复杂扰动背景下时空特征动态融合的视频显著性检测

陈昶安[1)], 吴晓峰[1)*], 王 斌[1,2)], 张立明[1)]

[1)] (复旦大学信息科学与工程学院智慧网络与系统研究中心 上海 200433)
[2)] (复旦大学电磁波信息科学教育部重点实验室 上海 200433)
(xiaofengwu@fudan.edu.cn)

摘 要: 现有的运动目标显著性提取算法对具有树枝摇晃、水波荡漾等复杂扰动背景的视频处理效果较差, 无法排除背景对显著目标提取的干扰. 针对此类视频, 提出一种基于时空显著性信息动态融合的目标提取算法. 在空间上, 利用简单线性迭代聚类(SLIC)超像素分割算法计算重建误差, 得到每帧图像上完整的显著目标; 在时间上, 考虑到显著目标内部各像素具有运动一致性的特点, 利用连续多帧图像的运动估计引入运动熵来表征, 同时利用中心周边差的机制来区分目标和背景的运动; 最后由于人的视觉系统对运动信息更敏感, 根据时间显著性的大小设置动态权重进行时空显著性融合, 得到最终能兼顾动静两种情况的视频显著图. 在 4 个视频数据库上的实验结果表明, 该方法能够较好地抑制复杂扰动背景对于运动显著目标提取的干扰, 优于对比方法.

关键词: 复杂扰动背景; 简单线性迭代聚类; 运动显著性; 运动一致性; 运动熵; 动态融合
中图法分类号: TP391.41

Video Saliency Detection Using Dynamic Fusion of Spatial-Temporal Features in Complex Background with Disturbance

Chen Chang'an[1)], Wu Xiaofeng[1)*], Wang Bin[1,2)], and Zhang Liming[1)]

[1)] (*Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University, Shanghai* 200433)
[2)] (*Key Laboratory for Information Science of Electromagnetic Waves* (*Ministry of Education*), *Fudan University, Shanghai* 200433)

Abstract: In recent years, most existing video saliency detection methods failed to find salient regions in complex background with disturbance (e.g. waving leaves, rippling water, etc.). In this paper, a framework based on dynamic fusion of spatial-temporal features for detecting video saliency is proposed. Firstly, the spatial saliency of current frame is computed by using the simple linear iterative clustering (SLIC) algorithm. Secondly the motion entropy in multiple continuous frames is calculated to represent the motion coherence of salient object over time. At the same time, center-surround difference is used to separate target motion from its neighbors. As the human visual system is more sensitive to the motion information, a dynamic fusion strategy is adopted to combine the spatial saliency map and temporal saliency maps in this paper, such that the final saliency map can take both static and moving objects into account. The experimental results in four video databases demonstrate that the proposed method performs better than traditional methods in video saliency detection.

Key words: complex background with destabilization; simple linear iterative clustering (SLIC); motion saliency; motion coherence; motion entropy; dynamic fusion

收稿日期: 2015-05-26; 修回日期: 2016-01-04. 基金项目: 国家自然科学基金(61572133); 高等学校博士学科点专项科研基金(20110071110018). 陈昶安(1989—), 男, 硕士研究生, 主要研究方向为图像处理、计算机视觉; 吴晓峰(1971—), 男, 博士, 高级讲师, 硕士生导师, 论文通讯作者, 主要研究方向为图像与信号处理、机器人智能信息处理与控制; 王 斌(1964—), 男, 博士, 教授, 博士生导师, 主要研究方向为信号与图像处理及其应用; 张立明(1943—), 女, 博士, 教授, 博士生导师, 主要研究方向为人工神经网络模型及其在图像识别中的应用.

人类时时刻刻接受着外界的海量信息，人的视觉系统通过对这些信息进行处理，提炼和增强重要信息，剔除和衰减不重要信息，从而能快速捕捉到显著的或感兴趣的区域．人眼的这一功能称为视觉注意力选择机制，在机器视觉领域对该机制的仿真被称为视觉显著性计算．

在实际生活中，人的视觉观测是连续的，并对运动目标具有跟踪的能力．生物学研究发现，人的视觉系统对运动信息相比其他图像信息，如纹理、亮度等更加敏感[1]．因此，在研究单幅图像的显著性基础上，人们更加关心在连续视频序列中运动信息对于目标显著性的影响．由于视频显著性的应用前景十分广泛，包括视频目标识别[2]、视频压缩[3]、视频合成和视频监控[4]等，本文主要着重讨论视频显著性．

视频中前景目标和背景的运动类型非常复杂．前景目标可以分为单目标运动、同尺度多目标运动、不同尺度多目标运动、目标不规律运动、目标运动暂停、目标沿景深方向运动等．背景又可以分为背景不动(即固定摄像头)，背景扰动(例如树叶随风摇摆，水面波纹等)，背景在视平面规则运动(即摄像头平动)，背景做景深方向运动(即摄像头沿深度方向上移动)等不同的情况．由于背景运动信息的引入以及目标运动复杂性的增加，运动目标的显著性处理的复杂度也相应越来越高，目前还没有一种算法可以很好地处理各类情形．本文算法较好地解决了在复杂扰动背景下的单目标及多目标的显著性检测问题，对于固定摄像头、摄像头平动以及摄像头抖动的视频类型都有较为理想的检测结果．

2005 年，Itti 等[5]提出了基于贝叶斯定理检测视频中异常事件的算法，其在最初的显著性模型[6]中引入了运动特征，用所有像素的颜色、亮度、方向和运动特征的先验概率密度分布和后验概率密度分布的 KL(Kullback-Leibler divergence)散度来计算视频显著图．2009 年，Achanta 等[7]提出了主要针对静态图像的基于图像频率域处理(frequency-tuned, FT)的算法，因其具有计算速度快的特点，在有些文献[8]中把它作为视频显著性的估计算法；这种算法对于简单背景中的较大目标处理效果较好，但是对于复杂背景的图像，计算结果与人类视觉系统的差异很大．Guo 等[3]认为，图像的显著性与图像频率域的相位谱相关，提出了基于相位谱的四元数傅里叶变换(phase spectrum of quaternion Fourier transform, PQFT)算法．针对视频序列，PQFT 算法将连续 2 帧之差作为四元数特征中的运动信息计算视频显著性；该算法运算速度很快，但是由于相位信息与局部特征相关，并且主要表征了亮度或颜色的变化，所以 PQFT 算法在大多数情况下仅突出了显著区域的边缘，无法较好地处理大目标．另一方面，由于其运动估计只考虑了连续 2 帧之间的差异，当背景也具有运动时会干扰运动显著性的准确性，无法准确估计运动目标．2014 年，Kim 等[8]利用视频帧的每个像素在空间和时间上的梯度分布建立方向一致性的算法，通过背景和前景运动目标的方向一致性对比，得到运动显著图，该算法不需要人为地调整时间和空间显著性在最终显著性中的权重，具有较好的自适应特点，能在一定程度上消除背景运动对前景目标运动显著性的干扰．Barnich 等[9]提出了一种“快速鲁棒的像素级背景建模”的视觉背景检测器(visual background extractor, ViBe)算法，其主要思想是为图像空间的每一个像素点构建一个样本集，样本集由该像素点在过去时刻的颜色空间特征以及其空间相邻像素的特征构成；在接下来的视频帧中对该像素新值与样本集进行比较，来判断其属于前景点还是背景点．ViBe 算法具有良好的实时性，检测效果也比较明显，算法的鲁棒性较高；但是由于该算法使用第一帧来初始化背景的样本集，如果运动的显著目标出现在第一帧中，将会把目标当作背景建立样本集，这样当在以后的视频帧中存在目标运动时，原先被遮挡的背景将被检测为前景目标，形成鬼影，严重影响检测效果．柳欣等[10]提出了利用背景低秩和运动目标稀疏性特点提取运动显著目标的方法，虽然此类方法的检测效果较好，鲁棒性也较高，但是对参数选取的依赖性较强，无法适应于多种类型场景．

通过以上分析可以看出，现有的算法在提取运动显著性上存在一些局限．虽然大部分算法都采用了空间和时间的显著性融合的方法[11]，但是融合的方式和权重的确定对结果的影响并没仔细分析．例如，Itti 的算法[5]和 PQFT 算法[3]都是把运动信息作为 4 个信息源通道中的一个通道来进行等权重的融合，没有考虑运动特征和空间特征的差异；Kim 的算法[8]将时间和空间作为一个整体来考虑，虽然不需要人为调节权重，事实上把时间和空间的信息同等考虑减弱了运动分量所起的作用，最终结果也不够理想．针对上述问题，根据生物学研究，人

的视觉系统对运动信息更敏感这一特点[1]，本文提出根据时间显著性的效果，动态设置权来自适应调节空间和时间的显著性融合，以达到合理地均衡两者的信息．此外，目前大多数视频显著性算法中仅使用相邻 2 帧[3]或 3 帧进行运动估计，这对带有扰动背景的处理效果很差．本文提出利用多帧的运动估计来代替仅使用相邻帧的估计，利用显著目标在运动的方向和速度大小上具有一致性的特点，能够克服在扰动背景下的时间显著性的提取．

综上所述，本文针对具有复杂扰动背景的视频提出一种基于时空特征动态融合的显著性提取算法(dynamic fusion of spatial temporal saliency, DFSTS)．采用现有在空间显著性计算中公认的最优算法——密集稀疏重建(dense and sparse reconstruction, DSR)算法[12]，计算每帧图像的空间显著性，保证了显著目标的空间完整性；在连续多帧图像的运动估计中，本文引入“运动熵”来衡量目标的运动一致性以区分显著目标与扰动背景.为了进一步改善算法性能，本文提出利用运动幅度随时间进行指数衰减加权来消除较早帧的运动对当前帧的影响，克服由于多帧处理带来的运动拖影；考虑到目标和背景的运动在方向和速度大小上都有明显的差异，模仿视网膜神经细胞的中心周边机制[13]，把运动矢量的中心周边差作为时间显著性提取的一个要素，进一步区分前景和背景运动，在一定程度上克服了摄像头平动和抖动情况造成的误差；动态调节权重的融合方法，同时保证目标的空间完整性和运动显著性．最后在 4 个视频数据库上进行了实验，验证 DFSTS 算法的实际效果．

1 时空显著性动态融合算法

与静态图像的显著性算法不同，在视频序列中单帧静态图像上具有显著性的目标，由于不一定存在时间上的运动信息，在视频序列中不一定是显著的[1]．但是，空间上的亮度、颜色等特征同样也对运动目标的显著性有所贡献，不能完全不考虑．因此，在计算视频显著性时，应该同时考虑目标的运动信息和空间特征．视频中的显著目标往往具有以下特点：前景目标完整且可描述；在连续帧中，前景目标的各部分运动规律具有一致性，且与背景的运动特征显著不同；其运动速度的大小和方向都与其周边区域显著不同，使得这个目标突出出来．

基于以上的假设，本文提出了如图 1 所示的时空显著性动态融合的算法框架，在空间显著性的计算方面采用了现有的算法，对空间显著性计算进行了简化；在时间显著性方面，主要细化了时间显著图的计算过程．首先对视频序列中的当前帧(时刻 t)通过超像素分割的方法计算其空间显著图，保证了显著目标的空间完整性．其次，在时间显著性方面，因目标运动在时间上存在一致性，考虑包含当前帧在内的连续 $N+1$ 帧的运动估计信息，算法对视频序列首先进行存储，当内存中的帧序列达到算法设定的连续 $N+1$ 帧数据时，计算 t 时刻的当前帧的时间显著图．对包含当前帧在内的连续 $N+1$ 帧的运动估计信息进行处理，获得第 t 帧的运动幅度熵显著图和运动方向熵显著图，而运动幅度中心周边差显著图和运动方向中心周边差显著图的计算则仅利用当前帧和前一帧的运动估计．最后根据时间显著图的效果设置动态权，对空间显著图与时间显著图进行自适应融合，得到目标既具有空间完整性同时又具有运动显著性的融合显著图．

1.1 空间显著性

人的视觉系统在感知显著目标时通常是捕捉一个概念上完整可定义的目标．本文选取了目前在图像显著性处理中性能突出的基于图像 SLIC 超像素分割[14]的 DSR 算法[12]计算当前帧的空间显著图，以获得空间完整的显著目标．

由于图像边缘部分更可能是背景，显著目标更可能出现在图像中心．DSR 算法用 SLIC 超像素概念将每帧分割成紧密的图像超像素，利用图像边缘超像素的颜色特征构造背景稀疏字典，并计算其他超像素利用该字典进行稀疏重建的误差；另一方面，对边缘超像素特征进行主元分析(principal component analysis, PCA)提取，利用提取的背景主元对其他超像素进行密集重建，并计算重建误差作为显著性的衡量；最后对多尺度计算的密集和稀疏重建误差显著图进行贝叶斯融合得到最终的空间显著图 $\boldsymbol{C}_{\mathrm{S}}$．空间显著性的计算详见文献[12]，这里就不再赘述．

1.2 时间(运动)显著性

计算时间显著性需要估计相邻 2 帧间目标的运动特征，即速度矢量．常用的运动信息估计方法

有连续帧差分法、块匹配法、光流法等. 连续帧差分法虽然速度最快，但是不能确定运动的方向；光流法的计算复杂度最大，很耗时；而块匹配法能够同时确定运动的方向和大小，计算复杂度处于前

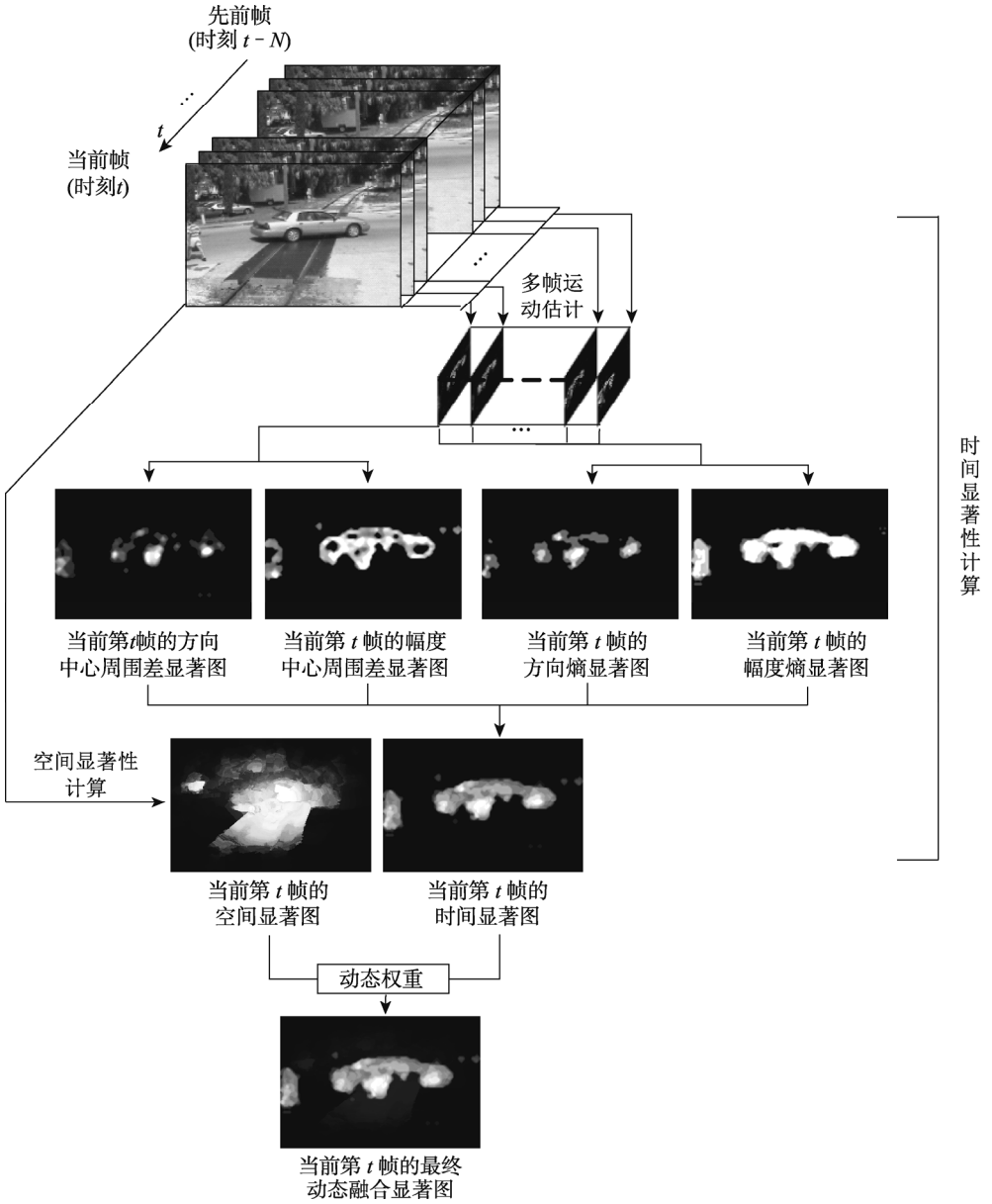


图 1　时空显著性动态融合算法框架

2 种算法之间，能比较全面地反映显著目标的运动特征. 本文采用计算复杂度比较折中的块匹配法. 运动估计的计算复杂度和所取块的尺寸成反比，本文通过实验分析选取适当子块的尺度(详见第 2.4 节). 另外，为了减少运算量提高运算速度，本文算法采用自适应十字搜索(adaptive rood pattern search, ARPS)的块匹配算法[15]估计运动，将图像划分为互不交叠的 $n\times n$ 的子块，将当前帧的子块与前一帧相应位置及其周边的子块进行比较，找到最相似的子块，子块偏移的位置矢量即为该子块运动估计. ARPS 算法的特点在于每次匹配都以前一帧相应位置的子块为中心，与该子块和它十字的 4 个方向的子块进行匹配计算，每次匹配用像素的绝对误差和作为相似性的度量，以绝对误差和最小的子块作为中心循环计算，直至所找到的匹配子块相似度达到设定的阈值为止，该方法的匹配速度较快.

通过相邻 2 帧间的子块匹配算法，得到了视频序列中所有帧的子块在 x, y 方向上的运动分量 M_x^t, $M_y^t, t\in\{1,2,\cdots,N_{\text{img}}-1\}$，其中 N_{img} 为视频序列的总帧数. 在时刻 t 某个子块运动的幅度及方向估计表示为

$$M=\frac{\sqrt{M_x^{t2}+M_y^{t2}}}{M_{\max}},$$

$$\theta_M=\arctan\frac{M_y^t}{M_x^t},\quad \theta_M\in[-\pi/2,\pi/2].$$

其中, $M_{\max}$ 为每相邻 2 帧运动估计时所有子块中出现的运动幅度最大值, 这样得到的是归一化后的运动幅度.

具有运动显著性的目标在运动幅度和方向上一般都具有一致性的特点, 从而能很好地与背景区分开来. 为了衡量运动一致性的特征, 计算运动时间显著图时, 通过一段时间连续的 $N+1$ 个运动估计的幅度 $M^{t-N},\cdots,M^t$, 可以得到目标在一段时间的运动特征, 这样比只取相邻 2 帧的运动更能反映目标的运动特点. 但是由于引入连续多帧的运动, 容易导致在早先帧中的显著目标出现在后续帧中, 形成运动拖曳. 为了减弱这种现象对显著性检测的影响, 本文对运动估计的幅度值随时间进行指数衰减, 以减小早先帧中的运动影响. 对每个子块的运动估计幅度重新定义为 $m^i=M^i\times e^{(i-t)}$, $i\in\{t-N,\cdots,t\}$, 得到重新定义的运动幅度 $m^{t-N},\cdots,m^t$, 当运动幅度小于阈值时可看作该子块是静止的. 对子块 s, 计算 2 个变量:

1) 子块 s 的运动幅度熵

$$M_{\text{ent}}(s)=-\sum_m p_m\lg p_m$$

其中, p_m 为运动幅度 m 在连续 $N+1$ 帧的所有子块中相应值的概率. 采用更多帧的运动估计能够较好地保证得到具有运动一致性的目标. 运动幅度熵反映了一个子块的运动在 $N+1$ 帧的所有子块的运动中出现的可能性的大小, 子块的运动幅度熵的值越大, 代表这个子块的运动幅度与其他子块的差异越显著, 显著性也就越高. 对一帧中所有的子块都计算其运动幅度熵, 并拼接为一张图, 最后对该图进行中值滤波和归一化得到运动幅度熵显著图. 注意, 以子块为单位的运动幅度熵显著图的尺寸比原图小 n 倍.

2) 计算子块与其相邻子块的差异. 根据 Itti 的中心周边理论[6], 一个子块和周边其他子块的运动差异越大, 这个子块也就越显著. 在静态图像的显著性计算中, 常采用颜色和亮度等特征计算中心周边差. 在运动估计中, 为了得到运动显著性的差异, 采用运动估计幅度的特征来计算中心周边差, 因此定义子块 s 的运动幅度的中心周边差

$$M_{\text{cs}}(s)=\frac{1}{8}\sum_{z\in N_s}|M_s-M_z|;$$

其中, M_s 为中心子块的运动幅度, N_s 为该子块在它所在帧的邻域, M_z 为邻域中子块的运动幅度. 图像边缘的子块以图像边缘为对称轴作镜像扩展处理. 子块运动幅度的中心周边差值越大, 这个子块的显著性也就越高. 同样, 对子块的中心周边差信息拼接为一张图并进行中值滤波和归一化, 得到中心周边差的显著图. 通过运动幅度熵和运动幅度中心周边差 2 张显著图, 就可以提取在运动幅度上显著的目标.

类似地, 计算子块 s 运动方向熵显著图(方向熵不需要随时间进行指数衰减)和运动方向中心周边差的显著图

$$\theta_{\text{ent}}(s)=-\sum_{\theta_M}p(\theta_M)\log p(\theta_M),$$

$$\theta_{\text{cs}}(s)=\frac{1}{8}\sum_{z\in N_s}|\theta_s-\theta_z|,$$

得到另外 2 张运动方向上的显著图. 同样, 在运动方向熵显著图和中心周边差显著图中, 如果一个子块的运动方向熵值越大, 说明该子块运动在方向上与其他子块的差异更加显著, 它的显著性也就越高. 子块的中心周边差值越大, 说明这个子块和周边子块的运动方向不同, 这个子块也更加显著.

1.3 动态融合

如前所述, 视频序列中的显著目标应该具有完整性和可定义的特点, 更重要的是在时间序列上, 显著区域在运动的幅度和方向上都与其周边的区域显著不同, 此时时间显著性的作用应更大些. 然而, 如果运动信息分布比较平均(如摄像头以一定速度平动), 此时空间显著性起的作用应重一些. 根据文献[11]对不同时空显著性融合方法的评价, 本文提出利用时间显著性的效果设置动态权, 采用动态加权的方法将空间和时间显著性进行融合, 计算出最终显著图. 最后的视频显著图(saliency map, SM)表示为

$$\mathbf{SM}=\omega_1\boldsymbol{C}_{\text{S}}+\omega_2\boldsymbol{C}_{\text{T}} \tag{1}$$

其中, 时间显著图 $\boldsymbol{C}_{\text{T}}$ 的计算是对运动幅度和方向上的显著图分别取最大值相加

$$\boldsymbol{C}_{\text{T}}=\max(\boldsymbol{M}_{\text{ent}},\boldsymbol{M}_{\text{cs}})+\max(\boldsymbol{\theta}_{\text{ent}},\boldsymbol{\theta}_{\text{cs}}).$$

需要注意的是, 因为运动显著性计算中的运

动估计是以块匹配方式进行的，为了保证空间和时间显著图尺寸大小一致，在进行融合前，需要将 4 幅时间显著图尺寸按照块匹配划分的比例 n 变换回原图计算.

式(1)中，ω_1, ω_2 分别为空间显著图 $\boldsymbol{C}_{\mathrm{S}}$ 和时间显著图 $\boldsymbol{C}_{\mathrm{T}}$ 的动态权重，

$$\begin{cases} \omega_1 = \dfrac{\mathrm{mean}(\boldsymbol{C}_{\mathrm{T}})}{\max(\boldsymbol{C}_{\mathrm{T}})} \\ \omega_2 = \dfrac{\max(\boldsymbol{C}_{\mathrm{T}}) - \mathrm{mean}(\boldsymbol{C}_{\mathrm{T}})}{\max(\boldsymbol{C}_{\mathrm{T}})} \\ \omega_1 + \omega_2 = 1 \end{cases}.$$

在背景静止的视频数据中，运动的显著目标在图像中的比例较小，使得时间显著图的均值接近于零，从而权重 ω_1 接近于零，这样运动显著性的权重增大，融合显著图包含更多的运动信息. 当时间显著图的运动信息较平均和分散时，时间显著图的均值也更接近最大值，这样提高空间显著性的权重来保证目标的完整性，对时间显著图进行修正.

图 1 所示的算法框架也展示了 Sheikh's 数据库中人车视频在计算视频显著性过程中每一个显著图. 在算法中，最终的融合显著图进行了归一化.

2　实验及讨论

2.1　实验

为了综合评价本文算法的性能，选取了 4 个测试数据库: Segtrack 数据库[16], ComplexBG 数据库[17], Sheikh's 数据库[18]和 Dataset2014 数据库[19]，这些数据库的视频涉及了目标大小不一、背景信息不同的各种情况. Segtrack 数据库共有 6 个视频序列，主要包含复杂背景以及摄像头运动的情况，显著目标尺寸较小，每个视频有 29~70 帧不等的图像帧，图像尺寸为 320×240 像素; ComplexBG 数据库共有 9 个视频序列，主要包含摄像头监控视频、复杂扰动背景情形下的多尺度多目标视频，每个视频有 3 000 多帧，图像尺寸为 160×128 像素; Sheikh's 数据库包含 3 个视频序列，其中只有一个序列提供了地面真实(ground truth, GT)，本文仅对这个 500 帧的视频序列进行实验，图像尺寸为 320×240 像素; Dataset2014 数据库共有 11 个视频序列，视频内容主要包含了复杂扰动背景和摄像头抖动等情况，涵盖了彩色、红外及湍流等多种类型，本文只针对其中 9 个彩色视频进行实验. 以上 4 个数据库都提供了通过人眼跟踪设备记录的实验观察者标记的连续显著区域 GT，以便对每个视频序列定量评价各算法的显著性提取结果. 虽然实验证明调整数据的图像尺寸可加快运算速度，但是为了防止由此可能引起的图像信息丢失，实验保持了输入视频的原始尺寸.

为了全面评估 DFSTS 算法性能，将之与目前一些主流算法进行了比较，这些算法包括 FT[7], PQFT[3], Kim[8]和 ViBe[9]. 实验中提出的算法选取 N=15 连续帧计算时间显著性，选取运动估计的图像块大小为 4×4 像素，关于图像块尺寸的选取原则详见第 2.4 节. 软件环境为 Matlab2013a，硬件条件为 4 GB RAM 的 Intel Core 2 Duo E8400，未进行并行加速处理.

2.2　可视化评价

作为算法性能评价的第一步，通过可视化的方法从定性的角度进行初步对比. 在 4 个数据库选取了具有不同信息特征的典型视频序列，以验证以上视频显著性算法的效果. 不同视频序列的具体信息如表 1 所示. 这些视频片段都是在室内或者室外环境中拍摄，涵盖了从简单到复杂、不同情景的运动状况，大部分视频都包含了不相关的复杂扰动背景，例如摇摆的树枝、降落的雪花、水波等等，还包含了环境亮度变化的情形. 其中的 Cheetah 视频包含了摄像头的运动，同时目标运动不规律，处理情况较复杂.

表 **1**　视频序列信息

视频名	所属数据库	帧数	尺寸	视频信息说明
Birdfall	Segtrack	30	240×320	一只鸟在复杂树林背景中向下飞行
Cheetah	Segtrack	29	320×240	一只羚羊在草原上奔跑，摄像头随羚羊运动
Shopping Mall	ComplexBG	1 285	320×256	一段购物商场的监控视频，其中有许多行人
Campus	ComplexBG	1 438	160×128	在树叶随风摆动的背景中，行人和汽车相继进入场景并离开的场景
Car	Sheikh's	200	360×240	行人和车辆相向穿插进入并离开
Skating	Dataset2014	180	280×180	大雪纷飞，两人由近向远滑雪
Boats	Dataset2014	113	320×240	在水波荡漾的湖面驾驶快艇驶过

图 2 所示为不同运动显著性算法在各个视频序列上的检测结果. 可以看出, 参与对比的算法在处理复杂带扰动背景下的运动目标显著性存在一些缺陷, 不能很好地抑制如同在 Campus 视频数据中存在的树叶摇晃这类非显著性运动; 对于背景静止的情况, 所有算法都可以得到较好的结果, 其中 DFSTS 算法和 ViBe 算法效果最好, 但是从 Birdfall

a. 视频当前帧

b. GT

c. FT 算法[7]

d. Kim 算法[8]

e. PQFT 算法[3]

f. ViBe 算法[9]

g. DFSTS 算法

图 2 运动显著性算法实验结果对比

视频可以看出, ViBe 算法容易引入“鬼影”, 显著性目标容易在后续帧的显著图上持续出现, 导致显著性提取出现错误.

在具有摄像头运动的 Cheetah 视频中, 对于摄像头移动引起的背景运动, ViBe 算法不能提取出显著目标; PQFT 算法不能很好地处理杂乱背景的图像, 因为将空间特征和运动信息共同作为四元数计算显著性, 并没有考虑空间特征和运动信息对于运动显著性的贡献程度, 所以显著图在得到运动显著目标的同时也引入了一些不相关的空间显著性; Kim 算法过于依赖梯度变化来表征特征变化以及运动变化, 所以突出了显著目标的边缘, 失去了目标的完整性; 本文提出的 DFSTS 算法效果优于其他算法, 不需要进行预处理即可去除摄像头运动引入的误差, 这是因为该算法采用运动方向和运动幅度的中心周边差作为运动显著性衡量的因素, 虽然背景和显著目标都在运动, 但是两者的运动方向和速度大小存在明显差异, 可以通过中心周边差进行区分, 在一定程度上消除了摄像头运动造成的运动估计误差. 另外, DFSTS 算法还能够处理相机抖动情况下引入的运动误差. 相机抖动引入的运动误差和摄像头平动类似, 其结果是对目标和背景都会产生同样的运动幅度和方向影响, 在抖动不大的时候, 对本文算法的影响主要

在于使得方向熵和幅度熵下降，无法获得较为理想的运动熵显著图，但是对中心周边差的影响不大，这样对时间显著性的计算进行了一定的弥补．同时，在时间显著性不够理想的情况下，通过动态加权融合的办法提高了空间显著性的权重，能够在一定程度上提高最终的显著性检测效果．

在运动目标存在遮挡的 Skating 视频中，虽然显著目标滑雪者被树干遮挡，但是 DFSTS 算法仍然检测出了运动目标，表明该算法对运动显著目标丢失、遮挡等情况具有一定的鲁棒性．这是由于不同于已有算法利用相邻帧的运动估计，DFSTS 算法利用多帧运动估计，只要保证在选取的多帧估计中运动目标没有全部丢失，仍能反映运动的一致性特点，就可以检测出运动目标．但是在 Skating 视频上 DFSTS 算法结果要略差于 ViBe 的算法，原因见第 2.5 节．

以上仅是在各个视频中选择了一帧作为代表展示不同算法的实验效果．为了更好地说明视频的显著性算法效果，选取部分视频的多个连续帧进行展示．图 3 所示为 Birdfall 视频的 3 帧画面，显示了一只鸟从复杂的树林背景中向下飞行的过程．可以看出，由于复杂的背景特征，FT 和 PQFT 算法都不能很好地提取显著目标；Kim 算法可以得到显著目标，但是目标的完整性和突出程度都比较差；DFSTS 算法和 ViBe 算法都得到了完整性良好的显著目标，但是 ViBe 算法的缺陷是导致了“鬼影”的产生，先前帧中的显著目标会在后续的显著图中出现，这样的结果并不是我们所期望的．图 4 所示为 Campus 视频结果，视频中的运动目标开始是行人，随后汽车进入场景，继而又有目标的进入和退出，同时加上背景树叶随风摆动的干扰，是一个复杂扰动背景的多目标视频．可以看出，其他算法不能很好地抑制由于树叶摇晃导致的复杂背景扰动，而 DFSTS 算法在抑制背景扰动的同时，很好地将显著目标提取了出来．

2.3　量化评价

为了定量地比较 DFSTS 算法和其他对比算法的性能，本文选用查准率 P 和查全率 R 这 2 个参数，并通过绘制 PRC (precision-recall curve) 曲线的方式评价算法的优劣．对视频序列每一帧的显著图进行阈值处理，将得到的二值图与 GT 进行比较，计算算法的查准率和查全率

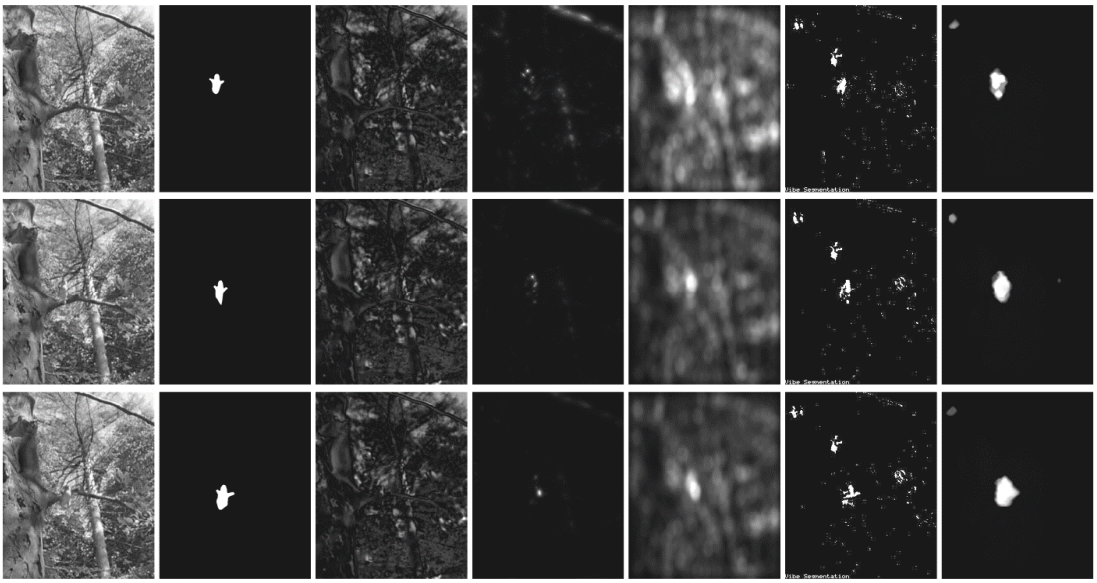
$$P=\frac{G\cap S}{S},$$

$$R=\frac{G\cap S}{G},$$

$$\mathrm{PRC}=\{(R_{\mathrm{thr}},P_{\mathrm{thr}})\mid \mathrm{thr}=0,1,\cdots,255\}$$

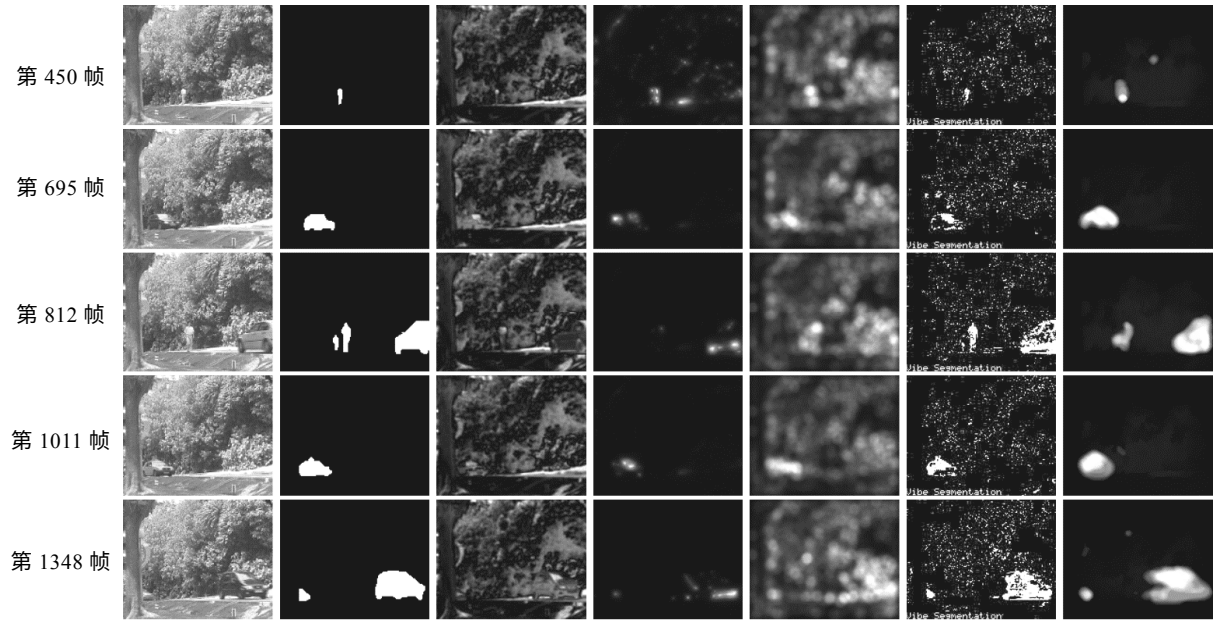
其中，G 为 GT，S 为显著图，$G\cap S$ 为被算法正确标记提取的显著区域；R_{thr}，P_{thr} 为在相应阈值时所得的 R 和 P．查准率考察的是正确计算为显著的像素数的比例，而查全率衡量的是检测到的显著像素数占 GT 中显著像素数的比例．PRC 曲线对于算法的评价准则是：在相同的查全率的情况下，查准率越高的算法效果越好．

为了综合考虑查全率和查准率表征的算法性能，定义综合评价指标(F-measure, F)值



a. 视频当前帧　b. GT　c. FT 算法[7]　d. Kim 算法[8]　e. PQFT 算法[3]　f. ViBe 算法[9]　g. DFSTS 算法

图 3　Birdfall 视频实验结果对比



a. 视频当前帧 b. GT c. FT 算法[7] d. Kim 算法[8] e. PQFT 算法[3] f. ViBe 算法[9] g. DFSTS 算法

图 4 Campus 视频实验结果对比

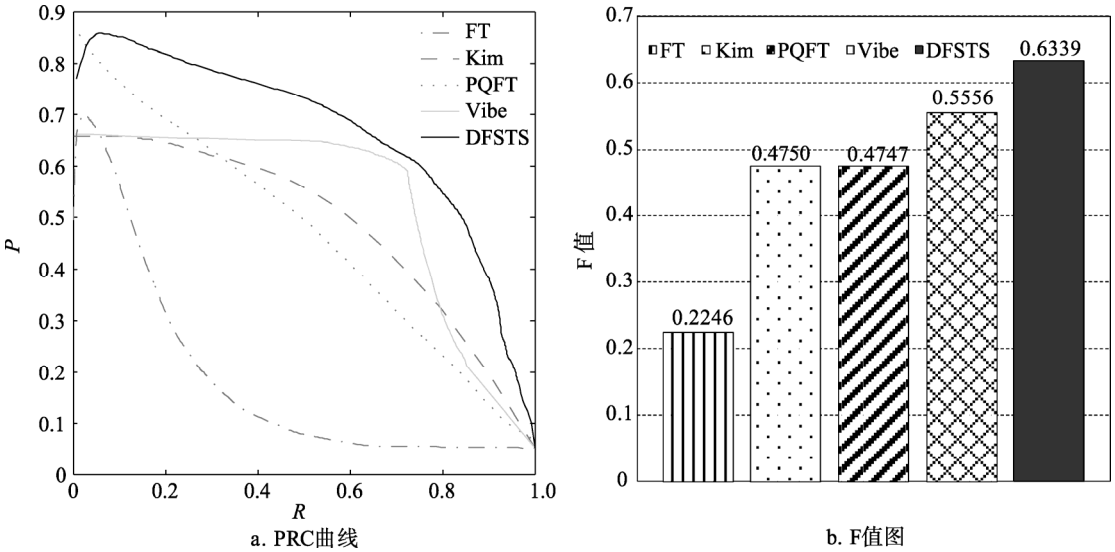
$$\mathrm{F}=\frac{2\times P_{\mathrm{thr}}\times R_{\mathrm{thr}}}{P_{\mathrm{thr}}+R_{\mathrm{thr}}},$$

其中 R_{thr}, P_{thr} 为在相应阈值时所得的 R 和 P. F 值有效地表征了显著性算法在提取显著目标的同时又抑制非显著目标的能力. F 值越高, 算法的显著性检测效果越好.

对每个数据库计算不同算法的 F 值, 以最大的 F 值作为展示, 得到相应的量化评价结果如表 2 所示. 可以看出, 在 3 个数据库上 DFSTS 算法显著性提取的平均性能要优于其他算法, 尤其是在处理具有复杂扰动背景的情况下效果更加突出. 针对所有的视频数据库, 绘制 DFSTS 算法和其他算法对比的 PRC 曲线及 F 值量化比较结果如图 5 所示. 从图 5b 可以看出, DFSTS 算法具有最高的 F 值, 表明该算法在提取视频显著性上效果最佳.

表 2 运动显著性算法的 F 值对比

视频名	FT	Kim	PQFT	ViBe	DFSTS
Segtrack	0.198 9	0.340 8	0.391 4	0.243 2	**0.476 7**
Sheikh's	0.104 2	0.617 1	0.557 0	0.738 6	**0.798 2**
ComplexBG	0.217 8	0.404 3	0.445 4	0.592 7	0.544 0
Dataset2014	0.377 4	0.537 8	0.505 2	0.648 1	**0.**716 7
平均值	0.224 6	0.475 0	0.474 7	0.555 6	**0.633 9**



a. PRC曲线 b. F值图

图 5 不同视频显著性算法的量化评价

2.4　图像子块的尺度选取讨论

使用块匹配算法进行运动估计时，块的大小不仅影响着算法的执行速度，而且也影响着算法的最终效果. 为此，本文以 Sheikh's 数据库的视频为例，选取不同的图像块尺寸，讨论块尺寸大小和计算速度及检测效果的关系，以确定选取最合适的尺寸用于 4 个数据库的整体评价. 实验中，选取块尺寸为 2×2, 4×4, 8×8, 12×12; 同时也加入了光流法作为参考，结果如表 3 所示. 可以看出，使用光流法进行运动估计的计算速度要明显慢于块匹配法，同时由于运动目标的亮度变化和运动噪声，检测效果也不是最好. 在块匹配法下，计算的复杂度和块的尺寸大小成反比，匹配块的尺寸选取影响着显著性算法的速度. 另一方面，因为将一个完整的图像块看成一个整体估计其运动矢量，会丢掉了面积小于设定图像块的目标所具有的运动特征，造成无法检测到较小的运动目标，检测效果随着块尺寸的增加也有一定的损失. 因此在选取算法的块尺寸时，本文进行了效率和效果上的平衡，最终选取运动估计的图像块大小为 4×4.

表 **3**　不同块尺寸及光流法性能实验结果

	光流法	块尺寸			
		2×2	4×4	8×8	12×12
F 值	0.776 2	0.799 2	0.798 2	0.775 8	0.739 4
图像处理速度/帧/s	25.810 1	8.451 9	5.318 8	4.636	4.151 9

在实验中还发现：视频中小目标的运动主要指背景中的扰动(如图 4 所示 Campus 视频的树叶随风摇摆)，正是显著性检测中需要抑制和消除掉的. 同时，当存在更大目标的运动时，人的视觉系统也会首先被更大目标的运动所吸引，大运动目标的显著性应会强于小目标的显著性，这从实验中各个结果显著图的目标亮暗程度可以看出. 所以，DFSTS 算法和实际视觉系统的处理是一致的. 当然，在没有大目标运动的情况下，对于小目标的运动估计是需要的，算法也可以调整块匹配的尺寸至 2×2 以适应这种要求.

2.5　算法局限性

虽然 DFSTS 算法能够较好地处理具有杂乱扰动背景的情况，但是对于背景中存在具有运动一致性的干扰时，运动显著性处理效果任然难以令人满意，具有一定局限性. 例如，在处理 Skating 视频中，DFSTS 算法在提取运动显著目标时，飞落的雪花在雪地中的运动进行块匹配估计时是无法估计出的，但是在草丛中的运动能够被估计出来；同时雪花的运动在运动幅度和方向上都具有一致性的特点，在画面中也占据较小的范围，因此被判定为较小的运动显著目标，从而无法从运动特征上和预期的显著目标区分开来，导致最后检测结果的失败. 在未来工作中，还需要进一步引入对前景和背景一致性特征区分的判别机制. DFSTS 算法程序目前没有进行优化，因此暂时还不能到达实时处理视频的速度.

本文虽然通过实验和分析后提出采用了 4×4 像素的块进行运动估计，但对于视频中没有大目标和运动的目标又小于 4×4 像素的情况下，算法会造成小目标不能准确估计的问题，调整块的尺度大小是一个办法，但目前尚无较好的自适应调整策略，有待进一步的改进.

3　结　语

本文提出了一种处理复杂扰动背景情况下提取视频的显著性算法. 该算法利用了在时间序列上显著目标具有的运动一致性，根据时间显著性调整空间和时间显著性融合的动态权重，得到最终的视频序列注意力选择目标. 为了保证显著目标的完整性，采用目前主流的 SLIC 分割方法计算空间显著性. 利用连续的多帧图像的运动估计计算运动特征的熵表征图像的一致性. 根据经典的中心周边理论，又可以很好地区分显著目标与背景的运动差异.

实验结果表明，DFSTS 算法对于具有复杂扰动背景的视频数据具有很好的实验效果，能够较好地解决在复杂扰动背景下的单目标及多目标的显著性检测问题，对于固定摄像头、摄像头平动以及摄像头抖动情况的处理效果也较为理想，在显著目标被遮挡或丢失的情况也具有一定的鲁棒性. 但是，对于检测小于选取的块匹配尺寸的显著目标，以及处理部分背景运动具有一致性的视频数据时，还存在一定局限性，可以进一步提高和改进.

参考文献**(References):**

[1] Ma Y F, Hua X S, Lie L, *et al.* A generic framework of user at-

tention model and its application in video summarization[J]. IEEE Transactions on Multimedia, 2005, 7(5): 907-919

[2] Gao D S, Vasconcelos N. Integrated learning of saliency, complex features, and object detectors from cluttered scenes[C] //Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2005, 2: 282-287

[3] Guo C L, Zhang L M. A novel multi-resolution spatiotemporal saliency detection model and its applications in image and video compression[J]. IEEE Transactions on Image Processing, 2010, 19(1): 185-198

[4] Mahadevan V, Vasconcelos N. Background subtraction in highly dynamic scenes[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2008: 1-6

[5] Itti L, Baldi P. A principled approach to detecting surprising events in video[C] //Proceedings of IEEE Computer Society, Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2005, 1: 631-637

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

[7] Achanta R, Hemami S, Estrada F, *et al.* Frequency-tuned salient region detection[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2009: 1597-1604

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

[9] Barnich O, Van Droogenbroeck M. ViBe: a universal background subtraction algorithm for video sequences[J]. IEEE Transactions on Image Processing, 2011, 20(6): 1709-1724

[10] Liu Xin, Zhong Bineng, Zhang Maosheng, *et al*. Motion saliency extraction via tensor based low-rank recovery and block-sparse representation[J]. Journal of Computer-Aided Design & Computer Graphics, 2014，26(10): 1753-1763(in Chinese)
(柳 欣, 钟必能, 张茂胜, 等. 基于张量低秩恢复和块稀疏表示的运动显著性目标提取[J]. 计算机辅助设计与图形学学报, 2014, 26(10): 1753-1763)

[11] Muddamsetty S M, Sidibé D, Trémeau A, *et al*. A performance evaluation of fusion techniques for spatio-temporal saliency detection in dynamic scenes[C] //Proceedings of the 20th IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2013: 3924-3928

[12] Li X H, Lu H C, Zhang L H, *et al.* Saliency detection via dense and sparse reconstruction[C] //Proceedings of IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2013: 2976-2983

[13] Packer O S, Dacey D M. Synergistic center-surround receptive field model of monkey H1 horizontal cells[J]. Journal of Vision, 2005, 5(11): 1038-1054

[14] Achanta R, Shaji A, Smith K, *et al.*Slicsuperpixels[R]. Lausanne: ÉcolePolytechniqueFédérale de Lausanne, 2010

[15] Yao N, Ma K K. Adaptive rood pattern search for fast block-matching motion estimation[J]. IEEE Transactions on Image Processing, 2002, 11(12): 1442-1449

[16] Tsai D, Flagg M, Nakazawa A, *et al*. Motion coherent tracking using multi-label MRF optimization[J]. International Journal of Computer Vision, 2012, 100(2): 190-202

[17] Li L Y, Huang W M, Gu I Y H, *et al*. Statistical modeling of complex backgrounds for foreground object detection[J]. IEEE Transactions on Image Processing, 2004, 13(11): 1459-1472

[18] Sheikh Y, Javed O, Kanade T. Background subtraction for freely moving cameras[C] //Proceedings of the 12th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 1219-1225

[19] Wang Y, Jodoin P M, Porikli F, *et al*. CDnet 2014: an expanded change detection benchmark dataset[C] //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2014: 393-400