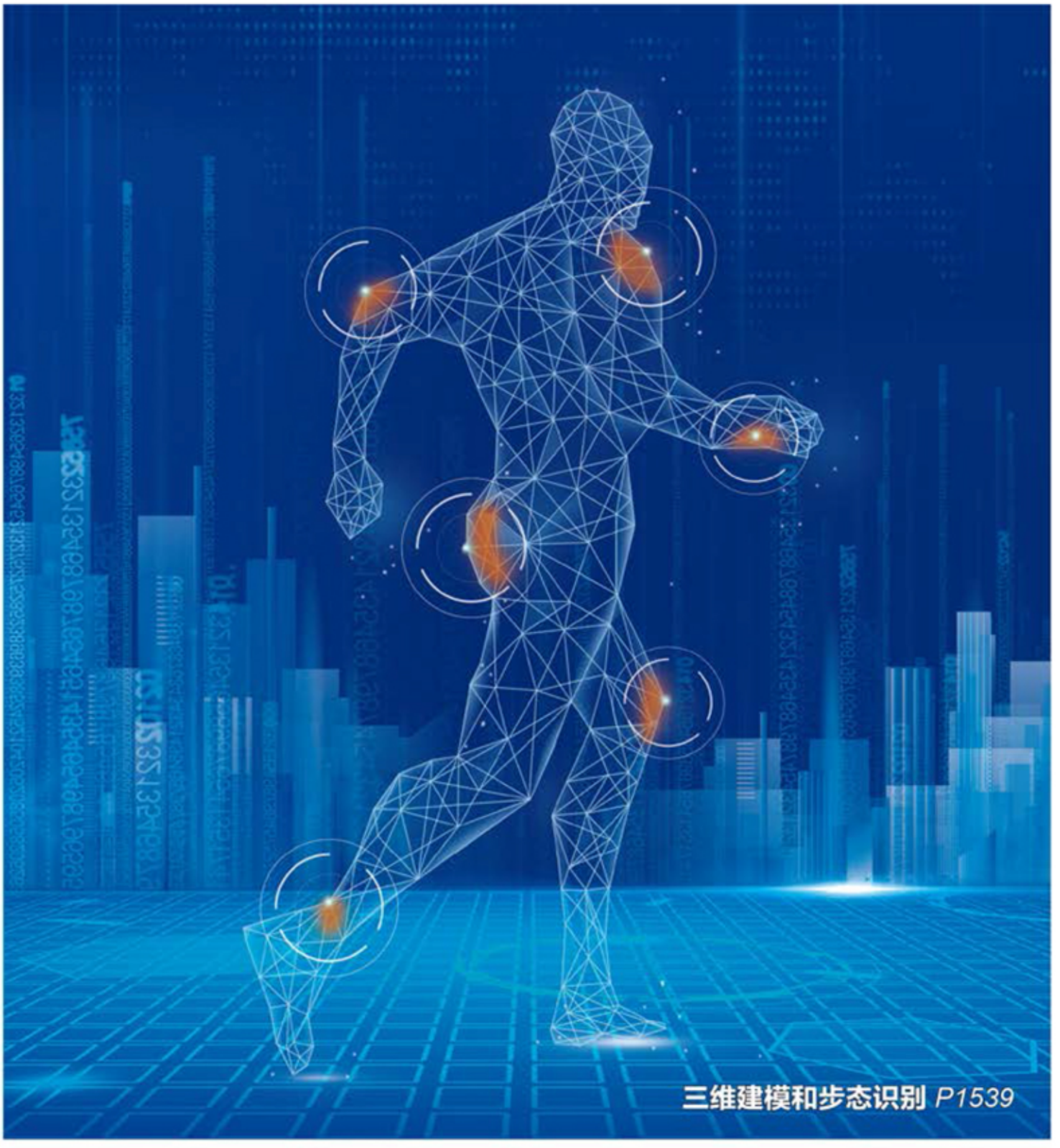
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
08
VOL.25

ISSN1006-8961
CN11-3758/TB



三维建模和步态识别 *P1539*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第8期（总第292期）
2020年8月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

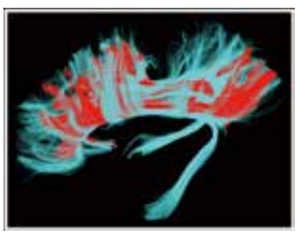
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

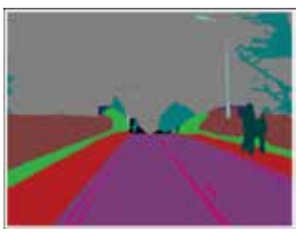
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年8月 第25卷 第8期（总第292期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

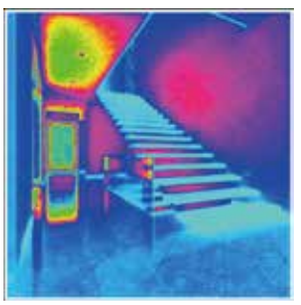
目次



神经纤维跟踪算法研究进展(第1513页)



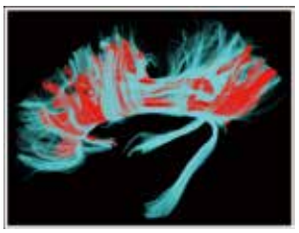
结合注意力机制的双路径语义分割(第1627页)



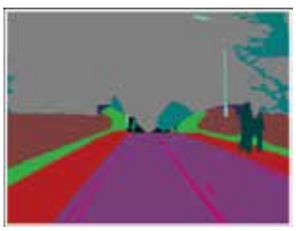
突变策略下多通路Metropolis光照传播(第1658页)

Volume 25, Number 8
Published August 16, 2020

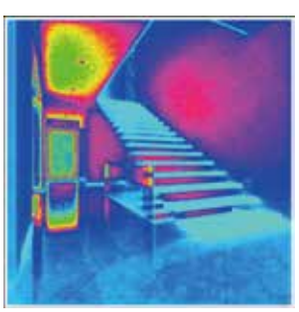
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Research progress of neural fiber tracking(P1513)



Two-path semantic segmentation algorithm combining attention mechanism(P1627)



Improved multiplexed Metropolis light transport based on fusion mutation strategy(P1658)

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2020)08-1558-09

论文引用格式:Zheng Y, Chen Y D and Hao C Y. 2020. Video object segmentation algorithm based on consistent features. Journal of Image and Graphics,25(08):1558-1566(征煜,陈亚当,郝川艳. 2020. 特征一致性约束的视频目标分割. 中国图象图形学报,25(08):1558-1566)[DOI:10.11834/jig.190571]

特征一致性约束的视频目标分割

征煜[1],陈亚当[1],郝川艳[2]

1. 南京信息工程大学计算机与软件学院,南京　210044;　2. 南京邮电大学教育科学与技术学院,南京　210023

摘　要:**目的**　视频目标分割是计算机视觉领域的一个重要方向,已有的一些方法在面对目标形状不规则、帧间运动存在干扰信息和运动速度过快等情况时,显得无能为力。针对以上不足,提出基于特征一致性的分割算法。**方法**　本文分割算法框架是基于马尔可夫随机场(Markov random field,MRF)的图论方法。使用高斯混合模型,对预先给定的已标记区域分别进行颜色特征的建模,获得分割的数据项。结合颜色、光流方向等多种特征,建立时空平滑项。在此基础之上,加入基于特征一致性的能量约束项,以增强分割结果的外观一致性。这项添加的能量本身属于一种高阶能量约束,会显著增加能量优化的计算复杂度。为此,添加辅助结点,以解决能量的优化问题,从而提高算法速度。**结果**　在DAVIS_2016(densely annotated video segmentation)数据集上对该算法进行评估与测试,并与最新的基于图论的方法进行对比分析,对比算法主要有HVS(efficient hierarchical graph-based video segmentation)、NLC(video segmentation by non-local consensus voting)、BVS(bilateral space video segmentation)和OFL(video segmentation via object flow)。本文算法的分割结果精度排在第2,比OFL算法略低1.6%;在算法的运行速度方面,本文算法领先于对比方法,尤其是OFL算法的近6倍。**结论**　所提出的分割算法在MRF框架的基础之上融合了特征一致性的约束,在不增加额外计算复杂度的前提下,提高了分割精度,提升了算法运行速度。

关键词:视频目标分割;特征一致性;马尔可夫随机场(MRF);辅助结点;能量优化

Video object segmentation algorithm based on consistent features

Zheng Yu[1], Chen Yadang[1], Hao Chuanyan[2]

1. School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing　210044, China;
2. School of Education Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing　210023, China

Abstract: **Objective** Video object segmentation is an important topic in the field of computer vision. However, the existing segmentation methods are unable to address some issues, such as irregular objects, noise optical flows, and fast movements. To this end, this paper proposes an effective and efficient algorithm that solves these issues based on feature consistency. **Method** The proposed segmentation algorithm framework is based on the graph theory method of Markov random field (MRF). First, the Gaussian mixture model (GMM) is applied to model the color features of pre-specified marked areas, and the segmented data items are obtained. Second, a spatiotemporal smoothing term is established by combining various characteristics, such as color and optical flow direction. The algorithm then adds energy constraints based on feature consistency to enhance the appearance consistency of the segmentation results. The added energy belongs to a higher-order energy constraint, thereby significantly increasing the computational complexity of energy optimization. The energy optimization problem is solved by adding auxiliary nodes to improve the speed of the algorithm. The higher-order constraint term comes

收稿日期:2019-11-19;**修回日期**:2020-03-10;**预印本日期**:2020-03-17

基金项目:国家自然科学基金项目(61702278,61802197);江苏省优势学科平台项目(PDPA)

Supported by: National Natural Science Foundation of China (61702278, 61802197)

from the idea of text classification, which is used in this paper to model the higher-order term of the segmentation equation. Each super-pixel point corresponds to a text, and the scale-invariant feature transform (SIFT) feature point in the super-pixel point is used as a word in the text. The higher-order term is modeled afterward via extraction and clustering. Given the running speed of the algorithm, auxiliary nodes are added to optimize the high-order term. The high-order term is approximated to the data and smoothing items, and then the graph cutting algorithm is used to complete the segmentation. **Result** The test data were taken from the DAVIS_2016 (densely annotated video segmentation) dataset, which contains 50 sets of data, of which 30 and 20 are sets of training and verification data, respectively. This dataset has a resolution of 854 × 480 pixels. Given that many methods are based on MRF expansion, $\alpha = 0.3$ and $\beta = 0.2$ are empirically set in the proposed algorithm to maintain a capability balance among the data, smoothing, and feature consistency items. Similar to extant methods, the number of submodels used to establish the Gaussian mixture model for the front/background is set to 5, $\sigma_h = \sigma_h = 0.1$. This paper focuses on the verification and evaluation of the proposed feature consistency constraint terms and sets $\beta = 0$ and $\beta = 0.2$ to divide the videos under the constraint condition. The experimental results show that the IoU score with higher-order constraints is 10.2% higher than that without higher-order constraints. To demonstrate its effectiveness, the proposed method is compared with some other classical video segmentation algorithms based on graph theory. The experimental results highlight the competitive segmentation effect of the proposed algorithm. Meanwhile, the average IoU score reported in this paper is slightly lower than that of the video segmenfation via object flow (OFL) algorithm because the latter continuously iteratively optimizes the optical flow calculation results to achieve a relatively high segmentation accuracy. The proposed algorithm takes nearly 10 seconds on average to segment each frame, which is shorter than the running time of other algorithms. For instance, although the OFL algorithm reports a slightly higher accuracy, its average processing time for each frame is approximately 1 minute, which is 6 times longer than that of the proposed algorithm. In sum, the proposed algorithm can achieve the same segmentation effect with a much lower computational complexity than the OFL algorithm. However, the accuracy of its segmentation results is 1.6% lower than that of the results obtained by the OFL algorithm. Nevertheless, in terms of running speed, the proposed algorithm is ahead of other methods and is approximately 6 times faster than the OFL algorithm. **Conclusion** Experimental results show that when the current/background color is not clear enough, the foreground object and the background are often confused, thereby resulting in incorrect segmentation. However, when the global feature consistency constraint is added, the proposed algorithm can optimize the segmentation result of each frame by the feature statistics of the entire video. By using global information to optimize local information, the proposed segmentation method shows strong robustness to random noise, irregular motions, blurry backgrounds, and other problems in the video. According to the experimental results, the proposed algorithm spends most of its time in calculating the optical flow and can be replaced by a more efficient motion estimation algorithm in the future. However, compared with other segmentation algorithms, the proposed method shows great advantages in its performance. Based on the MRF framework, the proposed segmentation algorithm integrates the constraints of feature consistency and improves both segmentation accuracy and operation speed without increasing computational complexity. However, this method has several shortcomings. First, given that the proposed algorithm segments a video based on super pixels, the segmentation results depend on the segmentation accuracy of these super pixels. Second, the proposed high-order feature energy constraint has no obvious effect on feature-free regions because the SIFT feature points detected in similar regions will be greatly reduced, thereby creating super-pixel blocks that are unable to detect a sufficient number of feature points, which subsequently influences the global statistics of front/background features and prevents the proposed method from optimizing the segmentation results of feature-free regions. Similar to traditional methods, the optical flow creates a bottleneck in the performance of the proposed method. Therefore, additional efforts should be devoted in finding a highly efficient replacement strategy. As mentioned before, the method based on graph theory (including the proposed method) still lags behind the current end-to-end video segmentation methods based on convolutional neural network (CNN) in terms of segmentation accuracy. Future works should then attempt to combine these two approaches to benefit from their respective advantages.

Key words: video object segmentation; feature consistency; Markov random field (MRF); auxiliary node; energy optimization

0 引 言

视频目标分割是指在视频帧序列中将前景物体与视频背景分离。目前该领域内已有许多方法解决这种二元分割问题(Chen 等,2019a,2019b),这些方法可以分为无监督方法和监督方法。前者不需要标记帧,直接输入视频数据;后者则要求人为提供额外的标记帧来进行初始化,通常是将视频帧序列的首帧分割结果作为已知数据输入算法。本文方法属于后者。

现有的视频对象分割方法中,基于图割的方法有效地保证了视频对象在帧间的传播。这类方法将每个帧分解为时空节点,从而将分割问题转化为在MRF(Markov random field)中的节点标记问题。大多数都是在寻找最优的前/背景分割标记,目的是使视频帧序列的 MRF 能量最小化。尽管取得了一定成效,但这些方法存在较严重的问题:比如当视频对象形状不规则、帧间光流信息存在明显干扰;或运动过快时,分割效果非常不理想。其根本原因在于,帧间局部连续性易受干扰,无法有效将标记帧监督信息传递给后续帧序列,因此本文提出引入视频全局特征来增强分割。具体来说,通过对原始分割能量增加一种全局一致性约束的高阶能量来提高算法的鲁棒性和健壮性,这种能量可以有效提取前景物体在视频整体中的表面一致性,并以此作为线索指引算法完成分割。

本文主要贡献是根据文本分类思想对分割方程的高阶项进行建模,即在原有的数据项和平滑项的基础上加入了一种具有全局特征一致性约束的高阶能量项来保证视频分割的全局一致性。并且受先前研究工作的启发(Yang 等,2016),通过对 MRF 模型添加辅助节点的方式来优化所添加的这项高阶能量,使其计算复杂度大大降低。

1 相关研究工作

1.1 无监督方法

无监督分割指在整个视频序列中提取对象而无需任何人工介入。HVS(hierarchical video segmentation)(Grundmann 等,2010)以及 Xu 和 Corso(2016)利用像素点之间的空间和颜色特征进行聚类,并通过不断合并,最终完成分割。但是此类算法没有充分考虑到前背景区域之间的差异性,很容易发生标签错分,最终导致分割失败。FST(fast object segmentation)(Papazoglou 和 Ferrari,2013)和 ACO(alternate convex optimization)(Jang 等,2016)根据在高级特征空间中测量分割块之间的距离,将过分割结果通过聚类形成前景和背景。SAG(saliency-aware geodesic)(Wang 等,2015),TS(track and segment)(Xiao 和 Lee,2016)和 ARP(primary object segmentation in videos based on region augmentation and reduction)(Koh 和 Kim,2017),则是产生了大量的分割块,并对所有块进行排序,将视频目标分割任务转化为一个选择问题。这些自动视频分割方法虽然具有优势,但只适用于有限的视频分割场景:只有当待分割目标与背景的差异较大时,才会得到较好的分割结果。所以多数情况下为了尽可能获得好的分割结果,仍然使用监督方法。

1.2 监督方法

一般情况下,监督方法要求提供标记帧。一些方法将这些标记帧通过光流运动信息扩散到其他未标记帧之上,从而得到分割结果。这种方法称为点跟踪,如 CUT(cost multicuts)(Keuper 等,2015)和 LTA(long term video analysis)(Ochs 等,2014)将目标帧中的每个像素与源帧中的像素匹配,如果满足设定的匹配阈值,将标签设置为一致,否则不一致。这种方法可以基于稀疏轨迹或特征点进行聚类,但当匹配包含旋转、遮挡和大运动等情况时,聚类是不可靠的。针对这些问题,Chen 等人(2018)提出将前景和背景的运动估计分离开来,以改善跟踪效果。然而,遮挡问题仍然没有得到解决。

另一种常见方法是基于概率图的图割方法。如 Perazzi 等人(2015)将整个视频看做一个 3D 分割图,以进行全局分割。BVS(bilateral space video segmentation)(Märki 等,2016)在帧之间建立像素级别的全连接图,并有效地在时空双边网格的空间中进行分割。OFL(object flow)(Tsai 等,2016)在像素和超像素上构建图模型,之后利用光流估计结果对分割结果进行迭代优化。JOTS(joint online tracking and segmentation)(Wen 等,2015)以相邻帧上同一对象的超像素点构建图模型以解决视频帧标记问题。这些方法仅参考了视频局部信息,并未考虑目标物体在视频全局中的一致性特征。

为了建立全局性的连接模型，NLC（non-local consensus）（Faktor 和 Irani，2014）采用随机漫步方法来扩散视频帧间的分割相似性。此外，TVS（transductive video segmentation）（Wang 等，2017）把视频按时间序列重组成一个多叉树型结构，以扩大相邻帧的范围。这些方法的局限性在于它们采用迭代方法优化，很容易陷入局部最优，同时也非常耗时。相比之下，本文使用完全不同的优化算法使得算法的时间代价显著降低。

基于卷积神经网络（convolutional neural network，CNN）的方法（Bao 等，2018；Maninis 等，2019）是一种新的视频目标分割方法，该方法主要通过将离线训练和在线训练过程结合起来，使得性能得到显著改善。但是网络模型需要使用大规模标记数据集进行预训练，这在很大程度上限制了其应用。因此，本文仍然采用基于图论的方法。

2　本文方法

在本文方法中，输入一个首帧经过标记的视频序列。和大多数视频分割方法类似，将分割的问题转化为在 MRF 中的节点标记问题。考虑到在视频处理中使用基于像素点的分割方法会导致巨大的内存和计算代价，采用超像素（Xu 和 Corso，2016）的方法进行分割。首先将视频帧序列中每一帧分割成若干个超像素点。然后利用 MRF 图模型解决分割问题，最后的分割结果 L^* 可以表示为

$$L^* = \arg\min_{\boldsymbol{L}} E(\boldsymbol{S}) \tag{1}$$

$$E(\boldsymbol{S}) = E_u(\boldsymbol{S}) + \alpha \times E_p(\boldsymbol{S}) + \beta \times E_f(\boldsymbol{S}) \tag{2}$$

式中，$\boldsymbol{L}$ 表示所有超像素标签的集合，有 $\boldsymbol{L} \in \{+1, -1\}$；$\boldsymbol{S}$ 表示所有超像素的集合，E_u 表示数据项，E_p 表示平滑项，E_f 表示特征一致性约束项。α 和 β 分别是线性组合的权重参数。

2.1　数据约束项

采用期望值最大算法，通过已标记帧（如首帧），分别建立前/背景的高斯混合模型（Gaussian mixed model，GMM）。假设 s 表示超像素，它的标签值为 $l(s) \in \{+1, -1\}$。如果 $l(s) = -1$，表示 s 为背景，如果 $l(s) = +1$，则表示 s 为前景。那么，在已建立的前/背景 GMM 模型中，通过计算 $p(c(s) \mid l(s) = +1)$ 和 $p(c(s) \mid l(s) = -1)$，求得 s 属于前景或背景的后验概率，其中 $c(s)$ 表示超像素点 s 的颜色直方图信息。

至此，每个超像素 s 的数据约束项可以表示为负对数形式，即

$$\phi_u(s) = -\ln p(c(s) \mid l(s) = +1)[l(s) = +1] - \ln p(c(s) \mid l(s) = -1)[l(s) = -1] \tag{3}$$

式中，$[\cdot]$ 为真值函数，即 $[\cdot]$ 为真时，该项为 1，否则为 0。因此，式（2）中的数据约束项可以表示为

$$E_u(\boldsymbol{S}) = \sum_s \phi_u(s) \tag{4}$$

由式（4）可以看出，如果一个超像素点在 GMM 中大概率属于前景，而被标记为背景，那么，将会导致该点产生较大的惩罚能量，反之亦然。

2.2　平滑约束项

平滑约束项常用于平滑相邻节点的标记值，一般包括空间平滑项和时间平滑项。空间相邻指的是在同一帧的两个超像素节点之间有共同的边，而时间相邻指的是相邻帧间的两个超像素之间有光流映射关系。首先，将所有的空间相邻节点对和时间相邻节点对表示为 ε_s 和 ε_t。于是，式（1）中的平滑约束项可以表示为

$$E_p(\boldsymbol{S}) = \sum_{(s,s')\in\varepsilon_s} E_p^s(s,s') + \sum_{(s,s')\in\varepsilon_t} E_p^t(s,s') \tag{5}$$

式中，s 和 s' 表示超像素，$E_p^s(s,s')$ 和 $E_p^t(s,s')$ 分别表示空间平滑约束和时间平滑约束。对于 $E_p^s(s,s')$，采用颜色特征和光流方向等梯度方向直方图特征（histogram of gradient，HOG）来计算节点之间的相似性，即

$$E_p^s(s,s') = [l(s) \neq l(s')] \cdot \exp(-\sigma_h^{-1} \|h(s) - h(s')\|^2) \cdot \exp(-\sigma_s^{-1} \|c(s) - c(s')\|^2) \tag{6}$$

式中，$[l(s) \neq l(s')]$ 为真时，该项为 1，否则为 0。$h(s)$ 表示该超像素的 HOG 特征，σ_h 和 σ_s 为对应的距离控制参数。

对于 $E_p^t(s,s')$，只通过两超像素的颜色差异来计算它们的相似性，即

$$E_p^t(s,s') = [l(s) \neq l(s')] \cdot \exp(-\sigma_s^{-1} \|c(s) - c(s')\|^2) \tag{7}$$

由于帧间的目标物体运动方向和速度可能不连续，所以在时间平滑项里面不考虑 HOG 特征。

2.3　特征一致性约束项

平滑项利用视频的时空连续性对分割结果进行

优化,但当物体形状不规则、帧间运动速度太快或存在明显干扰时,这种局部信息的约束并不能有效应对。因此,考虑一种基于视频整体的全局信息作为分割能量,将该项称为“特征一致性约束”,即$E_f(\boldsymbol{S})$。

特征一致性约束的思想来源于文本分类。定义$\boldsymbol{x}$表示某个文本,$l(\boldsymbol{x}) \in \{-1, +1\}$表示该文本的标签,$n_x$表示其文本长度,$x_i$表示$\boldsymbol{x}$中的第$i$个词。假设$x_i$的取词范围来源于一部容量为$|D|$的字典,那么$x_i$可以表示为当前词在字典中的索引号,并且,索引号范围在$\{1,\cdots,k,\cdots,|D|\}$之中。现在,对应两个不同的文本类型,即+1和-1,构造两个直方图$\boldsymbol{H}^{+1}$和$\boldsymbol{H}^{-1}$,维度(bin)均为$|D|$,H_k^{+1}和H_k^{-1}分别表示该类文本直方图中第k个bin的数量,即$H_k^{+1} = \sum_{\boldsymbol{x}} \sum_{i=1}^{n_x} [l(\boldsymbol{x}) = +1][x_i = k]$和$H_k^{-1} = \sum_{\boldsymbol{x}} \sum_{i=1}^{n_x} [l(\boldsymbol{x}) = -1][x_i = k]$。$H_k$则表示总数量,即$H_k = H_k^{+1} + H_k^{-1}$。那么,该字典中第$k$个词是属于+1或-1类文本的概率就分别是$p_k^{+1} = H_k^{+1}/H_k$和$p_k^{-1} = H_k^{-1}/H_k$。根据朴素贝叶斯定理,对于任一个未知类型文本$\boldsymbol{x}$,它的标签属于+1或-1的概率就可以通过计算$p(+1 \mid \boldsymbol{x}) = \prod_{i=1}^{n_x} p_{x_i}^{+1}$和$p(-1 \mid \boldsymbol{x}) = \prod_{i=1}^{n_x} p_{x_i}^{-1}$得到。最后通过比较$p(+1|x)$和$p(-1 \mid \boldsymbol{x})$的大小,确定该文本的预测类型。

借助该思想,把每一个超像素s看做一个文本,超像素中包含的SIFT(scale-invariant feature transform)特征点作为该“文本”中的一个词,所以,每个$\boldsymbol{x}$中共有n_s个词。在分割开始前,首先对整个视频进行SIFT特征点提取,把所有特征点聚类成$|D|$类(通常将SIFT特征点聚类成100类,实验表明当该数值过大或者过小都会对最终的分割结果产生负影响),这样就构造了一个容量为$|D|$的字典。$E_f(\boldsymbol{S})$定义为

$$E_f(\boldsymbol{S}) = -\sum_s \ln p(l(s)|s) = -\sum_s \ln \prod_{i=1}^{n_s} p_{s_i}^{l(s)} \tag{8}$$

式中,$p_{s_i}^{l(s)} = H_{s_i}^{l(s)}/H_{s_i}$,该项为本文算法提供了关键的全局约束作用,使得其对于局部噪音有很好的鲁棒性。

之后,对式(1)进行能量最小化,可以直接使用α-扩展算法(Boykov和Kolmogorov,2004)对其进行迭代优化。但由于$E_f(\boldsymbol{S})$在MRF中本身属于一种高阶能量,其优化过程所耗费的计算量十分庞大,会大大降低算法的性能。为此,借助Yang等人(2016)思想,通过添加辅助结点的方式来解决对式(1)的优化问题。定义v_s^k表示在超像素s中,特征点属于k的数量(见图1)。式(8)中的$p(l(s)|s)$可以重新被定义为

$$P(l(s)|s) = \prod_{k=1}^{|D|} (p_k^{l(s)})^{v_s^k} \tag{9}$$

因此,式(8)转为

$$\begin{aligned} E_f(\boldsymbol{S}) &= \sum_s -\ln \prod_{k=1}^{|D|} (p_k^{l(s)})^{v_s^k} = \\ &-\sum_s \sum_{k=1}^{|D|} v_s^k \ln p_k^{l(s)} = -\sum_{k=1}^{|D|} \ln p_k^{l(s)} \sum_s v_s^k = \\ &-\sum_{k=1}^{|D|} \ln p_k^{l(s)} H_k = -\sum_{k=1}^{|D|} (H_k^{+1} \ln p_k^{+1} + H_k^{-1} \ln p_k^{-1}) = \\ &-\sum_{k=1}^{|D|} \left(H_k^{+1} \ln \frac{H_k^{+1}}{H_k} + H_k^{-1} \ln \frac{H_k^{-1}}{H_k} \right) \end{aligned} \tag{10}$$

由于$H_k^{+1} + H_k^{-1}$是固定值,所以当$H_k^{+1} = H_k^{-1}$时,该对称函数有最小值,因此$E_f(\boldsymbol{S})$可近似为

$$E_f(\boldsymbol{S}) \approx -\sum_{k=1}^{|D|} |H_k^{+1} - H_k^{-1}| \approx \sum_{k=1}^{|D|} \min(H_k^{+1}, H_k^{-1}) \tag{11}$$

与之前相比,$E_f(\boldsymbol{S})$现在可由一种更简单的方法对其进行优化。该方法只需在原有MRF框架中添加一些辅助结点即可,而完全不需要增加多余的计算,如图1所示,词典的容量$|D| = 4$,词汇分别为$\{a,b,c,d\}$。连接超像素s和辅助结点$\boldsymbol{k}$的边权值等于s中所包含$\boldsymbol{k}$的数量,即v_s^k。具体来说,首先添加$|D|$个辅助结点在原分割图中,其中每个结点表示对应的H_k。如图1所示,把辅助结点和其对应

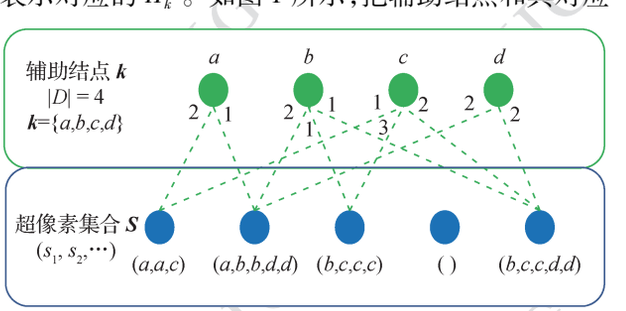


图1 MRF中添加辅助结点示意图

Fig. 1 Example illustrating how the auxiliary nodes and edges are added to the original MRF graph

的超像素结点进行连接，并计算边权值。最后，设置所有辅助结点的数据约束项都为 0，设置连接 s 和 k 的平滑约束项为 $[l(s) \neq l(k)] \cdot v_s^k$ 。

通过上述方法，本文分割图已建立完毕，并且可以用标准的图割方法（Boykov 和 Kolmogorov，2004）对其分割。由于添加了辅助结点，在分割时，那些含有同类 SIFT 特征的超像素，往往更容易被分割为一类。

3　实验结果与分析

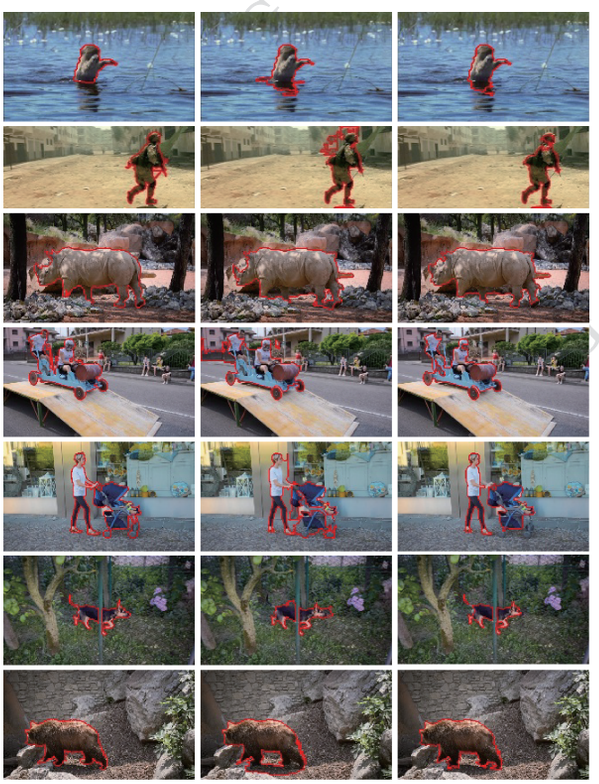
采用 MATLAB 工具编写本文算法，其中一些矩阵运算函数则用 C/C++ 实现。硬件环境是基于一台四核的 3.40 GHz CPU 个人电脑。测试数据选用 DAVIS_2016（densely annotated video segmentation）（Perazzi 等，2016）数据集，其中共有50 组数据，其中包括 30 组训练数据和 20 组验证数据，其分辨率均为 854 × 480 像素。根据经验设置 $\alpha = 0.3$ ，$\beta = 0.2$ ，以保持数据项、平滑项以及特征一致项这三者之间的能量平衡。另外，和大多数方法类似，用于建立前/背景的高斯混合模型的子模型数量设置为 5，$\sigma_h = \sigma_s = 0.1$ 。

采用 IoU（intersection over union）通用标准去测量视频分割结果的准确度。其可以表示为：IoU = $(C \cap G)/(C \cup G)$ ，其中 C 为最后分割结果，G 为真实值。

3.1　能量控制评估

数据项和平滑项在大多数类似工作中都出现过，因此不再过多测试。重点对本文所提出的特征一致性约束项进行单独验证与评估。在式（2）中，先后设置 $\beta = 0$ 和 $\beta = 0.2$ ，这样可以分别在有无该项约束的条件下，对视频进行分割。本文分别对有无特征一致性约束的分割结果计算了 20 组验证视频数据的平均 IoU 得分，结果如图 2 所示，有 $E_f(\boldsymbol{S})$ 约束得到的结果平均比无 $E_f(\boldsymbol{S})$ 约束的分割结果提高了 10.2% 的 IoU 得分，可见该约束项确实大大提高了分割结果的准确度。

从结果可以看出，本文所提出的特征约束项明显提高了分割结果的准确度。具体来讲，在没有使用该约束项时，单帧的分割结果往往会被局部约束的平滑项所干扰，尤其当前/背景颜色不够分明时，前景目标和背景会经常被混淆，使得算法产生误分割。但是当加入全局的特征一致性约束时，算法能够通过对视频整体进行特征统计，来优化每一帧的分割结果。正是由于这种利用全局信息来优化局部信息的方式，使得本文分割方法对于视频中出现的随机噪音、无规则运动、前\背景不分明等问题都有较强的鲁棒性。相比于其他一些利用局部信息进行分割的方法，本文方法会显得更加有效。例如图 2 中，士兵身上的迷彩服和背景黄土建筑十分相似，数据约束项会得到不明确的判断结果，即背景和前景的概率基本相同。此时，如果只有平滑项参与计算，那么平滑项会更加鼓励这些概率近似的超像素块被分配相同的标签，因此很容易产生前景物体被误分割成背景，而背景又被误分割成前景。而有了 $E_f(\boldsymbol{S})$ 的约束后，算法可以从视频整体进行分析，使具有明显前景物体特征的像素块和具有明显背景特征的像素块区分开来，从而达到有效分割这些近似区域的目的。



(a) 真实轮廓　(b) 无特征约束　(c) 有特征约束

图 2　部分实验效果

Fig. 2　Some experimental results((a) ground truth; (b) without feature consistency constraint; (c) with feature consistency constraint)

3.2　算法比较

为了说明本文方法的有效性，将本文方法与目

前一些经典的基于图论的视频分割算法进行比较，如 HVS(Grundmann 等,2010),NLC(Faktor 和 Irani, 2014),BVS(Märki 等,2016),OFL(Tsai 等,2016)。表 1 中给出了详细的分割结果 IoU 得分值比较,该值越大表示分割结果越准确。因为大部分所比较的工作都只给出了 DAVIS_2016 数据集中 20 组验证数据的分割结果得分,因此,表 1 也只给出这 20 组数据的相关比较。

表 1　与目前一些经典基于图论的方法的比较

Table 1　Accuracy comparison with other methods

序列	算法				
	本文	HVS	NLC	BVS	OFL
blackswanBVS	92.5	91.6	87.5	94.3	**94.7**
bmx-Trees	**38.7**	17.9	21.2	38.2	14.9
breakdance	**77.3**	55.0	67.3	50.0	52.2
camel	56.0	**87.6**	76.8	66.9	86.7
car-roundabout	79.5	77.7	50.9	85.1	**90.3**
car-shadow	82.2	69.9	64.5	57.8	**84.5**
cows	47.9	77.9	88.3	89.5	**91.1**
dance-twirl	**84.6**	31.8	34.7	49.2	53.9
dog	44.4	72.2	80.9	72.3	**89.7**
drift-chicane	**75.5**	33.1	32.4	3.3	10.4
drift-straight	**58.9**	29.5	47.3	40.2	33.8
Goat	57.4	58.0	1.0	66.1	**86.5**
horsejump-high	71.2	76.5	83.4	80.1	**86.3**
kite-Surf	**74.1**	40.5	45.3	42.5	70.3
libby	66.1	55.3	63.5	**77.6**	55.3
motocross-jump	**76.6**	9.9	25.1	34.1	60.7
paragliding-launch	**74.8**	53.7	62.8	64.0	63.7
parkour	72.1	24.0	**90.1**	75.6	85.9
scooter-black	76.4	62.4	16.2	33.7	**80.7**
soapbox	**90.0**	68.4	63.4	78.9	68.7
平均	66.4	54.6	55.1	60.0	**68.0**

注:加粗字体表示每行最优结果。

本文算法的分割效果具有一定竞争力,在近一半的视频测试序列上,本文算法得到最好的分割效果。本文在平均 IoU 得分上略低于 OFL(Tsai 等, 2016),原因是 OFL 对光流计算和分割计算这两部分结果不停地进行迭代优化,从而获得比较高的分割精度。这就导致 OFL 算法在性能方面表现不佳,具体分析见 3.3 节。

就目前研究而言,基于图论的方法(包括本文算法)与目前基于 CNN 的端到端视频分割方法(Bao 等,2018;Maninis 等,2019)相比,在分割准确度上还是有一定差距。但是基于 CNN 的方法的使用前提是需要采用大规模标记数据集进行线下训练,以获得较好的网络模型初始化参数。在很大程度上限制了该方法的普及与应用。为此,本文考虑到两种方法完全不同的技术特点和应用环境,所比较的方法还是以基于图论的方法为主。

3.3　算法运行时间

硬件环境是基于一台 i7—四核的 2.90 GHz CPU,内存容量为 8 GB 的个人电脑。测试所用到的视频分辨率均为 854 × 480 像素。本文算法各环节耗费的时间如表 2 所示。表 3 为不同方法运行时间对比。

表 2　本文算法运行时间分析

Table 2　Detailed running times

算法	运行时间/(s/帧)
超像素分割	0.76
颜色直方图	0.31
HOG 特征	0.57
SIFT 特征	1.13
光流计算	5.78
建立分割图	0.89
能量优化	0.54
总计时间	9.98

表 3　不同算法运行时间对比

Table 3　Comparison of running time of different algorithms

算法	平均运行时间/(s/帧)
本文	9.98
HVS	10.65
OFL	59.6

从表 2 和表 3 可以看出,本文方法分割每帧平均需要近 10 s 时间,明显优于其他工作中给出的运行时间。尤其对于精确度略高于本文算法的 OFL 算法,平均每帧的处理时间需要 1 min 左右,而本文

方法速度约为该算法的 6 倍。由此可见本文方法可以用低很多的运算复杂度达到与其同样的分割效果(注:这里关于所比较算法的运行时间,都直接来自于相关论文中给出的数据,其硬件平台条件存在差别,但差别不大,比较结果仅作为参考)。

另外,本文方法大部分时间耗费在光流计算之上(Ochs 等,2014),针对这个问题,后续可以用更加高效的运动估计算法将其替代。尽管如此,该方法与其他分割算法相比,在性能方面已有很大优势。

4 结　论

提出一种基于特征一致性的视频目标分割算法。该算法框架是基于马尔可夫随机场(MRF)的图论方法。首先使用高斯混合模型,对预先给定的已标记区域分别进行颜色特征的建模,获得分割的数据项。其次,结合颜色、光流方向等多种特征,建立时空平滑项。本文算法的核心是加入了基于特征一致性的能量约束项,以增强分割结果的外观一致性。这项添加的能量本身属于一种高阶能量约束,会显著增加能量优化的计算复杂度。为此,通过添加辅助结点的方式,来解决能量的优化问题,从而提高算法性能。在公用测试数据 DAVIS_2016 数据集上对该算法进行全面的评估与测试,并与多种最新基于图论的方法进行比较。本文所得到的分割结果精度排在第 2,略比 OFL 算法低了 1.6%,但在算法的运行速度方面,本文算法全面领先于其他列出的方法,尤其是 OFL 算法的近 6 倍。结果表明,本文算法在不增加额外计算复杂度的前提下,提高了分割精度,提升了算法的运行速度。

本文方法也有一些缺陷,主要有:1)由于本文算法采用的是基于超像素的分割方法,因此分割结果依赖于超像素的分割准确度,如果预处理阶段,超像素无法保证正确的物体轮廓信息,那么后续处理必然会产生分割误差。2)本文所提出的高阶特征能量约束对于无特征区域效果不明显,因为在类似区域所检测出的 SIFT 特征点会大大减少,会导致有些超像素块检测不到足够的特征点。这样会影响对前/背景特征的全局统计,并且所提出的 $E_f(\boldsymbol{S})$ 无法对无特征区域分割结果进行优化。3)和传统方法类似,光流法依然是该方法在性能上的一个瓶颈。因此在后期阶段,将更加着重于为该部分找到更加高效的替换策略。4)正如之前提到,基于图论的方法(包括本文方法)与目前基于 CNN 的端到端视频分割方法相比,在分割准确度上还是有一定差距。因此,将寻求一种新的方法,试图结合这两种方法,发挥出它们各自的优点,这也是现今视频分割领域的一个重要研究方向。

参考文献(References)

Bao L C, Wu B Y and Liu W. 2018. CNN in MRF: video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 5977-5986 [DOI: 10.1109/CVPR.2018.00626]

Boykov Y and Kolmogorov V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9): 1124-1137 [DOI: 10.1109/TPAMI.2004.60]

Chen Y D, Hao C Y, Liu A X and Wu E H. 2019a. Appearance-consistent video object segmentation based on a multinomial event model. ACM Transactions on Multimedia Computing, Communications, and Applications, 15(2): 1934-1945 [DOI: 10.1145/3321507]

Chen Y D, Hao C Y, Liu A X and Wu E H. 2019b. Multilevel model for video object segmentation based on supervision optimization. IEEE Transactions on Multimedia, 21(8): 1934-1945 [DOI: 10.1109/TMM.2018.2890361]

Chen Y D, Hao C Y, Wu W and Wu E H. 2018. Efficient frame-sequential label propagation for video object segmentation. Multimedia Tools and Applications, 77(5): 6117-6133 [DOI: 10.1007/s11042-017-4520-5]

Faktor A and Irani M. 2014. Video segmentation by non-local consensus voting//Proceedings of the British Machine Vision Conference. Nottingham, UK: BMVA Press: #21 [DOI: 10.5244/C.28.21]

Grundmann M, Kwatra V, Han M and Essa I. 2010. Efficient hierarchical graph-based video segmentation//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE: 2141-2148 [DOI: 10.1109/CVPR.2010.5539893]

Jang W D, Lee C and Kim C S. 2016. Primary object segmentation in videos via alternate convex optimization of foreground and background distributions//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 696-704 [DOI: 10.1109/CVPR.2016.82]

Keuper M, Andres B and Brox T. 2015. Motion trajectory segmentation via minimum cost multicuts//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3271-3279 [DOI: 10.1109/ICCV.2015.374]

Koh Y J and Kim C S. 2017. Primary object segmentation in videos based on region augmentation and reduction//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii: IEEE: 7417-7425 [DOI: 10.1109/CVPR.2017.784]

Maninis K K, Caelles S, Chen Y, Pont-Tuset J, Leal-Taixé L, Cremers D and Van Gool L. 2019. Video object segmentation without temporal information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(6): 1515-1530 [DOI: 10.1109/TPAMI.2018.2838670]

Märki N, Perazzi F, Wang O and Sorkine-Hornung A. 2016. Bilateral space video segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 743-751 [DOI: 10.1109/CVPR.2016.87]

Ochs P, Malik J and Brox T. 2014. Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(6): 1187-1200 [DOI: 10.1109/TPAMI.2013.242]

Papazoglou A and Ferrari V. 2013. Fast object segmentation in unconstrained video//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 1777-1784 [DOI: 10.1109/ICCV.2013.223]

Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M and Sorkine-Hornung A. 2016. A benchmark dataset and evaluation methodology for video object segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 724-732 [DOI: 10.1109/CVPR.2016.85]

Perazzi F, Wang O, Gross M and Sorkine-Hornung A. 2015. Fully connected object proposals for video segmentation//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3227-3234 [DOI: 10.1109/ICCV.2015.369]

Tsai Y H, Yang M H and Black M J. 2016. Video segmentation via object flow//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 3899-3908 [DOI: 10.1109/CVPR.2016.423]

Wang B T, Fu Z H, Xiong H K and Zheng Y F. 2017. Transductive video segmentation on tree-structured model. IEEE Transactions on Circuits and Systems for Video Technology, 27(5): 992-1005 [DOI: 10.1109/TCSVT.2016.2527378]

Wang W G, Shen J B and Porikli F. 2015. Saliency-aware geodesic video object segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 3395-3402 [DOI: 10.1109/CVPR.2015.7298961]

Wen L Y, Du D W, Lei Z, Li S Z and Yang M H. 2015. JOTS: joint online tracking and segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 2226-2234 [DOI: 10.1109/CVPR.2015.7298835]

Xiao F Y and Lee Y J. 2016. Track and segment: an iterative unsupervised approach for video object proposals//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 933-942 [DOI: 10.1109/CVPR.2016.107]

Xu C L and Corso J J. 2016. LIBSVX: a supervoxel library and benchmark for early video processing. International Journal of Computer Vision, 119(3): 272-290 [DOI: 10.1007/s11263-016-0906-5]

Yang J, Price B, Shen X H, Lin Z and Yuan J S. 2016. Fast appearance modeling for automatic primary video object segmentation. IEEE Transactions on Image Processing, 25(2): 503-515 [DOI: 10.1109/TIP.2015.2500820]

作者简介

征煜,1995 年生,男,硕士研究生,主要研究方向为图像视频分割、视频目标跟踪、语义分割、深度学习等计算机视觉技术。
E-mail: 17865138982@163.com

陈亚当,通信作者,男,副教授,主要研究方向为视频分割、视频场景重建、视频编辑、深度学习。
E-mail: cyd4511632@126.com

郝川艳,女,博士,讲师,主要研究方向为纹理合成与分析、图像处理、图像分割、视频融合、3 维动画。
E-mail: hcy@njupt.edu.cn