计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190600148

基于3D全时序卷积神经网络的视频显著性检测

王教金[1] 蹇木伟[1] 刘翔宇[1] 林培光[1] 耿蕾蕾[1] 崔超然[1] 尹义龙[2]
1 山东财经大学计算机科学与技术学院 济南 250014
2 山东大学软件学院 济南 250101
(125453468@qq.com)

摘 要 视觉是人类感知世界的重要途径之一。视频显著性检测旨在通过计算机模拟人类的视觉注意机制,智能地检测出视频中的显著性物体。目前,基于传统方法的视频显著性检测已经达到一定的水平,但是在时空信息一致性利用方面仍不能令人满意。因此,文中提出了一种基于全时序卷积神经网络的视频显著性检测方法。首先,利用全时序卷积对输入视频进行空间信息和时间信息的时空特征提取;然后,利用3D池化层进行降维;其次,在解码层中用3D反卷积和3D上采样对前端特征进行解码;最后,通过把时空信息有机地提取与融合,来有效地提升显著图的质量。实验结果表明,所提算法在3个广泛使用的视频显著性检测数据集(DAVIS,FBMS,SegTrack)上的性能优于当前主流的视频显著性检测方法。

关键词:显著性检测;时空特征;全时序卷积;神经网络
中图法分类号 TP391

Video Saliency Detection Based on 3D Full ConvLSTM Neural Network

WANG Jiao-jin[1],JIAN Mu-wei[1],LIU Xiang-yu[1],LIN Pei-guang[1],GEN Lei-lei[1],CUI Chao-ran[1] and YIN Yi-long[2]
1 School of Computer Science and Technology,Shandong University of Finance and Economics,Jinan 250014,China
2 School of Software Engineering,Shandong University,Jinan 250101,China

Abstract Video saliency detection aims to mimic human's visual attention mechanism of perceiving the world via extracting the most attractive regions or objects in the input video. At present,it is still a challenge for video saliency detection. Traditional video saliency-detection models have reached a certain level,but exploiting the consistency of spatio-temporal information is unsatisfactory. In order to solve this issue,this paper proposes a video saliency-detection model based on 3D full ConvLSTM neural network. Firstly,the full-time convolution is utilized to extract spatio-temporal features from the input video,and then the 3D pooling layer is explored for dimensionality reduction. Secondly,the extracted features are decoded by 3D deconvolution in the decoding layer,and the interpolation algorithm is applied to restore the saliency map to the original size of the original image. The proposed method extracts the time and space information jointly so as to effectively enhance the completeness of the saliency map. Experimental results show that the performance of the proposed algorithm is superior to state-of-the-art video saliency detection methods based on three widely used data sets (DAVIS,FBMS,SegTrack) for video saliency detection.

Keywords Saliency detection,Spatio-temporal feature,ConvLSTM,Neural network

视频显著性检测是指在一段连续的序列帧中自动、可靠地提取出引人注意的物体目标或用户感兴趣的图像区域。相比于图像序列,视频序列往往包含连续的运动线索和更丰富的外观形态信息,可以被用来更好地串联显著性目标和特有的动态特征。同时,视频序列中通常存在多变的背景、复杂的运动特征和各异视角的转变,从而为检测出视频显著性内容的一致性带了巨大挑战。近年来,视频显著性检测已成为计算机视觉领域研究者所关注的热点问题。

视频显著性检测的目的是通过空间和时间线索,从给定的视频序列中连续不断地找到相应的显著性运动目标。由于空间特征和时间特征难以做到一致性,并且资源消耗较大,本文提出了一种基于深度学习的全时序卷积神经网络模型,通过充分利用空间和时间的一致性特征来有效并快速地检测在视频中出现的显著性区域。

到稿日期:2019-06-26 返修日期:2019-10-25 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61601427,61976123,61771230);泰山学者青年专家支持计划
This work was supported by the National Natural Science Foundation of China (61601427,61976123,61771230),Taishan Young Scholars Program of Shandong Province.
通信作者:蹇木伟(jianmuweihk@163.com)

本文第 1 节介绍相关工作；第 2 节对算法模型进行整体介绍；第 3 节详细介绍算法模型中的对应模块，并给出模型图示；第 4 节在不同数据集上将所提方法与当前已有方法进行实验对比分析；最后总结全文。

1 相关工作

从建模方式上看，目前对视频显著性检测问题的研究方法可以分为两大类：自顶向下的显著性检测模型和自底向上的显著性检测模型。从具体使用的技术和特征学习方式来看，视频显著性检测方法可以分为基于传统手工特征的显著性检测方法和基于深度学习的显著性检测方法。

传统的视频显著性检测方法主要关注如何利用和设计手工特征的低层视觉线索（如颜色对比度、运动线索、中心先验约束、方向性等）进行显著性值的估计[1-8]。其中，帧内低层特征用来表示空间显著性，可以用图像处理中的一些经典方法[9-17]（如低秩分析、对比度先验和背景先验等）进行计算；而时间显著性一般用帧间目标的运动线索进行表示。文献[18]首先提取超像素级运动和颜色直方图以及全局运动直方图作为显著性计算特征；然后构建具有全局运动特征的超像素级虚拟背景节点结构图，并且利用图论上的最短路径算法迭代估计运动显著性，生成粗糙显著图，再使用帧间相似性矩阵计算前向和后向时间传播的显著性值，以获得更好的时间显著性图；最后使用帧内相似性矩阵计算局部和全局空间传播的显著性值，得到具有时空特征的显著性图。LGFOGR（Consistent Video Saliency Using Local Gradient Flow Optimization and Global Refinement）[19]方法通过利用梯度流场和能量优化来估计视频中的显著性区域。此方法结合帧内边界信息和帧间运动信息来索引显著性区域，并通过引入局部对比度和全局对比度来增强目标的显著性；同时，为了提高显著性计算的精确度，该方法进一步利用能量函数优化得到最终显著性图。文献[6]首先将每一帧转化为超像素图，然后分别在超像素级与帧级上提取局部特征和全局特征的运动直方图和颜色直方图。其中，每个超像素空间显著图是用全局对比度和空间稀疏来估计产生的，超像素级别的时间显著图是通过运动直方图计算产生的，而像素级的显著图是由超像素级的空间和时间显著图产生的，最后通过自适应融合方法产生最终像素级的时空显著性图。RWRV（Spatiotemporal Saliency Detection for Video Sequences Based on Random Walk with Restart）[20]方法应用随机重游走分布机制，利用运动特征、目标与背景对比度特征等计算出时间显著性图。然后，在一个类似的转换矩阵中用亮度、颜色和疏密空间特征进行游走，从而产生时空显著性图。FLRC（Video Saliency Detection via Spatialtemporal fusion and low-rank Coherency Diffusion）[21]方法用颜色对比度生成颜色显著性图，进而获得空间显著图；然后通过光流图的对比度来估计时间显著图；最后通过低秩一致性来引导时空显著性的扩散，并利用时间平滑策略提高显著性图的精度。

基于深度学习的视频显著性检测方法主要通过端到端的学习来提取显著性目标。文献[22]设计了一个基于全卷积神经网络的视频显著性检测模型。全卷积静态网络为每张单独的帧生成静态显著图，动态网络以帧对和静态显著图作为输入得到最终的显著图。文献[23]提出的视频显著性检测网络使用时空域中的 3D 滤波器来直接学习空间和时间信息以获得 3D 深度特征，并且将 3D 深度特征传递到输入图像像素中来检测显著性的物体边界；同时将监督学习应用于隐藏层来改善中间显著性细节，使得显著性块逐渐被重新定义，细化了显著性区域。文献[18]通过使用显著性引导的堆叠自动编码器进行显著性检测，首先提取像素和超像素时空相邻区域的显著性值作为高维特征向量，然后以无监督的方式学习堆叠的自动编码器以获得初始显著图，最后使用一些后处理操作来改善显著对象并抑制背景干扰物。文献[24]先通过 FlowNet[25]产生光流图，然后利用回归神经编码器（Recurrent Neural Encoder）优化并生成最终的显著图。

总体来说，目前基于深度学习的方法需要大量标注的样本进行训练，复杂度高，且更多地关注空间特征信息的提取，对时间上相对应的信息一致性关注较少，导致其实际应用效果不理想[19,26-30]。已有研究学者提出利用 FCN 网络结构模型解决复杂场景中视频显著性检测的问题，然而，其实时处理速度较慢，应该通过考虑视频序列中的多帧对来挖掘更多的时空信息。基于以上研究，本文提出了一种基于 FCN 框架，利用 ConvLSTM[27]学习 3D 特征进行端到端学习的视频显著性检测模型。本文的主要创新点如下：

1）针对全卷积网络过程中网络模型需要重复输入单一帧对，导致实时处理速度慢且时空信息不连续的问题，设计了一种基于全时序卷积网络的视频显著性模型，其能够有效地提取时空一致性信息，并在降低资源消耗的同时提升实时处理的速度；

2）为了保证时空一致性传播，提出了一种时序卷积层和 3D 池化层有机结合的编码网络模型，其能够在捕捉传递帧内信息特征的同时提取帧间信息特征；

3）针对时空特征在解码过程中信息易丢失的缺陷，设计了反卷积层和隐藏层相结合的解码网络模型，使得 3D 特征图经过每个隐藏层时能够被修正，进而生成与之对应的像素级时空显著图。

2 模型框架

基于 3D 全时序卷积模型主要包括 3 个模块：图像预处理、编码网络和解码网络。编码网络和解码网络的结构框架如图 1 和图 2 所示。

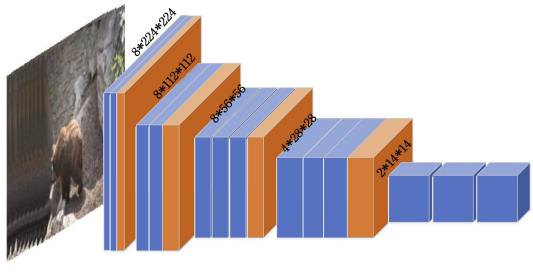

图 1 时序全卷积模型中的编码网络

Fig. 1 Model of encoder network in full ConvLSTM network

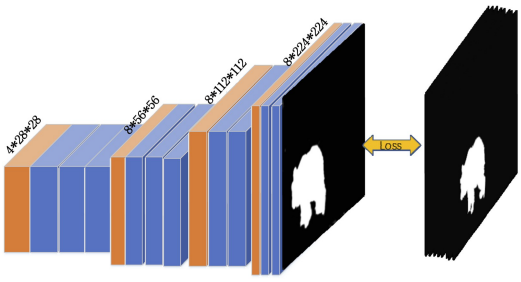

图 2 时序全卷积模型中的解码卷积层

Fig. 2 Model of decoder network in full ConvLSTM network

1)在图像预处理模块，本文通过双线性内插值法把输入图像处理为与编码网络输入端所需输入图像大小一致的序列帧图片。

2)编码网络模块主要由时序卷积层和 3D 池化层组成，主要功能是从输入的序列帧图片中提取含有空间和时间信息的 3D 深度特征。

3)解码网络模块主要由反卷积层和隐藏层组成。通过编码网络生成的特征图在每个隐藏的 3D 反卷积层中通过监督学习逐步改善 3D 深度特征图，进而生成像素级的时空显著图，最终再利用双线性内差值算法生成与原图像大小一致的显著图。

3 基于全时序卷积神经网络的模型行设计

目前，基于深度学习的方法进行视频显著性检测的框架有很多，但由于数据集相对较少且大多数网络框架是分开提取空间特征信息和时间特征信息，再通过时空信息融合的方法进行视频显著性检测，因此其检测性能较低。另一方面，基于深度学习方法的全卷积神经网络框架在实时处理速度上不尽人意。基于此，本文提出了一种基于 3D 全时序卷积的视频显著性检测模型。考虑到时间信息特征，该模型在初始的卷积运算中运用时序卷积进行时空特征的提取，为提高检测精度和实时处理速度，用多帧对输入的神经网络结构来产生像素级的显著图。在高层特征的基础上，所提的模型将视频帧序列提供给神经网络，并且使网络连续输出显著图。

3.1 图像预处理

使用 FCN 网络框架时，网络允许任意大小的图像作为输入，并保留空间信息。利用本文网络框架时，给定一段视频帧序列后，为节省训练资源，首先需要对序列帧图像进行预处理，以在保持图像内容与原图像一致的情况下修改图像的尺寸。为此，采用双线性内插的方法把原图像的尺寸修改为 224×244 大小的序列帧图像。

为求得缩放后的图像在 (i,j) 位置的像素值，然后需要根据该位置周围的 4 个像素点的像素值进行双线性插值，得到该点的像素值，然后对缩放后的图像的每一个像素点进行遍历，最后得到缩放后的整幅图像。

3.2 编码网络

在编码网络模块中，本文用预处理后的连续 8 帧图像作为输入序列，同时产生与输入大小一致的连续 8 帧显著图，之后通过双线性内插值算法恢复与原图像大小一致的显著图作为最终的显著性图像。本文模型中的编码网络可被认为全卷积网络底部的卷积层，如图 2 所示，本文模型框架的左侧是多层时序卷积层，时序卷积层共享的权重向量参数和偏置参数在架构上具有平移不变性。在输入端，图像大小是 $l*h*w*c$，其中 l 是时序的长度，本文采用固定的时序长度，即 l 为 8；h 是图像的高；w 是图像的宽；c 是图像的通道数。对于输入的序列帧图像，每一帧的图像大小为 h，w 和 c(RGB 3 个颜色通道)。整个序列帧图像在输出时，通过将输入图像与可训练的卷积核加入可训练的偏置项参数来获得显著性图。本文用 X 表示输入图像，其时序卷积滤波器由权重 W 和偏置项 b 确定。编码网络输出的信息特征图可由式(1)表示：

$$f_s(X;W,b)=W*_sX+b \tag{1}$$

其中，$*_s$ 是时序卷积符号，s 是卷积的步长。在每一个卷积层中，本文使用逐点非线性激活函数(tanh)。同时，在时序卷积层后添加非线性的下采样池化层，对于卷积操作，本文采用的卷积方式为时序卷积(ConvLSTM)。时序卷积在保存空间信息的同时，保持着时间上的一致性。ConvLSTM 是 FC-LSTM 的拓展，保留了 FC-LSTM 的优点，固有的卷积结构更适合时空信息的提取。ConvLSTM 同样拥有记忆细胞 C_t、输入门 i_t、输出门 O_t 和遗忘门 f_t，最终状态为 H_t。在 Conv-LSTM 各门的协同工作中，细胞输出信息的选择由输出门 O_t 所决定。在新数据输入时，若输入门 i_t 被激活，新数据将会被积累到存储器单元中；若遗忘门 f_t 被激活，则原有的单元状态 C_t 将会被忘记。输出门 O_t 会再次控制最新存储器单元的 C_t 值是否被传输到最终状态 H_t 上。根据以上定义，式(2)给出 ConLSTM 的表示。

$$\begin{aligned}
i_t&=\sigma(W_{xi}*\chi_t+W_{hi}*H_{t-1}+W_{ci}\circ C_{t-1}+b_i)\\
f_t&=\sigma(W_{xf}*\chi_t+W_{hf}*H_{t-1}+W_{cf}\circ C_{t-1}+b_f)\\
C_t&=f_t\circ C_{t-1}+i_t\circ\tanh(W_{xc}*\chi_t+W_{hc}*H_{t-1}+b_c)\\
o_t&=\sigma(W_{xo}*\chi_t+W_{ho}*H_{t-1}+W_{co}\circ C_{t-1}+b_o)\\
H_t&=o_t\circ\tanh(C_t)
\end{aligned} \tag{2}$$

其中，$*$ 代表卷积运算符，o 代表 Hadamard 乘积；在 t 时刻，输入门 i_t、遗忘门 f_t、输出门 O_t、记忆细胞 C_t、隐藏状态 H_t、学习权重 W 都均是三维 tensor。简单起见，暂时不考虑偏置项。如图 3 所示，此时的输入与各个门之间的连接由原先的前馈式换成了卷积形式，同时各个状态之间也换成了卷积运算。

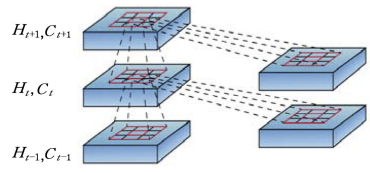

图 3 时序卷积示意图

Fig. 3 Schematic diagram of ConvLSTM

3.3 解码网络

本文的解码网络包含上采样层和卷积层。解码网络需要对编码网络所产生的特征图像进行上采样，多层反卷积网络作为最后的输出序列图像。该网络通过卷积层中相应的前向

和后向通道来实现反卷积操作。所有的时序卷积层、反卷积层、隐藏层的参数都将由训练学习得到。通过卷积运算和特征池化运算，输出的特征图是相对粗略的，分辨率也有所下降，为了获取像素级的显著性预测图，本文采用 3D 上采样层和多层反卷积层进行显著图的估计。本文中的解码网络如式(3)所示：

$$\mathbb{Y}=\mathbb{D}_p(F_p(I;\Phi_F);\Phi_D) \tag{3}$$

其中，I 是输入图像；$F_p(\cdot)$是通过编码网络在步长等于 S 时产生的特征图；$D_p(\cdot)$是在上采样后的反卷积层产生的最终显著图；S 因子由最后输出的图像 $\mathbb{Y}$ 与输入的图像 I 一致确定的；Φ 是在正向和反向卷积层中的所有参数，均可通过训练学习到。最后，在网络的最右侧，本文采用卷积核为 11 的卷积层通过 sigmoid 激活函数将标注特征图 $\mathbb{Y}$ 映射到训练出的显著性预测图 P。我们用 sigmoid 层输出每个像素在 0 和 1 范围内的实数值。

为了产生更好的显著图，同时获得更加有效的模型，本文运用一个融合损失函数。为平衡预测显著图 $P[0,1]^{224\times224}$ 和标注显著图 $G\{0,1\}^{224\times224}$ 之间的差异性，定义损失函数，如式(4)所示：

$$\mathscr{L}(P,G)=l_{\text{predict}}(\theta,W_{\text{predict}})+\sum_{t=1}^{T}l_{\text{UpSmapling3D}}(\theta,W_{\text{UpSmapling3D}}) \tag{4}$$

其中，T 是 UpSampling3D 的层数，θ 是所有网络层的参数集合，W 是相对应层的权重，l 是融合损失函数，如式(5)所示：

$$l(S,G)=l_{\text{cross\_entropy}}(S,G)+l_{\text{mae}}(S,G) \tag{5}$$

在融合函数中，$l_{\text{cross_entropy}}$是二元交叉熵损失，如式(6)所示：

$$l_{\text{cross\_entropy}}=-\frac{1}{N}\sum_{i=1}^{N}[g_i\log(p_i)+(1-g_i)\log(1-p_i)] \tag{6}$$

l_{mae}是平均损失函数，如式(7)所示：

$$l_{\text{mae}}(S,G)=-\frac{1}{N}\sum_{i}^{N}|g_i-p_i| \tag{7}$$

其中，$N=H\times W\times L$，表示在视频序列帧中图像持续的时间长度和单帧图像大小($8\times224\times224$)。

4 实验与结果

本节利用 3 个常用的公共基准视频数据集 DAVIS，FBMS，SegTrack2 来评估本文方法的显著性检测性能，并将实验结果与不同类别的其他算法进行比较。实验结果表明，本文提出的基于 3D 全时序卷积的视频显著性检测方法能获得较好的识别率。软件环境为 Python3.6，硬件采用 32 GB RAM，Inter Core i7 处理器，主频为 3.6 GHz 和英伟达 1080GPU。

4.1 数据集与实验设置

DAVIS (Densely Annotated Video Segmentation)数据集[31]是一个常用的且具有挑战性的视频数据集，它包含 50 个高质量的视频序列，总共 3455 帧，每个帧都具有完全注释的像素级标注图像。本文将其中的 25 个视频帧序列作为训练集，总共约 1668 帧，来训练所提模型。测试集包含 15 个视频帧序列，总共约 886 帧。

FBMS (Freiburg-Berkeley Motion Segmentation)数据集[2]包含 59 个自然场景视频序列，涵盖各种挑战，例如前景和背景大比例的外观变化、明显的形变和摄像机大尺度的运动等。此数据集最初被应用于运动分割，其中无显著性的运动目标被标记成了前景。在此数据集中，本文只用了标注比较精确的部分。FBMS 数据集分为训练集和测试集，其中训练集包括 35 个视频序列，测试集包括 20 个视频序列。

SegTrack2 数据集[3]提供了不同场景下的多种类型的运动视频，本文挑选了其中的士兵和猴子两种情景来代表运动过程中外形变化较大、背景复杂、光照干扰严重及运动速度较快 4 种情况，能够更全面地验证本文所提出的视频显著性检测算法的鲁棒性及有效性。

4.2 实验结果的视觉对比

本节对显著性检测的视觉进行对比，将所提方法与当前较好的多种方法分别在 3 个数据集上进行测试与比较，结果如图 4—图 6 所示。图中的每个像素点属于前景的概率均由该点的像素值大小代表。其中，SGSP[6]算法和 LGFOFR[19]算法均采用了基于超像素的显著性检测方法，其余方法都是基于像素级的处理结果。

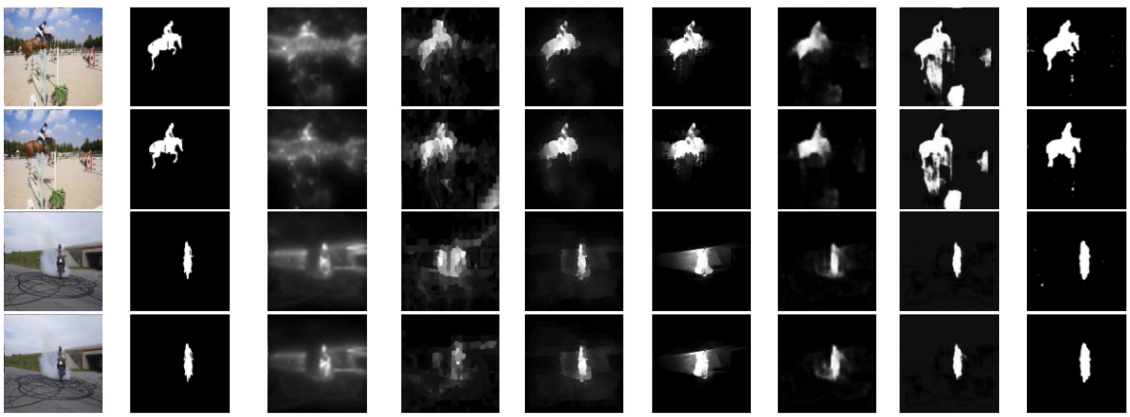

图 4 在 DAVIS 数据集上主流显著性算法检测性能的对比

Fig. 4 Comparison of different state-of-the-art saliency detection models performed on DAVIS database

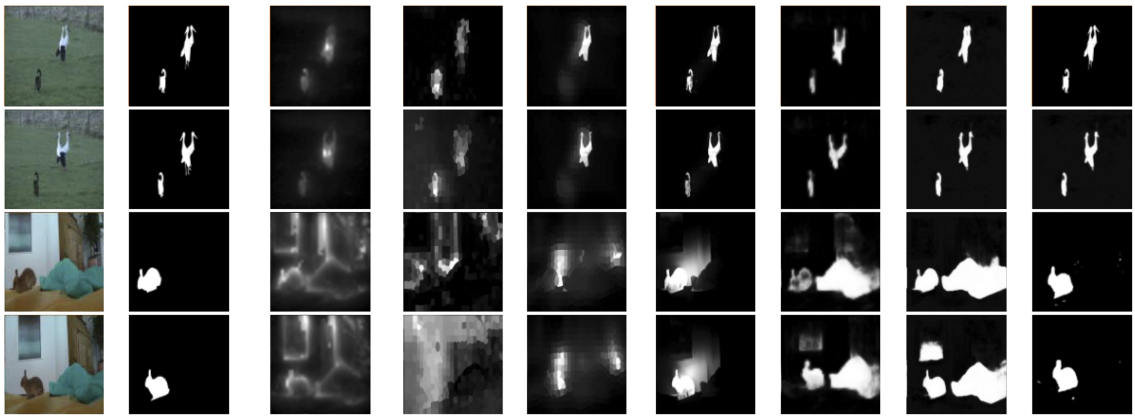

图 5 FBMS 数据集上主流显著性算法检测性能的对比

Fig. 5 Comparison of different state-of-the-art saliency detection models performed on FBMS database

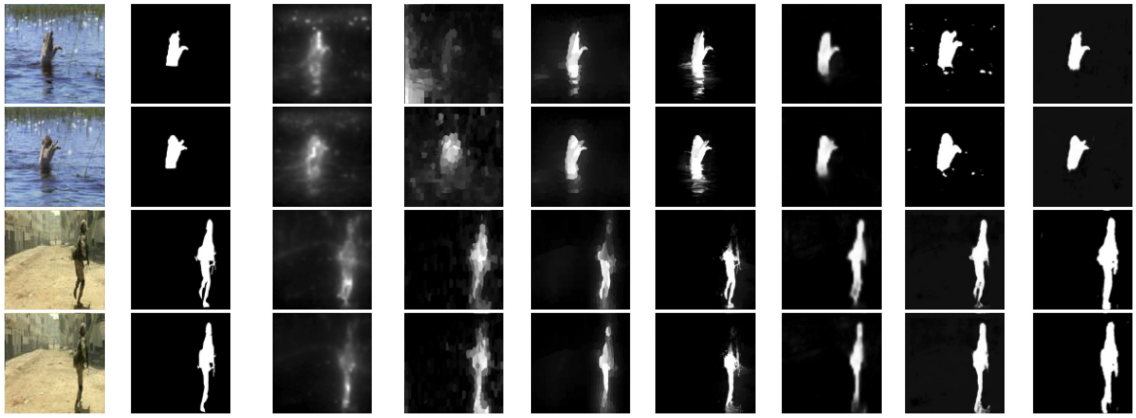

图 6 SegTrack 数据集上主流的显著性算法检测性能对比

Fig. 6 Comparison of different state-of-the-art saliency detection models performed on SegTrack database

从图 4—图 6 中可以看出，对于背景变化较大的情景，基于超像素的显著性检测结果均不理想，这充分说明了对背景变换较大的情景，运动特征和时间特征一致性对视频显著性检测的影响更为明显。从 RWRV[20] 算法的结果可看出，其对背景的抑制能力比较薄弱。SGSP[18] 算法由于无法确定目标的轮廓及其准确位置，因此算法的准确性受到影响。传统的视频显著性方法的基本思想多是采用颜色、亮度以及轮廓等低层视觉特征与周围背景做对比后确定显著性区域，而在纹理复杂且背景有大幅度变化的情况下，仅考虑相邻帧之间的时空域变化很容易导致将背景误判为前景。FLRC[21] 的视频显著性算法得到了相对比较好的结果，但是此方法对计算的负荷性能要求比较高，并且在复杂的背景下难以得到令人满意的性能。FCN[26] 和 DSSC[32] 是基于深度学习的视频显著性方法，实验结果表明其能够产生相对理想的检测结果，说明深度学习模型能够在视频显著性检测中取得较好的性能。从图 4(i)、图 5(i)和图 6(i)中可以看出，在时间变化较为多样的情况下，本文给出的基于深度学习的算法给予了前景更多的关注，甚至在一些具有挑战性的场景中也能取得良好的性能。实验结果表明，本文方法能够同时捕获帧内的空间信息和帧间的时间信息，因此可以准确地检测出视频中的显著性目标。

4.3 定量评估

为了证明本文方法的准确性和有效性，采用两种常规的评价指标即 P-R 曲线(准确率-召回率)和 F_β 值对实验结果进行有效的评估[23,28-30,33-37]。在显著性检测结果的评价中，准确率 P 表示得到的显著区域中实际目标物体所占的比例，召回率 R 表示检测结果中的目标物体占真值图(Ground Truth)中目标物体的比例。

$$P=\frac{TP}{TP+FP} \tag{8}$$

$$R=\frac{TP}{TP+FN} \tag{9}$$

其中，TP 表示被正确检测到的目标像素的个数；FP 表示背景被错误地检测成运动目标的个数，FN 表示该区域属于显著性物体但是被检测为背景区域的像素个数。

实验中对不同的视频显著性检测算法得到的显著图进行二值化处理，并以每帧图像的 Ground Truth 为标注图像的基准，通过式(8)、式(9)分别计算准确率 P 和召回率 R，并绘制 P-R 曲线。图 7 为不同方法在 3 个数据集上的 P-R 曲线图。从图 7 可以看出，本文方法在 DAVIS 和 SegTrack 数据集上的性能超过了当前主流的视频显著性检测算法，在 FBMS 数据集上也取得了与当前最好方法近似的性能。本文设计的模型由于在深度学习的过程中对帧内的空间信息和帧间的时间

信息同时进行了捕获，因此能够有效地检测出视频中的显著性区域，在实时检测和实时处理效率上也表现出较大的优势，证明了基于全时序卷积神经网络的视频显著性检测方法的有效性。

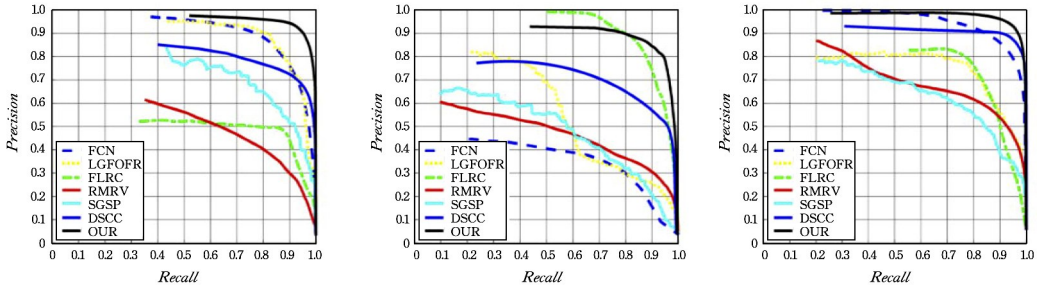

图 7 P-R 曲线图

Fig. 7 P-R graph

当两种算法的 P-R 曲线发生交叉时，不管是查准率还是查全率，对显著性的评估均是有偏差的，故本文引入了 F_β 值这一综合评价指标。F_β 在数值上表示 P 和 R 的加权平均数，计算公式如下：

$$F_\beta=\frac{(1+\beta^2)Precision\times Recall}{\beta^2\ Precision+Recall} \tag{10}$$

通常，为了更注重准确度，在算法的性能评估中将权重系数设置为 $\beta^2=0.3$[10-11,38-40]。

图 8 给出了不同视频显著性检测模型在所测试的 3 个数据集上的平均准确率 P、平均召回率 R 和平均 F_β 值。从图 8 中可以看出，本文模型取得了当前最优的性能。实验结果证明了本文所设计的基于全时序卷积神经网络的视频显著性检测算法是可行且高效的。

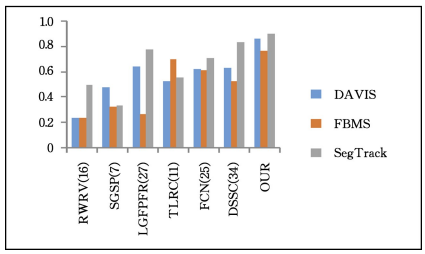

图 8 F_β-measure 值

Fig. 8 F_β-measure value

结束语 在视频显著性检测中，视频中的显著性目标在时间轴上会发生变化，而当前大多方法在分别提取空间特征和时间特征之后进行特征的融合，导致其空间和时间的一致性有所偏差。针对视频显著性检测的这些挑战性问题，本文提出了一种基于全时序卷积深度网络的视频显著性检测方法，其在卷积过程中同时考虑了帧内的空间信息和帧间的时间信息，保持了时空特征的一致性；同时，当把序列帧图像输入编码网络时，编码网络通过时序卷积层和 3D 池化层提取包含空间信息和时间信息的特征图，然后将其输出到解码网络进行反卷积操作，最终生成与序列帧相对应的显著图。实验结果表明，本文提出的方法在视频显著性检测的三大数据集上可以有效地检测出显著性区域，其可行性与有效性得到验证。本文的显著性检测模型对实时性的处理还需要进一步提升，接下来将考虑结合注意力模型和传统方法来提升目标检测的效率和网络学习的能力。

参 考 文 献

[1] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. Internationl Journal of Computer Vision, 2015, 115(3): 211-252.

[2] BROX, MALIK J. Object segmentation by long term analysis of point trajectories[C]// Proc. Eur. Conf. Comput. Vis.. 2010: 282-295.

[3] LI F, KIM T, HUMAYUN A, et al. Video segmentation by tracking many figure-ground segments[C]// Proc. IEEE Int. Conf. Comput. Vis.. 2013: 2192-2199.

[4] LI J, XIA C, CHEN X. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection [J]. IEEE Trans. Image Process., 2018, 27(1): 349-364.

[5] GALASSO F, NAGARAJA N S, CARDENAS T, et al. A unified video segmentation benchmark: Annotation, metrics and analysis[C]// Proc. IEEE ICCV. 2013: 3527-3534.

[6] LIU Z, ZHANG X, LUO S, et al. Superpixel-based spatiotemporal saliency detection[J]. IEEE TCSVT, 2014, 24(9): 1522-1540.

[7] FANG Y, WANG Z, LIN W, et al. Video saliency incorporating spatiotemporal cues and uncertainty weighting[J]. IEEE TIP, 2014, 23(9): 3910-3921.

[8] WANG L, WANG L, LU H, et al. Saliency detection with recurrent fully convolutional networks[C]// ECCV. 2016: 825-841.

[9] LIU Z, LI J, YE L, et al. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation[J]. IEEE Trans. Circuits Syst. Video Technol., 2017, PP(9): 1-17.

[10] WANG W, SHEN J, PORIKLI F. Saliency-aware geodesic video object segmentation[C]// IEEE CVPR. 2015: 3395-3402.

[11] CHENG M M, MITRA N J, HUANG X, et al. Global contrast based salient region detection[J]. IEEE TPAMI, 2015, 37(3): 569-582.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting [C]// NIPS. 2015.

[14] CONG R, LEI J, FU H, et al. Co-saliency detection for rgbd ima-

ges based on multi-constraint feature matching and cross label propagation[J]. IEEE TIP,2018,27(2):568-579.
[15] FU H,XU D,ZHANG B,et al. Object-based multiple foreground video co-segmentation via multi-state selection graph [J]. IEEE TIP,2015,24(11):3415-3424.
[16] HE K,ZHANG X,REN S,et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE TPAMI,2015,37(9):1904-1916.
[17] KOH Y J,KIM C S. Primary object segmentation in videos based on region augmentation and reduction[C]// IEEE CVPR. 2017:7417-7425.
[18] LIU Z,LI J,YE L,et al. Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation[J]. IEEE TCSVT,2017,27(12):2527-2542.
[19] WANG W,SHEN J,SHAO L. Consistent video saliency using local gradient flow optimization and global refinement[J]. IEEE TIP,2015,24(11):4185-4196.
[20] KIM H,KIM Y,SIM J Y,et al. Spatiotemporal saliency detection for video sequences based on random walk with restart[J]. IEEE Trans. Image Process. ,2015,24(8):2552-2564.
[21] CHEN C,LI S,WANG Y,et al. Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion[J]. IEEE Trans. Image Process. ,2017,26(7):3156-3170.
[22] CHENG M M,MITRA N J,HUANG X L,et al. Salient shape: group saliency in image collections[J]. The Visual Computer, 2014,30(4):443-453.
[23] FANG Y,LIN W,CHEN Z,et al. A video saliency detection model in compressed domain[J]. IEEE Trans. Circuits Syst. Video Technol. ,2014,24(1):27-38.
[24] LI G,XIE Y,WEI T,et al. Flow guided recurrent neural encoder for video salient object detection[C]// IEEE CVPR. 2018:3243-3252.
[25] ILG E,MAYER N,SAIKIA T,et al. Flownet 2. 0:Evolution of optical flow estimation with deep networks[C]// IEEE CVPR. 2017:2462-2470.
[26] WANG W,SHEN J,SHAO L. Video salient object detection via fully convolutional networks[J]. IEEE TIP,2018,27(1):38-49.
[27] SHI X,CHEN Z,WANG H,et al. Convolutional LSTM network:A machine learning approach for precipitation nowcas-ting [C]// NIPS. 2015.
[28] YANG C,ZHANG L,LU H,et al. Saliency detection via graph-based manifold ranking[C]// IEEE CVPR. 2013:3166-3173.
[29] ZHANG P,WANG D,LU H,et al. Amulet:Aggregating multi-level convolutional features for salient object detection[C]// IEEE ICCV. 2017:202-211.
[30] LE T N,SUGIMOTO A. Deeply supervised 3D recurrent FCN for salient object detection in videos[C]// BMVC. 2017:1-13.
[31] PERAZZI F,PONT-TUSET J,MCWILLIAMS B,et al. A benchmark dataset and evaluation methodology for video object segmentation[C]// Proc. CVPR. . 2016:724-732.
[32] HOU Q,CHENG M M,HU X,et al. Deeply supervised salient object detection with short connections[C]// Proc. IEEE Conf. Comput. Vis. Pattern Recognit. . 2017:5300-5309.
[33] FANG Y,WANG Z,LIN W,et al. Video saliency incorporating spatiotemporal cues and uncertainty weighting[J]. IEEE Trans. Image Process. ,2014,22(9):3910-3921.
[34] XI T,ZHAO W,WANG H,et al. Salient object detection with spatiotemporal background priors for video[J]. IEEE Trans. Image Process. ,2017,26(7):3425-3436.
[35] FAN D P,CHENG M M,LIU Y,et al. Structure-measure:A new way to evaluate foreground maps[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:4548-4557.
[36] FAN D P,GONG C,CAO Y,et al. Enhanced-alignment measure for binary foreground map evaluation[J]. arXiv:1805. 10421, 2018.
[37] FAN D P,CHENG M M,LIU J J,et al. Salient objects in clutter:Bringing salient object detection to the foreground[C]// IEEE ECCV. 2018:186-202.
[38] JIAN M,LAM K M,DONG J,et al. Visual-patch-attention-aware Saliency Detection[J]. IEEE Transactions on Cybernetics,2015,45(8):1575-1586.
[39] JIAN M,QI Q,DONG J,et al. Integrating QDWD with Pattern Distinctness and Local Contrast for Underwater Saliency Detection[J]. Journal of Visual Communication and Image Representation,2018,53:31-41.
[40] JIAN M,ZHOU Q,CUI C,et al. Assessment of Feature Fusion Strategies in Visual Attention Mechanism for Saliency Detection,Pattern Recognition Letters[OL].

WANG Jiao-jin, born in 1993, postgraduate. His main research interests include image processing and visual significance detection.

JIAN Mu-wei, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include image processing, pattern recognition, multimedia computing.