DOI:10.11918/j.issn.0367-6234.201812005

基于3D-LCRN视频异常行为识别方法

胡薰尹[1],管业鹏[1,2]

(1.上海大学 通信与信息工程学院,上海 200444; 2.新型显示技术及应用集成教育部重点实验室(上海大学),上海 200072)

摘 要: 自动准确识别监控视频中的异常行为在安防领域具有广泛的应用前景.本文提出一种基于3D-LCRN(3D Long-short-term Convolutional Recurrent Network)视觉时序模型的视频异常行为识别方法.首先,基于视频图像帧间的结构相似性,结合光照感应与光照补偿机制进行背景建模,获取对光照突变与背景运动不敏感的矫正光流场与矫正运动历史图.同时,针对异常与正常行为视频数据失衡问题,计算三通道矫正光流运动历史图COFMHI(corrected optical flow motion history image),随机提取视觉词块进行聚类,对样本数量与维度进行双向扩充,充分获取样本的微分和积分运动信息.在此基础上,采用3D-CNN深度学习网络模型对COFMHI进行学习,获取局部短时序时空-域特征,结合可学习贡献因子加权的LSTM网络以压制无关、冗余、具有混淆性的视频片段,进一步提取由短时序-长时序,由局部-全局的多层次时-空域特征用于异常行为识别.通过与同类方法的客观定量对比,实验结果表明,本文方法在光照突变与背景运动等复杂场景下具有优异的异常行为识别性能,进一步表明该方法有效、可行.

关键词: 矫正光流运动历史图;样本扩充;3D-LCRN;3D-CNN;LSTM;异常行为识别

中图分类号: TP391.7 **文献标志码:** A **文章编号:** 0367-6234(2019)11-0183-11

3D-LCRN based Video Abnormal Behavior Recognition

HU Xunyin[1], GUAN Yepeng[1,2]

(1. School of Communication & Information Engineering, Shanghai University, Shanghai 200444, China; 2. Key Laboratory of Advanced Display and System Application (Shanghai University), Ministry of Education, Shanghai 200072, China)

Abstract: Automatically anomaly recognition in surveillance videos is a crucial issue for social security. A 3D-LCRN visual time series model was proposed for abnormal behavior recognition on video surveillance. Firstly, a structural similarity background modeling method was proposed to obtain corrected optical flow and corrected motion history image, which was insensitive to illumination variation and background moving against background interference in complex scenes. Secondly, a new sample expansion method was proposed to solve the imbalance between normal training samples and abnormal ones, which enriched the spatial and temporal information of samples from both dimensionality and quantity. On dimensionality, the method stacked corrected optical flow and corrected motion history image to generate the corrected optical flow motion history image. In quantity, COFMHI was randomly cropped and clustered into center visual words by K-means. Finally, COFMHI was used as 3D-CNN input to extract local short-time spatial-temporal features of behavior. In order to suppress irrelevant, redundant and confusing video clips, a learnable contribution factor weighted LSTM was used to deeply extract the global long-time spatial-temporal features for abnormal behavior recognition. Through 3D-LCRN, abundant spatial-temporal features were extracted from both local to global and short-time to long-time levels. Experimental results show that the proposed method has excellent performance of abnormal behavior recognition in complex scenes such as illumination variation and background moving in comparison with the state-of-art methods.

Keywords: corrected optical flow motion history image; sample expansion; 3D-LCRN; 3D-CNN; LSTM; abnormal behavior recognition

异常行为的研究关乎人身财产安全,视频监控系统已成为预防犯罪行为和识别安全威胁的流行方式.但是目前用人力来分析海量视频信息非常昂贵和低效,因此需要自动检测和定位可疑异常行为并及时预警.由于人类行为的模糊性和歧义性,异常行为的精准识别具有一定的挑战性.

早期的工作主要提取运动区域的手工特征如光流OF[1]、方向梯度直方图HOG[2]、运动历史图MHI[3]等来对视频进行编码.Hyukmin等[4]通过融合MHI与HOG来对人体行为进行建模与识别.

收稿日期: 2018-12-04
作者简介: 胡薰尹(1994—),女,硕士研究生;
管业鹏(1967—),男,教授,博士生导师
通信作者: 管业鹏,ypguan@shu.edu.cn

Shiyang 等[5]通过光流提取前景目标上的稀疏粒子轨迹来计算运动不稳定性,以实现异常行为识别. 光流场与运动历史图在运动识别领域作为常用特征,在受约束场景下取得了较好效果,但在复杂场景下,上述方法均容易受到图像噪声、光照变化和背景抖动影响. 与提取浅层手工特征相反,大量研究致力于从海量标记视频数据中自动学习深层特征[6]. Christoph 等[7]结合 RGB 图像光流场来训练双流 2D-CNN 网络,达到了较好的行为识别性能. 但是 2D-CNN 容易丢失连续帧间时域运动信息的相关性,而这通常是行为识别的关键特征. 杨天明等[8]提出基于 3D-CNN 的时-空双流网络来进行动作识别. 但是 3D-CNN 只能对短时序间的运动结构进行建模. AMIN 等[9]提出了基于 2D-CNN 与 DB-LSTM 的卷积循环神经网络,可以对长时序间的运动结构进行建模. 考虑到人体运动行为之间具有很强的时间依赖性,需要同时对短时序与长时序间的运动区域进行建模. 此外,公开的训练数据集如 UMN[10]、CAVIAR[11]、Web[12]中正常行为视频片段数量远多于异常行为,使得模型容易陷入过拟合,很难从有限失衡样本中学习到其行为模式.

综上,视频异常行为识别的主要挑战有 3 点:1) 如何在光照变化、背景运动等复杂场景下压制背景干扰,提取出丰富的前景信息用于视频分析. 2) 如何提取多帧间的时-空域结构信息并保留上下文间的时-空相关性用于视频理解. 3) 如何通过有限且失衡的训练样本来训练神经网络. 针对上述问题,本文提出了基于 3D-LCRN 网络的视频异常行为识别方法. 该方法先建立包含光照感应与补偿机制的结构相似性背景模型,用于矫正光流场与运动历史图. 接着,融合多模态特征获得矫正光流运动历史图 COFMHI,并通过聚类扩充样本. 在此基础上,结合 3D-CNN、贡献因子加权的 LSTM 网络,提取 COFMHI 片段的多尺度时-空域特征用于异常行为判别.

1 光流运动历史图

1.1 结构相似性背景建模

光照突变、背景运动都会产生光流场与运动历史图,这些运动信息对异常行为的识别造成了一定的干扰. 为了解决上述问题,本文提出了结构相似性背景建模方法来提取前景,在此基础上矫正光流场与运动历史图,以对抗运动背景和光照突变的干扰.

结构相似性[13]可以在一定程度上反映两张图像的纹理差异. 即使背景是动态的,诸如树叶抖动、水纹波动、电梯运动也不会给背景造成较大的结构性改变,即帧间结构相似性基本保持不变. 而当前景目标运动或光照突变时,帧间结构相似性会降低. 基于背景图像具有结构相似性,背景更新模型定义为

$$B_t(x,y)=(1-S_t(l_{t-1}(x,y),l_t(x,y))\cdot\alpha)B_{t-1}(x,y)+S_t(l_{t-1}(x,y),l_t(x,y))\cdot\alpha\cdot I_t(x,y).$$

式中:α 为学习因子,设置为经验值 0.01,$I_t(x,y)$ 为 t 时刻输入图像在像素 (x,y) 处的像素值,S_t 是表征帧间结构突变程度的抑制因子,定义为

$$S_t(l_{t-1}(x,y),l_t(x,y))=\left(\frac{2\mu_1\mu_2+c_1}{\mu_1^2+\mu_2^2+c_1}\cdot\frac{2\sigma_{1,2}+c_2}{\sigma_1^2+\sigma_2^2+c_2}\right)^2,$$

$$S_t(l_{t-1}(x,y),l_t(x,y))\in[0,1].$$

式中:$l_{t-1}(x,y)$ 为背景图像 $\boldsymbol{B}_{t-1}$ 在点 (x,y) 处的亮度,$l_t(x,y)$ 为输入图像 $\boldsymbol{I}_t$ 在点 (x,y) 处的亮度,$\boldsymbol{\mu}_1$ 和 $\boldsymbol{\mu}_2$ 分别是 $l_{t-1}(x,y)$ 和 $l_t(x,y)$ 的局部均值,$\boldsymbol{\sigma}_1$ 和 $\boldsymbol{\sigma}_2$ 分别是 $l_{t-1}(x,y)$ 和 $l_t(x,y)$ 的局部方差,$\boldsymbol{\sigma}_{1,2}$ 是 $l_{t-1}(x,y)$ 和 $l_t(x,y)$ 之间的协方差(上述参数可通过与 3×3 大小的高斯滤波器卷积获得),c_1 和 c_2 是常数,分别设为 6.5 和 58.5.

部分实验结果如图 1 所示. 其中,图 1(a) ~ (f)依次为 $\boldsymbol{B}_{t-1}$、$\boldsymbol{l}_{t-1}$、$\boldsymbol{\mu}_1$、$\boldsymbol{\sigma}_1$、$\boldsymbol{\sigma}_{1,2}$ 与 $\boldsymbol{B}_t$;图 1(g) ~ (l)依次为 $\boldsymbol{I}_t$、$\boldsymbol{l}_t$、$\boldsymbol{\mu}_2$、$\boldsymbol{\sigma}_2$、$\boldsymbol{S}_t$ 与最终分割得到的前景目标 $\boldsymbol{F}_t$.

为了使模型能尽快地感知场景光线变化,以便做出相应的光照补偿,需加入光照突变感应机制. 当

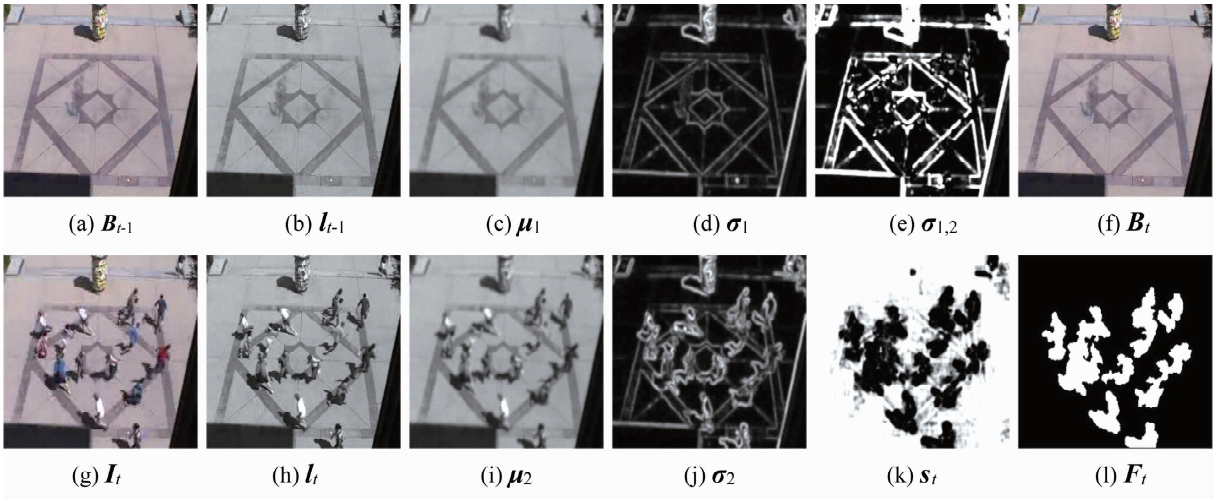


图 1 结构相似性建模中间结果展示

Fig. 1 Experimental results during structural similarity modeling

光照改变时，背景结构会发生变化，结构相似性会降低. 因此，使用最小结构相似性映射的均值 μ_s 来反应环境亮度的变化，定义为

$$\mu_s = \frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}\min(S_{t-1}(l_{t-2}(x,y),l_{t-1}(x,y)),S_t(l_{t-1}(x,y),l_t(x,y))).$$

式中：m、n 分别为图像的长和宽.

图像亮度变化越大，μ_s 越小. 为了降低光照突变对前景分割产生的干扰，当环境光照突变时，即当满足式(1)时，可依据式(2)更新背景 $\boldsymbol{B}_t$：

$$\mu_s < T_u, \tag{1}$$

$$B_t = \begin{cases} B_{\text{bright}}, \text{if } |u_b - u_t| < |u_d - u_t|; \\ B_{\text{dark}}, \text{otherwise.} \end{cases} \tag{2}$$

式中：T_u 是背景变化阈值，实验中设为 0.1[14]，$\boldsymbol{B}_{\text{bright}}$ 为明候选背景，$\boldsymbol{B}_{\text{dark}}$ 为暗候选背景，μ_t 为当前帧 $\boldsymbol{I}_t$ 亮度均值，μ_b 为明背景亮度均值，μ_d 为暗背景亮度均值.

若相邻两帧结构相似性的差异性 Δl_t 在光照突变后达到了历史最小值，则需更新 $\boldsymbol{B}_{\text{bright}}$ 与 $\boldsymbol{B}_{\text{dark}}$：

$$\Delta l_t = \sum_{x=0}^{m-1}\sum_{y=0}^{n-1}\frac{|l_{t-1}(x,y) - l_t(x,y)|}{S_t(l_{t-1}(x,y),l_t(x,y))+1},$$

$$\begin{cases} B_{\text{bright}} = I_t, u_b < u_t; \\ B_{\text{dark}} = I_t, u_d > u_t. \end{cases}$$

部分实验结果如图 2 所示. 其中，(a) ~ (d) 分别为第 500、821、1 193 与 1 400 帧输入图像，(e) 为 μ_s、T_u 曲线，(f) 为 Δl_t 曲线，(g) 为 μ_t、μ_b、μ_d 曲线. 可以看出，当场景光线变化时，μ_s 降低，Δl_t 升高. 模型感知到了光线变化，调整 μ_t、μ_b 与 μ_d 的大小，做出相应的光照补偿.

根据模型计算得到的背景图像，对视频帧和背景图像进行差分和形态学滤波[15]，提取出前景目标

$$F_t = D(F(I_t - B_t)),$$

式中：$D(.)$ 为图像的膨胀运算，$F(.)$ 为图像的腐蚀运算.

部分实验结果如下图. 其中，图 3(a) ~ (d) 为视频原图，图 3(e) ~ (h) 为背景图像，图 3(i) ~ (l) 为前景图像.

1.2　光流运动历史图与样本扩充

1) 光流运动历史图

经上述结构相似性背景建模后，为进一步压制光照变化和背景抖动影响，分别进行运动历史图 MHI 和光流场 OF 矫正：

$$H_t(x,y) = \begin{cases} \tau, \text{if } F_t(x,y) \neq 0; \\ \max(0, H_t(x,y) - \delta), \text{otherwise.} \end{cases}$$

$$d_t(x,y) = \begin{cases} (\sum_N wA^{\mathrm{T}}A)\sum_N wA^{\mathrm{T}}\Delta B, \text{if } F_t(x,y) \neq 0; \\ 0, \text{otherwise.} \end{cases}$$

式中：$H_t(x,y)$ 为第 t 帧像素 (x,y) 处矫正后的运动历史图，$F_t(x,y)$ 为第 t 帧前景图像 (x,y) 处的像素值，τ 为持续时间，δ 为衰退参数(实验中分别设为 50，1)，$d_t(x,y)$ 为像素 (x,y) 处的矫正光流场，w 是像素 (x,y) 的邻域 N 的权重函数[16]，$\boldsymbol{A}$ 与 $\Delta\boldsymbol{B}$ 为扩展系数[16].

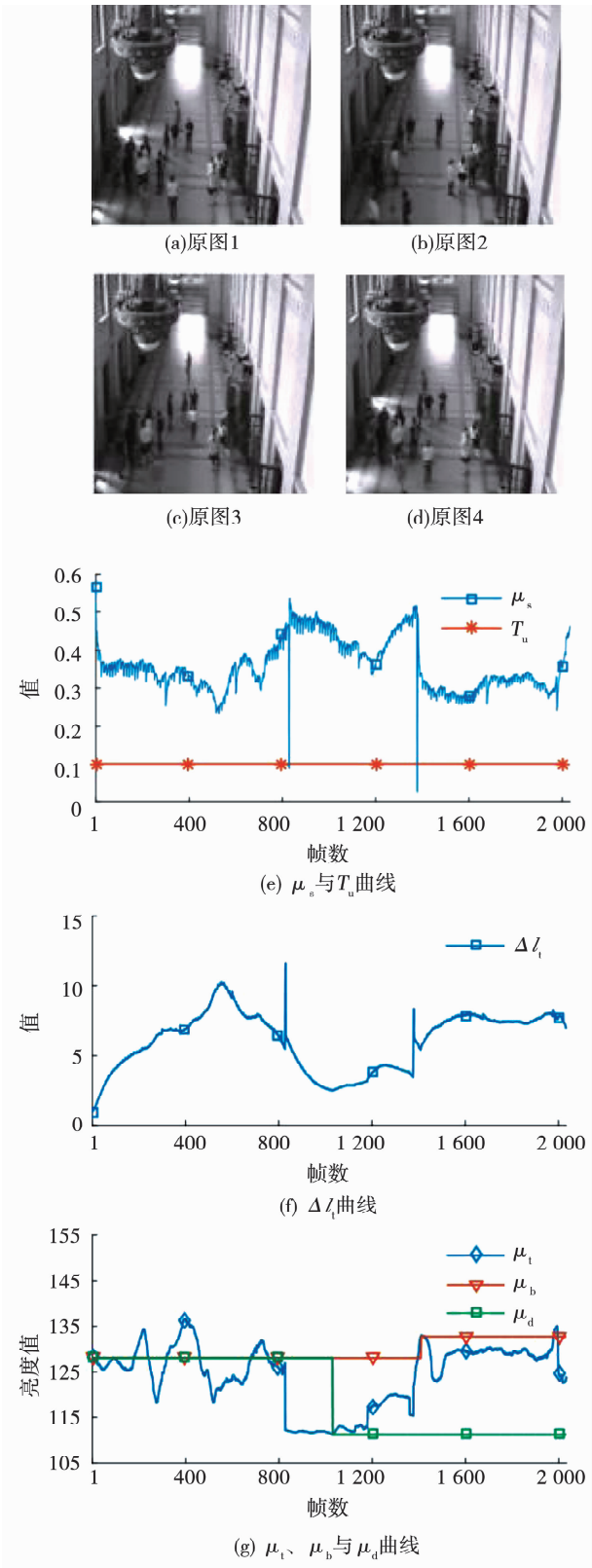


图 2　光照突变感应与补偿

Fig. 2　illumination sensing and compensation

部分实验结果如图 4 所示. 其中,图 2(a) ~ (b)分别为前、后帧输入灰度图像(以前景图像 $\boldsymbol{F}$ 为掩码得到);图 2(c) ~ (e)分别为权重函数 $\boldsymbol{w}$、扩展系数 $\boldsymbol{A}$、扩展系数 $\Delta\boldsymbol{B}$;图 2(f) ~ (h)分别为矫正光流场 $\boldsymbol{d}$ 在 x 方向的分量、矫正光流场 $\boldsymbol{d}$ 在 y 方向的分量与矫正运动历史图 $\boldsymbol{H}$.

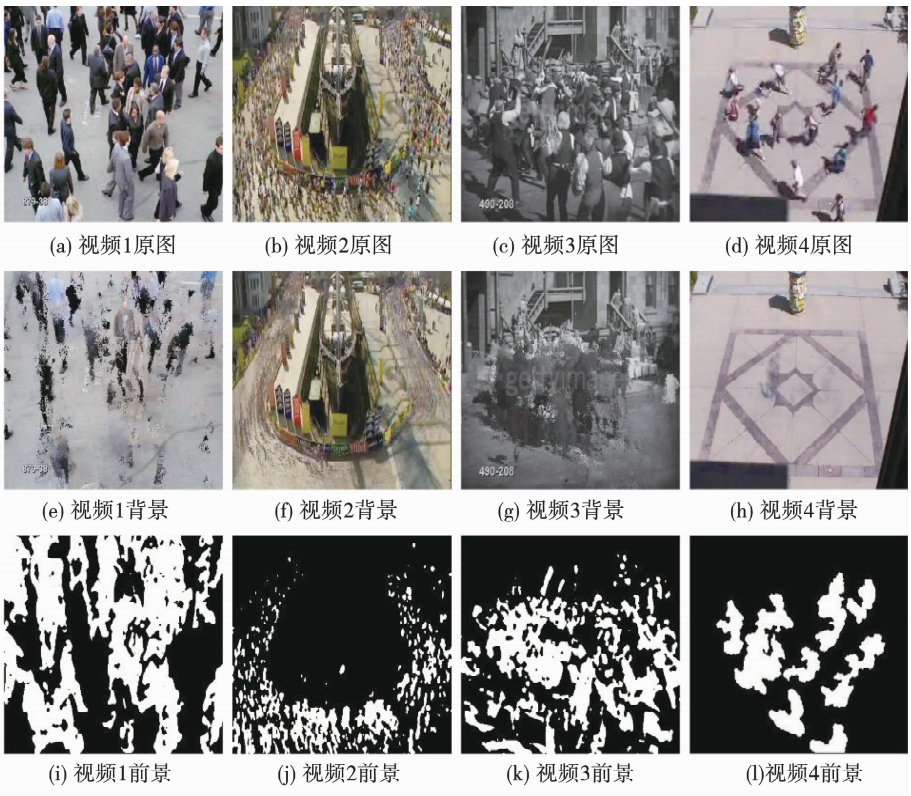


图 3 结构相似性建模得到的背景与前景图像

Fig. 3 Background and foreground images obtained from structural similarity modeling

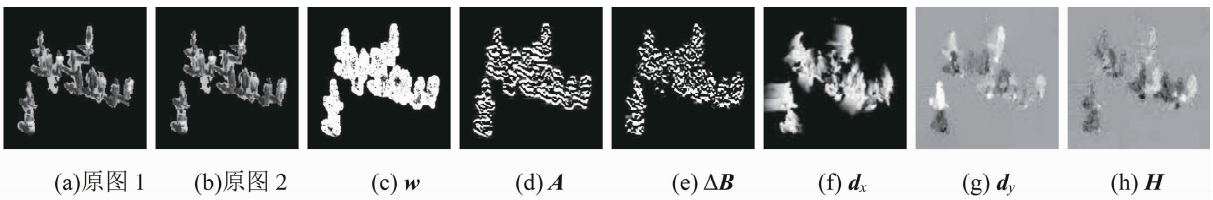


图 4 光流场与运动历史图矫正过程

Fig. 4 The correction of OF and MHI

在获取上述矫正运动历史图 CMHI 和矫正光流场 COF 的基础上,将 COF 分解为水平方向光流图 COFx 与垂直方向光流图 COFy,并对上述图像分别进行归一化后将 CMHI 作为图像的 R 通道,COFx 作为图像的 G 通道,COFy 作为图像的 B 通道. 对三通道进行堆叠形成矫正光流运动历史图 COFMHI.

部分实验结果如图 5 所示. 其中,图 5(a)、(f)、(k)为视频原图,图 5(b)、(g)、(l)为 CMHI,图 5(c)、(h)、(m)为 COFx,图 5(d)、(i)、(n)为 COFy,图 5(e)、(j)、(o)为 COFMHI.

2)样本扩充

由于实际视频监控中正常行为数量往往远超出异常行为,因此,为后续的基于深度学习方法进行视频异常行为识别,需进行相应的异常行为样本扩充. 具体方法与策略如下:对异常视频片段计算 COFMHI,将连续的 COFMHI,称为剪辑的片段,片段间隔设置为 T. 从每个剪辑片段 T_i 中随机提取 N 个 $n \times n \times 3 \times T$ 大小的区域,称为视觉词块. 对所有剪辑片段进行处理后,在剔除平均像素值较小的视觉词块的基础上,采用 K-means 聚类[17]形成 K 个聚类中心,获取聚类中心的视觉词块. 对聚类获得的 K 个 $n \times n \times 3 \times T$ 大小的扩充块进行尺度变换,转换成 $224 \times 224 \times 3 \times T$ 大小的视觉词块. 扩充后,将得到 $K \times T$ 帧 $224 \times 224 \times 3$ 大小的 COFMHI(计算 COFMHI 和提取扩充样本的过程如图 6 所示).

本文选取欧几里得距离 $d(i,j)$ 来度量样本间的相似性,误差平方和 S_E 作为聚类的目标函数:

$$d(i,j) = \sqrt{\sum_{x=1}^{n}\sum_{y=1}^{n}\sum_{t=1}^{T}\sum_{RGB}(S_i - K_j)^2},$$

$$S_E = \sum_{j=1}^{K}\sum_{S_i \in K_j} d(i,j)^2,$$

式中：S_i 为第 i 个聚类样本，K_j 为第 j 个聚类中心.

实验中，T 为 16，N 为 20，$K = [FN/(3T)]$，$n = [\sqrt{wh/4}]$. 式中：F 为异常视频剪辑片段数，w、h 为分别图像宽度和高度，[.] 为向下取整.

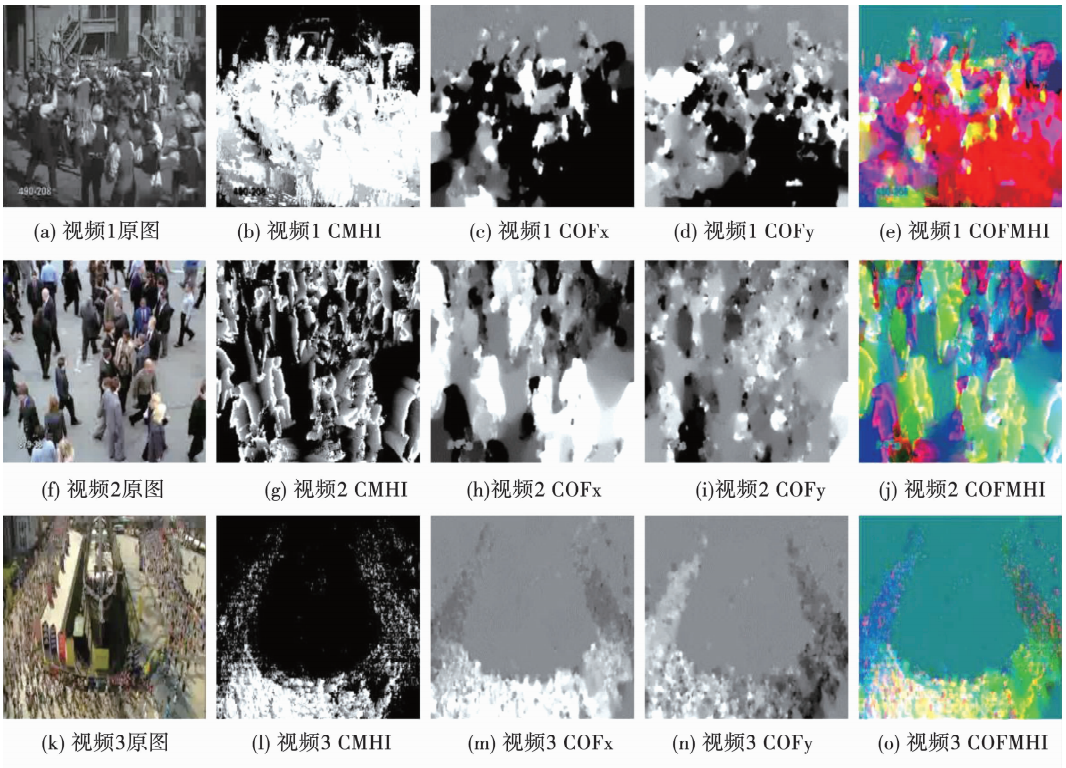


图 5　矫正光流运动历史图

Fig. 5　Corrected optical flow motion history image

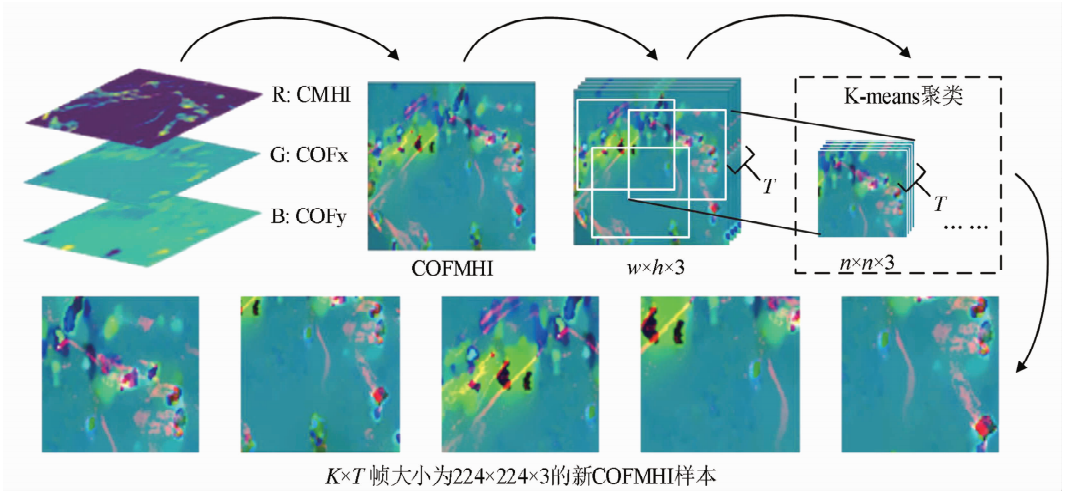


图 6　矫正光流运动历史图样本扩充

Fig. 6　Sample expansion of the corrected optical flow motion history images

2　基于 3D-LCRN 异常行为识别

2.1　3D-LCRN 网络结构

人体行为通常由一系列的子行为组成，子行为间有强烈的时间相关性. 例如打架斗殴包含挥动手臂、奔跑等子行为. 连续帧间的运动关联性比单帧图像更能区分行为. 因而，本文采用 3D-CNN 对短时序视频片段的运动信息进行建模，捕获局部时-空域特征. 此外，先前发生的行为在一定程度上会影响后续行为，例如跌倒后一般都会平躺然后弯腰起身. 因而，本文采用 LSTM 桥接短时序时-空域特征，进行长时序深层次全局时-空域特征提取.

正常行为或异常行为中的某些片段是无关、冗余或具有混淆性的,例如空白的街道,上下运动的电梯与飞驰而过的车辆等. 这些视频片段的主要内容从行人本身转移到了一些无关的运动物体上,会对网络训练造成一定的干扰. 因此,本文提出了可学习的贡献因子 α_t,使得每个视频片段的重要性有所不同. α_t 由 t 时刻 3D-CNN 输出 x_t 与 $t-1$ 时刻 LSTM 输出 $\boldsymbol{h}_{t-1}$ 计算所得

$$\alpha_t = \exp(\tanh(w_{x\alpha}x_t + w_{h\alpha}h_{t-1} + b_\alpha)).$$

式中:$w_{x\alpha}$、$w_{h\alpha}$ 为线性变换的权重,b_α 为偏置.

t 时刻 LSTM 输入 x'_t 由特征 x_t 和贡献因子 α_t 加权所得

$$x'_t = \alpha_t x_t.$$

在此基础上,本文构造了结合长-短时序的多层次网络模型 3D-LCRN,以正确地对行为间的时间结构进行建模,如图 7 所示. 3D-CNN 模块基于 ResNets[18],用于捕获连续动作帧间的局部短时序时-空结构信息. 本文剥离了 ResNets 最后的全连接层,增加了 256-d、2-d 两层全连接层,用于微调 3D-CNN 网络以适应后续建模. 贡献因子 α_t 加权的特征 x'_t 与长短时记忆网络 LSTM 相连,用于调整不同时刻时-空域信息的重要性. 3D-LCRN 结合 LSTM 的门控制记忆细胞(如图 8 所示)来存储过去态,当前态依据当前输入、输出和存储在该记忆细胞中的过去态进行更新,见式(3)~(7). 最后,基于 LSTM 输出计算每个时刻的类别概率分布 $P(y_t)$,通过对重叠片段的所有预测结果求平均值来获得每帧图像的所属类别,以实现正常与异常行为识别,见式(8)、(9). 3D-LCRN 的结构特性,使其能够在长时序间桥接重要信息,保留记忆,实现由短时序-长时序,由局部-全局的多层次时-空域特征提取.

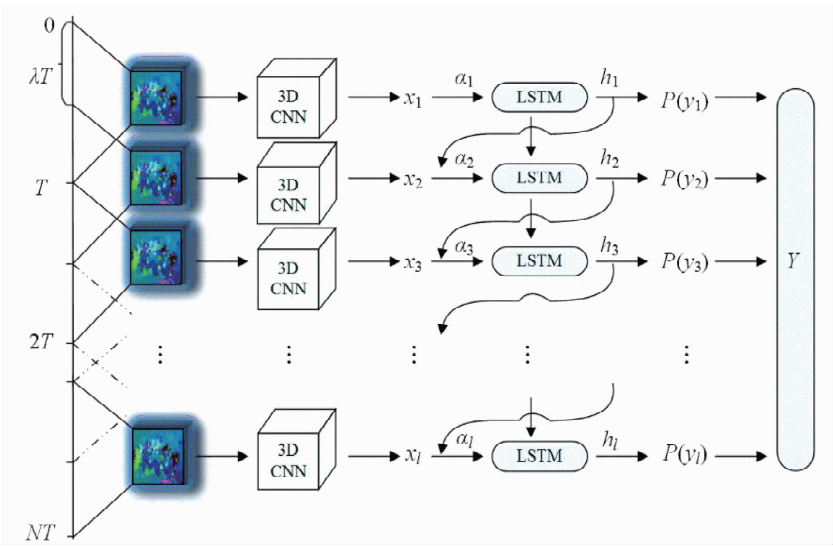


图 7 3D-LCRN 网络结构

Fig. 7 The structure of the 3D Long-short-term Convolutional Recurrent Network

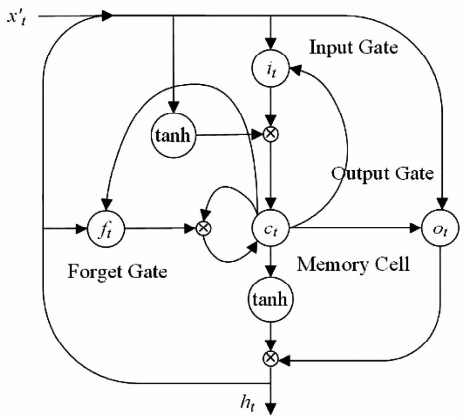


图 8 LSTM 记忆细胞结构

Fig. 8 The structure of the long short-term memory cell

$$i_t = \sigma(w_{xi}x'_t + w_{hi}h_{t-1} + w_{ci}c_{t-1} + b_i), \quad (3)$$

$$f_t = \sigma(w_{xf}x'_t + w_{hf}h_{t-1} + w_{cf}c_{t-1} + b_f), \quad (4)$$

$$o_t = \sigma(w_{xo}x'_t + w_{ho}h_{t-1} + w_{co}c_t + b_o), \quad (5)$$

$$c_t = f_t c_{t-1} + i_t \tanh(w_{xc}x_t' + w_{hc}h_{t-1} + b_c), \quad (6)$$

$$h_t = o_t \tanh(c_t). \quad (7)$$

式中:σ 为 sigmoid 激活函数,$\boldsymbol{w}_{x*}$、$\boldsymbol{w}_{h*}$、$\boldsymbol{w}_{c*}$ 是线性变换的权重,b_* 是偏置,$\boldsymbol{i}_t$ 为输入门,$\boldsymbol{f}_t$ 为忘记门,$\boldsymbol{o}_t$ 为输出门,$\boldsymbol{c}_t$ 为记忆细胞的状态,$\boldsymbol{h}_t$ 为 LSTM 输出.

$$P(y_t = z) = \frac{\exp(w_{hz}h_{t,z} + b_z)}{\sum_{z' \in Z}\exp(w_{hz}h_{t,z'} + b_z)}, \quad (8)$$

$$y_f = \max\left(\frac{\sum_{t \in f}P(y_t)}{\sum_{t \in f}}\right). \quad (9)$$

式中:$P(y_t = z)$ 是模型预测 t 时刻输入视频片段属于类别 z 的概率,y_f 为第 f 帧图像所属类别,w_{hz} 为权重,b_z 为偏置.

2.2　3D –LCRN 网络训练

3D-LCRN 网络训练包含两个阶段，即 3D-CNN 训练和贡献因子 α_t 加权的 LSTM 训练. 3D-CNN 模块基于 UMN、CAVIAR 与 Web 数据集对在 UCF101[19] 与 HMDB51[20] 数据集上预训练好的模型[21]进行微调，结构如图 9(a)所示. 模型输入为连续的 16 帧 COFMHI，大小为 3 ×16 ×224 ×224. 输出 2-d 向量，表示正常或异常行为. 微调后的模型剥离最后的 2-d 全连接层，抽取出 256-d 的特征向量.

LSTM 模块初始输入为多个 256-d 特征向量的平均值，这些特征向量由随机抽取的一段正常视频通过预训练好的 3D-CNN 获得. 基于该初始化，可以计算第 1 时刻的贡献因子 α_1. 在后续的每个时刻，LSTM 将依据上一时刻的输出计算新的贡献因子. 本文将训练样本通过 3D-CNN 提取的特征加权后作为输入馈送到 LSTM 中训练整个 3D-LCRN 网络，如图 9(b)所示. 训练时，通过输出类别与真实类别计算出的误差反向传播来对贡献因子与 LSTM 权重进行训练，而 3D-CNN 的权重保持不变. 图 4 中，3D-LCRN 网络滑动步长为 λT，时间步长为 l. 实验中，λ 与 l 分别设为 0. 25 与 40[22]，T 为 16. 学习率为 0. 003，在每 150 000 次迭代后减半. 对于所有非循环连接，dropout 设为 0. 5[23].

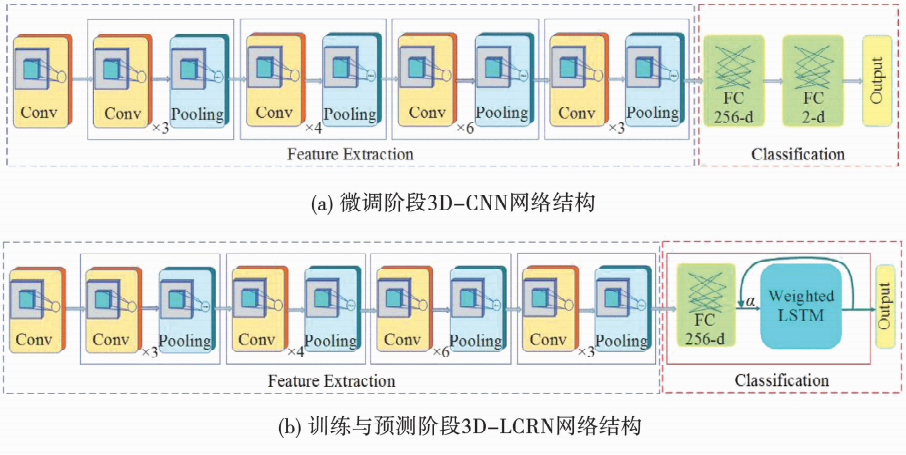


图 9　微调、训练与预测阶段 3D – CNN 与 3D – LCRN 网络结构

Fig. 9　The structure of 3D-CNN and 3D-LCRN during the fine tuning, training and predict stage

3　实验结果与分析

基于上述方法，本文选用多个公开数据集进行实验验证，包括 UMN、CAVIAR 与 Web. UMN 包含不同场景(草坪、室内和广场)中拍摄的 11 段视频，图像大小为 320 ×240 像素. 每段视频都包含正常部分，几十人随机地四处走动，然后发生异常事件，人们惊慌地逃离. CAVIAR 包含人们独自散步、与他人见面、逛街、进出商店、打架和昏迷，图像大小为 384 ×288 像素. Web 数据集由 8 个具有异常行为(恐慌逃逸、抗议者冲突和人群斗殴)的序列和 12 个具有正常行为(步行、马拉松跑步)的序列组成，图像大小不等. UMN 与 CAVIAR 场景相对简单，部分视频含有局部或全局的光照变化. Web 数据集主要包含针对城市场景的纪录片和视频，通常包含复杂的背景，对识别系统具有挑战性[24].

本文实验软件运行环境为 Windows 7 64 位，平台为 Python3. 6 + Opencv3. 3. 1 开源视觉库 + Tensorflow1. 8. 0 开源机器学习框架，硬件配置为 Intel® Core™i5-4440 3. 10GHz CPU，8G RAM 内存.

3.1　异常行为识别

图 10 为 OFMHI 与 COFMHI 部分实验结果对比. 其中，图 10(a)、(f)、(k)为视频原图，分别选自 UMN、CAVIAR 与 Web 数据集；图 10(b)、(g)、(l)为背景图像；图 10(c)、(h)、(m)为前景图像；图 10(d)、(i)、(n)为 COFMHI；图 10(e)、(j)、(o)为 OFMHI.

从图 10 中可以看出，OFMHI 包含大量由光照、抖动引起的背景干扰(透明绿色、红色)；而本文所提前景图像压制了背景抖动，对光照变化不敏感，矫正所得 COFMHI 几乎无背景干扰. 实验证明，本文所提 COFMHI 在复杂场景下仍具有较好的鲁棒性，能够有效压制背景干扰.

样本扩充时的聚类参数如表 1 所示. 其中，误差平方和为聚类完成后得到的结果. 由于 Web 数据集计算得到的 COFMHI 图像尺寸不一，本文统一归一化为 224 ×224 ×3 像素. 聚类范围为去除平均像素值小于 0. 2 的候选样本后总样本数量.

部分实验结果如图 11 所示. 由于每个聚类中心为 4D 视觉词块，这里选取 T = 1 通道进行展示. 其

中,图 11(a) ~ (d)为初始聚类中心,图 11(e) ~ (h)为最终得到的聚类中心,即扩充后的新样本. 可以看出,扩充后的新样本与原始样本不同,但是具有一定的相似性.

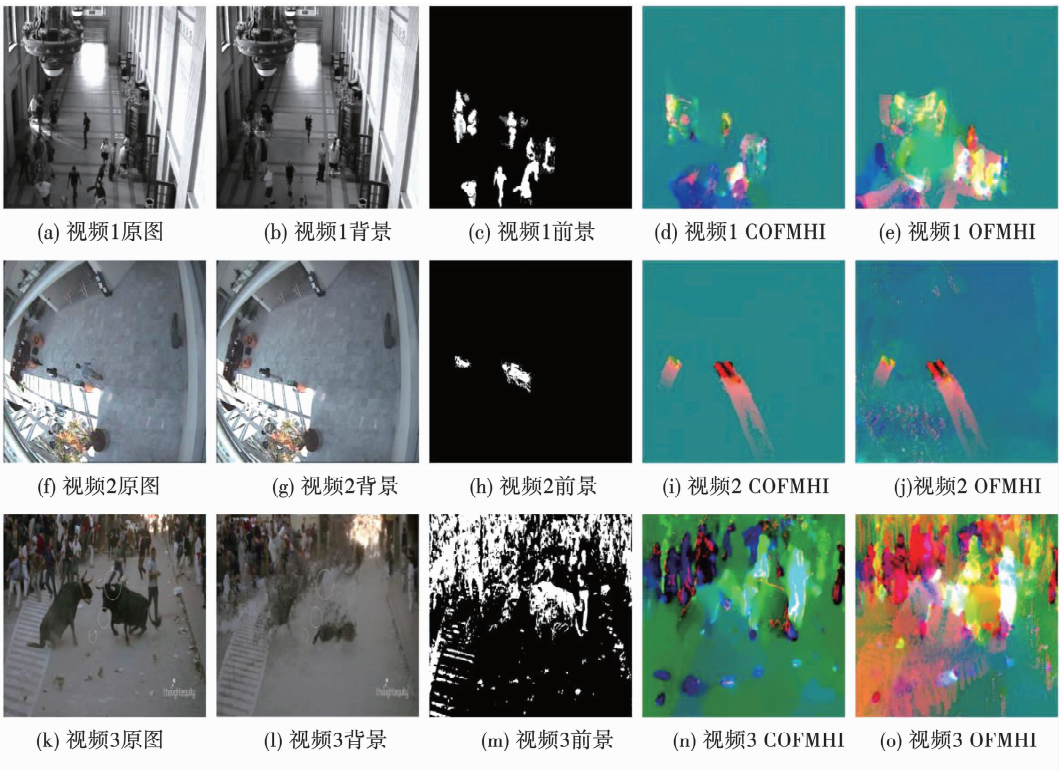


图 10 COFMHI 与 OFMHI 对比实验

Fig. 10 Comparisons between corrected optical flow motion history image and optical flow motion history image

表 1 K-means 聚类参数

Tab. 1 Parameters of K-means clustering

数据集	图像大小 $w \times h \times 3$	异常片段 F	候选样本 FN	样本大小 $n \times n \times 3 \times T$	聚类范围	聚类中心 K	误差平方和 S_E
UMN[10]	320 × 240 × 3	95	1 900	138 × 138 × 3 × 16	1 749	39	2.623×10^7
CAVIAR[11]	384 × 228 × 3	68	1 360	148 × 148 × 3 × 16	1 239	28	1.754×10^7
Web[12]	224 × 224 × 3	158	3 160	112 × 112 × 3 × 16	2 923	65	4.467×10^7

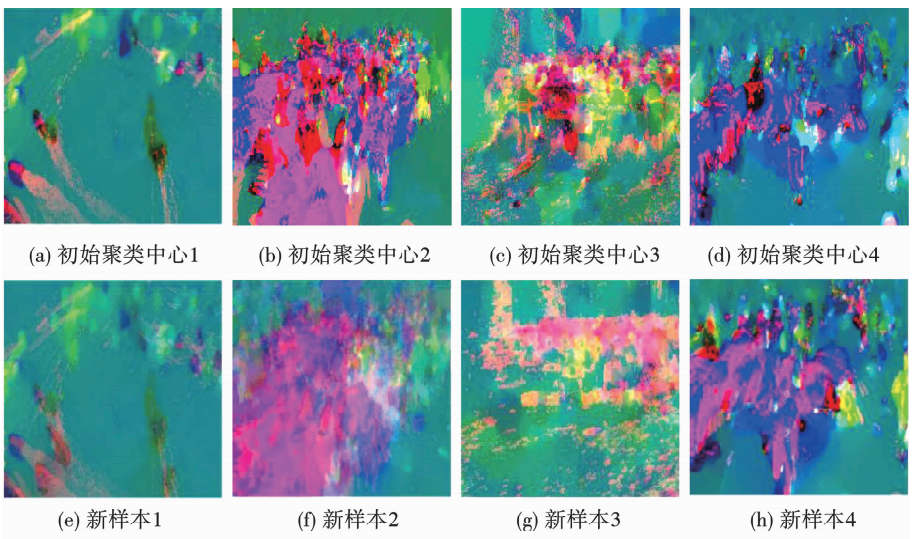
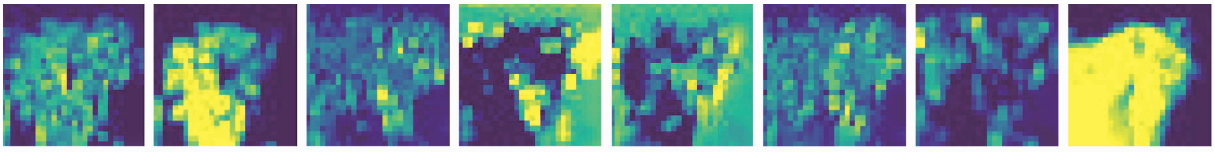


图 11 初始聚类中心与产生的新样本

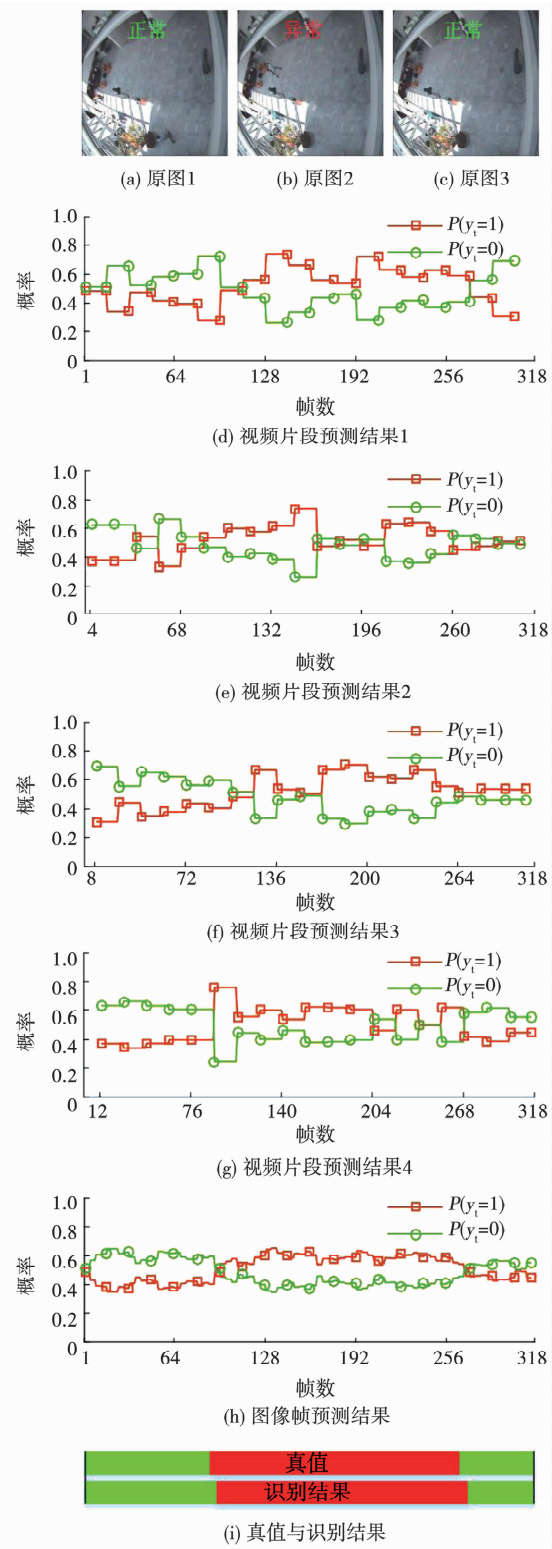
Fig. 11 Initial centers and new samples



(a) 特征1　(b) 特征2　(c) 特征3　(d) 特征4　(e) 特征5　(f) 特征6　(g) 特征7　(h) 特征8

图 12　3D-CNN 提取的时–空域特征

Fig. 12　Spatial-temporal feature maps extracted from 3D-CNN



(a) 原图1　(b) 原图2　(c) 原图3

(d) 视频片段预测结果1

(e) 视频片段预测结果2

(f) 视频片段预测结果3

(g) 视频片段预测结果4

(h) 图像帧预测结果

(i) 真值与识别结果

图 13　LSTM 预测概率分布

Fig. 13　Probability distribution predicted by LSTM

经过聚类扩充后，本文得到了 132 个聚类中心，即 2112 帧新 COFMHI 图像. 接着把原始样本与扩充产生的新样本一起送入 3D-CNN 进行网络微调，最后送入 3D-LCRN 网络进行分类训练，以实现异常行为识别. 部分实验结果见图 12 与图 13 所示.

图 12 为微调后 3D-CNN 第 2 层卷积层输出的部分时–空域特征. 该层共有 64 个特征，每个特征为 56 × 56 × 8 像素的 3D 特征块. 其中，图 12(a) ~ (h)为 8 个特征块的第一通道.

图 13 为 LSTM 预测所得类别概率分布. 其中，图 13(a) ~ (c)分别为第 20、175 与 296 帧输入图像，选自 CAVIAR 数据集；图 13(d) ~ (g)为视频片段预测结果，$P(y_t=1)$ 为 t 时刻当前片段属于异常行为的概率，$P(y_t=0)$ 为 t 时刻当前片段属于正常行为的概率；(h)为图像帧预测结果，$P(y_f=1)$ 为第 f 帧属于异常行为的概率，$P(y_f=0)$ 为第 f 帧属于正常行为的概率；(i)为真值与识别结果对比. 从图中可以看出，本文所提 3D-LCRN 网络有效、可行，异常行为识别结果准确、稳定.

3.2　客观定量评价对比

为进一步评价本文所提 COFMHI 的有效性，分别将光流图(矫正前后)、运动历史图(矫正前后)、光流运动历史图(矫正前后)与 3D-LCRN 网络结合，基于 3 个公开数据集进行 5 折交叉验证对比实验. 实验使用 6 种不同的图像数据来训练 6 个 3D-LCRN 网络，结果如表 2 所示. 其中，每一列的平均值与标准差由该列的 3 个数据计算得到. 由表 2 可知，光流图与运动历史图相结合后识别效果提升，与光流图相比提高了 0.7%，与运动历史图相比提高了 2.4%. 究其原因为光流场包含了运动目标的瞬态运动信息与表观结构，运动历史图包含了运动目标的轮廓轨迹和运动能量的空间分布，两相结合可以在一定程度上丰富时–空域特征的表现形式. 并且，三类图像经过本文所提方法进行矫正后识别效果均有提升，COFMHI 较 OFMHI 识别效果提高了 2.0%. 究其原因为本文所提方法能够在一定程度上对抗光照变化与背景抖动，压制了场景中部分背景干扰. 实验表明，本文方法有效可行.

为定量评价本文所提贡献因子的有效性,将 COFMHI 分别与有、无贡献因子的 3D-LCRN 结合,在 3 个公开数据集上进行 5 折交叉验证,实验结果如表 3 所示. 其中,每一行的平均值与标准差由该行的三个数据计算得到. 由表 3 可知,含有贡献因子的 3D-LCRN 模型识别精度较高,较不含贡献因子的 3D-LCRN 相比提高了 1.9%. 究其原因为本文所提贡献因子能让每个输入视频片段的重要性有所不同,通过自适应学习能够在一定程度上压制冗余、混淆或无关视频片段,提高异常行为识别精度.

表 2 基于不同预处理图像的异常行为识别性能对比

Tab. 2 Performance comparisons for different preprocessed images

数据集	输入图像					
	光流图		运动历史图		光流运动历史图	
	OF	COF	MHI	CMHI	OFMHI	COFMHI
UMN[10]	0.961	0.967	0.947	0.955	0.971	0.990
CAVIAR[11]	0.941	0.950	0.925	0.937	0.945	0.964
Web[12]	0.839	0.846	0.818	0.829	0.846	0.869
平均值 ± 标准差	0.914 ±0.053	0.921 ±0.053	0.897 ±0.056	0.907 ±0.056	0.921 ±0.054	0.941 ±0.052

表 3 有无贡献因子 α 的 3D-LCRN 识别性能对比

Tab. 3 Performance comparisons for 3D-LCRN with and without α

方法	UMN[10]	CAVIAR[11]	Web[12]	平均值 ± 标准差
3D-LCRN (withoutα)	0.988	0.941	0.837	0.922 ±0.063
3D-LCRN	0.990	0.964	0.869	0.941 ±0.052

为客观定量评价本文方法的有效性,选取方法[5][8-9]基于 3 个公开数据集进行 5 折交叉验证,实验结果如表 4 所示. 其中,每一行的平均值与标准差由该行的三个数据计算得到. 由表 4 可知,本文所提方法异常行为识别时性能最优. 究其原因在于,文献[5]基于轨迹计算运动不稳定性来判别异常行为. 在复杂场景下,行人间存在大量交叉遮挡,该方法难以跟踪并提取目标的完整运动轨迹,因而异常行为识别精度不高. 文献[8]基于输入为原始图片与光流图片的双流卷积神经网络来进行行为识别. 但是光流和 3D-CNN 提取的都是短时序特征,针对长视频,上下文间的相关性容易流失,并且在复杂场景下无法压制光线变化与背景运动等干扰,因而在简单场景下识别效果较优,但是在复杂场景下性能不如本文所提方法. 文献[9]通过 2D-CNN 提取 RGB 图像特征,送入双向 LSTM 网络进行深层特征提取,从而识别行为. 由于 2D-CNN 容易丢失连续视频帧间运动信息的时间相关性,并且视频片段具有一定的冗余与混淆性,因而识别精度不高. 实验表明,本文方法具有优异的异常行为识别性能.

表 4 不同异常行为识别方法性能对比

Tab. 4 Performance comparisons among the proposal and others

方法	UMN[10]	CAVIAR[11]	Web[12]	平均值 ± 标准差
Motion Instability[5]	0.981	0.951	0.821	0.918 ±0.069
Spatiotemporal-3D-CNN[8]	0.992	0.946	0.834	0.924 ±0.066
2D - CNN + LSTM[9]	0.987	0.957	0.841	0.928 ±0.063
COFMHI + 3D - LCRN	0.990	0.964	0.869	0.941 ±0.052

4 结 论

提出了一种基于 3D-LCRN 的异常行为识别方法. 1)通过结构相似性背景模型获取复杂场景下能够压制光照突变与背景运动的矫正光流场与矫正运动历史图. 2)提出样本维度与数量双向聚类扩充方法有效丰富了 COFMHI 样本的时-空域信息,在一定程度上克服了样本有限且失衡的问题. 3)提出结合可学习贡献因子的 3D-LCRN 网络对 COFMHI 进行分类识别,能够压制冗余,提取局部-全局、短时序-长时序的多层次时-空域特征,进一步提高了异常行为识别精度. 该方法在 UMN、CAVIAR 与 Web 公开数据集上平均识别准确率达到 94.1%,与现有的行为识别方法相比,本文方法能够在光照变化、背景抖动等复杂场景下保留视频上下文间的时-空相关性,准确、有效识别异常行为,具有优异的识别性能与一定的实用价值.

参考文献

[1] BROX T, BRUHN A, PAPENBERG N, et al. High accuracy optical flow estimation based on a theory for warping [C]//Proc 8th European Conference on Computer Vision. Prague: Springer, 2004: 25

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA: IEEE Press, 2005: 886

[3] MURTAZA F, YOUSAF M H, VELASTIN S A. Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description[J]. IET Computer Vision, 2017, 10(7): 758. DOI:10.1049/iet-cvi.2015.0416

[4] EUM H, YOON C, LEE H, et al. Continuous human action recognition using Depth-MHI-HOG and a spotter model [J]. Sensors, 2015, 15(3): 5197. DOI:10.3390/s150305197

[5] XIE Shiyang, GUAN Yepeng. Motion instability based unsupervised online abnormal behaviors detection [J]. Multimedia Tools & Applications, 2016, 75(12): 7423. DOI:10.1007/s11042-015-2664-8

[6] IJJINA E P, CHALAVADI K M. Human action recognition using genetic algorithms and convolutional neural networks[J]. Pattern Recognition, 2016, 59(11): 199. DOI:10.1016/j.patcog.2016.01.012

[7] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE Press, 2016: 1933

[8] 杨天明，陈志，岳文静. 基于视频深度学习的时空双流人物动作识别模型[J]. 计算机应用，2018，38(3)：895. DOI:10.11772/j.issn.1001-9081.2017071740
YANG Tianming, CHEN Zhi, YUE Wenjing. Spatio-temporal two-stream human action recognition model based on video deep learning [J]. Journal of Computer Applications, 2018, 38(3): 895. DOI: 10.11772/j.issn.1001-9081.2017071740

[9] ULLAH A, AHMAD J, MUHAMMAD K, et al. Action Recognition in video sequences using deep bi-directional LSTM With CNN features[J]. IEEE Access, 2018, 6(99): 1155. DOI:10.1109/ACCESS.2017.2778011

[10] UMN: Unusual crowd activity dataset of University of Minnesota [DB/OL]. 2006. http://mha.cs.umn.edu/Movies/CrowdActivity-All.avi

[11] Caviar: EC funded caviar project [DB/OL]. 2004. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[12] The Web Datasets[DB/OL]. 2009. http://www.vision.eecs.ucf.edu/projects/rmehran/cvpr2009/Abnormal_Crowd.html

[13] BRUNET D, VRSCAY E R, Wang Zhou. On the mathematical properties of the structural similarity index[J]. IEEE Transactions on Image Processing, 2012, 21(4): 1488. DOI:10.1109/TIP.2011.2173206

[14] LUO Yong, GUAN Yepeng. Motion objects segmentation based on structural similarity background modelling [J]. IET Computer Vision, 2015, 9(4): 476. DOI:10.1049/iet-cvi.2014.0261

[15] 冯宝，张绍荣，陈业航，等. 结合小波能量和汉森形状指数的肺结节分割[J]. 仪器仪表学报，2018，39(11)：240. DOI:10.19650/j.cnki.cjsi.J1803951
FENG Bao, ZHANG Shaorong, CHEN Yehang, et al. Nodule segmentation combining wavelet energy and hessian shape index [J]. Chinese Journal of Scientific Instrument, 2018, 39(11): 240. DOI:10.19650/j.cnki.cjsi.J1803951

[16] FARNEBACK G. Two-frame motion estimation based on polynomial expansion[C]//13th Scandinavian Conference on Image Analysis. Halmstad: Springer, 2003: 363

[17] 高国琴，李明. 基于 K-means 算法的温室移动机器人导航路径识别[J]. 农业工程学报，2014，30(7)：25. DOI:10.3969/j.issn.1002-6819.2014.07.004
GAO Guoqin, LI Ming. Navigating path recognition for greenhouse mobile robot based on k-means algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2014, 30(7): 25. DOI:10.3969/j.issn.1002-6819.2014.07.004

[18] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE Press, 2018: 6546

[19] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human action classes from videos in the wild: CRCV-TR-12-01 [R]. UCF Center for Research in Computer Vision, 2012

[20] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a largevideo database for human motion recognition[C]//2011 IEEE International Conference on Computer Vision. Barcelona: IEEE Press, 2011: 2556

[21] HARA K, KATAOKA H, SATOH Y. Learning spatio-temporal features with 3D residual networks for action recognition[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE Press, 2017: 3154

[22] Lu Na, Wu Yidan, Feng Li, et al. Deep learning for fall detection: 3D-CNN combined with LSTM on video kinematic data[J]. IEEE Journal of Biomedical and Health Informatics, 2019, 23(1): 314. DOI:10.1109/JBHI.2018.2808281

[23] SRIVASTAVA N, HINTON G, KRIZHEYSKY A, et al. Dropout: a simple way to prevent neural networks from over-fitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929

[24] 仉长崎，管业鹏. 基于动态粒子流场的视频异常行为自动识别[J]. 光电子·激光，2015，26(12)：2375. DOI:10.16136/j.joel.2015.12.0563
ZHANG Changqi, GUAN Yepeng. Dynamic particle flow field based automatic recognition of video abnormal behavior[J]. Journal of Optoelectronics·Laser, 2015, 26(12): 2375. DOI:10.16136/j.joel.2015.12.0563

（编辑　苗秀芝）