基于帧间高级特征差分的跨场景视频前景分割算法

张 锦[1,2],李 阳[2],任传伦[3],黄 炼[4],王帅辉[2],段晔鑫[1,2],潘志松[2],谢 钧[2]
(1. 陆军军事交通学院镇江校区,江苏镇江 212003; 2. 陆军工程大学指挥控制工程学院,江苏南京 210007; 3. 华北计算技术研究所,北京 100083; 4. 海军装备部驻上海地区军事代表局,上海 200129)

摘 要: 当前基于深度学习的有监督前景分割方法得益于大量待分割场景的标注信息,其性能大幅超越传统的无监督方法.然而,获取高精度的像素级标注需要耗费大量的人力和时间成本,这严重限制了有监督算法在无标注场景的部署应用.为解决对场景监督信息依赖的问题,设计了一种与传统的帧间差分法相融合的跨场景深度学习架构,即帧间高级特征差分算法.该算法重点围绕时域变化等跨场景共性知识的迁移,在不依赖待分割场景监督信息的前提下实现高精度分割.面向五类不同模式的困难场景开展实验,本文算法的平均F值达到0.8719,超越了当前最高性能的有监督算法FgSegNet_v2(相同的跨场景条件下)和最佳的无监督算法SemanticBS.本文算法对QVGA视频(320×240)的处理速度达到35帧/s,具有较好的实时性.

关键词: 前景分割; 迁移学习; 帧间差分法; 跨场景学习; 深度学习
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2021)10-2032-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200620

Cross-Scene Foreground Segmentation Algorithm Based on High-Level Feature Differencing Between Frames

ZHANG Jin[1,2], LI Yang[2], REN Chuan-lun[3], HUANG Lian[4], WANG Shuai-hui[2], DUAN Ye-xin[1,2], PAN Zhi-song[2], XIE Jun[2]
(1. *Zhenjiang Campus, Army Military Transportation University of PLA, Zhenjiang, Jiangsu* 212003, *China*;
2. *Command and Control Engineering College, Army Engineering University of PLA, Nanjing, Jiangsu* 210007, *China*;
3. *North China Institute of Computer Technology, Beijing* 100083, *China*;
4. *Shanghai Military Representative Bureau, Navy Equipment Department of PLA, Shanghai* 200129, *China*)

Abstract: Benefiting from large amounts of ground-truths of to-be-segmented scenarios, deep-learning based and supervised foreground segmentation algorithms generally outperform conventional unsupervised methods. However, pixel-wise annotation is a tedious task, especially when it comes to the annotation of foreground moving objects. This seriously limits the deployment of a supervised algorithm in a wide range of scenes without ground-truths. To address the dependence on supervised information of to-be-segmented unseen scenes, we design an inter-frame high-level feature differencing algorithm with a deep learning architecture via integrating the traditional frame differencing method. The proposed algorithm leverages the transfer of cross-scene common knowledge, such as temporal changes, so as to achieve high performance for the scene in the absence of supervised information of to-be-segmented scenes. We evaluate our method on five challenging scenes with different patterns. The average F-Measure of our algorithm is 0.8719, which surpasses the current highest-performance (supervised) algorithm (FgSegNet_v2) under the cross-scene learning condition and the best unsupervised algorithm SemanticBS. Our method which can process a QVGA (320 × 240) video at 35 frames per second shows favorable real-time performance.

Key words: foreground segmentation; transfer learning; frame differencing algorithm; cross-scene learning; deep learni

收稿日期:2020-06-28;修回日期:2020-12-03;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61806220)

1　引言

前景分割是指将视频序列中的背景和前景运动目标相分离,属于典型的像素级二分类问题,在智慧交通、智能安防等领域应用广泛. 由于视频场景往往存在动态背景、硬阴影、光照变化、摄像机抖动等诸多干扰,获取一个鲁棒的时域–空域特征表示是前景分割算法的关键.

深度卷积神经网络具有强大的特征提取能力,能够从数据中学习语义表示[1],特别是基于迁移学习的全卷积神经网络在前景分割、语义分割等像素级任务中大幅超越了传统算法[2-7]. 容易被忽视的一个重要的前提是,基于深度学习的前景分割方法(Deep-learning based Foreground Segmentation, DFS)普遍依赖大量待分割场景的监督信息. 例如FgSegNet[5,6]、Cascad-DFS[4]等高性能模型普遍使用了200帧带有高精度像素级标注(ground-truths)的样本. 然而,现实中的待分割场景通常是无标注信息且无法提前预知的,而获取待分割场景的ground-truths需要耗费大量的人力和时间成本. 此外,实验表明,高性能DFS在不使用待分割场景监督信息(仅使用非待分割场景监督信息)的情况下性能大幅下降,甚至低于一些高性能的传统无监督算法. 一个值得思考的问题由此产生:能否在不使用待分割场景ground-truths的前提下,获得一个面向待分割场景的DFS,且其精度接近或超越最佳无监督算法?

幸运的是,公开数据集[8,9]提供了大量带标注信息的视频序列. 如何利用非待分割场景的监督信息来训练得到一个面向待分割场景的高精度模型,是本文要研究的问题. 根据使用待分割场景ground-truths的数量,本文将不使用待分割场景ground-truths的问题定义为跨场景视频前景分割(Cross-Scene DFS, CS-DFS). 跨场景问题,在本质上属于迁移学习的范畴. 本文观察到,在跨场景视频序列之间存在大量的共性知识,这为迁移提供了重要的前提和基础. CS-DFS能够从大量有标注信息的非待分割场景中学到有效的特征表示,从而辅助模型在待分割场景进行前景/背景语义推断. 归纳起来,跨场景视频序列之间通常存在以下几个方面的共性知识.

(1)前景共性知识. 视频场景中的前景目标主要由人、车、船等可运动目标组成,不同场景中的前景目标具有较强的特征相关性.

(2)背景共性知识. 不同视频场景中的背景类往往由具有较高相似性的房屋、道路等固定物组成.

(3)噪声和干扰共性知识. 不同视频场景中的阴影、运动背景(晃动的水面、摇动的树枝)等干扰同样具有相似的特点.

(4)帧间时域变化共性知识. 这具体表现为视频中前后帧(不必相邻)之间存在前景目标的对应区域会发生改变,比如前景目标移动位置、改变姿态等,而没有变化的区域一般为背景. 这种时域变化提供了十分重要的先验信息:两帧之间保持时域稳定的区域一般为背景;而在发生变化的区域,至少某一帧在该区域有前景目标. 这对于算法性能的提升效益明显.

然而,大部分DFS[5-9]算法都忽略了跨场景共性知识. 有的DFS[3,10]虽然利用了跨场景共性知识,但仅利用了前3项共性知识,且未探讨跨场景条件下的算法性能. 而实验表明(见4.2.1节),仅对前3项共性知识进行迁移,在跨场景条件下难以超越先进的无监督算法. 本文系首次围绕如何有效利用帧间时域变化共性知识而开展的研究.

本文主要贡献如下:

(1)首次从帧间时域变化共性知识迁移的角度探讨跨场景视频前景分割问题,并通过实验验证了时域变化共性知识对于算法性能提升的巨大增益;

(2)围绕时域变化共性知识的利用,提出了3种典型的跨场景前景分割架构,并特别将传统帧间差分法的思想融入深度学习框架,提出了双流结构的高级特征差分对比架构;

(3)在多种不同类型困难场景下的实验表明,本文算法优于当前最高性能的DFS(在跨场景条件下)和最佳无监督学习方法.

2　相关工作

前景分割也叫背景减除,是计算机视觉领域的经典研究问题之一. 在过去30年间,学者们提出了众多前景分割算法[11]. 根据是否利用标注信息,这些算法可被分为2类,即深度有监督前景分割算法和无监督前景分割算法.

2.1　深度有监督前景分割算法

深度学习方法能够有效提取高等级特征表示. 特别是具备平移不变性的卷积神经网络,它能够有效解决动态背景难题,已成为研究前景分割问题的重要基石.

2016年,Braham[12]等将改进后的LeNet应用于前景分割,首次提出了基于深度学习的前景分割算法. 它以27×27的像素块(Patch)为输入来预测Patch中心位置的像素属于前景的概率. Patch-DFS模型较小,易于训练,但固定的Patch尺寸限制了高等级特征的提取能力.

为克服这一固有缺陷,以便有效利用图像级空域信息,基于全卷积网络的图像级模型(Image-DFS)被提出,比如近年出现的具有单流、编码-解码结构的FgSegNet_v2[6]和FgSegNet_S[5]模型. 它们以完整分辨率的图

像为输入来预测整幅图像的前景概率掩膜(mask),在基于200帧ground-truths的基础上获得了CDnet2014数据集的最高精度.尽管该网络结构简单,性能优异,但仍然存在以下2点不足:①将视频分割当成图像分割问题来处理,忽略了视频序列间的时序相关性;②针对每个场景独立训练前景分割模型,未能利用场景之间的共性知识.

利用3D卷积来提取时空特征,是视频前景分割算法的另一种典型框架.3D-DFS[3]将数据集所有场景的样本整合在一起训练,以此利用跨场景共性知识,最终得到一个通用的前景分割模型.该模型用到了所有场景70%的ground-truths,虽然得到了较高的精度,但未能超越FgSegNet.它以连续9帧为输入(其中8帧为参考帧,1帧为目标帧),仅产生目标帧的前景mask,因而模型训练和推断效率偏低.此外,相关文献没有分析该模型在跨场景条件下的性能,也未探讨帧间间隔对模型性能的影响.

上述有监督DFS虽然在数据集中取得了较高的精度,但是普遍存在2点不足:一是用到了大量待分割场景的groud-truths;二是所选择的训练样本和测试样本在序列空间中存在交错(比如从整个视频序中按相对均匀的间隔挑选训练样本,而不是连续选取选前p%作为训练集),这导致测试集中的前景目标很可能在训练样本中"出现过".而现实应用中,待分割场景是无法提前预知的,前景目标更不太可能"提前"出现在训练样本中.相比之下,CS-DFS虽然同属于有监督方法,但是其和无监督方法一样,都具有不依赖待分割场景监督信息的特点,适用性更强,也更具有研究价值.

2.2 无监督前景分割算法

混合高斯模型(Gaussian Mixture Model)是最经典的背景减除方法.它利用多个高斯分布来估计背景颜色随时间产生的变化[13].但是,它作为一种像素级模型,无法利用整个图像的空域信息.相反,鲁棒主成分分析法(Robust Principal Component Analysis)能够利用整个视频序列的时域、空域信息将具有低秩特性的背景和稀疏特性的前景相分离[14~16].不过,该算法对运动的硬阴影比较敏感.另外,它的求解过程包含对大规模矩阵进行SVD分解,这会导致运算效率偏低.帧间差分法(Frame Differencing, FD)是一种快速高效的方法,它通过前后帧中对应像素值的差分来估计运动的前景目标[17].尽管该方法能够利用时序信息,但在处理动态背景时不够鲁棒.

由于大多数传统的无监督算法仅利用了颜色信息,难以应对光照变化等复杂场景.作为改进,Bilodeau等人[18]引入局部二值相似模式来比较当前帧和背景的时空特征相似度.PAWCS[19]和SuBSENSE[20]分别将颜色特征和纹理特征与局部二值相似模式相结合,从而进一步提升模型鲁棒性.近来,利用深度学习来辅助传统前景分割模型的做法也逐渐受到关注[21~23].Braham等人[21]利用语义分割模型产生语义mask作为先验信息来降低传统方法的误检率,取得了当前CDnet数据集(无监督方法)的最高精度.Sultana等人[22]利用光流法产生运动mask,再利用生成对抗网络来补全mask中的运动区域,生成无前景目标的完整背景帧,然后采用类似FD的做法,将当前帧和估计的背景帧做差分得到前景mask.

尽管部分无监督方法利用了深度学习(语义分割)技术,但它们仅仅将深度学习用于数据预处理,其无监督型的主体架构没有改变.本文的不同之处在于,算法的主框架是深度卷积网络,只是在网络架构的设计中融入了传统FD算法的思想.另外,相比于目前的有监督DFS,本文算法能够高效利用时域变化的特征.相关网络每次接受2帧输入,并同时产生2个前景mask,结构简洁、运行高效.

3 本文方法

对$n(n\geqslant 2)$帧相邻图像进行特征提取和融合的方式,是设计具有时域变化共性知识学习能力的DFS的核心.假设n帧之间所包含的时域变化信息的总量等于C_n^2个2帧之间时序变化信息量的叠加.这意味着多次选取2帧进行训练能够达到同时使用n帧$(n\geqslant 2)$训练的效果.而当$n>2$时,网络架构空间复杂度的增加会导致训练和运行效率偏低,且不利于泛化能力的提升.因此,本文算法主要考虑在$n=2$的情况下,进行DFS网络设计.

3.1 原理描述

图1展示了时域变化共性知识为前/背景类别判断提供的重要的先验信息.对于视频中的前后帧(不必相邻),时域变化共性知识能够用于检测2帧间的差异,进而将整个图像空间$\mathcal{R}$划分为发生变化区域($\mathcal{C}$)和稳定的静态区域($\mathcal{S}$),即

$$\mathcal{R}=\mathcal{C}+\mathcal{S} \tag{1}$$

其中,$\mathcal{S}$通常在2帧中均为背景,对应图1(c)黑色区域;而$\mathcal{C}$在2帧中的类别可以根据前/背景共性知识决定.

如图1所示,$\mathcal{C}=\mathcal{C}_1\cup\mathcal{C}_2\cup\mathcal{C}_3$.对于前一帧(图1($a$)),对应的$\mathcal{C}_2\cup\mathcal{C}_3$区域的语义是车辆(属于典型的前景共性知识),因而应当归为前景类,而$\mathcal{C}_1$对应的道路则可基于背景共性知识作出判断.在后一帧(图1(b))中,$\mathcal{C}_1\cup\mathcal{C}_2$属于前景,而$\mathcal{C}_3$是背景.本节接下来先探讨几种典型的可利用时域变化共性知识的网络架构,然后基于架构给出具体的实例化网络,最后对网络训练进行介绍.

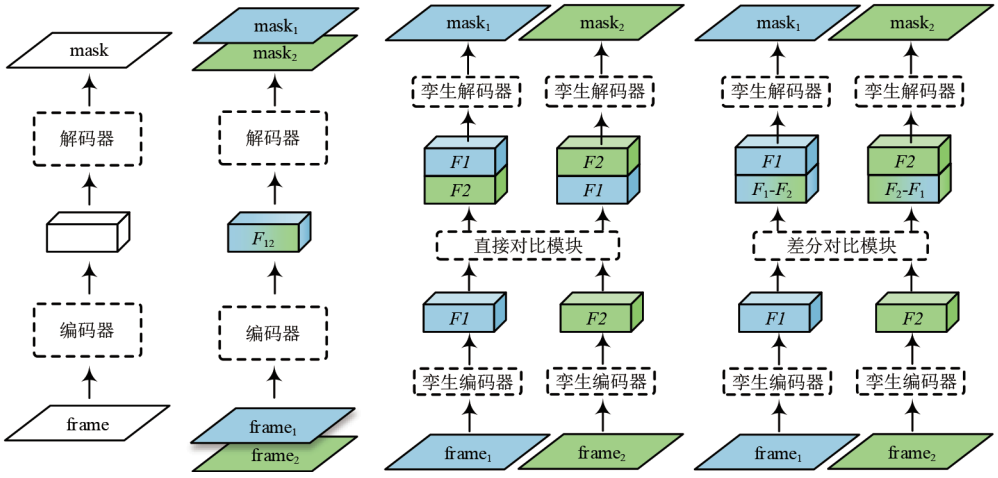
(*a*)前一帧　(*b*)后一帧　(*c*)NON-ROI为不感兴趣区域

图1　帧间时域变化示意图

型架构.由于每次只对单帧进行处理,该架构无法利用时域变化共性知识.考虑到单流架构和双流架构均能实现一次处理2帧,基于我们前期的相关研究[24],本文提出2种双流架构(图2(*c*)和图(*d*)),以及一种单流架构(图2(*b*)).不同的是,单流架构和双流架构分别采用低级特征聚合和高级特征聚合的方式来利用时域变化的共性知识.本文最终选取的是双流差分对比架构(图2(*d*)).

3.2　网络架构

FgSegNet_v2等有监督DFS[5~7]采用了图2(*a*)



(*a*)单流单帧型架构　(*b*)单流双帧型架构　(*c*)双流直接对比架构　(*d*)双流差分对比架构

图2　DFS架构示意图

3.2.1　单流架构

单流架构由编码器和解码器组成(图2(*b*)).编码器以同一场景的2帧沿通道维度聚合产生的低级时域对比特征作为输入,输入维度为$H\times W\times 6$(H/W表示帧高/宽,使用BGR彩色模式).在编码器中,先通过卷积操作对低级时域对比特征进行时域–空域特征编码,然后经解码器产生相同分辨率的前景概率图(维度为$H\times W\times 2$).最后,以阈值为0.5进行二值化操作后得到2幅前景mask.

3.2.2　双流架构

双流架构包括2条分支,由编码器、解码器和时域特征对比模块组成(图2(*c*)和图(*d*)).不同分支的编码器/解码器采用孪生结构,即网络结构相同、参数共享,以此降低一半网络参数,防止过拟合.下面以分支1为例进行介绍.编码器对frame$_1$进行特征编码,获取高级特征映射F_1.而时域特征对比模块则用于产生高级时域对比特征F_{1c},因为在有"对比"的情况下更容易获取有鉴别力的特征表示,提升解码精准度.本文考虑2种对比模式,即直接对比和差分对比.

$$F_{1c}=F_1\oplus F_2 \tag{2}$$

$$F_{1c}=F_1\oplus\left(F_1-F_2\right) \tag{3}$$

其中,$\oplus$表示特征沿通道维度聚合(concatenate);式(2)采用直接对比模式(图2(*c*)),直接使用F_2作为F_1的"对比"特征;差分对比模式(图2(*d*))则使用F_1和F_2的差分作为对比特征,见式(3)."对比"的目的是让解码器更好地区分变化区域($\mathcal{C}$)和静态区域($\mathcal{S}$)(图1),以提升解码准确度.由于$\left(F_1-F_2\right)$是frame$_1$相对于frame$_2$是否发生变化的直接表征,这一"直截了当"的表达形式能够降低网络学习时域变化模式的难度,有利于模型性能的提升.

不同于使用低级像素特征的FD算法,本文算法使用的是高级特征.高维的高级特征更有利于表征前/背景目标的本质,不易受噪声干扰,因而性能更鲁棒.

3.3　网络实例化

上述3种架构具有较好的灵活性,方便与当前高性能的DFS网络进行适配,实例化为具体的CS-DFS网络.本文选择结构简洁、单流且无分支的FgSegNet_S[5]作为实例化基准.它是CDnet2014数据集上排名第2的算法,其性能和排名第1的算法非常接近但结构更简洁优雅.它的网络架构由编码器和解码器组成,一次处理单帧图像并生成前景mask(图3).

单流架构的实例化网络(记为CS-DFS$_1$)和原始的

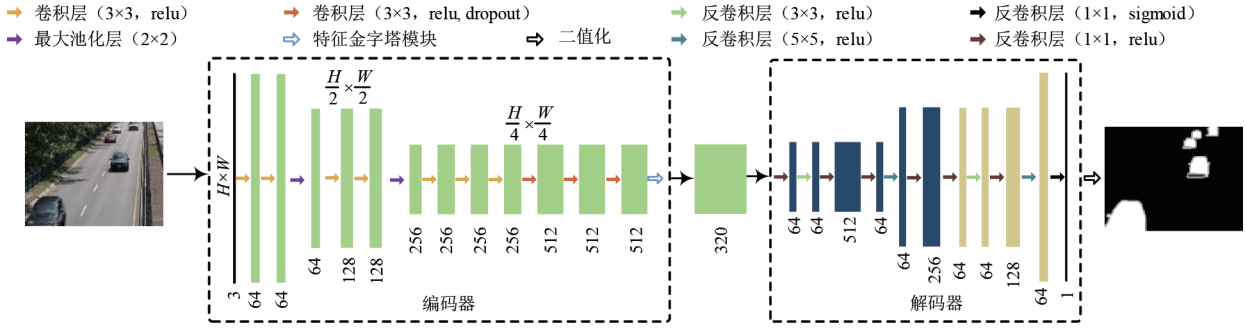


图3 CS-DFS网络实例化基准(FgSegNet_S网络结构,矩形表示特征映射,数字表示特征通道数)

FgSegNet_S中间层结构完全一致,仅输入层和输出层存在区别(图2(*b*)). FgSegNet_S输入为3通道(单帧),输出1幅前景mask,而CS-DFS$_1$输入为6通道(双帧),输出2幅前景mask. 双流架构网络(直接对比和差分对比2种模式的实例化网络分别简记为CS-DFS$_2$和CS-DFS$_3$)的每条分支相比原始FgSegNet_S增加了直接/差分对比模块,其他层(包括输入输出层)完全一致.

FgSegNet_S是单流网络,每次只对单帧进行分割,具体结构(图3)如下:编码器由卷积层、最大池化层、dropout层以及特征金字塔模块(Feature Pyramid Module, FPM)组成. 其中,FPM中采用不同膨胀率的空洞卷积[25,26]来提取多尺度特征,具体细节可参照文献[5]. 解码器主要由反卷积层和二值化操作层等组成. 图3中矩形块表示特征映射,数字代表特征通道的数量. 编码器通过不断降低分辨率同时增加特征映射的数量来实现高等级语义编码. 解码是编码的逆操作,解码器通过不断减小特征映射的数量同时提升分辨率来实现前景/背景语义解码. 解码器最后的输出层采用Sigmoid激活函数将特征映射到0到1之间,以此估计对应位置像素属于前景的概率值. 最后,经二值化操作(阈值0.5)得到前景mask.

总体而言,单流网络(图2(*b*))和双流网络(图2(*c*)和图(*d*))采用不同的工作机制. 单流网络采用时空融合的方式,网络中的每个卷积层都在同时处理(两帧的)时域和空域信息. 双流网络的编码器只对单帧的空域信息进行编码,然后通过显式的产生对比特征的方式来聚合时域信息,最后在解码器中进行时域–空域特征融合并解码.

3.4 网络训练

现实场景往往存在严重的前景/背景分布不均衡问题[5],这容易导致模型在训练时过多地"关注"数量上占主导的背景像素,而忽视前景特征的学习. 为抑制大量易学习的背景样本产生的损失,防止分类器被误导,本文采用Focal loss[27]作为损失函数,具体定义如下:

$$\text{focal loss}=\frac{1}{N}\sum_{i\in P}\left(\omega_i\cdot \text{CE}_i\right) \tag{4}$$

$$\text{CE}_i=-y_i\log\left(\hat{y}_i\right) \tag{5}$$

$$\omega_i=\left|y_i-\hat{y}_i\right|^2,\ y_i\in[0,1] \tag{6}$$

其中,CE表示交叉熵损失;P是有效像素i的集合,大小为N;y_i表示像素i的类别,前/背景分别用1/0表示;$\hat{y}_i\in(0,1)$为像素属于前景的预测概率;$\hat{y}_i$为调节因子,当预测误差较大,即预测值$\hat{y}_i$远离真实值y_i时,$\hat{y}_i$接近1,focal loss退化为标准的交叉熵损失,反传的梯度几乎不受影响,当预测误差较小时,$\hat{y}_i$趋向于0,易学习的背景样本产生的损失被大大降低. 通过这一损失调节机制,防止大量背景像素产生的小梯度的累积误导分类器.

本文基于keras深度学习框架,采用Adam优化器开展模型训练,进行批次梯度更新,批量大小设为10. 初始学习率设为10^4,基于动态学习率策略,每轮(epoch)迭代后,学习率按以下规律衰减,计算方式如下:

$$\text{LR}=0.0001\times 0.95^{\text{epoch}/10} \tag{7}$$

需要强调的是,模型训练时只对感兴趣区域(Region Of Interest, ROI)的有效像素产生的损失进行梯度反传. 另外,本文算法通过增加L2正则化项(系数为5e-4)来降低模型的结构性风险. 对于双帧输入型网络,帧间间隔G对时域变化共性知识的学习影响较大. 训练时在[0.5G, 1.5G]之间随机确定帧间隔来挑选一对训练帧,测试时G保持不变,一次输出两帧的前景mask.

4 实验及结果分析

选取5种不同类型的困难场景对算法性能进行综合评估,通过消融实验以及对比试验,验证本文算法的有效性和先进性.

4.1 实验说明

4.1.1 数据集构建

对CS-DFS进行训练需要用到大量有标注场景的视频序列. 本文融合2个大规模前景分割数据集来构建一个综合的训练、测试环境. CDnet数据集[8]是当前规模最大、模式类型最全的前景分割数据集,共计53个场景. 本文选取其中5种具有不同类型挑战的场景类,

并根据先进的无监督方法IURIS-5在些场景类中的表现,确定其中精度最低的5个场景(称为困难场景)作为测试集(表1). 其余48个场景加上SBI2015数据集[9]中的14个场景,共计62个有标注场景视频序列作为训练集. 为统一不同场景的视频尺度,通过补零和随机采样,分辨率统一为320×240. 模型训练迭代1000轮,每轮迭代在每个训练场景中随机选择5对视频帧. 为进一步提升模型的泛化性,本文通过仿射变换、饱和度和亮度调整以及添加高斯噪声等方式进行数据增广.

表1　实验场景说明

实验视频	大小×帧数	描述
PETS.	[576×720]×1200	室内,弱阴影,小尺度前景
boulev.	[240×352]×2500	室外,相机抖动
boats	[240×320]×7999	室外,动态背景
bungal.	[240×360]×1700	室外,硬阴影,大尺度前景
park	[288×352]×600	室外,红外视频,前/背景灰度相似度大

4.1.2　评价指标

在前景分割领域,常用的评价指标有召回率(R)、准确率(P)、特异度(Sp)、假阳率(FPR)、假阴率(FNR)、误检率(PWC)和F-Measure(F值). 由于F值能综合评估模型的总体性能,并且与CDnet数据集算法官方排名相关度最高,常被用作算法性能评价最重要的指标之一,因此本文主要根据F值对不同算法的性能进行比较.

$$F = 2RP/(R+P) \tag{8}$$

$$R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{9}$$

$$P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{10}$$

其中,TP和FP分别表示正确的正样本和错误的正样本;FN表示错误的负样本.

4.2　实验结果及分析

4.2.1　消融实验

本实验对提出的3种架构进行效果验证. 所有实验在相同的设置环境下进行,实验结果如表2所示. 差分对比网络CS-DFS$_3$具有最佳性能,F值比直接对比网络高出5个点以上. 高级特征直接对比网络CS-DFS$_2$和低级特征直接对比网络CS-DFS$_1$总体性能接近. 就具体场景而言,在具有动态背景的场景中(boats),CS-DFS$_3$和CS-DFS$_2$的F值大大超过CS-DFS$_1$. 这是因为前两者是基于高级特征对比的方式来学习时域变化共性知识,CS-DFS$_1$利用的是低级特征. 高级特征具有更好的语义鲁棒性,抗动态背景干扰的能力更强. 此外,CS-DFS$_1$以完整分辨率的对比特征为输入,而CS-DFS$_3$和CS-DFS$_2$利用的对比特征的分辨率是原始图像的四分之一. 分辨率降低造成的空间信息损失会带来不利影响,这在小尺度目标场景表现最为明显. 这也是CS-DFS$_3$在PETS和park两个前景目标较小的场景中的精度略低于CS-DFS$_1$的原因.

表2　不同网络架构的性能对比

Methods	Metrics	Overall	PETS.	boulev.	boats	bungal.	park
CS-DFS$_3$	R	0.8053	0.9221	0.8303	0.6901	0.7523	0.8318
	Sp	0.9985	0.9995	0.9992	0.9998	0.9982	0.9958
	P	0.9346	0.9593	0.9809	0.9628	0.9641	0.8060
	FPN	0.0015	0.0005	0.0008	0.0002	0.0018	0.0042
	FNR	0.1947	0.0779	0.1697	0.3099	0.2477	0.1682
	PWC	0.7344	0.1521	0.8725	0.2111	1.6880	0.7484
CS-DFS$_3$	F	0.8719	0.9334	0.8452	0.9472	0.8879	0.7457
CS-DFS$_2$	F	0.8039	0.9307	0.7530	0.8202	0.7844	0.7311
CS-DFS$_1$	F	0.8155	0.9514	0.8183	0.6660	0.8767	0.7651
CS-DFS$_0$	F	0.6786	0.9005	0.6175	0.4679	0.7218	0.6852

CS-DFS$_0$指相同配置下的单流单帧型网络(图2(a)). 尽管无法学习时域变化共性知识,但其在5个场景的平均F值也达到了0.6786,特别是在光照条件较好、前景/背景区分度明显的PETS场景,能较好地克服弱阴影的干扰,获得了0.9005的F值. 这表明前景、背景、干扰3方面共性知识能够提供重要的前/背景预测先验知识. 然而,CS-DFS$_0$的精度相比CS-DFS$_3$下降了近0.2,这也证明了时域变化共性知识对于算法性能提升的巨大效用.

图4为不同架构算法对同时输入网络的相隔300帧的2幅图像的处理效果. 由于图中白色车辆和路旁白色街道特征接近,不采用特征对比的方式(如CS-DFS$_0$)容易将其误判为背景. 采用差分对比的架构的CS-DFS$_3$能够显式地进行2帧对比分析,因而取得了最佳的检测效果.

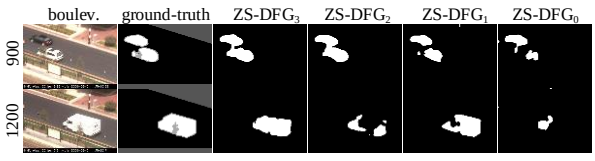


图4　不同架构网络的分割效果示意图(前/后帧序号分别为900/1200;NON-ROI为不感兴趣区域)

4.2.2　帧间间隔试验

本实验在CS-DFS$_3$的基础上进行. CS-DFS$_3$作为成对输入网络,2帧间间隔决定了前景运动目标的变化程度,是关乎时域变化共性知识学习的关键性指标.

如图5(a)所示,在训练帧间隔G_{train}等于测试帧间隔G_{test}的情况下,间隔越大,模型的精度越高. 当G=2时,间隔太小导致前后帧变化不明显,因而难以有效学习时域变化共性知识. 此时CS-DFS$_3$的F值降为0.6712,与未利用时域变化信息的算法(CS-DFS$_0$)接近.

如图5(b)所示,当G_{train}=300时,减少G_{test}会导致模

型性能下降. 同样,当G_{train}=5时,G_{test}太大或太小都会造成网络性能下降,只有二者接近时性能最优. 这是因为当测试帧间隔和训练帧间隔接近时,两种情况下的时域变化特征分布相同,有助于模型泛化性能的提升.

以上试验均采用一次输出2帧的"对比"工作模式. 除此之外,本文还尝试了每次只输出1帧的"背景减除"工作模式,具体做法如下:首先对待分割视频的前500帧做时域中值滤波得到"纯净"的背景帧,然后固定$frame_1$(图2(*d*))为该背景帧,$frame_2$为每次输入的待分割视频帧. 在此情况下,CS-DFS$_3$平均F值提升为0. 8765. 尽管精度有了微小提升,但是其带来的缺点是运行效率减半(18帧/s).

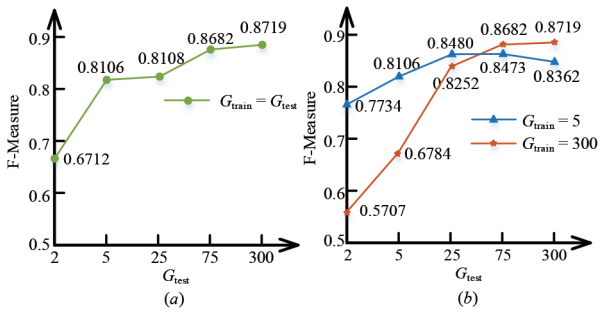


图5 帧间隔对算法性能影响示意图

4.2.3 性能对比及分析

当前有监督DFS算法大都使用了待分割场景的ground-truths. 为了比较的公平性,本文将CDnet排名第1的算法FgSegNet_v2[6]放在相同的跨场景条件下开展实验. FgSegNet_v2(采用图2(*a*)的架构)由于无法利用时域变化共性知识,其综合性能仅为0. 7010,和CS-DFS$_0$算法接近,远低于本文算法CS-DFS$_3$.

SemanticBS[21]、IURIS-5[28]、PAWCS[19]、STBM[29]和MBS[30]是CDnet数据集[8]中排名最优的无监督算法. 需要强调的是,与另外4个方法纯无监督方法不同,SemanticBS以有监督语义分割网络(Pyramid Scene Parsing Network, PSPNet)产生的语义概率映射来辅助前景分割. 该算法对于常见的运动语义类别(人、船、车)场景(如PETS、boats、bungal),语义概率映射提供的先验信息效用明显. 算法中的时域判断模块对相机抖动的敏感性造成在场景boulev中的F值偏低. PSPNet是在RGB图像上进行的预训练,因而该算法对红外场景(park)的优势不明显. 此外,STBM算法不仅利用了RGB信息,还融合了区域级纹理特征,而纹理特征更有利于对红外图像的特征辨识,故该算法在park场景中精度最高. 本文算法在5个场景中的平均精度最高. 图6中的效果图进一步展示了本文算法和这些最先进的无监督算法在5个场景上的前景分割视觉效果. 在PETS场景,CS-DFS$_3$是唯一检测出有前景目标即将从左侧进入画面的算法. 在bungal场景中,本文算法和SemanticBS能够较好地排除硬阴影的干扰,而其他算法受硬阴影影响严重. 在boulev场景中,本文算法的漏检和错检相对较少. 在park场景中,相较于其他算法,本文算法能够更清晰地检测出人物轮廓. 总体而言,本文算法在5个场景上取得了最佳的检测效果,总体性能略高于排名第1的无监督算法(SemanticBS),具体指标如表3所示.

速度方面,DFS类算法可以通过GPU并行计算,因而其运行速度通常高于无监督算法. 基于Nvidia Titan XP GPU,CS-DFS$_3$算法对QVGA视频(320×240)的视频处理速度能达到35帧/s,实时性较好.

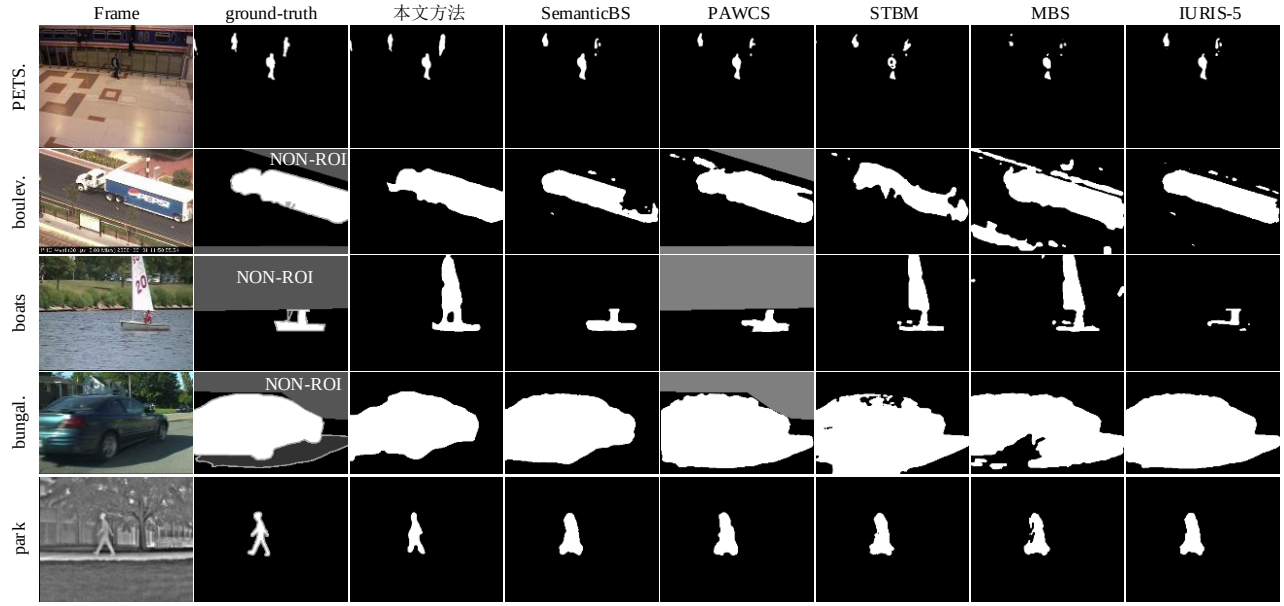


图6 本文方法和高性能无监督算法的前景分割视觉效果图(NON-ROI表示不感兴趣区域)

表3　本文算法和无监督算法的性能F值对比

场景	PETS	boulev	boats	bungal	park	Overall	Speed
IURIS-5[28]	0. 9354	0. 7680	0. 7532	0. 8392	0. 7652	0. 8122	-
MBS[30]	0. 8648	0. 8672	0. 9041	0. 7475	0. 7099	0. 8187	-
STBM[29]	0. 8610	0. 7423	0. 8760	0. 7927	0. 8323	0. 8208	12
PAWCS[19]	0. 9315	0. 8444	0. 8416	0. 8387	0. 8286	0. 8570	-
SemanticBS[21]	0. 9442	0. 7372	0. 9795	0. 9259	0. 7694	0. 8712	**7**
FgSegNet_v2[6]	0. 8833	0. 7313	0. 5581	0. 7896	0. 5426	0. 7010	**38**
CS-DFS$_3$(本文)	0. 9334	0. 8452	0. 9472	0. 8879	0. 7457	0. 8719	35

5　结论

本文利用跨场景共性知识迁移机制,针对跨场景条件下的视频前景分割问题开展研究,重点围绕时域变化共性知识的学习,提出了3种高效、灵活的跨场景前景分割架构,在基于FgSegNet_S的基础上进行了实例化. 实例化网络在5种不同模式的困难场景下进行实验,验证了时域变化共性知识能够大幅提升跨场景前景分割算法的性能. 其中,融合传统帧间差分法思想的高级特征差分对比网络CS-DFS$_3$在5个场景的平均F值为0. 8719,超越了最佳的有监督算法(跨场景条件下)和高性能的无监督前景分割算法. CS-DFS$_3$的分割速度(35帧/s)能够满足实时性要求. 如何通过自监督学习[29, 30]策略来有效利用待分割场景的无标注视频帧信息,以进一步提升算法精度,是下一步研究的目标.

参考文献

[1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[A]. Proceedings of the IEEE[C]. USA: IEEE, 1998. 2278 – 2324.

[2] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [C]. Long Beach, CA, USA: IEEE, 2019. 3141 – 3149.

[3] Sakkos D, Liu H, Han J G, et al. End-to-end video background subtraction with 3d convolutional neural networks [J]. Multimedia Tools and Applications, 2018, 77(17): 23023 – 23041.

[4] Wang Y, Luo Z M, Jodoin P M. Interactive deep learning method for segmenting moving objects[J]. Pattern Recognition Letters, 2017, 96: 66 – 75.

[5] Lim L A, Yalim Keles H. Foreground segmentation using convolutional neural networks for multiscale feature encoding[J]. Pattern Recognition Letters, 2018, 112: 256 – 262.

[6] Lim L A, Keles H Y. Learning multi-scale features for foreground segmentation[J]. Pattern Analysis and Applications, 2020, 23(3): 1369 – 1380.

[7] Babaee M, Dinh D T, Rigoll G. A deep convolutional neural network for video sequence background subtraction[J]. Pattern Recognition, 2018, 76: 635 – 649.

[8] Wang Y, Jodoin P M, Porikli F, et al. CDnet 2014: An expanded change detection benchmark dataset[A]. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops[C]. Columbus, OH, USA: IEEE, 2014. 393 – 400.

[9] Maddalena L, Petrosino A. Towards benchmarking scene background initialization[A]. New Trends in Image Analysis and Processing – ICIAP 2015 Workshops[C]. Cham, GER: Springer, 2015. 469 – 476. DOI: 10.1007/978 – 3 – 319 – 23222 – 5_57.

[10] Mandal M, Dhar V, Mishra A, et al. 3DFR: A swift 3D feature reductionist framework for scene independent change detection[J]. IEEE Signal Processing Letters, 2019, 26(12): 1882 – 1886.

[11] Bouwmans T, Javed S, Sultana M, et al. Deep neural network concepts for background subtraction: A systematic review and comparative evaluation[J]. Neural Networks, 2019, 117: 8 – 66.

[12] Braham M, Van Droogenbroeck M. Deep background subtraction with scene-specific convolutional neural networks[A]. 2016 International Conference on Systems, Signals and Image Processing (IWSSIP)[C]. Bratislava, Slovakia: IEEE, 2016. 1 – 4.

[13] Chen M L, Wei X, Yang Q X, et al. Spatiotemporal GMM for background subtraction with superpixel hierarchy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1518 – 1525.

[14] Shi G M, Huang T, Dong W S, et al. Robust foreground estimation via structured Gaussian scale mixture modeling [J]. IEEE Transactions on Image Processing, 2018, 27 (10): 4810 – 4824.

[15] 常侃, 张智勇, 陈诚, 等. 采用低秩与加权稀疏分解的视频前景检测算法[J]. 电子学报, 2017, 45(9): 2272 – 2280. Chang K, Zhang Z Y, Chen C, et al. Video foreground detection by low-rank and reweighted sparse decomposition [J]. Acta Electronica Sinica, 2017, 45(9): 2272 – 2280. (in

Chinese)
[16] 秦晓燕, 袁广林, 李从利, 等. 一种快速鲁棒的视频序列运动目标检测方法[J]. 电子学报, 2017, 45(10): 2355 – 2361. Qin X Y, Yuan G L, Li C L, et al. An approach to fast and robust detecting of moving target in video sequences[J]. Acta Electronica Sinica, 2017, 45(10): 2355 – 2361. (in Chinese)
[17] Paul N, Singh A, Midya A, et al. Moving object detection using modified temporal differencing and local fuzzy thresholding[J]. The Journal of Supercomputing, 2017, 73(3): 1120 – 1139.
[18] Bilodeau G A, Jodoin J P, Saunier N. Change detection in feature space using local binary similarity patterns[A]. 2013 International Conference on Computer and Robot Vision[C]. Regina, SK, Canada: IEEE, 2013. 106 – 112.
[19] St-Charles P L, Bilodeau G A, Bergevin R. A self-adjusting approach to change detection based on background word consensus[A]. 2015 IEEE Winter Conference on Applications of Computer Vision[C]. Waikoloa, HI, USA: IEEE, 2015. 990 – 997.
[20] St-Charles P L, Bilodeau G A, Bergevin R. SuBSENSE: A universal change detection method with local adaptive sensitivity[J]. IEEE Transactions on Image Processing, 2015, 24(1): 359 – 373.
[21] Braham M, Piérard S, Van Droogenbroeck M. Semantic background subtraction[A]. 2017 IEEE International Conference on Image Processing (ICIP)[C]. Beijing, China: IEEE, 2017. 4552 – 4556.
[22] Sultana M, Mahmood A, Javed S, et al. Unsupervised deep context prediction for background estimation and foreground segmentation[J]. Machine Vision and Applications, 2019, 30(3): 375 – 395.
[23] Zeng D D, Zhu M, Kuijper A. Combining background subtraction algorithms with convolutional neural network[J]. Journal of Electronic Imaging, 2019, 28(1): 013011.
[24] Zhang J, Li Y, Chen F Q, et al. X-net: A binocular summation network for foreground segmentation[J]. IEEE Access, 2019, 7: 71412 – 71422.
[25] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[A]. European Conference on Computer Vision[C]. Cham, GER: Springer, 2018. 833 – 851. DOI: 10.1007/978 – 3 – 030 – 01234 – 2_49.
[26] Yang M K, Yu K, Zhang C, et al. DenseASPP for semantic segmentation in street scenes[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT, USA: IEEE, 2018. 3684 – 3692.
[27] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[A]. 2017 IEEE International Conference on Computer Vision (ICCV)[C]. Venice, Italy: IEEE, 2017. 2999 – 3007.
[28] Bianco S, Ciocca G, Schettini R. Combination of video change detection algorithms by genetic programming[J]. IEEE Transactions on Evolutionary Computation, 2017, 21(6): 914 – 928.
[29] Chen M L, Yang Q X, Li Q, et al. Spatiotemporal background subtraction using minimum spanning tree and optical flow[A]. European Conference on Computer Vision[C]. Cham, GER: Springer, 2014. 521 – 534. DOI:10.1007/978 – 3 – 319 – 10584 – 0_34.
[30] Sajid H, Cheung S C S. Universal multimode background subtraction[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3249 – 3260.

作者简介

张 锦 男, 1985 年生, 湖南邵东人. 2009 年获海军工程大学核科学与技术专业硕士学位, 其后在陆军军事交通学院镇江校区工作. 目前在陆军工程大学攻读博士学位. 研究方向为人工智能、图像处理.
E-mail:zhang_jin_1115@163.com

李 阳 男, 1984 年生, 河北廊坊人. 2007 年、2010 年、2018 年分别在北京航空航天大学、解放军理工大学、陆军工程大学获学士、硕士和博士学位. 现为陆军工程大学讲师, 主要研究方向为人工智能、机器视觉与图像检索.
E-mail:solarleeon@outlook.com

潘志松(通信作者) 男, 1973 年生, 江苏南京人. 2003 年获南京航空航天大学博士学位. 现为陆军工程大学教授、博士生导师. 主要研究方向为人工智能、模式识别.
E-mail:panzs@nuaa.edu.cn

谢 钧 男, 1973 年生, 四川成都人. 2005 年获南京大学博士学位. 现为陆军工程大学教授、博士生导师. 主要研究方向为智能信息处理、无线网络.
E-mail:xiejun73@189.cn