DOI:10.11918/201905181

一种整合语义对象特征的视觉注意力模型

李 娜[1,2], 赵歆波[1,2]

(1. 西北工业大学 计算机学院,西安 710129; 2. 陕西省语音与图像信息处理重点实验室,西安 710129)

摘 要: 视觉注意力建模作为预测人类在观察场景时注意力分布的关键技术,在计算机视觉的众多领域均有广泛应用.传统的视觉注意力模型着重研究人眼注视点,计算出的显著图更多的是反映眼动信息,并未将大脑的感知出的语义信息反映出来.针对这一问题,本文提出了一种整合了语义对象特征的视觉注意力模型.首先,本文建立了眼动跟踪数据库 VOC2012-E,研究并记录普通人在观察自然场景时的眼动数据.然后,受语义分割启发,利用全卷积神经网络(Fully Convolutional Networks, FCN)提取语义对象特征,同时用激活函数 PReLu 和优化函数 Adam 改进 FCN 网络使其更有效地提取的语义对象特征,来模仿大脑对语义对象特征的感知.接着,提取在人类潜意识层吸引人注意力的如方向,颜色,强度特征等 28 个低级特征.最后利用支持向量机(Support Vector Machine, SVM)将之前提取的语义对象特征及低级特征映射到人类视觉空间,同时引入真实眼动数据进行有监督的训练,得到可以预测人眼视觉显著图的视觉注意力模型.实验结果表明,在 VOC2012-E 及 MIT300 数据库上与其他 8 种经典模型及 4 种先进模型相比,本文提出的视觉注意力模型性能更好,更有生物学优势.

关键词: 视觉注意力模型;语义对象特征;FCN;SVM;深度学习

中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2020)05-0099-07

Incorporating semantic object features into a visual attention model

LI Na[1,2], ZHAO Xinbo[1,2]

(1. School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China;
2. Shaanxi Provincial Key Laboratory of Speech and Image Information Process, Xi'an 710129, China)

Abstract: Visual attention modeling is a key technique for predicting the distribution of human attention when people are observing scenes, which is widely used in the fields of computer vision. Traditional visual attention models focus on the human eyes fixation points to reflect the eye movement information by calculating saliency maps, while they cannot reflect the perceived semantic information of the brain. To solve this problem, a visual attention model was proposed based on extracting semantic features. First of all, the eye tracking database VOC2012-E was established to study and record the eye movement data of human while observing natural scenes. Then, inspired by image semantic segmentation, the Fully Convolutional Networks(FCN) was used to extract the semantic object features. In order to extract the semantic object features more effectively, the FCN8s network was improved by activation function PReLu and optimization function Adam to mimic the brain's perception of semantic object features. Next, 28 low-level features such as direction, color, and intensity characteristics were extracted, which attract attention in the human subconscious layer. Finally, Support Vector Machine(SVM) was used to map the previously extracted semantic object features and the low-level features into the human visual space. The real eye movement data was introduced for supervised training, and a visual attention model was obtained which can predict the human visual saliency map. Experimental results showed that the visual attention model proposed in this paper had better performance and biological advantages over the other eight classical models and four advanced models on the VOC2012-E and MIT300 databases.

Keywords: visual attention model; semantic object features; FCN; SVM; deep learning

视觉注意力机制是人类视觉快速扫描场景后获取高度相关信息的大脑信号处理机制.视觉注意力模型是指计算机利用视觉搜索得到的各种特征,估计人类注意力显著信息的技术.对于许多计算机视觉任务,人类视觉注意力建模是重要的科学问题.例如,对于视频压缩,人眼感兴趣区域(Region of Interesting, RoI)是需要重点保留的关键信息.对人机交互和广告设计来讲,准确了解人眼在场景中哪些信息感兴趣,可以更好设计出满足用户需求的产品.生物认知学为人类视觉注意机制研究提供了生物学基础[1].当人眼观察视觉信息时,经大脑特定部位选择性地感知的信息会在人眼视网膜的黄斑上

收稿日期: 2019-05-24
基金项目: 国家自然科学基金(61231016,61871326)
作者简介: 李 娜(1992—),女,博士研究生;
赵歆波(1970—),男,教授,博士生导师
通信作者: 赵歆波,xbozhao@nwpu.edu.cn

成像. 例如,功能磁共振成像(Functional Magnetic Resonance Imaging, FMRI)显示人类大脑的梭形面部区域感知面部信息[2],大脑的旁海马区感知地方和建筑物[3-4]信息. 受认知研究启发,本文记录并分析了人眼在观察场景时的眼动数据,得出语义对象区域为人类关注的高意识区域,这些语义对象特征会更吸引视觉注意力.

近年来,众多科研工作者对人眼运动进行了研究,涌现出大量基于不同理论的视觉注意力模型. 一些研究人员在低级注意力模型上取得了进展. 其中最有影响力的是由 Itti[5] 等提出的基于特征整合理论[6]的自下而上注意力模型. 该理论利用颜色,方向和强度等低级特征预测图像中吸引人眼注意的显著区域. 基于神经反应去相关的思想,Diaz 等[7] 提出了自适应白化显著性模型(Adaptive Whitening Saliency, AWS). Zhang 等[8] 提出了自然统计显著模型(Saliency Using Natural statistics, SUN),该模型将视觉特征的自身信息作为自下而上的显著性. Torralba [9] 提出了一种用于视觉搜索的贝叶斯框架用于显著性计算. Harel[10] 等提出了基于图论的显著模型(Graph Based Visual Saliency, GBVS). Vig 等[11]利用机器学习方法计算图像区域的显著性. Tavakoli 等[12]利用无监督模型提取的特征用于显著性计算. 尽管这些模型表现良好,但却忽略了语义对象特征对人类视觉注意力的吸引.

Judd 等[13]和赵等[14]将低级特征和语义特征整合到了学习框架中. Kummerer 等[15]和 Mahdi 等[16]利用预训练的深度学习模型提取低级特征及语义特征用于显著性预测. 但是,这些模型使用的语义特征比较有限,而实验结果表明,视觉注意力模型的性能通过语义特征的引入获得了极大地提高. 虽然这些模型使用了语义特征作为自上而下的指导信息,但提取的语义信息类别很有限.

人类感知语义信息的过程涉及了大量大脑感知神经,其过程极其复杂. 然而,深度学习在图像语义分割领域成果斐然,涌现了各种性能优良的网络,如(FCN[17], DeepLab 系列等[18-19]),深度学习网络本质就是人脑的仿生结构,而图像语义分割就是计算机自动从图像中识别并分割出对象区域,这与本文提取出在人眼观察场景时人脑感知的语义对象特征的目的不谋而合,因此,本文将性能优异的语义分割网络迁移到视觉注意力模型中,来提取语义对象特征,大大增多了语义信息的类别数量.

此外,除了由人类认知控制的自上而下的视觉注意外,还有一种潜意识的机制,即自下而上的视觉注意力,吸引人眼注意一些低级特征[6]. 因此,本文经过分析,提取了 28 个吸引人眼的低级特征,除 RGB 颜色,亮度,强度等常见的低级特征之外,由于 Lab 颜色模型是基于人对颜色的感觉,所以本文同时提取了 Lab 颜色空间的显著性特征.

本文利用深度学习网络提取语义对象特征,将其与刺激人眼的低级特征通过支持向量机(Support Vector Machine, SVM)进行整合,训练这些特征与眼动跟踪技术获取的真实注视点之间的映射关系,得到能预测人眼注视点的视觉注意力模型.

1 眼动数据采集与分析

为了研究普通人在观看场景时的视觉行为,本文从用于语义分割的 VOC2012 数据库[20]里带有语义分割标签的 2 913 张图像里挑选出研究用的图像. 图像选取时最主要的原则是图像中语义对象尺寸不能过大,因为当语义对象尺寸占了图像绝大部分面积时,统计出的眼动注视点落入语义对象区域的数量是没有意义的. 同时,兼顾数据库语义对象的多样性,尽可能地均衡了各类语义对象的数量,最终选择了 2 000 张图像进行研究. 然后记录人眼观察这 2 000 张图像时的眼动跟踪数据,形成新的数据库 VOC2012-E. 该数据库用于对注视点的定量分析,并为显着性模型的研究提供基准图像. 与其他公开语义对象分割的数据集相比,本文建立新数据集的主要动机是分析注视点信息.

本文利用 Tobii TX300 眼动仪记录受试者的眼球运动,其可以进行高精度和高准确性的眼动追踪. 实验中,为保证数据的有效性,每次只播放 100 张图像,每张图像播放 5s,在播放下一张图像前自动进行快速校准,每次采集大致进行 10 ~ 15 min,共有 10 名受试者参与实验.

如表 1 所示,本文统计了每张图像的眼动注视点落入语义对象区域的总数,然后计算出落入语义对象区域的注视点占全部注视点的平均占比,数值超过 83.53%,这表明语义对象特征吸引了受试者大部分的注意力. 因此,为了提高视觉注意力模型预测的准确性,语义对象特征的引入意义重大.

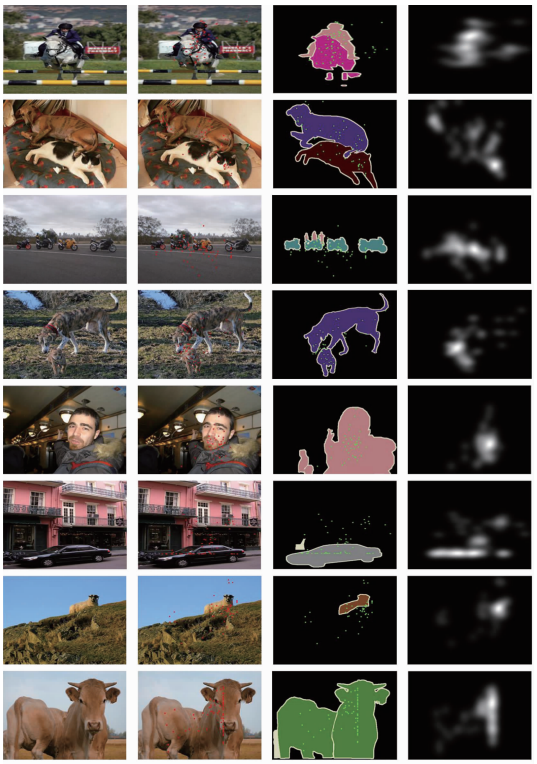
表 1 眼动注视点统计

Tab. 1 Statistics of fixation points

眼动注视点总数	落入显著区域的眼动注视点数	比率/%
556 395	464 766	83.53

如图 1(b)所示,汇总实验中记录的所有受试者的真实眼动注视点,叠加到原图 1(a)上,得到眼动点的分布情况. 图 1(c)重点突出语义对象与眼动点分布的关系,经过进一步分析,无论目标较大的图像

（第 8 列图像牛），还是目标较小的图像（第 9 列图像羊），语义对象都吸引了受试者大部分的注意力. 图 d 为将真实眼动注视点进行高斯滤波得到的基准图像，用于第 4 节的注意力模型的训练.



(a) 原图　(b) 眼动注视点　(c) 语义对象与眼动注视点　(d) 基准图像

图 1　眼动跟踪实验

Fig. 1　Eye tracking experiment

2　语义对象特征的提取

根据实验分析得出的语义特征提取的必要性，利用 FCN 网络从图像语义分割的角度出发提取图像的语义对象特征. 与语义分割任务不同，本文提取的语义对象特征是否具有精确边缘，并不影响视觉注意力模型的预测人的注视点，但特征提取的时间和硬件成本却对视觉注意力模型的训练至关重要. 因此，综合比较了常见语义分割模型，本文利用综合性能最好的 FCN-8s 网络提取语义特征，同时为了保证特征提取的鲁棒性和有效性，对其进行了改进：采用了参数线性整流（Parametric Rectified Linear Unit，PReLu）函数取代了线性整流函数（Rectified Linear Unit，Relu）；使用适应性矩估计（Adaptive monent estimation，Adam）优化网络的学习率.

图 2 为本文利用的 FCN-8s 的网络结构. 在本文使用的网络结构中，在卷积之后不再使用 Relu 激活函数. 虽然 Relu 激活函数由于自身只有线性关系，收敛速度很快，但是当输入为负数时，Relu 的输出被设为 0，导致对应权重无法更新，即该神经元坏死，将会对任何数据都无法响应. 为解决这一问题，本文采用 PReLu 激活函数，其计算公式如下

$$\mathrm{PRelu}(x_i)=\begin{cases} x_i, & x_i>0; \\ a_i x_i, & x_i\leqslant 0. \end{cases} \tag{1}$$

式中：a 的取值是在 0 ~ 1 之间变化的数，i 为不同的通道. 如式（1）所示，与 Relu 函数不同的是，当 PReLu 的输入为负数时，它的输出非 0，从而避免了神经元坏死，而且 PReLu 只增加了很少的参数，所以只增加了很少的网络的计算量.

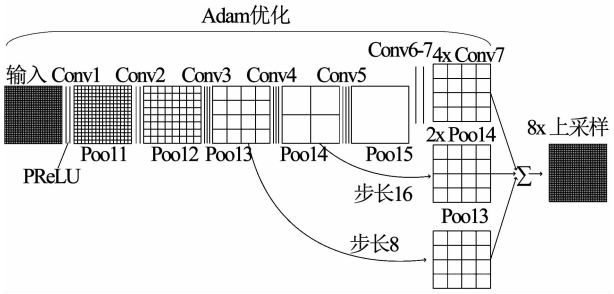


图 2　本文使用的 FCN-8s 网络结构

Fig. 2　The FCN-8s network structure used in this paper

此外，在网络的训练过程中，本文不再采用梯度下降优化算法，而是利用 Adam 优化算法调整网络更新权重和偏差参数. 梯度下降法虽然是最常用的优化算法，但是为其选择合适的学习率比较难，学习率太小会导致网络收敛过于缓慢，而学习率太大可能会影响网络收敛. 而且不同特征应采用不同的更新率，比如，出现频率较小的特征，应有更大的更新率，但是梯度下降法的学习率是固定的. 而 Adam 优化算法可以为每个特征计算自适应学习率，同时改进了梯度下降法的缺点，如学习率消失、收敛过慢等. 在本文提取语义特征的实际过程中，Adam 优化算法效果良好，收敛速度比梯度下降法更快，使特征提取更为有效. 图 3 为本文改进后的 FCN-8s 架构提取的语义对象特征 F_S 示例. 图 3（a）为原图，图 3（b）为提取的语义对象特征，图 3（c）为人工标记的语义对象.

3　低级特征的提取

人类的早期视觉通路会利用视网膜及初级视觉皮层提取如强度，颜色和方向等若干低级特征. 因此为提高视觉注意力模型的性能，本节经过反复实验及结果比较，选取了包含了 28 个低层特征的特征集 $F_L=\{f_1,f_2,\cdots,f_{28}\}$：图 4 中（1）~（13）图为 13 个亮度特征，其通过金字塔滤波器对 3 尺度的多分辨率亮度图像进行 4 方向的滤波得到；图 4 中（14）~（16）图为利用 ITTI[5] 模型计算得到的颜色，强度，方向 3 个特征；由于 Lab 颜色模型是基于人对颜色

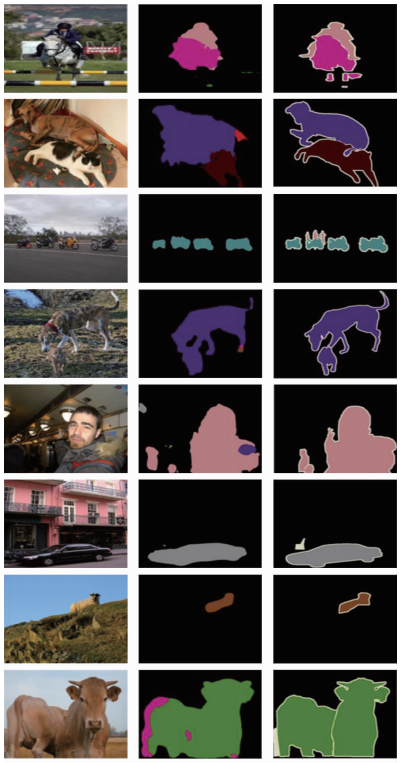


图 3 本文利用改进的 FCN-8s 提取的语义对象特征

Fig. 3 Semantic object features extracted by the improved FCN-8s

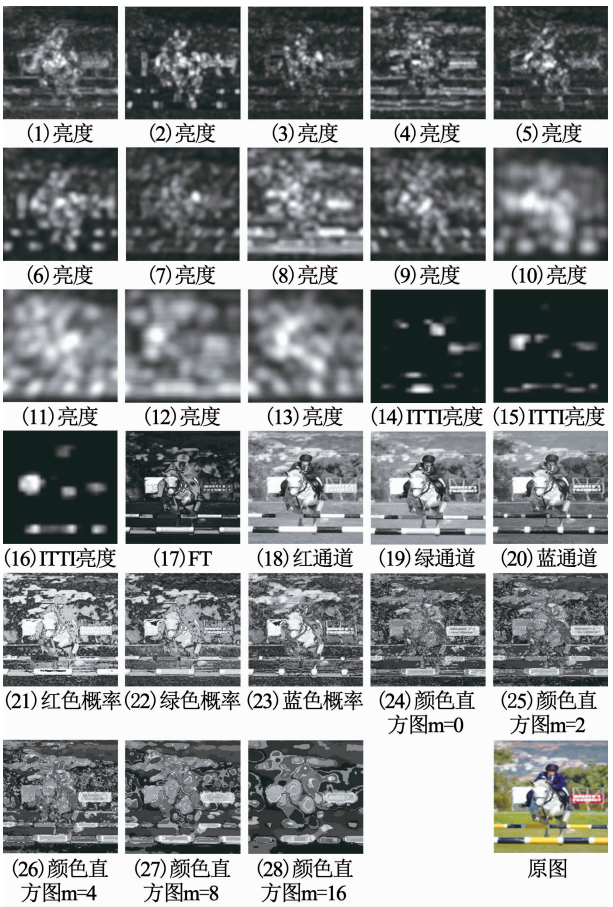


图 4 28 个低级特征

Fig. 4 Twenty eight low-level features

的感觉,因此本文利用 FT[21] 模型提取 Lab 色彩空间特征(图 4 中(17)图);图 4 中(18)~(23)图为计算红、绿、蓝三颜色通道值及概率值,分别得到的 3 个色度特征及 3 个色度概率特征;图 4 中(24)~(28)图为利用中值滤波器对 6 尺度的彩色图像进行滤波,并计算三维色度直方图而得到的 5 个色度直方图特征.

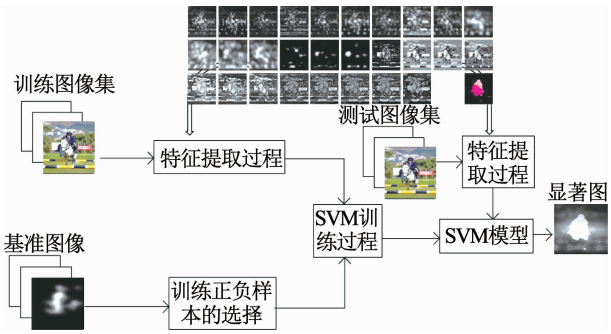


图 5 视觉注意力模型训练过程

Fig. 5 Training process of visual attention model

4 视觉注意力模型的训练

提取语义对象特征 F_S 和低级特征集合 F_L 之后,得到本文提取的特征集合 $F=\{F_S,F_L\}$. 为获取提取的特征集 F 与本文构建的 VOC2012-E 数据库里记录的眼动追踪数据的映射关系,本文引入机器学习理论,利用非线性分类器模拟人脑神经系统的非线性映射. 在统计学习理论中发展起来的 SVM 方法是一种通用学习方法,其在非线性分类应用中有非常好的表现. 因此,本文利用 SVM 理论,训练得到每个特征与视觉注意力之间的关系,并在训练过程中,本文基于特征整合理论对特征集合 F 里的特征进行并行处理[5],最后利用训练得到的映射关系生成视觉显著图.

具体地,训练过程中,取样本集 $S\subset T$,T 为基准图像的训练集. 样本 $s\in S$. 令

$$L(s)=\begin{cases}+1, & s\text{ 为显著样本;}\\ -1, & s\text{ 为非显著样本.}\end{cases}$$

设 P 为基准图像的像素集,$P=\{p_1,p_2,\cdots,p_N\}$,N 为基准图像中像素的个数. $O(p_i)$ 表示像素的显著度,$i=1,2,\cdots,N$. 对像素集 P 进行排序得到有序集合 $P_o=\{p_{o_1},p_{o_2},\cdots,p_{o_N}\}$,其中 $O(p_{o_1})\geqslant O(p_{o_2})\geqslant\cdots\geqslant O(p_{o_N})$. 在利用 SVM 模型进行训练时,本文选择 $S_p\subset S$ 作为正样本,$S_n\subset S$ 作为负样本,其中 $S_p=\{p_{o_1},p_{o_2},\cdots,p_{o_m}\}$,$m=0.05N$,$S_n=\{p_{o_l},p_{o_l},\cdots,p_{o_N}\}$,$N-l=0.3N$,最终预测出显著图,训练过程的流程图如图 5 所示.

5　实验及结果分析

为评测本文视觉注意力模型的性能，将其与 8 种经典的视觉注意力模型在 VOC2012-E 数据库上进行比较，这 8 种模型分别是 AIM [22]，AWS [7]，Judd[13]，ITTI [5]，GBVS[10]，SUN[8]，STB[23] 和 Torralba[9]，然后在 MIT300 测试数据集上在线测试性能，除以上 8 种模型之外，本文模型同时与 4 种先进的视觉注意力模型在 MIT300 数据库上相比较，这 4 种模型分别是 eDN [11]，UID[12]，Deep Gaze2[15]，DeepFeat[16]. 评价函数选取受试者工作特征曲线下面积（Area Uner Curve，AUC），线性相关系数（Correlation Coefficient，CC）及归一化扫描路径显著性（Normalized Scan Path Saliency，NSS）. AUC 是度量视觉注意力模型预测出的显著图与基准图像的差异的一个评价函数，通常，AUC 的值介于 0.5 ~ 1.0 之间，AUC 越大，模型的表现与基准图像更相近. CC 用来衡量视觉注意力模型预测的显著图与基准图像的线性相关性. NSS 为人眼凝视位置在模型输出显著图上的归一化显著值为

$$N_{NSS}(x,y)=\frac{1}{\sigma_s}(S(x,y)-\mu_s), \tag{2}$$

$$\sigma_s=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i-\mu_s)^2}. \tag{3}$$

式中：S 为模型的输出显著图，σ_s 和 μ_s 为模型输出显著图上的均值与标准差.

图 6 为 9 种模型预测显著图的实验结果图的示例，输入图片来自于 VOC2012-E 数据库. 从实验结果可以看出，与其他 8 种模型相比，本文预测的显著图中语义对象特征是高显著性的，这说明本文提出的视觉注意力模型显然更接近于人类视觉认知.

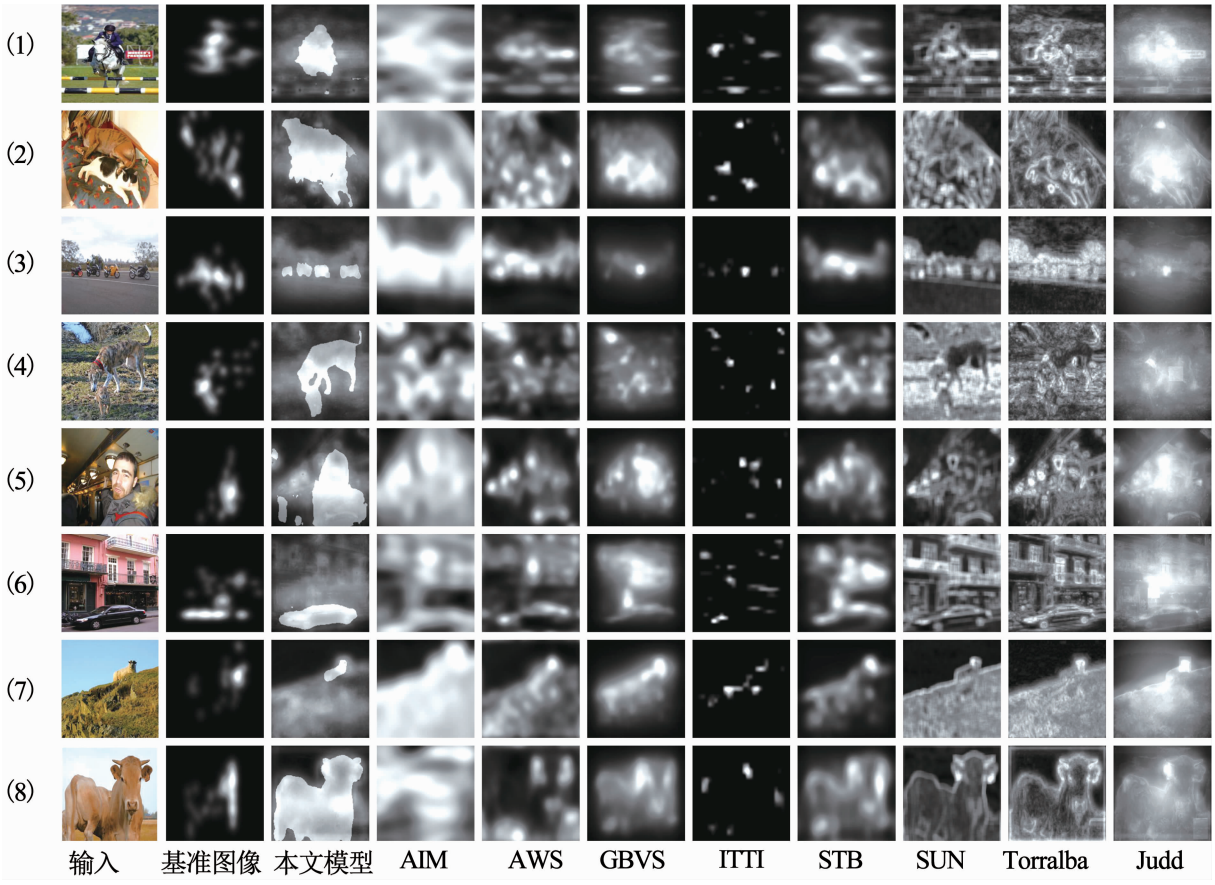


图 6　视觉注意力模型预测的显著图对比

Fig. 6　Comparison of saliency maps predicted by visual attention models

表 2 为 9 种模型在 VOC2012-E 数据库上的评价函数结果比较. 从表 2 可以明显看出，三种评价函数 AUC，NSS 和 CC 评价结果都表明本文模型性能优于其他模型. 本文模型的 AUC 值最高（0.823），其次是 Judd（0.822）和 GBVS（0.810），而 ITTI 模型的 AUC 值仅为 0.531，这意味着本文算法与其他模型相比，更接近于基准图像. 本文模型的 NSS 值为 1.360，其次是 GBVS （1.263）和 Judd （1.242），而 STB 模型的 NSS 值只有 0.399，在 9 种模型中最低，仅为本文模型的 1/4. 本文模型的 CC 值（0.557）也

最大,而 CC 值越高,注意力模型预测的显著图与注视点之间的相关性越高. 因此,本文模型预测的显著图更接近于人眼注视点,这是因为本文模型的提出是受人类自由观察自然场景的感知过程的启发,在模型训练过程中融合了语义对象特征和典型的低级特征.

此外,为了进一步评估本文提出的模型,本文在公开数据库 MIT300 进行评测,该数据库包含了 9 个受试者观察 300 张自然图像时人眼注视点的基准数据,评测结果如表 3 所示,除了前文比较的 8 种模型之外,同时比较了 4 种先进的模型:Deep Gaze2 及 Deep feat 同样整合了高级特征与低级特征;eDN 同样利用 SVM 整合特征;及无监督模型 UID.

根据表 3 结果,AUC 最高值为 Deep Gaze2 模型(0.84),而本文模型的 AUC 值与 GBVS,Judd, eDN 及 DeepFeat 模型的 AUC 值相近,但是本文模型的 NSS 值(1.27)在 12 种模型中最高,而且本文模型的 CC 值与 DeepFeat 模型的 CC 值并列最高(0.49). 其中,整合了同样高级特征与低级特征的 Deep Gaze2 模型与 Deep feat 模型与本文模型总体表现最优,而本文模型略胜一筹,这说明了本文整合语义特征的先进性. 虽然同样利用 SVM 整合特征的 eDN 模型的 NSS 值较低(1.14),但其 AUC 值与本文相同,CC 值也较高,说明了本文利用 SVM 整合特征的有效性. 因此本文模型在 MIT300 数据库上性能表现良好,可以有效地预测人眼注视点.

表 2 9 种模型在 VOC2012-E 数据库上的实验结果比较

Tab. 2 Comparisons of experiment results of nine models on VOC2012-E database

评价函数	本文模型	AIM	AWS	GBVS	ITTI	STB	SUN	Torralba	Judd
AUC	**0.823**	0.664	0.718	0.810	0.531	0.779	0.619	0.679	0.822
NSS	**1.360**	0.595	0.892	1.263	0.674	0.399	0.460	0.681	1.242
CC	**0.557**	0.241	0.351	0.521	0.145	0.453	0.184	0.275	0.518

表 3 12 种模型在 MIT300 数据库上的实验结果比较

Tab. 3 Comparisons of experiment results of twelve models on MIT300 database

评价函数	本文模型	AIM	AWS	GBVS	ITTI	STB	SUN	Torralba	Judd	eDN	Deep Gaze2	UID	Deep Feat
AUC	0.82	0.77	0.74	0.81	0.60	0.63	0.63	0.68	0.81	0.82	**0.84**	0.78	0.81
NSS	**1.27**	0.79	1.01	1.24	0.43	0.97	0.97	0.69	1.18	1.14	1.16	1.24	1.26
CC	**0.49**	0.31	0.37	0.48	0.14	0.37	0.37	0.25	0.47	0.45	0.45	0.45	**0.49**

6 结 论

本文提出了一种整合了语义对象特征的视觉注意力模型. 通过建立眼动跟踪数据库 VOC2012-E,研究并记录普通人在观察自然场景时的眼动数据,经过分析得出语义对象特征在吸引人们的注意力时有重要作用. 然后,受语义分割启发,利用深度学习网络 FCN-8s 提取语义对象特征,同时用激活函数 PReLu,优化函数 Adam 改进网络使其更有效地提取的语义对象特征. 接着,提取在人类潜意识层吸引人注意力的如方向,颜色,强度特征等 28 个低级特征. 最后训练机器学习分类器 SVM 将之前提取的语义对象特征及低级特征映射到人类视觉空间,训练后的模型可以预测自然场景的人眼视觉显著图. 经过在 VOC2012-E 及 MIT300 数据库的测试,与其他 8 种经典模型及 4 种先进模型相比,本文提出的视觉注意力模型性能更好,更符合人类观察图像时的视觉习惯,即语义对象高显著.

参考文献

[1] SMOLA A J, MIKA S, SCHOLKHOPF B, et al. Regularized principal manifolds[J]. Journal of Machine Learning Research, 2001, 1: 179. DOI: 10.1007/3-540-49097-3_17

[2] KANWISHER N, MCDERMOTT J H, CHUN M M, et al. The fusiform face area: A module in human extrastriate cortex specialized for face perception[J]. The Journal of Neuroscience, 1997, 17(11): 4302. DOI: 10.3410/f.717989828.793472998

[3] EPSTEIN R A, KANWISHER N. A cortical representation of the local visual environment[J]. Nature, 1998, 392(6676): 598. DOI: 10.1016/S1053-8119(18)31174-1

[4] EPSTEIN R A, HARRIS A, STANLEY D, et al. The parahippocampal place area: Recognition, navigation, or encoding?[J]. Neuron, 1999, 23(1): 115 DOI: 10.1016/S0896-6273(00)80758-8

[5] ITTI L, KOCH C, NIEBUR E, et al. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254. DOI: 10.1109/34.730558

[6] KOCH C, ULLMAN S. Shifts in selective visual attention: Towards the underlying neural circuitry[J]. Human Neurobiology, 1987, 4(2): 115. DOI: 10.1007/978-94-009-3833-5_5

[7] GARCIADIAZ A, FDEZ-VIDAL X R, PARDO X M, et al. Decorrelation and distinctiveness provide with human-like saliency [M]. Advanced Concepts for Intelligent Vision Systems. Berlin, Heidelberg: Springer, 2009: 343. DOI: 10.1007/978-3-642-04697-1_32

[8] ZHANG Lingyun, TONG M H, MARKS T K, et al. SUN: A Bayesian framework for saliency using natural statistics[J]. Journal of Vision, 2008, 8(7): 32. DOI: 10.1167/8.7.32

[9] TORRALBA A. Modeling global scene factors in attention[J]. Journal of the Optical Society of America A—Optics Image Science and Vision, 2003, 20(7): 1407. DOI: 10.1364/JOSAA.20.001407

[10] HAREL J, KOCH C, PERONA P, et al. Graph-based visual saliency [C]//Proceeding of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006: 545

[11] VIG E, DORR M, COX D. Large-scale optimization of hierarchical features for saliency prediction in natural images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 2798. DOI: 10.1109/CVPR.2014.358

[12] TAVAKOLI H R, LAAKSINEN J. Bottom-up fixation prediction using unsupervised hierarchical models[M]. Asian Conference on Computer Vision. Cham: Springer, 2016: 287. DOI: 10.1007/978-3-319-54407-6_19

[13] JUDD T, EHINGER K, DURAND F, et al. Learning to predict where humans look[C]//Proceedings of International Conference on Computer Vision. Piscataway: IEEE, 2009: 2106. DOI: 10.1109/ICCV.2009.5459462

[14] ZHAO Qi, KOCH C. Learning a saliency map using fixated locations in natural scenes[J]. Journal of Vision, 2011, 11(3): 9. DOI: 10.1167/11.3.9

[15] KUMMERER M, WALLIS T S A, GATYSL A, et al. Understanding low-and high-level contributions to fixation prediction [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4789. DOI: 10.1109/ICCV.2017.513

[16] MAHDI A, QIN Jun, CROSBY G. DeepFeat: A bottom-up and top-down saliency model based on deep features of convolutional neural nets[J]. IEEE Transactions on Cognitive and Developmental Systems, 2019. DOI: 10.1109/TCDS.2019.2894561

[17] LONG J, SHELHAMER E, DARRELL T, et al. Fully convolutional networks for semantic segmentation[J]. Computer Vision and Pattern Recognition, 2015: 3431

[18] CHEN Liangchien, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(4): 834. DOI: 10.1109/TPAMI.2017.2699184

[19] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv e-prints, 2017

[20] EVERINGHAM M, GOOL V L, WILLIAMS C K I, et al. The {PASCAL} {V} isual {O} bject {C} lasses {C} hallenge 2012 {(VOC2012)} {R} esults [DB/OL]. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[21] ACHANTA R, HEMAMI S S, ESTRADA F J, et al. Frequency-tuned salient region detection[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 1597. DOI: 10.1109/CVPRW.2009.5206596

[22] BRUCE N D, TSOTSOS J K. Saliency based on information maximization[C]//Proceedings of Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2005: 155

[23] WALTHER D, KOCH C. Modeling attention to salient proto-objects [J]. Neural Networks, 2006, 19(9): 1395. DOI: 10.1016/j.neunet.2006.10.001

（编辑　苗秀芝）