基于深度融合的显著性目标检测算法

张冬明[1)] 靳国庆[2)] 代 锋[2)] 袁庆升[1)] 包秀国[1)] 张勇东[3)]

[1)](国家计算机网络应急处理协调中心 北京 100029)

[2)](中国科学院计算技术研究所智能信息处理实验室 北京 100190)

[3)](中国科学技术大学信息科学技术学院 合肥 230026)

摘 要 自然图像往往包含各种复杂的内容,基于单一特征的显著性检测算法很难从复杂场景中提取符合人类视觉的显著性目标.虽然多种显著图的融合能够弥补或者纠正单一特征带来的检测缺陷,但是不合理的显著图融合方式可能会进一步降低算法的检测性能.为了解决多种显著图的有效融合问题,作者提出了一种基于深度卷积神经网络的特征图深度融合模型.算法使用四种低层显著图作为网络的输入,采用前融合和后融合的双通道卷积网络学习图像的显著目标.前融合通道利用一个多层的全卷积网络生成对目标物体边缘敏感的显著图,后融合通道使用权重共享的浅层网络分别获得四种目标对象位置保持的高层语义显著图.两个通道的特征图再通过一个四层的全卷积网络进行优化,从而获得最终的显著图.在公开数据集上的大量实验证明了本文提出的显著图深度融合算法的有效性.

关键词 显著目标检测;人工特征;深度融合;深度学习;显著图

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2019.02076

Salient Object Detection Based on Deep Fusion of Hand-Crafted Features

ZHANG Dong-Ming[1)] JIN Guo-Qing[2)] DAI Feng[2)] YUAN Qing-Sheng[1)]

BAO Xiu-Guo[1)] ZHANG Yong-Dong[3)]

[1)](*The National Computer Network Emergency Response Technical Team Coordination Center of China*, *Beijing* 100029)

[2)](*Intelligent Information Processing Key Lab*, *Institute of Computing Technology*, *Chinese Academy of Sciences*, *Beijing* 100190)

[3)](*School of Information Science and Technology*, *University of Science and Technology of China*, *Hefei* 230026)

Abstract Visual saliency detection is an important and fundamental research problem in neuroscience and psychology, which investigates the mechanism of human visual systems in selecting regions of interest from complex scenes. Recently it has also been a 5 active topic in computer vision, due to its applications to object detection, video summarization, image editing techniques, image retrieval, face detection, and fine-grained visual categorization. Saliency detection is commonly interpreted as a process that includes two stages: (1) detecting the most salient regions in accordance with human visual attention and (2) segmenting the accurate boundary of that regions. In general, saliency detection methods can be categorized as either bottom-up or top-down. The former focuses on stimulus-driven stage of attention which is of main interest in computer vision community. Contrasted with bottom-up methods, top-down approaches usually require supervised learning with manually labeled ground truth. However, natural images often contain a variety of complex content. Saliency detection method based single visual feature hardly extract salient object that are consistent with human visual system from complex scenes. Although the

收稿日期:2018-07-02;在线出版日期:2019-04-30.本课题得到国家重点研发计划(2018YFB0804202)、国家自然科学基金项目(61672495,61771458,61525206)资助.**张冬明**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为视频编码、多媒体内容分析与理解、模式识别.E-mail:dmzhang@ict.ac.cn.**靳国庆**(通信作者),博士,助理研究员,主要研究方向为模式识别.E-mail:jinguoqing@ict.ac.cn.**代 锋**,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为视频编码.**袁庆升**,博士研究生,高级工程师,主要研究方向为信息安全.**包秀国**,博士,正高级工程师,主要研究方向为信息安全.**张勇东**,博士,教授,主要研究领域为多媒体内容分析.

fusion of various saliency maps is able to compensate or correct the defects of single visual feature, the irrational fusion may further degrade the performance. In order to solve the problem of effective fusion of saliency maps, we propose a deep fusion model based on deep convolutional neural network. We use four hand-crafted feature maps, which include Local Contrast (LC), Global Contrast (GC), Spatial Variance (SV) and Center Variance (CV), as the input of the network and learn the saliency values in a dual-estimation process. The anterior fusion estimation is fed with fused maps to learn values with precise boundaries. The posterior fusion estimation is fed with feature map to learn the object conception saliency separately. Four feature maps are exploited dependent to detect salient region, and as well as six fusion schemes including HF, HF_p, HF_a, proposed in the paper, and CRF, CSVM and WA. The evaluations show that HF gains the highest performance, which shows the superiority of anterior and posterior. Therefore, HF is selected as the fusion scheme and the two features are concatenated and integrated into a jointly optimized network for final saliency detection. Extensive experiments are carried out on four benchmark datasets including ASD, PASCALS, ECSSD, HKU-IS. Two measurements including *MAE* and *F*-Score are calculated to evaluate the proposed method's performance and the existing methods including HS, GMR, DSR, DRFI, LEGS, MDF, MCDL and ELD, among which, HS, GMR, DSR and DRFI are traditional methods based on low-level features, while LEGS, MDF, MCDL and ELD are based on deep learning. The PRCs of the methods are also used to illustrate their performances. The results show that the proposed method gain significant and consistent improvements over the representative deep learning framework based saliency detection methods. We believe that the proposed method's success comes from the following factors. (1) Four hand-crafted feature maps come from different level features, which are complementary. (2) Deep network is efficient to locate the correlation among the feature maps. (3) Contrasted with shallow fusion model SVM and CRF, HF can fuse the different level feature maps and get the better performance.

Keywords　salient object detection; hand-crafted features; deep fusion; deep learning; saliency map

1　引　言

显著性目标检测旨在根据人的注意力来选择图像的关键目标，它不仅需要输出显著目标的位置和尺度，而且需要给出每个像素的显著值[1]. 传统的显著性目标检测方法主要依赖于三个方面：有效的特征表示、隐形的显著性先验和优秀的特征融合策略. 传统算法往往基于人工特征的对比度（例如：颜色、纹理和边缘梯度[2-3]）和人工设计机制（例如：频域[4]和低秩矩阵[5]）来计算显著度. 受人类视觉注意力机制的启发，Itti 等人[6]最早提出了基于中心环绕机制的显著目标检测模型. 其它常用的显著性先验包括背景先验[7-8]和物体性先验[1,9].

神经学研究[10]表明，视觉神经元的显著度是由不同刺激的最强响应决定的. 对应到视觉显著目标检测算法上，不同的视觉特征就是不同的刺激，图像某个位置的视觉显著度是由该位置的多个显著特征的最强响应决定. 由于自然图像中往往包含多种复杂纹理背景，每种特征[3]都有一定的局限性，只能在某些方面凸显目标的显著性，单独依赖一种特征的显著性检测模型不能获得很好的检测效果. 为了提升检测效果，很多方法试图将多种显著特征进行融合来描述物体的显著性. GC（Global Contrast）算法[11]将全局颜色对比度和颜色空间分布进行线性融合得到图像的显著性目标. PD（Pattern Distinctness）算法[12]采用模式差异、颜色独特性和高层先验作为目标的显著特征，通过简单的相乘运算计算目标的显著值. SF（Saliency Filters）方法[13]利用像素的独特性和空间分布特征作为其显著特征，并采用高维滤波器将二者进行融合，最后通过图像上采样获取像素点的显著性度量. UFO（Uniqueness, Focusness and Objectness）算法[9]采用一种启发式搜索策略，将像素的独特性、聚焦性和对象性三种重要特征进

行融合构建显著目标检测模型. Yan 等人[14]提出了多尺度融合的显著性检测模型,该模型提取不同尺度图像的局部对比度特征和空间分布特征,然后使用多层次推理机制,将不同尺度的显著图进行融合得到最终的显著图. Zhang 等人将显著性检测建模为图模型,并利用吸收马尔可夫链进行节点汇聚来获得显著图[15]. 上述算法虽然能够显著提升模型检测性能,但是均未考虑特征之间的交互抑制关系,因此还有很大的提升空间.

人类视觉系统的物体知觉阶段把相互独立的各种视觉特征融合成物体的表征,从而形成对感兴趣物体的定位. 显著特征融合与视觉系统的知觉阶段类似,把多种视觉显著特征关联起来,形成互补的优势,从而更好的凸显显著性目标. 基于不同的特征图融合函数,可以将显著特征融合模型划分为不同的类型. 线性感知模型[6]是一个自底而上的检测过程,通过线性加权的方式将不同的特征图结合在一起. 贝叶斯模型[16]是一个自底而上与自顶而下相结合的融合模型,把感知特征、先验分布和后验概率按照贝叶斯规则进行概率融合,具有很强的特征融合能力. 决策论模型[17]把视觉显著性目标检测过程解释演化成目标与背景分割的最优化决策问题,而信息论[18]则把计算模型定义为场景中最大化信息抽样. 特征图融合可以看成在一个时间序列上眼球的空间运动,因此可以使用隐马尔科夫模型(Hidden Markov Models,HMM)[7]、条件随机场(Conditional Random Fields,CRF)[19]等图模型在空间和时间上对特征图进行更加复杂的融合.

近年来,深度学习极大促进了显著性目标检测技术的发展,与传统的算法相比,基于深度学习的显著性模型在公开数据集上的检测结果远远超过了传统算法. 如全网化边缘检测[20](Holisitcally-Nested Edge Detector,HED)利用多尺度多水平的信息融合获得了更准确的语义分割. 进一步,在 HED 的不同尺度的边信息上利用短连接,可以获得更丰富的特征来细化语义分割[21]. 分阶段细化模型(Stage-wise Refinement Model,SRM)[22]将显著性检测分为粗检测和细化两个步骤,细化时利用金字塔池化(Pyramid Pooling Module,PPM)并与初始显著图进行融合. 文献[23]提出了一种无监督的显著性检测学习方法,以降低对数据标注的要求.

自然图像图包含各种复杂的背景,单一特征往往很难从复杂背景中提取出显著物体. 每种特征都有一定的局限性,各种特征融合可以弥补或者纠正单一特征的缺陷,从而获得更符合人类视觉的显著目标. 然而,不恰当的特征图融合方式可能会进一步降低显著目标检测的准确性,因此需要设计更加有效的模型将多种显著图进行融合.

基于以上分析,我们提出了一种基于深度学习的特征图融合模型,通过有监督的学习机制将低层特征显著特征进行自动融合. 算法采用局部颜色对比度、全局颜色对比度、空间紧致度和中心先验特征四种低层特征生成显著图. 低层特征图融合过程包括前融合网络、后融合网络和特征图优化网络三个组成部分,如图 1 所示. 我们使用四种低层显著特征作为网络的输入,采用双通道的卷积网络学习图像的显著目标. 上面的前融合网络利用全卷积编码网络生成对物体目标边缘敏感的显著图,下面的后融合网络用权重共享的浅层网络分别获得四种目标对象位置保持的高层显著图. 前融合阶段,四种显著图被拼接后输出到一个多层的卷积网络,前融合的编码网络能够保证输出的显著图对物体的边缘敏感. 后融合阶段,各种显著图被分别输出到一个共享权重的浅层网络中,该网络能够保证生成的显著图更

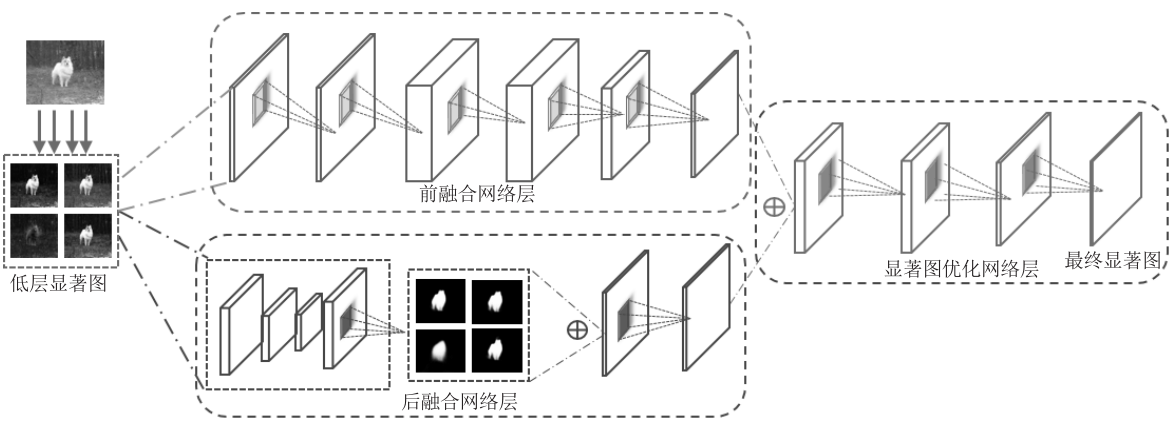

图 1 基于深度网络的人工显著特征深度融合模型框架图

好地保持其物体特性.最后,前融合特征图和后融合特征图经过一个四层的编码网络进行特征图优化,从而获得最终的特征图.

与之前工作相比,本文基本思路是利用已有的显著图特征,着重于解决多显著图如何进行有效融合,以提升显著性检测的性能,实验结果表明深度学习是一种有效的显著图融合的方法.

2　显著特征计算

2.1　颜色对比度

高对比度的视觉信号是人类视觉定位感兴趣目标的重要线索,也是显著性目标检测模型中重要的度量特征.基于此,我们提出了基于颜色对比度的图像像素显著性特征.具体来讲,一个像素的显著度等于该像素与图像中其它像素颜色的对比度之和,那么像素 I_k 的显著性定义为

$$S_{cc}\propto\sum_{\forall I_i\in I}D(I_k,I_i)\cdot\omega(p_k,p_i)\tag{1}$$

其中,I_i 表示图像的其它像素点,$D(I_k,I_i)$ 表示像素 I_k 和像素 I_i 的颜色距离.由于距离像素点 k 近的像素对其影响较大,因此还需要考虑位置因素的影响,加入权重 $\omega(p_k,p_i)$,其定义如下:

$$\omega(p_k,p_i)=\frac{1}{Z}\exp\left(-\frac{\|p_i,p_j\|^2}{2\sigma_p^2}\right)\tag{2}$$

其中 $Z=\sum_{j=1}^{N}\omega(p_k,p_i)$ 是归一化因子,σ 是距离调节参数,控制着 S_{cc} 的特征尺度,σ 取值较小时,$\omega(p_k,p_i)$ 趋近于 0,则 S_{cc} 成为一个局部对比度算法,反之,S_{cc} 表示一个全局对比度算子.

为了使距离度量更符合人眼对色彩的感知,我们定义了基于人眼感知的色彩差异度量方法:

$$D(c_i,c_j)\propto 1-\exp\left(-\frac{d(c_i,c_j)}{2\sigma}\right)\tag{3}$$

其中,$d(c_i,c_j)$ 表示颜色之间的欧式距离,σ 表示图像颜色的方差.引入图像颜色的方差度量使得图像中的颜色对比度具有全局特征,两种颜色的对比度不仅与其自身相关,而且与其所在的图像整体颜色分布相关.在图像对比度比较大的图像中,即使两种颜色的差值较大,其对比度可能会很低;在颜色单调的图像中,即使两种颜色的差值较低,其在整幅图像中的对比度可能会很高.因此,颜色值相同的像素显著度相等,我们可以将式(1)转换为

$$S_{cc}(I_k)\propto f_iD(c_i,c_j)\cdot\omega(p_k,p_i)\tag{4}$$

其中像素 I_k 的像素值为 c_j,f_i 表示像素值 c_j 的概率密度,n 为图像的像素点个数.得到图像的对比度后,将其值域归一化到[0,1]之间得到对比度显著图.

2.2　空间紧致度

颜色分布的空间信息是显著性度量的重要特征.通常情况下,显著区域的颜色分布相对比较紧凑,而背景区域的颜色分布则具有很高的空间变换性,即背景颜色一般会分布在整个图像区域.我们使用颜色在空间分布的方差衡量其紧致性,方差越小,颜色越集中分布在图像的某一块区域,方差越大,颜色的分布越松散,颜色的紧致度定义为

$$S_{sv}(I_k)\propto 1-\frac{1}{N}\sum_{i=1}^{N}\|p_i^{c_l}-\mu^{c_l}\|^2\tag{5}$$

其中,c_l 表示像素点 I_k 的颜色值,N 表示图像中 c_l 颜色的像素点个数,$p_i^{c_l}$ 表示 c_l 颜色的像素点的位置坐标,μ^{c_l} 表示 c_l 颜色的质心,定义为

$$\mu^{c_l}=\frac{1}{N}\sum_{i=1}^{N}p_i^{c_l}\tag{6}$$

为了更高效地计算空间紧致度,我们将颜色进行了量化[3].得到的空间紧致度显著图被量化到[0,1]之间.

2.3　中心先验

视觉研究[24]表明,人类倾向于关注图像中心位置的物体,距离图像中心越近,物体目标越显著.由此可以看出,中心先验是显著性分析的重要特征,充分利用这一特征可以获得更加有效的视觉显著度计算模型.我们定义中心先验的显著度检测模型:

$$S_{cv}(I_k)\propto 1-\frac{1}{N}\sum_{i=1}^{N}\|p_i^{c_l}-p_c\|^2\tag{7}$$

其中 p_c 表示图像的中心位置,$p_i^{c_l}$ 表示像素值为 c_l 的像素点坐标值.

3　深度融合网络

采用上节的显著特征计算方法,我们生成局部颜色对比度(Local Contrast,LC)、全局颜色对比度(GC)、空间紧致度(Spatial Variance,SV)和中心先验显著度(Center Variance,CV)四种低层显著特征,构建了四个相互独立且相互补充的显著图.下面我们将详细描述显著图的深度融合算法.

3.1　网络结构

深度融合网络首先实现显著图的前融合.显著图被拼接成 $128\times128\times4$ 的四通道图像,输入到前融合网络,如图 2 所示.深度网络将四个通道的显著

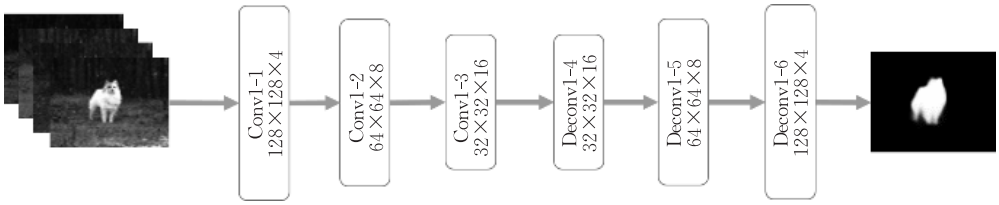

图 2 前融合网络架构图

特征逐层抽象,得到高级语义的显著图.前融合网络包括 3 层编码器和 3 层解码器,通过编码器获得融合层的显著图,即 $x_s^0 \rightarrow x_s^1 \rightarrow x_s^2 \rightarrow z$,之后通过解码器还原特征图,即 $z \rightarrow \hat{x}_s^2 \rightarrow \hat{x}_s^1 \rightarrow \hat{x}_s^0$.

整个深度图的后融合过程如图 3 所示,前四层是权重共享的独立通道,随后是一个两层编码融合网络层.四种不同的显著图作为四个通道的输入,通过深度网络的学习,将四个通道的显著特征逐层抽象,最终得到带有高层语义的显著特征表达.经过深度网络逐层抽象后的视觉特征更适合显著图的融合,这个显著图融合过程是由特征编码网络和特征解码网络组成.首先通过并行独立的两层特征编码网络和两层特征解码网络,得到四个独立的初始显著图 $\hat{x}_{s_i}$,然后通过一个双层的编码网络将四个独立的显著图进一步融合,即 $(\hat{x}_{s_0}^0 \rightarrow \hat{x}_{s_1}^0 \rightarrow \hat{x}_{s_2}^0 \rightarrow \hat{x}_{s_3}^0) \rightarrow (\hat{x}_{s_0}^1 \rightarrow \hat{x}_{s_1}^1 \rightarrow \hat{x}_{s_2}^1 \rightarrow \hat{x}_{s_3}^1) \rightarrow (\hat{x}_{s_0}^2 \rightarrow \hat{x}_{s_1}^2 \rightarrow \hat{x}_{s_2}^2 \rightarrow \hat{x}_{s_3}^2)$.显著特征图的编码方程和解码方程用公式表示为

$$x_{s_i}^{l+1} = \sigma(\boldsymbol{W}_{s_i}^l x_{s_i}^l + b_{s_i}^l) \tag{8}$$

$$\hat{x}_{s_i}^l = \sigma(\boldsymbol{W}_{s_i}^l x_{s_i}^{l+1} + c_{s_i}^l) \tag{9}$$

其中 $x_{s_i}^l$ 是独立显著图通道在第 l 层的参数,$\hat{x}_{s_i}^l$ 是显著图融合后的结果.$\boldsymbol{W}_{s_i}^l$ 是连接第 l 层与第 $l+1$ 层的权重矩阵,$b_{s_i}^l$ 和 $c_{s_i}^l$ 分别是编码器和解码器的偏移量.

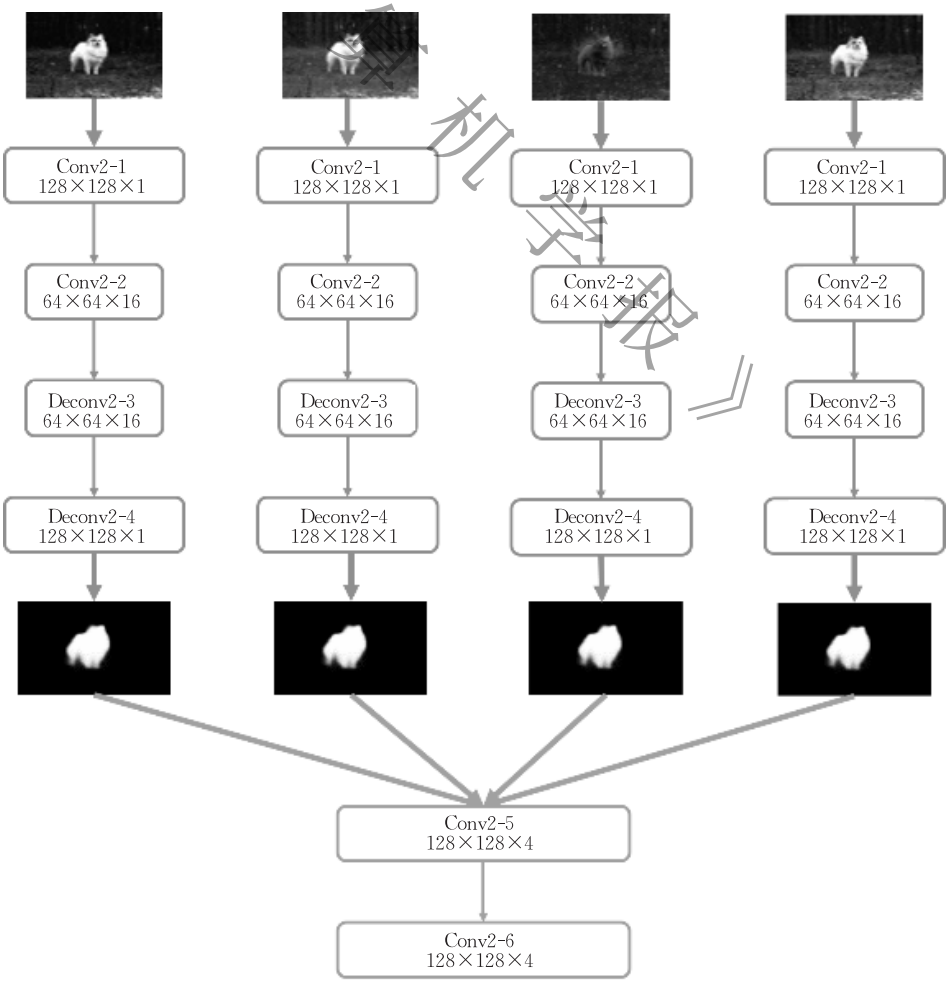

图 3 后融合网络架构图

由于初始显著图的维度不变，因此我们将此特征图定义为紧致的低层显著特征图. 紧致的低级显著特征图一方面能够找到最佳的特征非线性组合，另一方面能有效地实现层级特征和高层特征之间的平衡，从而提高算法的融合性能.

最后，前融合网络和后融合网络的输出被拼接成 128×128×8 的特征图. 这里，我们使用“级联(concate)”方法，这个操作只改变特征的深度，不改变特征的宽和高. 尽管两种融合方法输出的显著图大小一样，但是他们使用不同的感受野和融合方法，所以两个融合方法获得的显著图特性不同.

如图 1 所示，我们使用另外的三层卷积网络对获得的显著图进一步优化. 第一层的卷积核为 3×3×8，第二层的卷积核为 3×3×4，最后一层的卷积核为 1×1×1，每个卷积层后面跟着一个 ReLu 层. 因此，优化网络能够获得显著图的非线性融合并生成最终的显著图. 特征融合网络没有改变显著图的大小，因此我们的显著图优化方法不会带来显著边缘模糊的问题.

3.2　模型训练

模型的参数可以通过最小化模型预测值和对应标注间的误差来优化. 我们使用了一个二分类网络来分割显著目标和非显著背景区域，网络的最终优化目标通过最小化 softmax 的交叉熵损失函数实现.

$$L(\Theta)=-\sum_{j=0}^{1}1_{(y=j)}\log\left(\frac{e^{z_j}}{e^{z_0}+e^{z_1}}\right) \tag{10}$$

其中，z_0 和 z_1 分别是训练样本标注的非显著值和显著值.

本文采用 xavier 初始化网络参数，使用随机梯度下降的反向传播算法最小化代价函数，从而实现对网络参数优化更新. 在训练时，我们从 MSRA10K 数据集中删除 ASD 的图像，然后从剩余图像中选择 9000 张作为训练样本，每次训练的过程中仅使用一张图像. 我们选取 Batchsize 为 1，即每 1 个样本都要进行一次迭代处理. 参数的学习率 $\varepsilon=1e-10$，动量选取为 $v_1=0.95$，权重衰减设置为 0.0005，参数 ω 的更新策略为

$$v_{i+1}=0.95\cdot v_i-0.0005\cdot\varepsilon\omega_i-\frac{\partial L}{\partial\omega}\bigg|\omega_i \tag{11}$$

$$\omega_{i+1}=\omega_i+v_{i+1} \tag{12}$$

其中 $i\geq1$ 表示训练迭代的次数.

为了解决训练样本不足的问题，我们使用了数据扩充和 Dropall 两种数据增强方法. 考虑到显著目标的不对称性，我们通过水平反射的方法将训练图像进行翻转，并将翻转后的图像和显著性标注加入训练集，从而将训练样本的数量增加一倍. 由于融合网络之间是权重共享的，我们使用结合了 DropOut 和 DropConnect 的 Dropall 方法，不仅丢弃部分节点，而且舍弃部分神经元的连接. 通过 Dropall 方法减少了神经元之间的依赖性，从而使得网络学习到更加鲁棒的特征表达，避免了模型的过拟合问题.

4　实　验

针对本文提出的基于深度融合的显著目标检测算法，本节给出了算法在公开数据集上的实验结果. 实验结果表明我们的算法能够有效的检测出图像的显著性目标，与其它典型显著性检测算法相比，我们的算法具有更高的准确率和召回率.

我们将本文提出的算法与 8 种经典的显著性目标检测模型进行了对比，包括 HS(Hierarchical Saliency)[14]、GMR(Graph-based Manifold Ranking)[7]、DSR(Dense and Sparse Reconstruction)[25]、DRFI(Discriminative Regional Feature Integration)[26]、LEGS(Local Estimation and Global Search)[27]、MDF(Multiscale Deep Features)[28]、MCDL(Multi-Context Deep Learning)[29] 和 ELD(Encoded Low-level Distance)[30]. 其中 HS、GMR、DSR 和 DRFI 是使用低层特征的传统检测算法，而 LEGS、MDF、MCDL 和 ELD 是基于深度学习的显著性目标检测算法. 为了保持对比实验的公平性，我们直接使用作者提供的默认配置的源代码和模型，部分数据集直接使用了作者发布的显著性检测结果.

我们在四种公开的数据集上进行了实验，包括 ASD[8]、PASCALS[1]、ECSSD[14] 和 HKU-IS[28]，这四个数据集都是显著性目标检测的典型数据集，特别是数据集 HKU-IS，非常具有挑战性.

为了保证对比实验的客观性，我们采用准确率(Precision)、召回率(Recall)、平均绝对误差(Mean Absolute Error，MAE)等显著性目标检测普遍使用的评价标准. 为了更形象地展示实验结果，我们绘制了实验结果的 PRC 曲线(Precision-Recall Curve). 为了综合评估准确率和召回率，我们使用 F-score 量化分析结果，准确率与召回率的加权调和平均 F-score 定义为

$$F\text{-score}=\frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall} \tag{13}$$

为了凸显准确率的重要性，实验设置 $\beta^2=0.3$.

为了统计显著目标检测的准确率与召回率，我

们首先将连续的显著图用固定阈值归一化到 0 到 255 的离散空间，然后与二值化的 groudtruth 标注图进行比较，计算得到检测结果的准确率和召回率. 最后，通过自适应阈值（显著图的均值）将显著图二值化，就可以统计出来检测结果的平均准确率、平均召回率和最大 F-score.

4.1 融合方法测评

为了验证基于人工特征的深度融合方法 HF 的有效性，我们在四种数据集上测试了不同的融合算法. 本文提出的融合算法与几种前融合和后融合策略进行了对比，包括如下方法：

(1) 前融合方法 HF_a. 前融合方法是指将四种人工特征拼接后送到前融合网络中，如图 2 所示，最后接一个 1×1 的卷积层生成融合显著图.

(2) 后融合方法 HF_p. 后融合方法是指将四种人工特征分别经过如图 3 的融合网络后，经过 1×1 的卷积网络融合四种显著图.

(3) 加权平均融合 WA(Weighted Averaging). 加权平均法是最简单直接的显著图融合方法，它能够提高显著图的信噪比，但是会削弱图像的细节信息，降低显著图的对比度.

(4) 特征级联 CSVM(Concatenated-SVM)[31]. 特征级联方法将待融合的特征进行级联，使用线性 SVM 进行分类识别. 我们将四种人工显著图进行级联，并使用 SVM 判别像素是否显著.

(5) 条件随机场融合 CRF[32]. 基于条件 CRF 的显著图融合模型充分地利用各类特征挖掘图像的显著性，稳定性高，对阈值不敏感.

表 1 给出了多种融合算法的检测结果. 将局部颜色对比度(LC)、全局颜色对比(GC)、空间紧致度(SV)和角点中心先验(CV)四种人工显著图融合后，显著性检查的性能得到了大幅度的提升. 简单的加权平均融合方法 WA 并不能显著提升显著图的融合效果，特征级联 SVM 和条件随机场 CRF 虽然能够提升显著图的检测效果，但是融合的精度远远低于我们提出的融合模型.

表 1 不同特征融合方法在四种数据集上的 F-score

数据集	HF	HF_p	HF_a	CRF	CSVM	WA	CV	SV	GC	LC
ASD	**0.959**	0.932	0.842	0.715	0.628	0.582	0.528	0.531	0.603	0.492
PASCALS	**0.838**	0.802	0.524	0.775	0.597	0.448	0.431	0.502	0.564	0.519
ECSSD	**0.874**	0.856	0.571	0.762	0.614	0.473	0.503	0.431	0.561	0.415
HKU-IS	**0.861**	0.841	0.584	0.793	0.538	0.415	0.382	0.473	0.515	0.437

本文提出的深度融合模型的优越性来源于以下几个方面：(1) 多种显著图经过深度网络逐层抽象后含有高层语义概念，因此更容易融合；(2) 本文提出的前融合网络能够发现显著图之间的相关性；(3) 与浅层的线性 SVM、CRF 相比，本文提出的融合方法将显著图的低层、中层和高层的多个语义层级特征融合在一起，对显著性目标检测更有利.

4.2 与现有方法对比

为了更客观地评价各种算法的性能，表 2 统计出了各种算法在 ASD、PASCALS、ECSSD、HKU-IS 四种数据集上生成显著图的 MAE 和 F-score 指标. 在 ASD 和 PASCALS 数据集上，本文提出的算法要明显优于其它算法，即在使用相同阈值分割显著图的情况下，我们的算法具有最高的准确率，得到的分割目标与真实标注最为相近. 在 PASCALS 数据集上，我们算法的 F-score 比 ELD 算法高了 0.131. 在 ECSSD 和 HKU-IS 数据集上，我们的算法也明显优于其它深度学习方法. 由于 ELD 使用了低层人工特征与高层深度语义特征的融合，因此其能够获得比其它显著性检测算法更加优秀的检测性能. 特别需要指出的是 MDF 算法在 HKU-IS 数据上获得了很高的 F-score，这是因为 MDF 算法在训练时使用了 HKU-IS 数据集中 3000 张图像作为训练样本.

表 2 不同特征融合方法在四种数据集上的 MAE 和 F-score

方法	ASD		PASCALS		ECSSD		HKU-IS	
	MAE	F-score	MAE	F-score	MAE	F-score	MAE	F-score
Our	0.021	0.959	0.079	0.838	0.079	0.874	0.086	0.861
ELD	0.025	0.930	0.101	0.784	0.080	0.868	0.076	0.839
MCDL	0.037	0.904	0.130	0.743	0.097	0.822	0.098	0.780
MDF	0.053	0.921	0.108	0.782	0.109	0.832	0.109	0.849
LEGS	0.064	0.895	0.115	0.772	0.102	0.827	0.106	0.788
DRFI	0.085	0.925	0.196	0.716	0.166	0.790	0.115	0.806
DSR	0.080	0.886	0.205	0.675	0.173	0.742	0.118	0.785
GMR	0.075	0.921	0.217	0.677	0.189	0.750	0.132	0.733
HS	0.111	0.904	0.262	0.669	0.228	0.734	0.157	0.741

图 4 给出了本文提出的算法与现有方法检测结果的 PRC 曲线. 从图中可以看出，我们的算法显示出了更高的准确率和召回率. 由于我们的算法通过深度网络融合了显著图的低级、中级和高级语义特征，因此，我们算法的检查性能超过了当前性能最好的深度学习方法 ELD. 尤其是在 PASCALS 数据集上，我们的算法给出了更高的 PRC 曲线.

图 4 同时给出了与其它算法的比较 F-score 曲线. 在所有的测试数据集上，我们算法的准确率和召回率更加的均衡. 与 LEGS、MDF、MCDL、ELD 等深度学习算法相比，我们的算法具有更好的泛化能力，在所有数据集上的 F-score 都给出了更优秀的结果. 此外，我们也观察到了本文所提出的方法在处理比较复杂的数据集时表现得更好，例如 HKU-IS 和 PASCALS，其中包含大量具有多个显著对象的图像. 这表明我们的方法能够检测和分割最显著的对象，而其他方法在这些方面存在着缺陷.

图 4 也给出了方法在平均准确率、平均召回率、和最大 F-score 的比较结果. 可以看出，我们的算法在所有的数据集上都获得了较好的准确率和召回

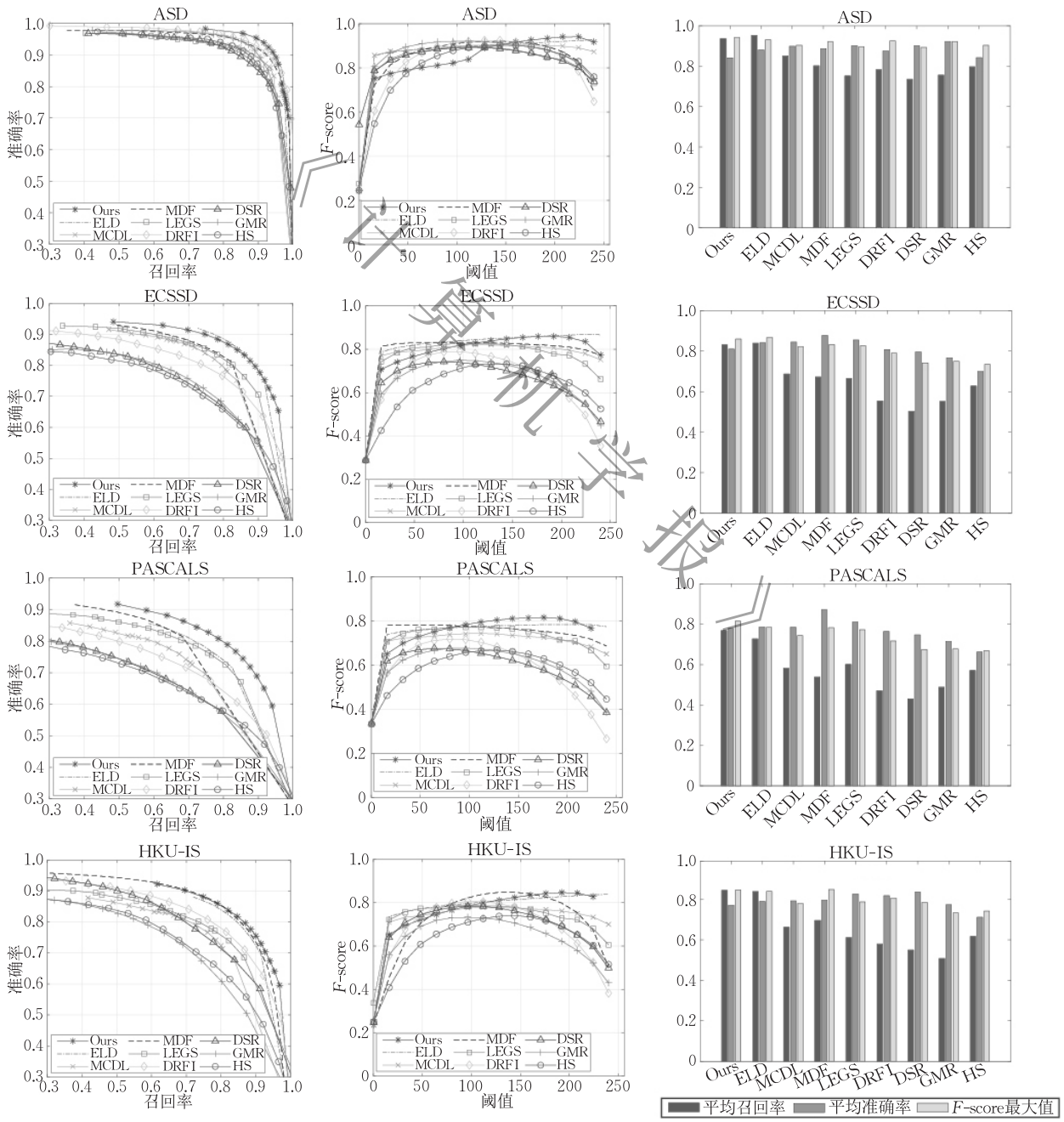

图 4　不同算法在四种数据集上的 PRC 曲线、F-score 以及平均召回率、平均准确率、最大 F-score 柱状图
（与 9 种现有最先进的方法进行了比较，我们的算法性能远远超过了传统方法和基于深度学习的算法）

率.由于我们的模型通过前融合和后融合方法将四种深度图融合后,经过一个编码网络对显著度进一步优化,这就使得我们获得的显著图具有更好的精度.

图 5 给出了我们的方法与上述方法的视觉比对.基于低层人工特征的算法虽然在一定程度上能够凸显显著性目标,但是往往不能够有效地过滤掉背景噪声,同时在低对比度的情况下表现不佳.HS 算法往往会丢失显著性目标的形状信息,GMR 方法对显著性目标的位置极为敏感,当显著性目标接触到图像边缘时,GMR 算法不能够很好地将显著性目标与背景区域分割.DRFI 算法通过有监督的学习融合了高层先验信息,其检测效果总体表现要好于其它传统算法.基于深度学习的显著性检测算法能够显著提升目标的精准度,不仅能够给出显著目标的精确定位,而且给出了更为可信的显著值.和基于深度学习的模型相比,我们的融合算法的检测结果仍不逊色.与性能最好的深度模型 ELD 相比,我们的检测结果对显著目标的边缘处理更加准确,与人工标注结果 GT 更相近.可以很容易地看到,本文提出的融合模型不仅明显突出了显著性目标,而且可以生成精准的目标物体边界.由于我们的模型经过了四种基本显著图的前融合和后融合,我们的方法生成的显著图更可靠.此外,由于使用了多特征融合,显著图不同区域会获得不同打分,即显著图的像素值介于[0,255]之间,来体现显著性的重要程度,而导致显著图会出现"模糊".在各种复杂的场景下,本文提出的融合算法都能够生成更准确的显著图,尤其能够凸显低对比度的目标对象.

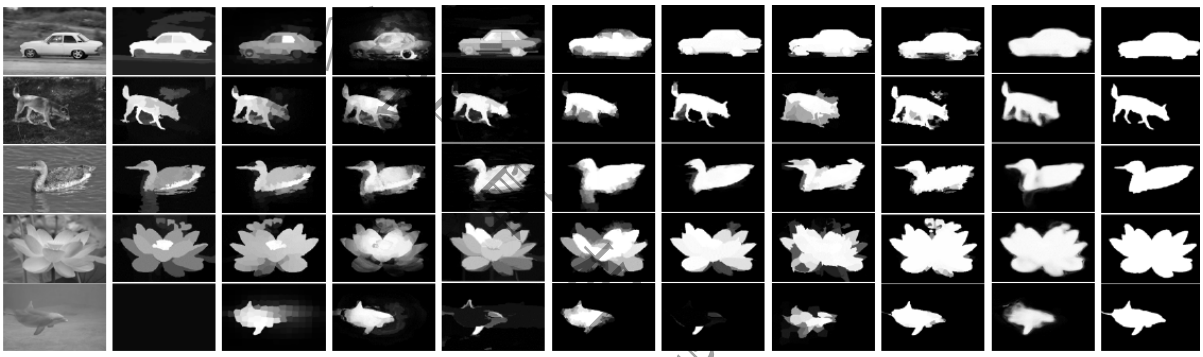

图 5 我们算法与其它显著性检查算法的视觉比对

5 总 结

为了解决多种显著图的高效融合问题,本文提出了一种基于深度融合的显著性目标检测算法.算法设计前融合、后融合和优化融合的多样特征融合网络,从而获得基于低层显著图融合的显著性目标.通过定性与定量实验比较验证了本文提出的基于深度融合的显著性目标检测算法能有效地提取图像中的显著性目标,证明了本文算法的有效性和鲁棒性.在未来的工作中,我们将进一步探索并优化网络结构,利用深度网络的多尺度特征提升特征融合的能力,从而提高显著性目标检测的精度.

参 考 文 献

[1] Li Yin, Hou Xiaodi, Koch C, et al. The secrets of salient object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 280-287

[2] Zhang Peng, Wang Rui-Sheng. Detecting salient regions on location shift and extent trace. Journal of Software, 2004, 15(6): 891-898(in Chinese)
(张鹏,王润生.基于视点转移和视区追踪的图像显著区域检测.软件学报,2004,15(6):891-898)

[3] Cheng Ming-Ming, Mitra N J, Huang Xiaolei, et al. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 569-582

[4] Hou Xiaodi, Zhang Liqing. Saliency detection: A spectral residual approach//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2007: 1-8

[5] Shen Xiaohui, Wu Ying. A unified approach to salient object detection via low rank matrix recovery//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 853-860

[6] Itti L, Koch C, Niebur E, et al. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259

[7] Yang Chuan, Zhang Lihe, Lu Huchuan, et al. Saliency detection via graph-based manifold ranking//Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2013: 3166-3173

[8] Achanta R, Hemami S, Estrada F, Sabine S. Frequency-tuned salient region detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1597-1604

[9] Jiang Peng, Ling Haibin, Yu Jingyi, Peng Jingliang. Salient region detection by UFO: Uniqueness, focusness and objectness //Proceedings of the IEEE International Conference on Computer Vision. Columbus, USA, 2013: 1976-1983

[10] Li Zhaoping, Snowden R J. A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of colour-orientation interference in texture segmentation. Visual Cognition, 2006, 14(4-8): 911-933

[11] Cheng Ming Ming, Mitra N J, Huang Xiaolei, Hu Shi Min. Salient shape: Group saliency in image collections. Visual Computer, 2014, 30(4): 443-453

[12] Margolin R, Tal A, Zelnik-Manor L. What makes a patch distinct//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2013: 1139-1146

[13] Perazzi F, Krahenbuhl P, Pritch Y, Hornung A. Saliency filters: Contrast based filtering for salient region detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 733-740

[14] Yan Qiong, Xu Li, Shi Jianping, Jia Jiaya. Hierarchical saliency detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2013: 1155-1162

[15] Zhang Lihe, Ai Jianwu, Jiang Bowen, et al. Saliency detection via absorbing Markov chain with learnt transition probability. IEEE Transactions on Image Processing, 2018, 27(2): 987-998

[16] Xie Y, Lu H, Yang M H. Bayesian saliency via low and midlevel cues. IEEE Transactions on Image Processing, 2013, 22(5): 1689-1698

[17] Gao Dashan, Vasconcelos N. Discriminant saliency for visual recognition from cluttered scenes. Advances in Neural Information Processing Systems, 2004, 17: 481-488

[18] Bruce N D B, Tsotsos J K. Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, 2009, 9(3): 5.1

[19] Ye T, Zhang D, Gao K, et al. Salient region detection: Integrate both global and local cues//Proceedings of the IEEE International Conference on Multimedia and Expo. Chengdu, China, 2014: 1-6

[20] Xie Saining Tu Zhuowen. Holistically-nested edge detection// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1395-1403

[21] Hou Qibin, Cheng Ming-Ming, Hu Xiaowei, et al. Deeply supervised salient object detection with short connections// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5300-5309

[22] Wang Tiantian, Borji A, Zhang Lihe, et al. A stagewise refinement model for detecting salient objects in images// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4039-4048

[23] Zhang Dingwen, Han Junwei, Zhang Yu. Supervision by fusion: Towards unsupervised learning of deep salient object detector//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 4068-4076

[24] Bruce N D B, Tsotsos J K. Saliency based on information maximization//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2005: 155-162

[25] Li Xiaohui, Lu Huchuan, Zhang Lihe, et al. Saliency detection via dense and sparse reconstruction//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 2976-2983

[26] Jiang Huaizu, Wang Jingdong, Yuan Zejian, Wu Yang. Salient object detection: A discriminative regional feature integration approach. International Journal of Computer Vision, 2014, 9(4): 1-18

[27] Wang Lijun, Lu Huchuan, Ruan Xiang, Yang Ming-Hsuan. Deep networks for saliency detection via local estimation and global search//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3183-3192

[28] Li Guanbin, Yu Yizhou. Visual saliency based on multiscale deep features//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5455-5463

[29] Zhao Rui, Ouyang Wanli, Li Hongsheng, Wang Xiaogang. Saliency detection by multi-context deep learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1265-1274

[30] Lee Gayoung, Tai Yu-Wing, Kim Junmo. Deep saliency with encoded low level distance map and high level features. arXiv preprint, 2016, arXiv:1604.05495

[31] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features//Proceedings of the IEEE Computer Society Conference on Computer Vision Pattern Recognition. Kauai, USA, 2001: 1-9

[32] Lafferty J D, Mccallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the 18th International Conference on Machine Learning. Williamstown, USA, 2001: 282-289

ZHANG Dong-Ming, Ph. D., professor. His research interests include video coding, multimedia content analysis and understanding, and pattern recognition.

JIN Guo-Qing, Ph. D., assistant professor. His research interests include multimedia content analysis and understanding.

DAI Feng, Ph. D., associate professor. His research interest is video coding.

YUAN Qing-Sheng, M. S., senior engineer. His research interest is information security.

BAO Xiu-Guo, Ph. D., professor. His research interest is information security.

ZHANG Yong-Dong, Ph. D., professor. His research interests include multimedia content analysis and video coding.

Background

This work is supported in part by the National Key Research and Development Plan of China under Grant No. 2018YFB0804202, in part by the National Natural Science Foundation of China under Grant No. 61672495, No. 61771458 and No. 61525206.

Saliency detection is considered to be a key attentional mechanism that facilitates learning and survival by enabling organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data. Therefore, saliency detection has been an important preprocessing in many visual detection and recognition tasks.

The difference between the various algorithms is in the way they compute distinctness. Some focus on the patterns, others on the colors, and several add high-level cues and priors. Potential saliency prior is one of the key factors of detecting salient information. Inspired that humans tend to gaze at the center of images, Itti et al. proposals one of the earliest saliency models based on center-surround mechanisms. Other commonly used prior knowledge includes background prior, foreground information and boundary prior. It is commonly observed that human gaze to objects of interest in an image tend to be saliency regions in focus. Therefore, the clue of a well identifiable object is very helpful to locate the salient regions. Although many works have been done to improve the saliency detection gradually, especially the methods based on deep learning, the performances often degrade badly on natural images, since they contain a variety of complex content. We observe that saliency detection method based single visual feature hardly extract salient object that are consistent with human visual system from complex scenes. Although the fusion of various saliency maps is able to compensate or correct the defects of single visual feature, the irrational fusion may further degrade the performance. In order to solve the problem of effective fusion of saliency maps, we propose a deep fusion model based on deep convolutional neural network. Extensive experiments on the typical open datasets show the superiority of the proposed method.

Applied into video retrieval, object recognition etc., this work can greatly reduce the influence due to the background, noise and improve the precision.