软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(8):2527−2544 [doi: 10.13328/j.cnki.jos.005560]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

视频显著性检测研究进展*

丛润民[1], 雷建军[1], 付华柱[2], 王文冠[3], 黄庆明[4], 牛力杰[1]

[1](天津大学 电气自动化与信息工程学院,天津 300072)

[2](Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632, Singapore)

[3](北京理工大学 计算机学院,北京 100081)

[4](中国科学院大学 计算机与控制学院,北京 100190)

通讯作者: 雷建军, E-mail: jjlei@tju.edu.cn

摘 要: 视频显著性检测是计算机视觉领域的一个热点研究方向,其目的在于通过联合空间和时间信息实现视频序列中与运动相关的显著性目标的连续提取.由于视频序列中目标运动模式多样、场景复杂以及存在相机运动等,使得视频显著性检测极具挑战性.对现有的视频显著性检测方法进行梳理,介绍相关实验数据集,并通过实验比较分析现有方法的性能.首先,介绍了基于底层线索的视频显著性检测方法,主要包括5类:基于变换分析的方法、基于稀疏表示的方法、基于信息论的方法、基于视觉先验的方法和其他方法.然后,对基于学习的视频显著性检测方法进行了总结,主要包括传统学习方法和深度学习方法,并着重对后一类方法进行了介绍.随后,介绍了常用的视频显著性检测数据集,给出了4种算法性能评价指标,并在不同数据集上对最新的几种算法进行了定性和定量的比较分析.最后,对视频显著性检测的关键问题进行了总结,并对未来的发展趋势进行展望.

关键词: 视频显著性检测;底层线索;机器学习;深度学习

中图法分类号: TP391

中文引用格式: 丛润民,雷建军,付华柱,王文冠,黄庆明,牛力杰.视频显著性检测研究进展.软件学报,2018,29(8):2527−2544. http://www.jos.org.cn/1000-9825/5560.htm

英文引用格式: Cong RM, Lei JJ, Fu HZ, Wang WG, Huang QM, Niu LJ. Research progress of video saliency detection. Ruan Jian Xue Bao/Journal of Software, 2018,29(8):2527−2544 (in Chinese). http://www.jos.org.cn/1000-9825/5560.htm

Research Progress of Video Saliency Detection

CONG Run-Min[1], LEI Jian-Jun[1], FU Hua-Zhu[2], WANG Wen-Guan[3], HUANG Qing-Ming[4], NIU Li-Jie[1]

[1](School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

[2](Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138632, Singapore)

[3](School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[4](School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: As a hot topic in computer vision community, video saliency detection aims at continuously discovering the motion-related salient objects from the video sequences by considering the spatial and temporal information jointly. Due to the complex backgrounds, diverse motion patterns, and camera motions in video sequences, video saliency detection is a more challenging task than image saliency detection. This paper summarizes the existing methods of video saliency detection, introduces the relevant experimental datasets, and

* 基金项目: 国家自然科学基金(61722112, 61520106002, 61332016, 61620106009, 61602344); 国家重点研发计划(2017YFB1002900)

Foundation item: National Natural Science Foundation of China (61722112, 61520106002, 61332016, 61620106009, 61602344); National Key Research and Development Program of China (2017YFB1002900)

收稿时间: 2017-10-30; 修改时间: 2018-01-04; 采用时间: 2018-01-22; jos 在线出版时间: 2018-02-08

CNKI 网络优先出版: 2018-02-08 11:56:06, http://kns.cnki.net/kcms/detail/11.2560.TP.20180208.1155.012.html

analyze the performance of some state-of-the-art methods on different datasets. First, an introduction of low-level cues based video saliency detection methods including transform analysis based method, sparse representation based method, information theory based method and visual prior based method, is presented. Then, the learning-based video saliency detection methods, which mainly include traditional methods and depth learning based methods, are discussed. Subsequently, the commonly used datasets for video saliency detection are presented, and four evaluation measures are introduced. Moreover, some state-of-the-art methods with qualitative and quantitative comparisons on different datasets are analyzed in experiments. Finally, the key issues of video saliency detection are summarized, and the future development trend is discussed.

Key words: video saliency detection; low-level cue; machine learning; deep learning

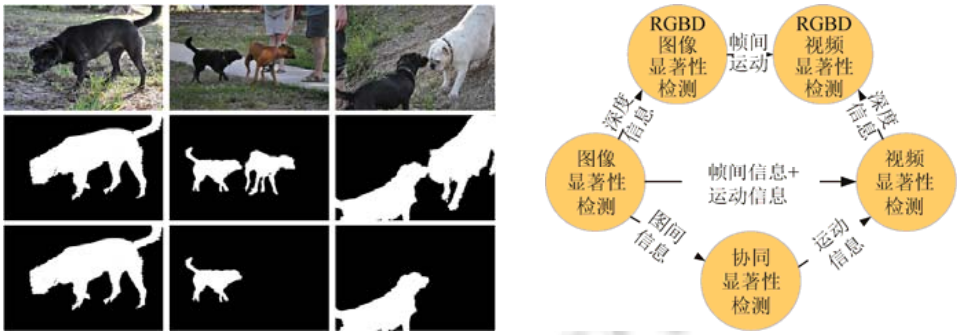
人类通过视觉感知系统捕获客观世界中的重要物体和场景信息,如景深、外貌、颜色、形状等属性.无论置身于简单场景或是复杂环境,人类都可以迅速定位场景中的感兴趣区域,抓住关键信息,快速、有效地完成信息的处理和综合.为使计算机系统也具备快速定位重要目标、感知场景重要信息的功能,视觉显著性检测任务应运而生.场景的显著性区域通常包含了人类感兴趣的重要目标或最能表达图像的内容,是能够在较短时间内吸引人的视觉注意力的区域,而显著性检测就是找出这些感兴趣目标或区域的过程.显著性检测作为一种有效的预处理技术已被广泛应用于检索[1]、识别[2]、分割[3]、重定向[4]、增强[5]、行人检测[6]、评价[7]、压缩[8]等众多计算机视觉任务.

根据处理对象的不同,显著性检测可以分为图像显著性检测、协同显著性检测和视频显著性检测方法等.经过十余年的发展,面向图像的显著性检测方法[9−31]已经形成了较为完善的检测体系,可以分为两大类:一类是由任务驱动的、慢速的、任务依赖的自顶向下(top-down)的检测方法,这类方法往往需要训练过程和特定的先验知识;另一类是由数据驱动的、快速的、下意识的自底向上(bottom-up)的检测方法,这类方法主要利用底层线索(颜色、形状、深度等)直接进行显著性模型构建.此外,随着成像设备的进步与发展,深度信息的获取方式越来越简单、越来越便捷,这为 RGBD 图像显著性检测算法的兴起和发展奠定了基础.相对于 2D 图像显著性检测的飞速发展,RGBD 图像显著性检测算法研究虽然起步较晚,也取得了一定的成果[32−36].但是,研究人员在深度信息对人类感知系统的作用机理、如何有效利用深度信息等方面还未达成共识,仍需进一步深入研究.

协同显著性目标(co-salient object)是指多张图像中重复出现的同一或近似的视觉显著性物体.与传统的图像显著性检测模型不同,协同显著性检测的目的在于提取图像组中共有的显著性目标.由于图像组中显著性目标的类别、内部特性和位置等因素是完全未知的,使得协同显著性检测成为一项更具挑战性的任务.基于此,协同显著性目标需同时具备两个特性:(1) 协同显著性目标在单张图像中应该是显著的;(2) 协同显著性目标在同组图像之间应该具有较高的相似性.协同显著性目标检测方法[37−42]已广泛应用于协同分割、近似目标检测、目标协同识别以及图像简报生成等众多领域.图 1(a)给出了图像显著性检测与协同显著性检测的区别,其中,第 1 行为输入的一组图像,第 2 行为单图显著性检测结果真图,第 3 行为协同显著性检测结果真图.从图中可以看出,如果将每幅图像单独进行显著性检测,那么两只狗都应该被检测出来;如果将 3 张图像看作 1 个图像组进行协同显著性检测,那么应该只有黑色的狗才是共有显著性目标.也就是说,在单一图像中显著的目标不一定为协同显著性目标,还需利用图间约束关系进一步判断,以确定显著性目标是否共有.

随着大数据时代的来临,数据形式发生了翻天覆地的变化,传统的图像数据已不足以满足人们日益增长的感官需求,视频数据量呈现出井喷式的增长,如何准确、一致地提取视频数据中的显著性目标成为亟待解决的新课题.鉴于视频显著性检测技术良好的可扩展性,已被广泛应用于视频目标检测、视频摘要、基于内容的视频检索等领域.不同于图像显著性检测,视频显著性检测需要同时结合时间信息和空间信息,连续地定位视频序列中与运动相关的显著性目标.与协同显著性检测相比,视频显著性检测还需考虑运动信息和时序特性,而且具有“相邻视频帧之间相关性较大”的先验.几种不同的显著性检测模型之间的联系如图 1(b)所示.因此,如何充分挖掘视频序列的运动信息和时序关系成为视频显著性检测研究的关键.由于视频数据量大、场景变化明显、目标大小不一致等问题,使得视频显著性检测研究难度较大,算法性能整体较低.视频显著性检测通常包含两个研究方向,即视频显著性目标检测(video salient object detection)和动态视觉显著性检测(dynamic visual saliency

detection)[43,44].近年来,视频显著性目标检测方向发展迅速,新算法层出不穷,算法性能不断被刷新.因此,本文重点关注视频显著性目标检测的相关研究,并希望通过对相关研究现状的梳理和提炼,为国内外同行提供一个可靠、完整的参考.



(a) (b)

Fig.1 Differences and relations between different saliency detection models

图 1 显著性检测模型的区别与联系

根据是否需要进行训练学习,本文将视频显著性方法分为基于底层线索的方法和基于学习的方法两类.其中,基于底层线索的视频显著性检测方法可以进一步划分为基于变换分析的方法、基于信息论的方法、基于稀疏表示的方法、基于视觉先验的方法和其他方法 5 类,而基于学习的方法可以分为传统学习方法和深度学习方法两类,具体分类方案如图 2 所示.本文第 1 节对基于底层线索的视频显著性检测方法进行介绍.第 2 节讨论基于学习的视频显著性检测方法.第 3 节介绍视频显著性检测常用的数据集、定量评价指标以及相关方法的对比实验.最后对视频显著性检测的关键问题进行总结,并对未来可能的研究方向进行展望.

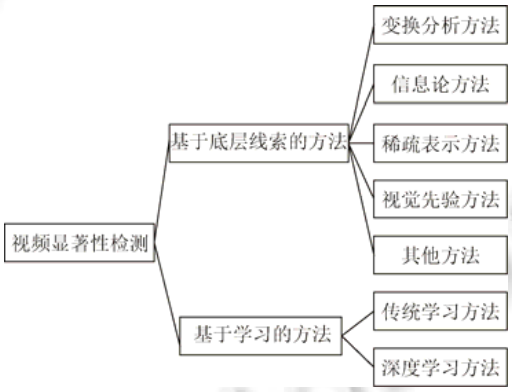


Fig.2 Classification chart of video saliency detection algorithms

图 2 视频显著性检测方法分类图

1 基于底层线索的视频显著性检测方法

从开始研究至今,基于底层线索的方法一直是视频显著性检测领域的主流方法,该类方法从底层线索出发,提取视频的运动特征,探索视频的视觉先验信息,挖掘视频的帧间关系,并结合视频的空时信息,建立显著性检测模型.该类方法不需要进行训练学习,操作简单、方便,是一类基础的检测方法.但是,由于运动场景的更新、目标尺寸的变化以及拍摄视角的切换,使得该类方法的检测准确率远远没有达到理想的要求,有待进一步的提高和完善.根据算法采用技术的不同,该类方法又可以进一步划分为基于变换分析的方法、基于稀疏表示的方

法、基于信息论的方法、基于视觉先验的方法和其他方法 5 类,下面将具体展开介绍.

1.1 基于变换分析的方法

基于变换分析的方法通过数学变换提取视频序列的有用信息,进而实现显著性检测.常用的数学变换有傅里叶变换、离散余弦变换等.在介绍具体方法之前,我们首先回顾一下两种主要变换方法的数学模型.

傅里叶变换(Fourier transform,简称 FT)是一种非常重要的数学分析工具和信号处理方法.图像或视频帧可以看作是二维的离散信号,因而,可以对其进行二维离散傅里叶变换,公式如下:

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\mathrm{e}^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{1}$$

其中,$f(x,y)$表示图像或视频帧的像素值;M,N 对应图像的长和宽.

离散余弦变换(discrete cosine transform,简称 DCT)是一种与 FT 相关的变换,它通过一组不同频率和幅值的余弦函数来近似一幅图像,其本质上就是傅里叶变换的实数部分.同样,如果将图像或视频帧看作是二维信号,那么其二维离散余弦变换的公式可以表示为

$$F(u,v)=\alpha(u)\alpha(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\cos\frac{(2x+1)u\pi}{2M}\cos\frac{(2y+1)v\pi}{2N} \tag{2}$$

其中,$\alpha(u)=\begin{cases}\frac{1}{\sqrt{M}},u=0\\ \sqrt{\frac{2}{M}},u\neq 0\end{cases}$,$\alpha(v)=\begin{cases}\frac{1}{\sqrt{N}},v=0\\ \sqrt{\frac{2}{N}},v\neq 0\end{cases}$,$M$ 和 N 对应图像的长和宽.

Hou 等人[14]提出了一种基于傅里叶变换的简单、快速的显著性检测方法,称为谱残差(spectral residual,简称 SR).该方法以谱残差的幅度谱来度量图像的显著性值,获得了较好的检测结果.经过实验分析,Guo 等人[45,46]发现傅里叶变换的相位谱(phase spectrum of Fourier transform,简称 PFT)可以获得更好的显著性检测结果,而且降低了算法的计算复杂度.将每个像素的值表示为由强度、颜色和运动特征组成的四元数,进而利用四元数傅里叶变换的相位谱(phase spectrum of quaternion Fourier transform,简称 PQFT)计算视频序列的时空显著性.实验结果表明,该方法获得了较好的检测性能,而且对白噪声具有较高的鲁棒性,满足实时处理要求,特别适合于工程应用.

Cui 等人[47]将 SR 方法扩展至视频领域,提出了一种快速的运动显著性检测方法——时间频谱残差(temporal spectral residual).首先,分别在视频片段的 X-T 和 Y-T 平面上利用时间频谱残差自动地从背景中分离显著性目标.然后,利用阈值选择机制对噪声进行抑制.最后,利用投票机制对实验结果进行修正.与传统的复杂背景建模不同,该方法仅基于傅里叶频谱分析,具有较好的实时性.

Fang 等人[48]在压缩域中提出了一种基于特征对比的视频显著性检测模型.首先,对于未预测视频帧(I 帧)进行离散余弦变换得到变换系数.然后,利用 DCT 系数提取视频帧的亮度、颜色和纹理特征,并通过特征对比方法计算得到静态显著性图.同时,提取视频预测帧(P 帧和 B 帧)的运动向量并计算得到运动显著性图.最后,通过加权融合方法得到视频显著性检测结果.该方法在压缩域中进行处理,可以更加方便地应用于基于网络的多媒体领域,如视频重定向、视频质量评价等.

受静态图像的谱残差显著性计算方法启发,刘宇光等人[49]提出了一种基于运动谱残差的视频显著性检测方法.该方法将视频序列相邻两帧所产生的运动矢量场分解为创新部分和先验部分,其表达式如下:

$$\boldsymbol{MV}(frame)=\boldsymbol{MV}(innovation)+\boldsymbol{MV}(prior) \tag{3}$$

其中,$\boldsymbol{MV}(innovation)$表示创新部分,对应运动显著性区域,$\boldsymbol{MV}(prior)$表示先验部分,对应静止的背景或场景的整体运动.研究发现,运动矢量谱残差反映了视频中运动反差强烈的区域,即视频中的显著性区域.因此,该方法首先利用光流法计算视频的运动矢量,其次分别计算水平运动矢量和垂直运动矢量的谱残差.然后,通过加权融合两个分量的运动谱残差.最后,通过顶帽变换对结果进行增强,得到最终的视频显著性检测结果.

基于变换的视频显著性检测方法得到的结果仅能确定视频序列中显著性目标的大体位置和主要轮廓,而

显著性目标的内部区域的均匀性和完整性较差,通常只能应用于检测、跟踪等领域.由于该类算法仅利用了一些简单的变换关系获取显著性信息,所以算法的运算速度较快,非常适合于大型的实时系统.

1.2 基于信息论的方法

基于信息论的方法通常利用自信息、信息熵、条件熵等概念研究信息量与显著性的关系,进而确定显著性目标.在日常生活中,极少发生的事件一旦发生就容易引起人们的关注,因而包含较多信息量.也就是说,小概率事件所包含的信息量多.相反,人们习以为常的事件包含的信息量相对较少.因此,事件信息量与其发生的概率成反比.一个事件 x 的信息量 $I(x)$可以表示为

$$I(x)=-\log\big(p(x)\big) \tag{4}$$

其中,$p(x)$表示事件 x 出现的频率.

此外,可以利用信息熵来描述信源的不确定度,其计算公式如下:

$$H(X)=-\sum_{x\in\aleph}p(x)\log\big(p(x)\big) \tag{5}$$

Qiu 等人[50]将信息论理论引入空时显著性检测中,建立了信息论与视觉显著性之间的关系.根据香农定理可知,如果一个事件是独特的,那么它将包含较多的信息量.相反,如果一个事件经常发生,那么它将包含较少的信息量.为了测量空时事件的独特性,该方法联合空间和时间的条件概率分布对时空事件进行建模,并计算得到空时显著性值.

Hou 等人[43]提出了一种基于特征稀有性的动态视觉注意力模型.该模型利用增量编码长度(incremental coding length,简称 ICL)来测量每个特征的透视熵增益,并以最大化采样视觉特征的熵作为目标.为了优化能量损耗,该模型根据其增量编码长度实现特征之间重新分配.通过选择具有较大增量编码长度的特征,模型可以实现静态和动态场景的显著性区域选择.

Liu 等人[51]将基于信息论计算得到空时显著性结果用于目标的运动检测.目标的显著性由空时视频卷(spatio-temporal volumes)的信息显著性图表示.首先,将视频序列进行划分,得到空间向量集合和时间向量集合,进而利用信息论方法计算得到空间显著性结果和时间显著性结果.然后,利用动态融合机制将两种显著性图融合得到空时显著性结果.在计算信息显著性图过程中,算法通过降维和核密度估计(kernel density estimation,简称 KDE)方法提取输入视频数据的信息量.

Li 等人[52]将信息论中的条件熵引入视觉显著性计算.人类视觉感知过程的一个基本原理是抑制频繁出现的输入响应,同时保持对新输入的敏感响应.因此,可以将中心区域的显著性定义为其周围局部区域的不确定性.为了描述这种不确定性,该方法引入了信息论中的条件熵.给定中心区域的周围区域,显著性被定义为考虑感知失真(perceptional distortion)的局部区域最小化条件熵(minimum conditional entropy).为了简化问题,该方法假设数据满足多元高斯分布,进而利用有损编码长度近似估计条件熵.最后,显著性图由像素显著性值的累积得到.该方法对帧间运动和仿射变换都不敏感,且不需要先验知识和预训练过程,同时适用于图像和视频数据.

1.3 基于稀疏表示的方法

基于稀疏表示的方法以稀疏表示为基础,通过挖掘显著性区域的稀疏系数特性计算显著性图.稀疏表示是一种有效的信号处理技术,其目的在于利用给定的超完备字典中尽可能少的非零系数来表示信号的主要信息,获得更简洁的信号表达方式,进而更简便地提取信号中所包含的信息.鉴于其优异的数据表达能力,该技术已广泛应用于分类、追踪、检测等领域.稀疏表示模型的一般形式如下:

$$\alpha^{*}=\arg\min_{\alpha}\|\boldsymbol{x}-\boldsymbol{D}\cdot\alpha\|_{k}+\lambda\|\alpha\| \tag{6}$$

其中,$\boldsymbol{x}$ 为观测数据向量,$\boldsymbol{D}$ 为字典,α为稀疏系数,λ为正则参数,k 为稀疏度量.通过求解公式 6 的能量函数,可以得到最优的稀疏系数向量,获得信号的稀疏表达形式.稀疏表示已成功应用于图像显著性检测中[53−55],其基本思想为:如果利用背景字典对图像区域进行稀疏重建,那么图像背景区域的重建误差将会较小,而前景显著性区域的重建误差则较大,因而可以利用重建误差来表征图像的显著性大小.若采用背景字典对图像区域进行稀疏重建,

那么区域的显著性值与稀疏重建误差成正比;如果采用前景字典对图像区域进行稀疏重建,那么区域的显著性值与稀疏重建误差成反比.具体表达式如下:

$$\begin{cases} S=\varepsilon_B=\left\|\boldsymbol{x}-\boldsymbol{D}_B\alpha^*\right\|_2^2 \\ S=\exp\left(-\beta\cdot\varepsilon_F\right)=\exp\left(-\beta\cdot\left\|\boldsymbol{x}-\boldsymbol{D}_F\alpha^*\right\|_2^2\right) \end{cases} \tag{7}$$

其中,S表示图像的显著性值,ε_B表示利用背景字典$\boldsymbol{D}_B$计算得到的重建误差,ε_F表示利用前景字典$\boldsymbol{D}_F$计算得到的重建误差,β为控制比例的加权系数,通常可设置为 10.

Li 等人[56]提出了一种可同时应用于图像和视频数据的基于稀疏表示的中心-周围显著性检测模型.首先将图像划分为多组中心块与其邻域块的组合,然后利用邻域块对中心块进行稀疏表示求得最优稀疏系数,最后利用增量编码长度作为显著性值的度量参数,得到显著性图.

Luo 等人[57]提出了一种空时约束下的基于特征选择的视频显著性估计模型.对于给定的视频,首先,利用视频片段的随机采样块去训练学习一个过完备字典,得到每个视频帧的良好表达.然后,利用该字典对每帧视频的进行稀疏表示,并根据其熵增益选择稀疏特征,构建视频帧的空间显著性图.最后,将所有选择的特征响应进行加权求和后得到最终的视频显著性图.值得一提的是,该方法在特征选择过程中引入了时间一致性和时间差异性.一方面,选择在连续视频帧中对信息增益有重复贡献的特征,并赋予其较高的权重,即时间一致性.另一方面,选择导致场景变化的特征,即时间差异性.

Ren 等人在稀疏表示的视频显著性应用研究中做了多项十分有意义的工作.文献[58]提出了一种基于鲁棒的时域对齐和局部-全局空间对比的视频显著性目标检测算法.经观察发现,连续视频帧中的显著性运动是稀疏的,而在变换域中的视频帧的背景是低秩.因此,视频帧可以通过仿射变换进行转化和对齐,并将其分解为一个对应对齐背景的低秩矩阵和一个表示显著性运动的稀疏矩阵.该方法利用稀疏分解和低秩分解同时估计显著的前景运动和相机运动,得到时间显著性结果.在空间显著性计算中,将全局信息引入中心-周围对比模型,并增加了高斯先验分布来修正区域的显著性值.最后,将空间显著性结果和时间显著性结果融合得到视频显著性图.文献[59,60]提出了一种基于稀疏表示的视频显著性检测模型.对于时间显著性,将目标块的运动建模为稀疏重建过程,利用相邻帧中的重叠块来重构目标块,并引入平滑项,以学习连贯的运动轨迹.因此,时间显著性模型结合了重建误差、稀疏正则化和局部轨迹对比度来测量显著性.对于空间显著性,利用重建误差和稀疏正则化来捕获每帧视频中具有高中心-周围对比度的区域.最后,通过对空间显著性图和时间显著性图进行乘积融合得到视频显著性结果.

不同于图像的显著性检测,视频中的移动目标比静态目标更容易引起人类的视觉关注.基于这一观察,Xue 等人[61]提出了一种基于低秩分解和稀疏表示的运动显著性检测方法.该方法利用 *X-T* 和 *Y-T* 平面上的视频片段的低秩分解和稀疏表示来实现前景移动目标与背景的分离.此外,该方法引入空间信息来保持检测到的运动对象的完整性.最后,通过自适应阈值选择和噪声消除技术进行修正得到最终结果.该方法无需任何背景建模过程,而且可以适用于不同的视频场景,并对低分辨率和含噪等情况鲁棒.

Chen 等人[62]提出了一种基于空时融合和低秩一致性扩散的视频显著性检测方法.首先,将原始视频划分多个短视频组,并利用局部对比方法分别计算颜色显著性和运动显著性.然后,利用上一个视频组的显著性结果生成的前景/背景模型以及运动显著性图来调整颜色显著性结果.考虑视频帧间的一致平滑特性,利用前后两帧的显著性结果更新当前帧的显著性结果,得到更新后的颜色显著性图和运动显著性图,并通过点乘融合得到初始空时显著性图.最后,为了保证显著性图的时间一致性,进一步提高检测的准确性,利用低秩一致性扩散策略对初始显著性结果进行优化,得到最终的空时显著性图.

1.4 基于视觉先验的方法

受人类视觉感知系统启发,许多视觉先验信息被应用于检测图像中的显著性目标,如局部对比先验、全局对比先验、目标先验、背景先验、中心先验、紧致性先验、稀有性先验、独特性先验等.这些视觉先验信息

符合人类的视觉感知机制,是十分有效、便捷的目标描述方式.目前,许多研究学者已将这些先验信息扩展至视频显著性检测领域,并进一步挖掘了视频数据特有的先验信息,如运动先验等.

人类具有无与伦比的快速定位复杂动态环境中重要事件的能力.研究发现,在视觉注意力机制的快速指导下,人们往往没有时间进行详细的视觉分析.因此,Itti 等人[63]认为启发式计算在准确、实时定位感兴趣事件时是十分必要的.他们提出了一种新的"感官惊喜"理论(theory of sensory surprise),为重要信息的获取提供了原理上的可计算方法.感官惊喜模型通过计算特征差异得到空时显著性结果,其中特征主要包括亮度、颜色、方向、闪烁和运动.

Seo 等人[44]提出了一种基于自相似性的空时显著性计算模型.对于给定的图像或视频,将局部回归核(local regression kernel)作为局部描述算子,用以测量像素与其周围邻域的相似度.然后,利用余弦相似性的泛化形式——矩阵余弦相似度描述其自相似性得到显著性值.该框架生成的显著性图中的每个像素值描述了给定周围特征矩阵情况下的特征矩阵的统计似然性,且取得了良好的检测效果,不仅可以自动定位给定视频中显著性移动目标,而且对噪声和其他系统扰动具有较好的鲁棒性.

受基于运动的感知分组的生物学机理启发,Mahadevan 等人[64]提出了一种基于中心-周围框架的无监督空时显著性检测算法.该算法将图像显著性检测中的中心-周围先验扩展至视频领域,并将视频区域块的显著性定义为预设特征集的能量强度.该扩展以空时视频块为基础,并将其建模为视频的动态纹理模型.视频域中心-周围先验与动态纹理建模的结合产生了强大的、多功能的、完全无监督的空时显著性检测算法,适用于高动态背景和移动摄像机等复杂场景.

许多视频显著性检测方法容易偏向于边缘或角落区域,这在统计学上具有一定的意义,但并不符合人类的视觉感知机制.此外,由于显著性区域和高度纹理结构背景之间的模糊性,使得现有方法往往不能在复杂场景中准确定位显著性区域.基于此,Kim 等人[65]提出了一种基于纹理对比的空时显著性检测模型.该方法的关键思想是通过利用视网膜和视觉皮层中捕获的两种对比信息来近似模拟自底向上的视觉注意的生物学机制.因此,他们利用亮度对比和方向连贯性对比来定义纹理对比模型,并将其扩展至具有时间梯度的多尺度时空域中,生成了可靠、鲁棒的显著性检测结果.值得一提的是,该方法即使在杂乱背景下也可以有效地检测图像和视频中的显著性区域.Kim 等人[66]还提出了一种基于空时方向一致性对比的视频显著性检测框架.该方法利用 3D 结构张量来分析视频序列空时梯度的分布,并结合空时方向一致性对比来确定不同视频中的显著性区域.对于各种不同环境下拍摄的视频,包括背景中的存在动态照明和纹理的情况,该方法都表现出了较好的性能.

Zhou 等人[67]提出了一种保持原有视频中重要时刻的将高帧率(high-frame-rate,简称 HFR)视频转换为常速、低帧率(low-frame-rate,简称 LFR)视频的方法.他们设计了一种鲁棒的空时显著性计算方法来评估视觉重要性.首先,将视频分割为多个颜色一致性较好的空时区域.然后,提取每个区域的颜色和运动特征,利用中心-周围对比先验计算得到特征对比图,并考虑区域的位置、速度、加速度和前景概率,计算区域的局部先验值.最后,将特征对比和局部先验在多尺度下进行融合得到区域的显著性度量.

以文献[44]为基础,Le 等人[68]结合底层特征和中层特征,提出了一种基于区域的多尺度空时显著性检测方法.首先,对每帧视频进行多尺度分割,得到不同尺度下的多个区域.然后,提取每个区域的底层特征和中层特征,其中底层特征主要包括颜色、强度、方向、光流方向和光流幅度,中层特征主要包括中心先验、目标先验、背景先验和运动先验.进而通过融合底层特征图和中层特征图得到每个尺度下的空间显著性结果.随后,将运动信息引入自适应时间窗口,并与对应尺度下的空间显著性结果融合得到多尺度的空时显著性图.最后,进行多尺度融合得到视频显著性检测结果.

Xi 等人[69]将图像显著性检测中的背景先验扩展至视频领域,提出了一种基于空时背景先验的视频显著性目标检测算法.首先,对每帧视频进行超像素分割,并利用 SIFT 流估计获取运动信息.然后,分别从时间和空间角度提取背景先验信息,融合后得到最终的空时背景先验.对于空间背景先验信息,仍采用传统的图像处理方法.对于时间背景先验,他们采用了能量函数优化方法,其主要流程如下.

- 根据"背景部分在视频中是刚性的,前景和背景具有相反的运动轨迹,且背景区域往往要比前景区域更大

一些”的先验知识,通过分析 SIFT 流确定初始背景先验.即不同视频帧中的同一位置像素的误差越小,则说明该像素是背景的可能性越大.

• 利用多对约束和一致性传播确定背景区域.由于相邻两帧视频的时间间隔较短,那么显著性目标可能只有一部分发生了明显的运动,而静止部分很可能会被当作背景种子.因此,作者通过累积多帧信息,提取一个中间状态,使得最终的背景选择结果尽可能地与这个中间状态接近,该过程称为“多对约束”机制;此外,为了完整、连续地提取背景区域,Xi 等人设计了“一致性传播”机制,要求已选背景区域的前后相邻帧中的对应区域也应该被选择为背景.

通过空时背景先验确定出背景种子后,首先,构建空间图模型,利用测地线距离计算得到空间显著性图.然后,对空间显著性图进行阈值分割得到前景种子点,在时间图模型上利用流形排序方法得到时间显著性图.最后,将空间显著性图和时间显著性图取平均后得到最终的视频显著性检测结果.

1.5 其他方法

除了上述介绍的几类方法外,其他一些图像处理技术如超像素分割、随机游走、能量函数优化等也被应用于视频显著性检测中,接下来将重点介绍几种典型算法.

Liu 等人[70]以超像素为基本处理单元,通过融合时间显著性图和空间显著性图得到了视频显著性结果.该方法主体在超像素级上进行计算,并辅助以帧级的单帧数据作为全局特征.首先,提取超像素级和帧级的运动直方图特征和颜色直方图特征.然后,利用这些特征计算超像素级的空间和时间显著性图.空间显著性通过全局对比特性和空间稀疏特性计算得到.时间显著性则同时结合了运动独特性和时间相关性.考虑到显著性超像素的运动直方图特征应该与帧级的全局运动直方图特征不同,他们将运动独特性表示为视频帧的全局差异性.此外,考虑到相邻帧的显著性目标的显著性值应该具有较高的一致性,他们利用上一帧的匹配超像素及其邻域的时间显著性值对当前帧的某个超像素的时间显著性值进行约束,以此保证其时间相关性.最后,利用一种自适应融合方法将空间和时间显著性图融合得到最终的结果.

Wang 等人[71]提出了一种基于局部梯度流估计和全局修正的视频显著性检测算法.研究表明,目标的边界特性和不连续性揭示了视频帧的重要内容.然而,对于某些复杂场景来说,颜色不连续特性不足以完全表示场景特性,还需要利用运动信息进行补充,比如在光流场中变化剧烈的像素更容易引起人们的注意等.因此,该方法首先结合颜色梯度幅度和运动梯度幅度计算每帧视频的空时梯度值,并利用梯度流场(gradient flow field)测度确定虚拟背景区域.然后,根据局部显著性线索和全局显著性线索得到初始显著性图.其中,局部显著性线索表示为区域与其周围背景区域的差异性,全局显著性线索定义为区域到虚拟背景点的最短距离.为了获得更加平滑、空时一致性更好的检测结果,他们利用能量函数进行空时显著性优化,能量函数主要包括数据项和平滑项.该方法可以在少量的外貌和运动模式约束下自动地定位视频中的前景区域.Wang 等人[72,73]提出了一种基于测地线距离的无监督视频显著性对象分割方法,并将视频显著性结果作为对象分割的先验.首先,将视频帧划分为多个超像素区域,并将颜色图像的空间边缘概率图和光流梯度幅度图融合得到空时边缘图.然后,构建帧内图模型,利用测地线距离度量帧内显著性,通过自适应阈值分割方法获得视频帧的初始显著性区域.随后,结合初始显著性目标区域,考虑不同帧的时域邻接关系,构建帧间图模型,并基于最短测地线距离计算空时显著性图.最后,将空时显著性结果、外貌模型和动态位置模型融入能量最小化框架,实现视频显著性目标分割.

Kim 等人[74]利用带重启过程的随机游走(random walk with restart,简称 RWR)框架实现了视频序列的空时显著性检测.首先,利用运动独特性、时间一致性和陡然变化特性等约束计算时间显著性分布,并将其作为随机游走的重启分布.然后,利用强度、颜色和紧致性等空间特征构造转移概率矩阵.最后,将 RWR 模型归一化后的稳态分布作为最终的空时显著性结果.时间显著性计算过程主要考虑了以下 3 个特征.

• 运动独特性(motion distinctiveness):将运动轮廓作为运动特征,并以此构建运动图模型,利用 Tanimoto 测度计算边权重,并通过传统的随机游走模型计算稳态分布得到运动特征图.

• 时间一致性(temporal consistency):利用前一帧视频的空时显著性结果计算当前帧的时间一致性,将当前帧某个块的时间显著性定义为在前一帧视频中的对应位置周围所有块中最匹配块内的所有像素的空时显著性

的平均值.

• 突变特性(abrupt change):主要用于解决突然出现新目标的问题,如果当前帧的某个块与前一帧中最匹配块的相似性小于特定阈值,则将其定义为突变块.

Liu 等人[75]提出了一种基于超像素和空时传播的视频显著性检测方法.首先,对视频帧进行超像素分割和光流估计,提取超像素的运动直方图和外貌直方图.然后,构建包含一个虚拟背景节点的图模型,利用最短路径算法计算运动显著性图.值得注意的是,他们以一种迭代的方式进行运动显著性计算,直到两次迭代的全局运动直方图特征的欧氏距离小于特定阈值才中止.紧接着,进行时域传播获取视频帧间信息,利用当前视频帧的前后 5 帧数据的显著性图对当前帧显著性结果进行更新,得到引入时间信息的显著性图.最后,在单个视频帧内进行空间传播,先后计算局部显著性和全局显著性,得到最终的视频显著性结果.

基于生物视觉特征和视觉信息学,方志明等人[76]提出了一种融合时间和空间显著性的视频显著性检测算法.受视觉皮层层次化感知特性和 Gestalt 视觉心理学的启发,他们提出了一种层次递进的空间显著性计算方法.在低层,提取视频帧的特征,确定多个候选显著性区域.在中层,利用矩阵的最小 F-范数特性选取竞争力最强的局部显著性区域.在高层,基于 Gestalt 理论进行视觉整体感知,对局部区域进行整合,得到空间显著性图.在时间显著性图计算中,结合光流点的位置、运动方向和运动幅值特征,对光流点进行二分类,排除噪声干扰,并利用运动幅度定义运动显著性值.最后,将单帧显著性结果在灰度颜色空间进行表示,将运动显著性图在 Munsell 颜色空间进行表示,提出了一种通用的、基于视觉敏感度的显著性可视化的表示方法.

2 基于学习的视频显著性检测方法

除了上述基于底层线索的视频显著性检测方法以外,基于学习的检测方法也受到了研究学者的广泛关注.特别是随着深度学习技术的发展和成熟,已有多项工作利用深度学习实现了视频的显著性检测,大幅度地提高了算法的性能.本节将介绍几种典型的基于学习的视频显著性检测方法,并着重讨论两种利用深度学习实现视频显著性检测的方法.

2.1 传统学习方法

Liu 等人[77]提出了通过有监督的学习来检测图像或连续图像中的显著性目标.首先,利用条件随机场(condition random field,简称 CRF)对显著性检测问题进行建模,该模型融合了基于 CRF 学习得到的一组显著性特征,并将分割结果也引入 CRF 模型中用以检测未知大小和形状的显著性对象.该方法提出了一组新颖的特征来分别描述显著性目标的局部、区域和全局特性,即多尺度对比、中心-周围直方图和颜色空间分布.此外,通过引入动态显著性特征将该模型扩展至视频领域.

通过主摄像机运动去除(dominant camera motion removal)技术,Huang[78]等人提出一种基于轨迹的视频显著性检测方法.该方法适用于任何固定或移动相机拍摄的视频,无需任何先验信息.首先,提取了视频中关键点的一组空间和时间相干的轨迹.然后,利用速度熵和加速度熵来对轨迹进行描述.这样,就可以利用长期的物体运动来滤除短期噪声,并以相同的方式表示各种时间长度的目标运动.此外,无论相机是否静止,背景区域的轨迹(即非显著性轨迹)往往都与主相机运动相一致.因此,他们提出了一种可以同时处理静止和运动相机的显著性计算方法,利用单类支持向量机(one-class SVM,简称 OCSVM)对轨迹进行分类并消除运动中的一致性轨迹,进而将每个轨迹的显著性扩散到其周围区域,产生空时显著性结果.

2.2 深度学习方法

近年来,深度学习技术蓬勃发展,已被广泛应用于诸如分类、检测、识别、检索、语音处理等多个领域,受到了学术界和工业界的广泛关注.目前,常用的深度学习网络有:AlexNet 网络、VGG 网络、GoogleNet 网络、ResNet 网络、全卷积网络(fully convolutional network,简称 FCN)、反卷积网络(deconvolution network,简称 DN)等.本小节将重点介绍两种基于深度学习的视频显著性检测算法.

Wang 等人[79]提出了一种基于全卷积网络的视频显著性目标检测算法.该算法是目前唯一正式发表的基于

深度学习的视频显著性算法.该算法重点解决了两个技术问题:(1) 深度视频显著性模型在训练过程中缺乏足够多的像素级的视频显著性标注数据;(2) 快速的视频显著性训练和检测.图 3 给出了算法的原理框图,可以看出,该方法主要由两个网络模型组成,即用于捕获空间信息的静态显著性网络和用于捕获时间信息的动态显著性网络.值得注意的是,动态显著性网络的输入包括7通道数据,分别为当前视频帧的RGB数据(3通道)、下一视频帧的RGB数据(3通道)和当前帧图像的静态显著性图(1通道).这样的操作可以直接在动态网络部分输出空时显著性结果,而不需要耗时计算光流信息,节省了大量的训练和检测时间.针对标注数据量不足的问题,该方法提出了一种新的数据增强方法,它根据现有的大量带标注的图像数据,仿真生成了视频训练数据,这使得算法网络能够学习多种显著性信息,防止了利用有限训练数据导致的过拟合问题.借助于大量的训练数据(包括 150K 的合成视频序列和真实视频数据),网络充分学习到了空间和时间的显著性线索,可获得准确的空时显著性估计结果.

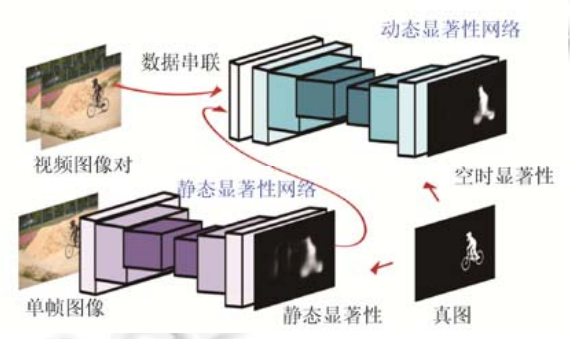


Fig.3 Flowchart of the algorithms in Ref.[79]

图 3 文献[79]算法流程图

另外一篇基于深度学习的视频显著性检测算法仅为预出版版本,由 Le 等人[80]提出,刊登于 arXiv 网站上.结合视频帧的局部和全局上下文信息,提出了一种基于空时深度(spatiotemporal deep,简称 STD)特征的视频显著性检测方法,其原理框图如图 4 所示.首先,将视频帧划分为多个尺度下的超像素区域,并提取 STD 特征.STD 特征由局部特征和全局特征堆叠而成,提取过程通过两个网络实现,一个是基于区域的卷积神经网络(region-based convolutional neural network,简称 region-based CNN),用以提取每帧视频的区域特征,经过时间一致性合并后可以得到局部特征;另一个是基于块的卷积神经网络(block-based CNN),用以提取视频序列的全局特征.然后,结合 STD 特征,利用空时条件随机场(spatiotemporal conditional random field,简称 STCRF)模型计算得到每个尺度下显著性值.STCRF 是 CRF 的时域扩展,同时考虑了帧内和帧间的邻域关系.STCRF 模型生成的显著性图具有较好的时间一致性,可以精确地检测显著性目标的边界和抑制噪声干扰.最后,将多尺度下得到的显著性图进行融合得到最终的空时显著性结果.相对于文献[76],该方法获得了更加优异的性能.

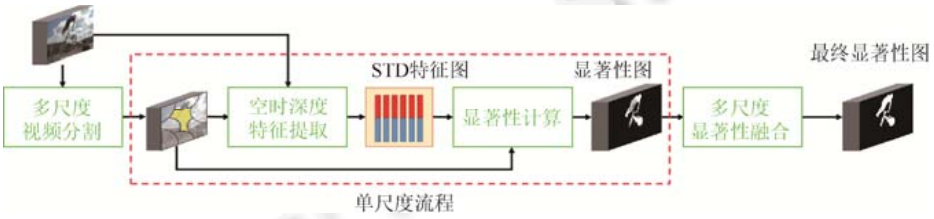


Fig.4 Flowchart of the algorithms in Ref.[80]

图 4 文献[80]算法流程图

3 实验与分析

本节将从实验的角度对几种最新的视频检测算法进行比较分析,首先对常用的视频显著性检测数据库进行介绍,然后给出算法性能的评价指标,最后在不同数据库上比较和分析几种最新的视频检测算法的性能.

3.1 实验数据集

为了科学、有效地评价各类视频显著性检测算法的性能,提出了许多标准的视频数据集供研究人员对算法进行测试和对比,常用的数据集主要有:

(1) SegTrack 数据集

SegTrack 数据集包含两个版本.2010 年,Tsai 等人建立了 SegTrackV1 数据集[81],起初用于目标跟踪实验,后被广泛用于视频分割、显著性检测等任务.该数据集包括 6 个视频,分别为鸟下落、猎豹、女孩、猴子与狗、跳伞和企鹅,视频的每一帧都标注了像素级的真图(ground truth).由于企鹅视频的真图不可用,因此通常利用其它 5 个视频进行算法实验.需要说明的是,该数据集中的鸟下落视频和跳伞视频的显著性目标都较小,而且跳伞视频还受到了较大强光干扰,这些无疑都增加了检测的难度.2013 年,Li 等人对 SegTrackV1 数据集进行了扩展,形成了包含 14 个视频序列的 SegTrackV2 数据集[82],囊括了多种场景和运动活动.

(2) ViSal 数据集

2015 年,北京理工大学组建了包括 17 个极具挑战性视频序列的 ViSal 数据集[71].该数据集中多数视频的分辨率为 320×240,每个视频的长度范围为 30 帧~100 帧.在数据集中,每隔 5 帧按照给定的类别手动标注了像素级的视频显著性真图.该数据集是一个难度较大的视频显著性检测数据集,主要表现在以下几个方面:复杂的颜色分布(如摩托车、牛等视频序列)、高度杂乱的背景(如男人、熊猫等视频序列)、多种目标运动模式(如慢速的船视频序列、快速的汽车视频序列)、快速的拓扑变化(如猫、摩托车等视频序列)以及存在相机运动(如摩托车等视频序列)等.

(3) MCL 数据集

2015 年,高丽大学组建了 MCL 数据集[74],其由 9 个分辨率为 480×270 的视频序列组成,每个视频序列包含 100 帧~400 帧数据,涉及室内和室外场景,且多数视频序列中包含多个快速运动的目标.而且,有 3 个视频序列(球、球场和大厅视频)还存在相机运动.该数据集每隔 8 帧数据给出了视频显著性目标的像素级标注.由于视频场景复杂、背景杂乱、纹理丰富、运动类型多样,使得该数据集十分具有挑战性.

(4) DAVIS 数据集

2016 年,苏黎世联邦理工学院公开了一个稠密标记的视频目标分割数据集——DAVIS 数据集[83].该数据集由 50 个高质量的视频序列组成,包括 480 p 版本和 1 080 p 版本,且对每帧视频进行了像素级的真图标注.该数据集是目前已知最大的带像素级标注的视频分割数据集.视频序列包含多种具有挑战性的情况,如遮挡、运动模糊、外貌变化等.该数据也被广泛用于对视频显著性检测算法进行评估.

(5) UVSD 数据集

2017 年,上海大学公布了一个新的视频显著性检测数据集——UVSD 数据集[75],包含 18 个视频,并对每帧视频数据进行了像素级的标注.该数据集中的视频分辨率以 320×240 为主,每个视频序列包含了 70 帧~300 帧数据.由于视频中的显著性目标较小、前景背景对比度低、背景杂乱、存在相机运动和视角变化等,使得该数据集难度较大.

3.2 评价指标

为了验证算法的有效性,除了直观的与真图进行视觉对比外,还需要利用评价指标定量分析算法的性能.本节将介绍 4 种常用的评价指标.

(1) 准确率-召回率

通过对比二值显著性图和真图,可以计算出准确率(precision)和召回率(recall).二值显著性图采用对显著性图进行固定阈值分割的方式得到.像素的显著性值范围在[0,255]之间变化,将分割阈值依次从 0 变化到 255,大于

等于阈值的像素值置1,小于阈值的像素值置0,进而生成256张二值显著性图.将每张二值显著性图与真图比较,就可以计算出每个阈值下的准确率和召回率:

$$\begin{cases} Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN} \end{cases} \tag{8}$$

其中,TP表示二值显著性图中的前景区域与真图的前景区域的重合像素个数,FP表示二值显著性图中的前景区域但对应于真图中的背景区域的像素个数,FN表示二值显著性图中背景区域但对应于真图中的前景区域的像素个数.准确率和召回率相互制约,较高的准确率说明有较多的显著性区域被正确检测为前景区域,而较高的召回率则说明检测到的显著性区域能尽可能覆盖真图中的前景区域,完整性好.以召回率为横轴,准确率为纵轴,可以绘制准确率-召回率曲线(PR曲线),曲线位置越靠近右上方,说明算法性能越好.

(2) F-measure

F-measure是综合准确率和召回率的评价指标,较为全面地反映了算法的性能,其定义式如下[16].

$$F_\beta = \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{9}$$

其中,β^2=0.3以更加强调准确率的作用.F_β数值越大,说明算法性能越好.

(3) ROC曲线下面积——AUC

ROC曲线(即受试者工作特性曲线)是以假阳性概率(false positive rate,简称FPR)为横轴,真阳性概率(true positive rate,简称TPR)概率为纵轴所组成的坐标图,FPR和TPR的定义如下:

$$\begin{cases} FPR = \dfrac{FP}{FP+TN} \\ TPR = \dfrac{TP}{TP+FN} \end{cases} \tag{10}$$

其中,TN表示二值显著性图中的背景区域且对应于真图中的背景区域的像素个数.ROC曲线越趋近于左上方,说明算法的性能越好.AUC即为ROC曲线下的面积,AUC数值越大,说明算法性能越好.

(4) 平均绝对误差——MAE

平均绝对误差描述了二值显著性图与真图的像素级的直接比较,数值越小,说明两张图像越接近,算法性能越好,其定义式如下:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| S(x,y) - G(x,y) \right| \tag{11}$$

其中,S表示二值显著性图,G表示真图,W和H对应图像的宽和高.

3.3 视频显著性检测算法的分析与比较

本节在不同数据库下进行实验,并对多种视频显著性检测的相关算法进行了比较和分析.实验所用视频显著性检测数据集包括SegTrackV1数据集、UVSD数据集和ViSal数据集.实验对比了9种显著性检测算法,包括单图显著性检测算法(如DSR[21]和RRWR[30])、协同显著性检测算法(CCS[37])以及最新的视频显著性检测算法(STBP[69]、SP[70]、CVS[71]、SG[73]、RWRV[74]和SGSP[75]).

图5给出了3组不同视频序列的部分可视化结果,从左到右依次为跳高视频序列、慢跑视频序列和山地自行车视频序列.图5的前两行图像分别表示输入的视频序列和对应的真图,可以发现,跳高视频和山地自行车视频序列中的显著性目标较小且背景较为杂乱.尤其是山地自行车视频,前景和背景的对比度较低,给检测带了较大的难度.通过对图5不同方法的比较分析,可以得出以下结论.

(1) 由于缺少运动信息和帧间信息约束,单图显著性检测方法不能有效地提取视频中的显著性目标.图5中第3行和第4行为单图显著性检测结果.对于跳高和山地自行车视频,该类方法完全无法定位出显著性目标.对

于慢跑视频,该类算法虽然基本可以大致定位显著性目标,但是背景区域存在较多的误检,背景抑制能力较差.

(2) 协同显著性检测算法虽然引入了帧间关系,但缺少运动信息,因而也不能获得较好地检测视频中的显著性目标.从图 5 的第 5 行结果可以看出,协同显著性检测算法无法定位视频中的显著性目标,检测效果较差.主要因为 3 个视频的背景较为复杂、前景背景对比度低,即使引入了帧间关系,对算法的提升也十分有限.此时,运动信息的作用显得尤为重要,相对于背景区域,显著性目标发生了十分明显的运动,而协同显著性算法并未考虑运动信息,因而导致检测性能较差.

(3) 与其他算法相比,视频显著性检测算法获得了相对较好的检测结果.对于相对简单的慢跑视频,CVS 算法获得了最好的检测结果,其次是 SG 算法.其余几种算法虽然可以确定显著性目标的大体位置,但也存在背景区域抑制能力较差、前景目标检测不完整等问题.对于另外两个较难的视频,多数视频显著性检测算法仅能大致定位显著性目标所在的区域,而不能准确、完整的提取显著性目标.此外,由于复杂背景区域的干扰,导致许多背景区域被误检为显著性区域,降低了算法的性能.换言之,现有算法仍远远没有达到理想的效果.

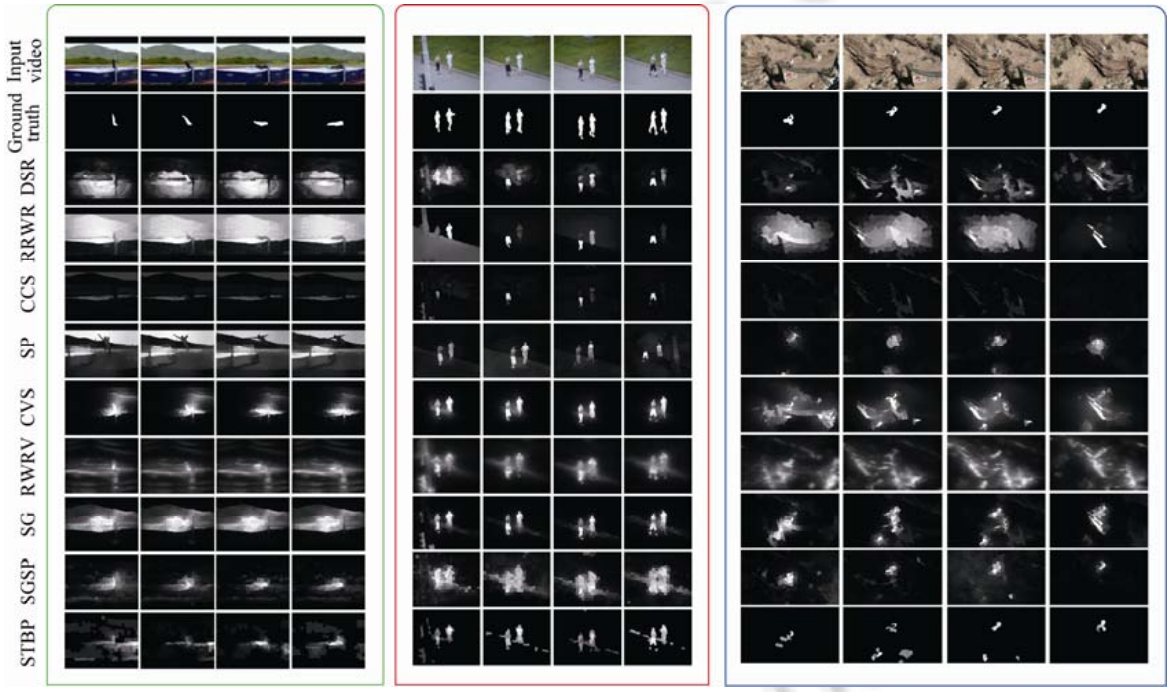


Fig.5 Visual results of different methods

图 5 不同方法的可视化结果

表 1 给出了不同算法的定量分析结果.从表中可以看出,视频显著性检测算法获得了更好的定量评估结果.在 SegTrackV1 数据集中,性能最优的视频显著性检测算法(STBP 算法)的 F 值可以达到 0.658 3,MAE 值可以达到 0.034 2.STBP 算法比单图显著性检测算法的 F 值至少提高了 21.38%,MAE 值至少获得了 73.79%的性能增益(由 0.130 5 下降至 0.034 2).同样地,在另外两个数据集中,视频显著性检测算法也获得了最优的检测结果,如 UVSD 数据集中 F 值达到了 0.602 5,ViSal 数据集中 F 值达到了 0.681 5.因此,直接移植现有单图或协同显著性算法进行视频显著性检测往往不能获得十分理想的结果,还需设计针对视频数据的显著性检测算法.虽然视频显著性检测算法获得最优的性能,但相关性能指标仍不理想.在单图显著性检测中,F 值已达到 0.9 以上,而视频显著性检测算法的 F 值仅为 0.6 左右,还有巨大的提升空间,需进一步探索更加高效、准确的视频显著性检测方法.

综上所述,相对于单图显著性检测算法和协同显著性检测算法,引入运动信息和帧间关系的视频显著性检测算法获得了较好的性能.这也间接说明了研究视频显著性检测算法的必要性,证明了运动信息和帧间关系在定位视频中显著性目标的重要作用.在未来,还需进一步深入挖掘视频运动信息,探索更好的帧间关系描述方式,进而辅助显著性目标判别,这也是视频显著性检测领域的一个研究重点和难点.

Table 1 Quantitative comparisons with different methods on three datasets

表 1 3 个数据集下的不同方法的定量结果比较

方法	数据库					
	SegTrackV1		UVSD		ViSal	
	F_β	MAE	F_β	MAE	F_β	MAE
DSR[21]	0.444 5	0.130 5	0.386 6	0.116 0	0.692 3	0.106 1
RRWR[30]	0.326 7	0.196 3	0.391 8	0.183 9	0.670 7	0.169 0
CCS[34]	0.148 6	0.143 7	0.312 5	0.110 3	0.531 7	0.142 7
STBP[69]	0.658 3	0.034 2	0.491 4	0.084 0	0.681 5	0.098 7
SP[70]	0.215 9	0.119 5	0.229 6	0.149 0	0.572 3	0.151 0
CVS[71]	0.537 0	0.108 5	0.513 5	0.102 9	0.667 6	0.113 9
SG[73]	0.621 8	0.081 0	0.484 7	0.105 0	0.664 0	0.112 9
RWRV[74]	0.445 8	0.151 1	0.315 2	0.177 9	0.466 2	0.190 3
SGSP[75]	0.627 5	0.125 8	0.602 5	0.157 4	0.622 6	0.177 2

4 总结与展望

视频显著性检测是计算机视觉领域的一项基础研究工作,可以作为后续许多研究的先导性操作,具有十分重要的理论研究意义和实际应用价值.本节将进一步总结梳理视频显著性检测的关键问题,并对未来的发展趋势进行展望.

4.1 视频显著性检测的关键问题

(1) 有效挖掘视频序列的运动信息,探索运动与显著性之间的关系,设计有效的运动显著性度量.现有方法往往借助光流信息描述物体的运动,但光流计算过程十分耗时,且获取的光流估计不准确,这将极大地降低运动信息提取的准确性.深度学习技术可以通过设计有效的网络结构避免光流估计过程,是一个值得考虑的研究切入点.此外,对于视频背景杂乱、显著性目标小、前景背景对比度低等难度较大的场景,运动信息的作用将更加重要.而且,当单帧视频中存在多个显著性目标时,还需要借助运动信息对目标进行筛选,提取与运动相关的显著性目标.

(2) 充分提取视频帧间对应关系,构建帧间约束机制,设计简单、有效的时间显著性模型.现有方法通常利用视频前后帧的信息,获取帧间关系.实际上,帧间关系的提取可以借鉴协同显著性检测中的图间关系提取方法,如相似性匹配、传播等技术.但需要注意的是,视频序列的相邻帧的外貌、背景等信息变化不大,这与协同显著性检测的处理场景是不一样的.

(3) 考虑视频显著性目标的一致性,获得更加完整、统一的视频显著性检测结果.现有方法往往忽略了视频显著性目标的全局一致性和帧间相关性,即显著性目标在整个视频序列中应该是反复出现的统一目标.因此,可以通过设计优化模型(如能量函数优化、传播优化等)进一步优化显著性检测结果.

(4) 视频序列中并非每一帧中都存在显著性目标,而且有可能单帧视频中的显著性目标并非是整个视频的显著目标,因此还需要处理如下几种特殊情况.

① 某些视频帧中没有显著性目标,可以通过设计一种判别机制来对视频帧进行预甄别来解决该问题.

② 某些视频帧中出现了新目标,进而可能存在遮挡问题,还需进一步对目标进行判别,此时应考虑视频显著性目标的全局一致性.

③ 某些视频序列的目标运动过快,这样容易产生运动模糊等问题,还需进一步研究解决方案.

(5) 设计高效的视频显著性检测系统,实现显著性区域的实时提取.作为前期预处理技术,通常需要算法具有较高的实时性,而现有方法在实时性和准确性方面往往不能兼得.因此,需要在进一步提升检测效果的同时,考虑进一步降低运算量,节省算法运行时间.

4.2 展 望

经过多年的发展,视频显著性检测技术已取得了一定的进展,但其检测精度还远远没有达到人们的预期,具有较大的发展空间.尤其是大数据时代的来临和深度学习技术的发展,为视频显著性检测指明了一条新的道路.现有研究表明,深度学习不仅可以获得更高的检测结果,还可以有效避免光流估计过程,实际测试环节表现出了较好的实时性,可谓一举两得.在未来,基于深度学习的视频显著性检测方法将会取得更大的进展.此外,通过进一步挖掘运动信息和帧间关系,探索融合底层线索和深度学习的视频显著性检测框架,也具有较好的发展前景.

References:

[1] Gao Y, Shi MJ, Tao D, Xu C. Database saliency for fast image retrieval. IEEE Trans. on Multimedia, 2015,17(3):359–369.

[2] Ren ZX, Gao SH, Chia LT, Tsang IWH. Region-Based saliency detection and its application in object recognition. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(5):769–779.

[3] Fu HZ, Xu D, Lin S, Liu J. Object-Based RGBD image cosegmentation with mutex constraint. In: Proc. of the CVPR. 2015. 4428–4436.

[4] Lei JJ, Wu M, Zhang CQ, Wu F, Ling N, Hou CP. Depth-Preserving stereo image retargeting based on pixel fusion. IEEE Trans. on on Multimedia, 2017,19(7):1442–1453.

[5] Lei JJ, Zhang CC, Fang YM, Gu ZY, Ling N, Hou CP. Depth sensation enhancement for multiple virtual view rendering. IEEE Trans. on Multimedia, 2015,17(4):457–469.

[6] Xiao DG, Xin C, Zhang T, Zhu H, Li XL. Saliency texture structure descriptor and its application in pedestrian detection. Ruan Jian Xue Bao/Journal of Software, 2014,25(3):675–689 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4438.htm [doi: 10.13328/j.cnki.jos.004438]

[7] Gu K, Wang SQ, Yang H, Lin WS, Zhai GT, Yang XK, Zhang WJ. Saliency-Guided quality assessment of screen content images. IEEE Trans. on Multimedia, 2016,18(6):1098–1110.

[8] Han S, Vasconcelos N. Image compression using object-based regions of interest. In: Proc. of the ICIP. 2006. 3097–3100.

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

[10] Ma YF, Zhang HJ. Contrast-Based image attention analysis by using fuzzy growing. In: Proc. of the ACM MM. 2003. 374–381.

[11] Zhang P, Wang RS. Detecting salient regions based on location shift and extent trace. Ruan Jian Xue Bao/Journal of Software, 2004,15(6):891–898 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/891.htm

[12] Harel J, Koch C, Perona P. Graph-Based visual saliency. In: Proc. of the ANIPS. 2006. 545–552.

[13] Zhai Y, Shah M. Visual attention detection in video sequences using spatiotemporal cues. In: Proc. of the ACM MM. 2006. 815–824.

[14] Hou XD, Zhang LQ. Saliency detection: A spectral residual approach. In: Proc. of the CVPR. 2007.

[15] Liu T, Sun J, Zheng NN, Tang XO. Shum HY. Learning to detect a salient object. In: Proc. of the CVPR. 2007.

[16] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-Tuned salient region detection. In: Proc. of the CVPR. 2009. 1597–1604.

[17] Cheng MM, Zhang GX, Mitra NJ, Huang X, Hu SM. Global contrast based salient region detection. In: Proc. of the CVPR. 2011. 409–416.

[18] Zhu WJ, Liang S, Wei YC, Sun J. Saliency optimization from robust background detection. In: Proc. of the CVPR. 2014. 2814–2821.

[19] Zhou L, Yang ZH, Yuan Q, Zhou ZT, Hu DW. Salient region detection via integrating diffusion-based compactness and local contrast. IEEE Trans. on Image Processing, 2015,24(11):3308–3320.

[20] Lei JJ, Wang BR, Fang YM, Lin WS, Callet PL, Ling N, Hou CP. A universal framework for salient object detection. IEEE Trans. on Multimedia, 2016,18(9):1783–1795.

[21] Li XH, Lu HC, Zhang LH, Ruan X, Yang MH. Saliency detection via dense and sparse reconstruction. In: Proc. of the ICCV. 2013. 2976–2983.

[22] Chen TS, Lin L, Liu LB, Luo XN, Li XL. DISC: Deep image saliency computing via progressive representation learning. IEEE Trans. on Neural Networks and Learning Systems, 2015,27(6):1135–1149.

[23] He SF, Lau RW, Liu WX, Huang Z, Yang QX. SuperCNN: A superpixelwise convolutional neural network for salient object detection. International Journal of Computer Vision, 2015,115(3):330–344.

[24] Lee G., Ta YW, Kim J. Deep saliency with encoded low level distance map and high level features. In: Proc. of the CVPR. 2016. 660–668.

[25] Li GB, Yu YZ. Deep contrast learning for salient object detection. In: Proc. of the CVPR. 2016. 478–487.

[26] Liu N, Han JW. DHSNet: Deep hierarchical saliency network for salient object detection. In: Proc. of the CVPR. 2016. 678–686.

[27] Zhang J, Dai YC, Porikli F. Deep salient object detection by integrating multi-level cues. In: Proc. of the WACV. 2017. 1–10.

[28] Hou QB, Cheng MM, Hu XW, Borji A, Tu ZW, Torr P. Deeply supervised salient object detection with short connections. In: Proc. of the CVPR. 2017. 5300–5309.

[29] Qin Y, Lu HC, Xu YQ, Wang H. Saliency detection via cellular automata. In: Proc. of the CVPR. 2015. 110–119.

[30] Li CY, Yuan YC, Cai WD, Xia Y, Feng DD. Robust saliency detection via regularized random walks ranking. In: Proc. of the CVPR. 2015. 2710–2717.

[31] Kim J, Han D, Tai YW, Kim J. Salient region detection via high-dimensional color transform and local spatial support. IEEE Trans. on Image Processing, 2015,25(1):9–23.

[32] Guo F, Shen JB, Li XL. Learning to detect stereo saliency. In: Proc. of the ICME. 2014. 1–6.

[33] Lei JJ, Zhang HL, You L, Hou CP, Wang LH. Evaluation and modeling of depth feature incorporated visual attention for salient object segmentation. Neurocomputing, 2013,120:24–33.

[34] Cong RM, Lei JJ, Zhang CQ, Huang QM, Cao XC, Hou CP. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Processing Letters, 2016,23(6):819–823.

[35] Ju R, Liu Y, Ren TW, Ge L, Wu GS. Depth-Aware salient object detection using anisotropic center-surround difference. Signal Processing: Image Communication, 2015,38:115–126.

[36] Feng D, Barnes N, You SD, McCarthy C. Local background enclosure for RGB-D salient object detection. In: Proc. of the CVPR. 2016. 2343–2350.

[37] Fu HZ, Cao XC, Tu ZW. Cluster-Based co-saliency detection. IEEE Trans. on Image Processing, 2013,22(10):3766–3778.

[38] Cao XC, Tao ZQ, Zhang B, Fu HZ, Feng W. Self-Adaptively weighted co-saliency detection via rank constraint. IEEE Trans. on Image Processing, 2014,23(9):4175–4186.

[39] Li YJ, Fu KR, Liu Z, Yang J. Efficient saliency-model-guided visual co-saliency detection. IEEE Signal Processing Letters, 2015, 22(5):588–592.

[40] Huang R, Feng W, Sun JZ. Saliency and co-saliency detection by low-rank multiscale fusion. In: Proc. of the ICME. 2015. 1–6.

[41] Song HK, Liu Z, Xie YF, Wu L, Huang MK. RGBD co-saliency detection via bagging-based clustering. IEEE Signal Processing Letters, 2016,23(12):1722–1726.

[42] Cong RM, Lei JJ, Fu HZ, Huang QM, Cao XC, Hou CP. Co-Saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation. IEEE Trans. on Image Processing, 2018,27(2):568–579.

[43] Hou XD, Zhang LQ. Dynamic visual attention: Searching for coding length increments. In: Proc. of the NIPS. 2008. 681–688.

[44] Seo HJ, Milanfar P. Static and space-time visual saliency detection by self-resemblance. Journal of Vision, 2009,9(12):1–27.

[45] Guo CL, Ma Q, Zhang LM. Spatio-Temporal saliency detection using phase spectrum of quaternion Fourier transform. In: Proc. of the CVPR. 2008. 1–8.

[46] Guo C, Zhang L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. on Image Processing, 2010,19(1):185–198.

[47] Cui XY, Liu QS, Metaxas DN. Temporal spectral residual: Fast motion saliency detection. In: Proc. of the ACM MM. 2009. 617–620.

[48] Fang YM, Lin WS, Chen ZZ, Tsai CM, Lin CW. A video saliency detection model in compressed domain. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(1):27–38.

[49] Liu YG, Chen YW. Video saliency detection algorithm based on motion spectral residual. Computer Engineering, 2014,40(12): 247–250,257 (in Chinese with English abstract).

[50] Qiu GP, Gu XD, Chen ZB, Chen QQ, Wang C. An information theoretic model of patiotemporal visual saliency. In: Proc. of the ICME. 2007. 1806–1809.

[51] Liu C, Yuen PC, Qiu GP. Object motion detection using information theoretic spatio-temporal saliency. Pattern Recognition, 2009, 42(11):2897–2906.

[52] Li Y, Zhou Y, Yan JC, Niu ZB, Yang J. Visual saliency based on conditional entropy. In: Proc. of the ACCV. 2009. 246–257.

[53] Lu HC, Li XH, Zhang LH, Ruan X, Yang MH. Dense and sparse reconstruction error based saliency descriptor. IEEE Trans. on Image Processing, 2016,25(4):1592–1603.

[54] Li NY, Sun BL, Yu JY. A weighted sparse coding framework for saliency detection. In: Proc. of the CVPR. 2015. 5216–5223.

[55] Yuan YC, Li CY, Kim J, Cai WD, Feng DD. Dense and sparse labeling with multi-dimensional features for saliency detection. IEEE Trans. on Circuits and Systems for Video Technology, 2018,28(5):1130–1143. [doi:10.1109/TCSVT.2016.2646720]

[56] Li Y, Zhou Y, Xu L, Yang XC, Yang J. Incremental sparse saliency detection. In: Proc. of the ICIP. 2009. 3093–3096.

[57] Luo Y, Tian Q. Spatio-Temporal enhanced sparse feature selection for video saliency estimation. In: Proc. of the CVPRW. 2012. 33–38.

[58] Ren ZX, Chia LT, Rajan D. Video saliency detection with robust temporal alignment and local-global spatial contrast. In: Proc. of the ACM ICMR. 2012. 1–8.

[59] Ren ZX, Gao SH, Rajan D, Chia LT, Huang Y. Spatiotemporal saliency detection via sparse representation. In: Proc. of the ICME. 2012. 158–163.

[60] Ren ZX, Gao SH, Chia LT, Rajan D. Regularized feature reconstruction for spatiotemporal saliency detection. IEEE Trans. on Image Processing, 2013,22(8):3120–3132.

[61] Xue YW, Guo XJ, Cao XC. Motion saliency detection using low-rank and sparse decomposition. In: Proc. of the ICASSP. 2012. 1485–1488.

[62] Chen CLZ, Li S, Wang YG, Qin H, Hao AM. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans. on Image Processing, 2017,26(7):3156–3170.

[63] Itti L, Baldi P. A principled approach to detecting surprising events in video. In: Proc. of the CVPR. 2005. 631–637.

[64] Mahadevan V, Vasconcelos N. Spatiotemporal saliency in dynamic scenes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(1):171–177.

[65] Kim W, Kim C. Spatiotemporal saliency detection using textural contrast and its applications. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(4):646–659.

[66] Kim W, Han JJ. Video saliency detection using contrast of spatiotemporal directional coherence. IEEE Signal Processing Letters, 2014,21(10):1250–1254.

[67] Zhou F, Kang SB, Cohen MF. Time-Mapping using space-time saliency. In: Proc. of the CVPR. 2014. 3358–3365.

[68] Le TN, Sugimoto A. Region-Based multiscale spatiotemporal saliency for video. arXiv:1708.01589, 2017.

[69] Xi T, Zhao W, Wang H, Lin WS. Salient object detection with spatiotemporal background priors for video. IEEE Trans. on Image Processing, 2017,26(7):3425–3436.

[70] Liu Z, Zhang X, Luo SH, Meur OL. Superpixel-Based spatiotemporal saliency detection. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(9):1522–1540.

[71] Wang WG, Shen JB, Shao L. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans. on Image Processing, 2015,24(11):4185–4196.

[72] Wang WG, Shen JB, Porikli F. Saliency-Aware geodesic video object segmentation. In: Proc. of the CVPR. 2015. 3395–3402.

[73] Wang WG, Shen JB, Yang RG, Porikli F. A unified spatiotemporal prior based on geodesic distance for video object segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,40(1):20–33.

[74] Kim H, Kim Y, Sim JY, Kim CS. Spatiotemporal saliency detection for video sequences based on random walk with restart. IEEE Trans. on Image Processing, 2015,24(8):2552–2564.

[75] Liu Z, Li JH, Ye LW, Sun GL, Shen LQ. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation. IEEE Trans. on Circuits and Systems for Video Technology, 2017,27(12):2527–2542. [doi: 10.1109/TCSVT.2016.2595324]

[76] Fang ZM, Cui RY, Jin JX. Video saliency detection algorithm based on biological visual feature and visual psychology theory. Acta Physica Sinica, 2017,66(10):1–14 (in Chinese with English abstract).

[77] Liu T, Yuan ZJ, Sun J, Wang JD, Zheng NN, Tang XO, Shum HY. Learning to detect a salient object. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(2):353–367.

[78] Huang CR, Chang YJ, Yang ZX, Lin YY. Video saliency map detection by dominant camera motion removal. IEEE Trans. on Circuits and Systems for Video Technology, 2014,24(8):1336–1349.

[79] Wang WG, Shen JB, Shao L. Video salient object detection via fully convolutional networks. IEEE Trans. on Image Processing, 2018,27(1):38–49.

[80] Le TN, Sugimoto A. Video salient object detection using spatiotemporal deep features. arXiv:1708.01447, 2017. 1–13.

[81] Tsai D, Flagg M, Rehg JM. Motion coherent tracking with multi-label MRF optimization. In: Proc. of the BMVC. 2010. 1–11.

[82] Li FX, Kim T, Humayun A, Tsai D, Rehg JM. Video segmentation by tracking many figure-ground segments. In: Proc. of the ICCV. 2013. 2192–2199.

[83] Perazzi F, Pont-Tuset J, McWilliams B, Gool LV, Gross M, Sorkine-Hornung A. A benchmark dataset and evaluation methodology for video object segmentation. In: Proc. of the CVPR. 2016. 724–732.

附中文参考文献:

[6] 肖德贵,辛晨,张婷,朱欢,李小乐.显著性纹理结构特征及车载环境下的行人检测.软件学报,2014,25(3):675–689. http://www.jos.org.cn/1000-9825/4438.htm [doi: 10.13328/j.cnki.jos.004438]

[11] 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. http://www.jos.org.cn/1000-9825/15/891.htm

[49] 刘宇光,陈耀武.基于运动谱残差的视频显著性检测算法.计算机工程,2014,40(12):247–250,257.

[76] 方志明,崔荣一,金璟璇.基于生物视觉特征和视觉心理学的视频显著性检测算法.物理学报,2017,66(10):1–14.

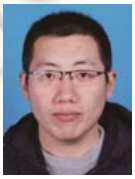
丛润民(1989－),男,山东招远人,博士生,主要研究领域为计算机视觉,显著性检测.

王文冠(1990－),男,博士,CCF 学生会员,主要研究领域为计算机视觉.

雷建军(1980－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为视频编码与处理,计算机视觉.

黄庆明(1965－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为计算机视觉,模式识别.

付华柱(1983－),男,博士,研究员,主要研究领域为计算机视觉,医学图像分析.



牛力杰(1993－),男,硕士生,主要研究领域为计算机视觉.