觉注意过程中倾向于关注对象中央位置的特性,即中央偏置(Center-bias)。使用二维高斯函数模拟对象显著区域分布函数,模型使用的二维高斯函数的表达式为:

$$G(x, y) = \exp\left\{-\frac{1}{2}\left(\left(\frac{x-x_0}{\sigma_x}\right)^2 + \left(\frac{y-y_0}{\sigma_y}\right)^2\right)\right\} \quad (1)$$

其中: (x_0, y_0) 为对象限定框中心点坐标; $\sigma_x = k * width$; $\sigma_y = k * height$; $width, height$ 分别为对象的长宽; k 为比例常数。

图 4 为实验结果,其中:图 4(a)为挖取的对象图像块,图 4(b)为针对图 4(a)中对象使用高斯生成方法获得的简单对象特征图。

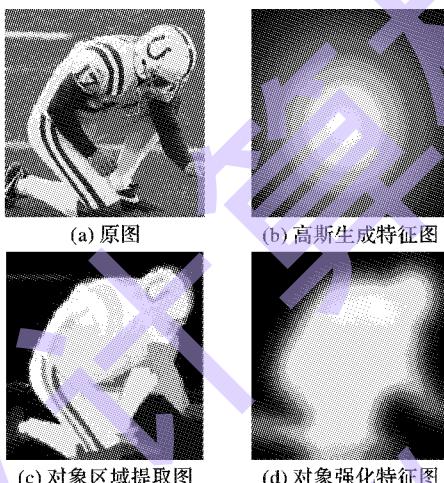


图 4 对象特征图生成

使用高斯生成方法获得对象特征图忽略了对象结构信息,而人眼感知对象的重要依据就是对象的边缘,因而突出对象区域、弱化背景区域有利于显著性预测。为了将对象整体区域突出,将每个检出对象对应的图像块挖出,使用文献[20]给出的算法进行处理。图 4(c)是处理后的结果,可以看出对象区域与背景可以明显区分。为了强化所有对象区域,对图像进行阈值分割和形态学腐蚀,根据分割结果拉升对象区域像素灰度级,然后进行一定程度的高斯模糊处理。图 4(d)为处理后的效果图,可以看到处理后的图像保留了原始对象的结构,对象区域得到了增强。

考虑到对象尺寸对显著性的影响,大尺度物体中央偏置的作用应越弱,局部显著区域的引导应越强。利用式(2),将高斯生成的中央偏置特征图与局部增强后的对象特征图整合。

$$S_{obj_k} = \left(1 - \frac{A}{A_{image}}\right)S_C + \frac{A}{A_{image}}S_D \quad (2)$$

其中: A 表示第 k 个对象的显著区域的面积; A_{image} 表示图像面积; S_C 为高斯生成的中央偏置特征图; S_D 为局部区域增强后的特征图; S_{obj_k} 为第 k 个对象的对象特征图。

2 对象级突出图生成

一幅图像中可能包含多个对象目标。为了突出所有对象,使用线性叠加的方式将多个对象特征图整合为一幅对象级突出图。考虑到对象个体属性影响对象受关注程度,本章将采用回归的方法确定对象特征整合的叠加权重,如图 5 所示。

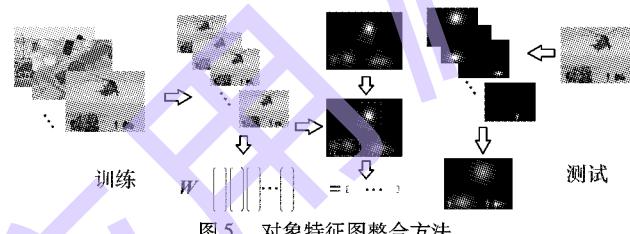


图 5 对象特征图整合方法

2.1 对象属性

影响视觉注意转移的对象属性主要分为两类,即对象的外观属性和对象的语义属性。外观属性描述对象基本结构,语义属性描述对象身份信息。本文中主要考察五种对象属性对视觉注意的影响:

- 1) 对象尺寸 f_1 。虽然对象尺寸对于显著性的影响机制并不明确,但是根据眼动跟踪实验分析,针对小尺寸物体,人眼倾向于视其为整体,而对大尺寸物体,人眼倾向于关注其部分显著区域。本文中尺寸属性为对象限定框面积。
- 2) 对象长宽比例 f_2 。一个细长的对象更难以获得较高的显著性。本文通过计算限定框长宽比获得。
- 3) 对象区域的比例 f_3 。指对象区域占整个对象限定框区域的面积比。对象区域面积通过 1.2 节的方法获得,该属性考察了对象的凸性。
- 4) 对象语义置信 f_4 。通过 1.1 节方法给出的属性值 $score$,来确定该目标属于某一类的概率大小,该属性值越大该物体语义越明确。
- 5) 对象类属出现的频率 f_5 。通过 1.1 节获取每个对象相应的类属信息 $class_i$,类属出现的频率通过对样本集中的各类对象出现次数进行统计获得。

2.2 对象特征整合

从每副对象级特征图中提取对象属性,使用之前分析的五种属性信息组成的向量 $F_i = (f_1, f_2, f_3, f_4, f_5)$,并对每维数据作归一化处理,通过线性回归模型训练获得对象特征图的线性叠加权重。

训练样本的标签由真实眼动跟踪数据给出,式(3)为具体计算形式:

$$l_i = \frac{fixations(obj_i)}{area(obj_i)} \quad (3)$$

其中: $fixations()$ 表示落入当前对象区域的正样本的数目; $area()$ 表示对象区域面积。 l_i 衡量当前对象单位面积受关注的程度,对象单位面积受关注程度越高,其在对象整合过程中的权重应越高,因而 l_i 与叠加权重成正比。

式(4)通过一个线性回归模型对已有样本数据进行训练,获得对象整合叠加权重 W :

$$L = WF \quad (4)$$

其中: $F = \{F_1, F_2, \dots, F_N\}$ 为训练样本数据集合; $L = \{l_1, l_2, \dots, l_N\}$ 为训练样本标签集合。

测试时根据式(5)~(6)获得对象级突出图:

$$\alpha = k \sum_{i=1}^d w_i f_i \quad (5)$$

$$S_{obj} = \sum_{i=1}^N \alpha_i S_{obj_i} \quad (6)$$

其中: k 为归一化系数; α_i 为当前对象整合的叠加权重; S_{obj} 为最终获得的对象级突出图。

3 显著图生成

视觉注意是自底向上和自顶向下两种机制作用的结果。完全使用自顶向下的对象特征进行显著区域预测有一定缺陷, 主要表现在以下几个方面: 首先, 知识是对训练样本数据的抽象表示, 由于神经网络的规模和训练样本中对象种类的限制, 场景中部分对象对应的特征没有被抽象在网络结构中; 其次, 部分不具有明确语义的区域被错误地认为是对象, 对视觉注意形成错误的引导; 另外, 人眼视觉注意转移的生理学机制并不清楚, 兴趣区可能落在不具有对象特征区域中。因此, 使用像素级特征给出低层次显著性信息是必要的。

视觉注意模型中常用的像素级特征有颜色、亮度、方向等^[2-3, 12]。本文直接使用 GBVS(Graph-Based Visual Saliency) 算法^[4]整合多种像素级特征获取像素级突出图 S_{pixel} 。式(7)给出了整合的方法:

$$S(i, j) = N(S_{pixel}(i, j) + \lambda S_{obj}(i, j)) \quad (7)$$

其中: $S(i, j)$ 为最终给出的视觉注意显著图; $N()$ 为归一化操作; λ 控制对象级突出图与像素级突出图的相对权重, 通过实验分析可知 $\lambda = 0.4$ 时效果较好。当图像中不存在显著物体或无法获得高置信度的对象信息时, 图像任意位置 $S_{obj}(i, j) = 0$, 此时完全由像素级特征驱动的视觉注意引导。

4 实验结果及分析

本次实验是以 Visual Studio 2012 为实验平台, 选取 OSIE 和 MIT 数据集作为实验数据。OSIE 数据集包含 700 张含有一个或多个明显语义对象的图片以及 15 名受试者的眼动跟踪数据, 此外该数据集还提供了语义对象统计及人工标注的精确对象区域。MIT 数据集包含 1003 张自然场景图片以及 15 名受试者的眼动跟踪数据。这两个数据集是当前视觉注意研究领域中较大的数据集。为了验证本文方法的准确率, 将本文算法与 GBVS^[4]、Itti^[2]、Judd^[3]、AIM^[10]、LG^[8] 等视觉注意方法进行对比。

对比实验中使用的评价指标为 ROC(Receiver Operating Characteristic) 曲线, 实现方法与文献[12, 15]相同。图 6~8 为实验对比结果, 显著区域百分比是通过对归一化显著图作阈值处理获得, 真正率(True Positive Rate)反映当前落入显著区域的样本占所有样本的比例。通过变化显著区域百分比获得 ROC 曲线。为了更直观比较算法效果, 实验结果图中标注了每种算法的 AUC(Area Under Curve) 值, AUC 值通过计算 ROC 曲线下的面积获得。AUC 值越大表示该方法给出的显著性预测结果越准确。

图 6 为利用对象级突出图作为显著图在 OSIE 数据集上的实验结果。相对于 RCNN 算法, fasterRCNN 算法使用了更

深层次的网络结构和更多对象类别的训练样本, 具有较高的对象位置预测准确率和对象检出率。实验分析可以看出, 使用 fasterRCNN 算法生成对象级突出图可以更好进行显著性预测。同时, 人脸特征(FACE)的引入进一步提升了预测准确性, 从一个侧面说明了对象性信息对视觉注意的转移具有引导作用。

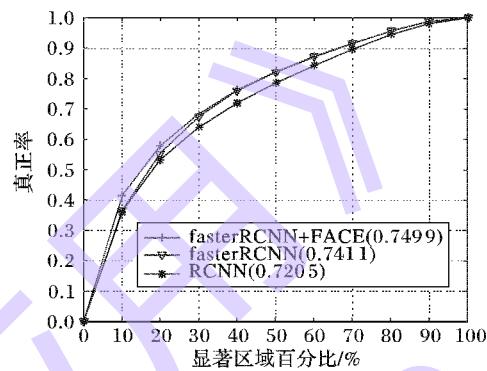


图 6 对象级突出图在 OSIE 数据集上的 ROC 曲线

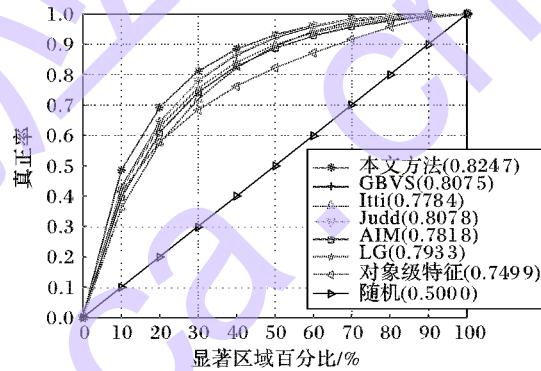


图 7 多种视觉注意算法在 OSIE 数据集上的 ROC 曲线

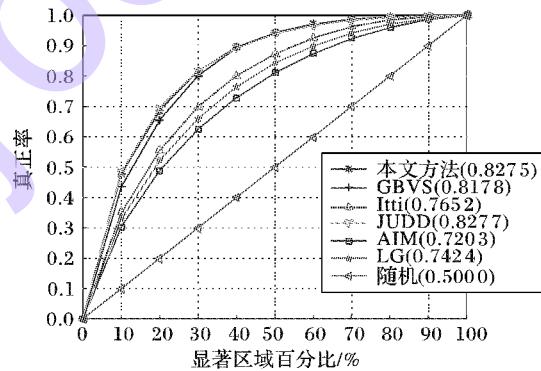


图 8 MIT 数据集上的 ROC 曲线

图 7 是多种视觉注意算法在 OSIE 数据集上的 ROC 曲线, 可以看出本文方法实验效果明显好于其他算法。仅次于本文算法的是 GBVS 和 Judd, Itti 的准确率较差。图中对象级特征曲线为使用 fasterRCNN 结合人脸特征生成对象级突出图获得, 由于该方法完全使用自顶向下的对象特征, 显著性预测准确率明显弱于其他方法, 因而证明了引入像素级特征必要性。图 8 为 MIT 数据集上的实验结果, 本文方法和 Judd 算法为最好的两种方法, 实验结果相差不大。AIM 和 LG 方法效果较差。本文方法和 Judd 方法均使用了对象特征, 可以看出整合了对象特征的方法相对于完全自底向上模型有明显优势。

图 9 中给出了多种算法显著图的直观对比。与其他方法

强调对象边缘不同,本文结合了对象信息的方法可以有效突

出图像中的完整对象区域。

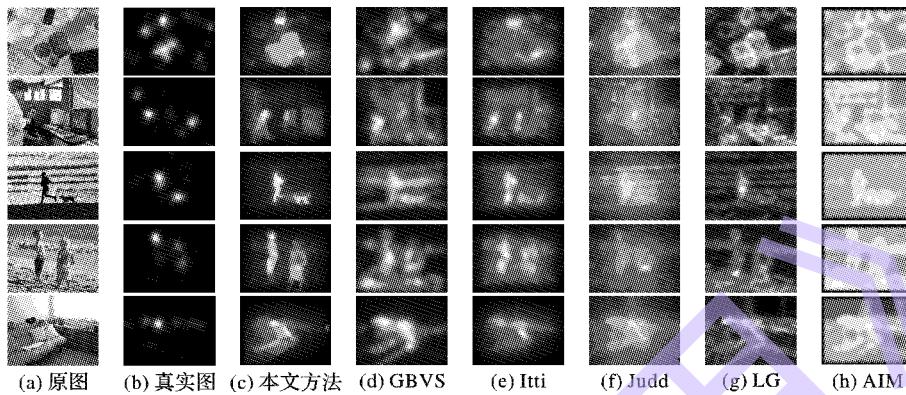


图 9 实验结果对比

5 结语

本文提出一种结合对象信息的视觉注意方法。与传统的视觉注意整合对象方法相比,该方法利用卷积神经网学到的对象特征,获取图像中对象位置等信息;然后通过一个线性回归模型将同一幅图像的多个对象加权整合,获得对象级突出图;最后,根据视觉注意的层次整合机制,将低层次特征和对象特征进行融合形成最终的显著图。本文方法在不同数据集上的准确率要高于现有模型。针对包含明显对象的图像,本文方法克服了部分现有模型由于边缘强化效果导致的显著区域预测不准的问题。本文方法仍然存在一定局限性,未来的工作将尝试非线性对象整合以及增大训练样本数量和网络规模以获取更多种对象特征。

参考文献:

- [1] BORJI A, ITTI L. State-of-the-Art in visual attention modeling [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 185 – 207.
- [2] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254 – 1259.
- [3] HAREL J, KOCH C, PERONA P. Graph - based visual saliency [C]// NIPS 2006: Proceedings of the 2006 Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2006: 545 – 552.
- [4] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks [EB/OL]. [2015-10-10]. <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- [5] BORJI A, ITTI L. Exploiting local and global patch rarities for saliency detection [C]// CVPR 2012: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 478 – 485.
- [6] VIG E, DORR M, COX D. Large-scale optimization of hierarchical features for saliency prediction in natural images [C]// CVPR 2014: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 2798 – 2805.
- [7] BRUCE N, TSOTSOS J. Saliency based on information maximization [EB/OL]. [2015-10-10]. <https://papers.nips.cc/paper/2830-saliency-based-on-information-maximization.pdf>.
- [8] LI J, LEVINE M, AN X, et al. Visual saliency based on scale-space analysis in the frequency domain [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(4): 996 – 1010.
- [9] BORJI A, SIHITE D, ITTI L, et al. Probabilistic learning of task-specific visual attention [C]// CVPR 2012: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 470 – 477.
- [10] ZHANG L, TONG M, MARKS, T, et al. SUN: a Bayesian framework for saliency using natural statistics [J]. *Journal of Vision*, 2008, 8(7): Article No. 32.
- [11] SIAGIAN C, ITTI L. Rapid biologically-inspired scene classification using features shared with visual attention [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(2): 300 – 312.
- [12] JUDD T, EHINGER K, DURAND F, et al. Learning to predict where humans look [C]// ICCV 2009: Proceedings of the 2009 International Conference on Computer Vision. Piscataway, NJ: IEEE, 2009: 2106 – 2113.
- [13] ZHAO Q, KOCH C. Learning a saliency map using fixated locations in natural scenes [J]. *Journal of Vision*, 2011, 11(3): Article No. 9.
- [14] BORJI A. Boosting bottom-up and top-down visual features for saliency estimation [C]// CVPR 2012: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2012: 438 – 445.
- [15] XU J, JIANG M, WANG S, et al. Predicting human gaze beyond pixels [J]. *Journal of Vision*, 2014, 14(1): Article No. 28.
- [16] YOSINSKI J, CLUNE J, NGUYEN A, et al. Understanding neural networks through deep visualization [EB/OL]. [2015-10-10]. <http://arxiv.org/abs/1506.06579>.
- [17] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// CVPR 2012: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 580 – 587.
- [18] UIJLINGS J, van de SANDE K E A, GEVERS T, et al. Selective search for object recognition [J]. *International Journal of Computer Vision*, 2013, 104(2): 154 – 171.
- [19] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [EB/OL]. [2015-10-10]. <http://arxiv.org/abs/1506.01497>.

(下转第 3228 页)

语音切分方法的切分准确度仍不及人工切分,更优秀的汉语语音音节切分方法尚待研究。

参考文献:

- [1] FARAJI N, AHADI S M, SHEIKHZADEH H. Sequential method for speech segmentation based on random matrix theory [J]. IET Signal Processing, 2013, 7(7): 625 – 633.
- [2] SATTAR F, NILSSON M, CLAESSEN I. Segmentation and its real-world applications in speech processing [C]// Proceedings of the 9th International Symposium on Signal Processing and Its Applications. Piscataway, NJ: IEEE, 2007: 1 – 4.
- [3] BROGNAUX S, DRUGMAN T. HMM-based speech segmentation: improvements of fully automatic approaches [J]. IEEE/ACM Transactions on Audio Speech and Language Processing, 2016, 24(1): 5 – 15.
- [4] KISS G, SZTAHO D, VICSI K. Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features [C]// Proceedings of the 4th International Conference on Cognitive Infocommunications. Piscataway, NJ: IEEE, 2013: 579 – 582.
- [5] MPORAS I, LAZARIDIS A, GANCHEV T, et al. Using hybrid HMM-based speech segmentation to improve synthetic speech quality [C]// Proceedings of the 4th Panhellenic Conference on Informatics. Piscataway, NJ: IEEE, 2009: 118 – 122.
- [6] ZIOLKO B, MANANDHAR S, WILSON R C, et al. Wavelet method of speech segmentation [C]// Proceedings of the 6th European Signal Processing Conference. Piscataway, NJ: IEEE, 2006: 1 – 5.
- [7] LEE K S. MLP-based phone boundary refining for a TTS database [J]. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(3): 981 – 989.
- [8] ABEL A K, HUNTER D, SMITH L S. A biologically inspired onset and offset speech segmentation approach [C]// Proceedings of the 2015 International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2015: 1 – 8.
- [9] MUSFIR M, KRISHNAN K R, MURTHY H A. Analysis of fricatives, stop consonants and nasals in the automatic segmentation of speech using the group delay algorithm [C]// Proceedings of the 2014 20th National Conference on Communications. Piscataway, NJ: IEEE, 2013: 1 – 6.
- [10] AKDEMIR E, CILOGLU T. HMM topology for boundary refinement in automatic speech segmentation [J]. Electronics letters, 2010, 46(15): 1086 – 1087.
- [11] 杜守栓. 方言口音普通话语音自动切分算法研究 [D]. 北京: 中国科学院计算技术研究所, 2006: 15 – 26. (DU S S. Research on robust automatic segmentation of dialectal speech [D]. Beijing: Chinese Academy of Sciences, Institute of Computing Technology, 2006: 15 – 26.)
- [12] 张继勇, 郑方, 杜术, 等. 连续汉语语音识别中基于归并的音节切分自动机 [J]. 软件学报, 1999, 10(11): 1212 – 1215. (ZHANG J Y, ZHENG F, DU S, et al. Merging-based syllables detection automaton in continuous Chinese speech recognition [J]. Journal of Software, 1999, 10(11): 1212 – 1215.)
- [13] 韩虎. 汉语连续语音的音节自动标注算法研究及实现 [D]. 哈尔滨: 哈尔滨工业大学, 2008: 21 – 44. (HAN H. Research and realization of the automatic syllable marking algorithm for Chinese continuous speech [D]. Harbin: Harbin Institute of Technology, 2008: 21 – 44.)
- [14] 张扬, 赵晓群, 王缔罡. 基于音节长度高斯拟合的汉语音节切分方法 [J]. 计算机应用, 2016, 36(5): 1410 – 1414. (ZHANG Y, ZHAO X Q, WANG D G. Chinese speech segmentation method based on Gauss distribution of time spans of syllables [J]. Journal of Computer Applications, 2016, 36(5): 1410 – 1414.)
- [15] FISHER E, TABRIKIAN J, DUBNOV S. Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model [J]. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(2): 502 – 510.
- [16] 金学成, 汪增福. 基于线性预测残差倒谱的基音周期检测 [J]. 模式识别与人工智能, 2008, 21(1): 104 – 110. (JIN X C, WANG Z F. A pitch detection algorithm based on linear predication residual cepstrum [J]. Pattern Identification and Artificial Intelligence, 2008, 21(1): 104 – 110.)
- [17] 党晓妍, 魏旋, 崔慧娟, 等. 声码器清浊音判决算法优化 [J]. 清华大学学报: 自然科学版, 2008, 48(7): 1119 – 1122. (DANG X Y, WEI X, CUI H J, et al. Improvement of voiced-unvoiced classification in vocoders [J]. Journal of Tsinghua University (Science and Technology), 2008, 48(7): 1119 – 1122.)

Background

ZHANG Yang, born in 1989, Ph. D. candidate. His research interests include processing of speech signals.

ZHAO Xiaoqun, born in 1962, Ph. D., professor. His research interests include processing of speech signals, coding theory.

WANG Digang, born in 1988, Ph. D. candidate. His research interests include coding theory.

(上接第 3221 页)

- [20] JIANG H, WANG J, YUAN Z, et al. Salient object detection: a discriminative regional feature integration approach [C]// CVPR 2012: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2083 – 2090.
- [21] 暴林超, 蔡超, 肖洁, 等. 一种用于复杂目标感知的视觉注意模型 [J]. 计算机工程, 2011, 37(13): 17 – 19. (BAO L C, CAI C, XIAO J, et al. Visual attention model for complex target perception [J]. Computer Engineering, 2011, 37(13): 17 – 19.)
- [22] 肖洁, 蔡超, 丁明跃. 一种图斑特征引导的感知分组视觉注意模型 [J]. 航空学报, 2010, 31(11): 2266 – 2274. (XIAO J,

CAI C, DING M Y. A novel visual attention model based on blob-guided perceptual grouping [J]. Acta Aeronautica et Astronautica Sinica, 2010, 31(11): 2266 – 2274.)

Background

This work is partially supported by the Huawei Innovation Fund (YJCB2010022IN).

YANG Fan, born in 1990, M. S. candidate. His research interests include visual attention, saliency object detection, deep learning.

CAI Chao, born in 1971, Ph. D., associate professor. His research interests include computer vision, object recognition, medical image processing, mission planning.