文章编号：1003-0530(2022)07-1497-10

基于拉格朗日场的多级运动特征暴力行为识别

娄 久 左德承 张 展 刘宏伟
（哈尔滨工业大学计算机科学与技术学院，黑龙江哈尔滨 150001）

摘 要：针对暴力行为识别过程中缺乏描述不同时间尺度下暴力行为运动变化的问题，本文提出了一种基于拉格朗日场的多级运动特征暴力行为识别算法。该算法将描述非线性粒子运动的拉格朗日场引入暴力行为分析过程中，首先通过构建基于光流的拉格朗日场来挖掘不同时间尺度下暴力行为运动特征，设计了基于拉格朗日场的多级运动模块，该模块可以根据输入光流序列长度，计算多级运动特征；然后构建了基于流量门控制机制的双流网络，将多级运动特征和RGB图像特征融合；最后，利用LSTM和全连接模型计算识别结果。实验证明，该方法在公共暴力识别数据集上取得了很好的效果，特别是在真实监控场景的RWF-2000数据集上，暴力行为识别正确识别率可以达到88.4%，优于其他算法。

关键词：暴力行为识别；多级运动特征；双流网；拉格朗日场；光流
中图分类号：TP3-05 **文献标识码：**A **DOI：**10.16798/j.issn.1003-0530.2022.07.016

引用格式：娄久，左德承，张展，等. 基于拉格朗日场的多级运动特征暴力行为识别[J]. 信号处理，2022，38(7)：1497-1506. DOI：10.16798/j.issn.1003-0530.2022.07.016.
Reference format： LOU Jiu，ZUO Decheng，ZHANG Zhan，et al. Violence recognition based on multilevel-motion features of Lagrange field[J]. Journal of Signal Processing，2022，38(7)：1497-1506. DOI：10.16798/j.issn.1003-0530.2022.07.016.

Violence Recognition Based on Multilevel-motion Features of Lagrange Field

LOU Jiu ZUO Decheng ZHANG Zhan LIU Hongwei
（School of Computer Science and Technology，Harbin Institute of Technology，Harbin，Heilongjiang 150001，China）

Abstract: In different time scales in the process of violence recognition，a multilevel-motion feature violence recognition algorithm based on Lagrange field is proposed in this paper. In this algorithm，the Lagrange field describing the nonlinear particle motion is introduced into the process of violence analysis. The opt Lagrange field based on optical flow is constructed to mine the motion characteristics of violence at different time scales，and a multilevel-motion module based on Lagrange field is designed. The module can calculate the multilevel-motion characteristics according to the length of the input optical flow sequence；Then，a dual flow network based on flow gate control mechanism is constructed to fuse multilevel-motion features and RGB image features；Finally，the recognition results are calculated by using LSTM and full connection model. Experiments show that this method has achieved good results in public violence identification data set，especially in RWF-2000 data set of real monitoring scene，the correct recognition rate of violence identification can reach 88.4%，which is better than other algorithms.

Key words: violence recognition；multilevel-motion features；two-streams network；Lagrange field；optical flow

收稿日期：2021-08-03；修回日期：2022-04-05
基金项目：国家自然科学基金项目(62171155)

1 引言

暴力行为是以人身、财产为侵害目标，采取暴力手段，对被害人的身心健康和生命财产安全造成极大的损害，直接危及人的生命、健康与自由的一种行为[1]。随着信息技术的发展，通过视频监控来识别暴力行为已经成为公共安全防护的研究重点。

暴力行为是一种攻击行为，该种行为在时间上具有突发性、行为动作上表现出急速猛烈和局部重复的特点，即一般由多个单独的不同时间尺度的运动模式序列组成，如整体持续纠缠、局部重复性踢打等。由于暴力行为具有持续性和局部重复的特点，因此，研究发现能够表征时空变化的运动特征是暴力行为识别的关键[2]。从时空粒度上来看，目前可以得到的最小粒度的运动特征是光流特征[3]，光流特征反应了微小时间内物体对应的图像像素的运动方向和速度，这里的微小时间是视频的帧长。为提高光流特征对行为的表述能力，学者们尝试在时空维度上进行粒度扩展，如在空间上将局部相似的像素聚合，构建基于光流的局部描述子特征，代表的有光流直方图（Histogram of Optical Flow Orientation，HOFO）[4-5]、混合动态纹理特征（Mixture of Dynamic Texture，MDT）[6]、运动边界直方图 MBH（Motion Boundary Histograms）特征[7]、加速度特征[8]等；Cong 等人将多尺度概念引入到光流直方图（Multi-scale Histogram of Optical Flow，MHOF）[9]，从不同空间粒度下研究物体的运动。在时间粒度上，主要利用连续光流进行轨迹提取，其中最有代表性的是 Wang 等人提出的改进版的稠密轨迹提取算法（Improved Dense Trajectories，IDT）[10]，该算法利用稠密轨迹提取（Dense Trajectories，DT）算法得到视频序列中的轨迹，然后沿轨迹提取 HOF、HOG 等特征，该方法是深度网络技术兴起前行为识别的主流方法。

随着深度网络技术的发展，可以利用深度网络自主学习表征行为的运动特征[11]，Simonyan 等人提出了基于双流卷积神经网络的行为识别算法[12]，该网络在空间流卷积神经网络基础上增加了时间流卷积神经网络，即将连续几帧光流图像堆叠在一起形成的图像块作为输入，利用多层叠加卷积网络（Convolutional neural network，CNN）提取视频中物体的运动信息。该模型的提出首次打破了 IDT 算法在行为识别领域的领先地位。此后，在双流网的基础上，又出现了时序分割网络（Temporal Segment Network，TSN）[13]、时序推理网络（the Temporal Relation Network，TRN）[14]等。Tran 等人提出了的三维卷积神经网络（Convolutional 3 Dimension，C3D）[15]，该网络在视频块或者堆叠后的视频帧形成的立方体中进行卷积操作。Quo Vadis 等人在 3D 网和双流网的基础上又发展出 I3D（Inflated 3D ConvNet）[16]，即双流膨胀 3D 卷积网络模型，该模型将原来双流网的 2D 卷积核扩展到 3D，从而实现更深层的运动特征提取。基于长短时记忆网络（Long Short Term Memory，LSTM）的 ConvLSTM 通常被串接在卷积模块后，主要基于卷积模块输出来捕捉输入视频之间的长距离依赖关系[17]。然而无论是双流网还是 C3D、I3D，或者 CNN+LSTM，这些方法都是通过增加网络层数来获取时间粒度较大运动特征。网络层数越深，模型的参数越多，很可能会导致过拟合。同时，在利用深层网络获得较大时间粒度的特征时，受到卷积特性的影响，小粒度时间范围内的运动特征会被平滑掉。为此，我们希望能够提取一种多级的运动特征，可以更全面的表征暴力行为特点，提高暴力行为识别结果。

拉格朗日场起源于动力学系统理论，利用非稳定场来描述非线性动力学系统中流体的变化[18]，能够很好地描述粒子运动规律。由于宏观世界物体运动都遵循物理规则，因此尝试将拉格朗日场引入计算机视觉分析中，Alexander Kuhn 等人提出了基于拉格朗日场的视频框架，证明拉格朗日场可以很好的表征非局部的、长期的运动信息[19]；Haller 等人提出的基于拉格朗日场的有限时间 Lyapunov 指数（Finite Time Lyapunov Exponents，FTLE）来度量流场中相邻粒子的运动轨迹[20]，根据相似轨迹进行运动分割，FTLE 已经被 Ali 等人成功地用于描述和分割人群视频片段[21]；Tobias Senst 等人在前人研究的基础上构建了一种基于拉格朗日方向场的运动特征来进行暴力行为识别，该特征主要是基于方向场变化轨迹计算得到，后期利用扩展词包（extend bag-of-words）融合分类方法，在 Hockey 等暴力识别公开数

据集上取得很好的识别结果[22]。通过以上分析可以看出，以往研究侧重于基于拉格朗日场捕捉到的粒子轨迹来构建长时运动特征，但是对于暴力行为来说，暴力行为具有短时重复和时空非局部性，需要在不同时间尺度下对行为模式进行表征，而目前缺乏相应的研究。

综上所述，目前缺乏一种能够表征不同时间尺度下暴力行为运动模式的暴力行为识别方法。为解决这一问题，本文提出了一种基于拉格朗日场的多级运动特征暴力行为识别方法。

首先构建了基于拉格朗日场（Opt-Lagrange Field）的 Multilevel-motion 模块，该模块能够将输入的连续光流信息转化成不同时间尺度下的多级运动特征，利用该特征来表征不同时间尺度下暴力行为运动模式。然后，采用双流网模型作为分类模型，保留原始的 RGB 空间特征输入，将原来的光流特征替代为多级运动特征，利用该方法在国际公开暴力识别数据库上进行了实验，取得了很好的实验结果，证明该方法能够有效提升暴力识别结果。

论文第 2 节讲述了基于拉格朗日场的多级运动特征提取方法，在第 3 节着重介绍了基于 Opt-Lagrange Field 的多级特征暴力识别模型，实验设置和结果分析在第 4 节，实验数据本文采用了四个数据集，分别是 Movies Fight、Hockey Fight、Crowd Violence 和 RWF-2000，第 5 节给出结论。

2　基于拉格朗日场的多级运动特征

2.1　拉格朗日场

拉格朗日场常用于描述非线性动力系统中的流体运动，拉格朗日场揭示了系统随时间演化的内在运动模式[18]。如果将视频中待跟踪物体看作质点，则目标的初始位置坐标作为参考点，那么沿时间轴跟踪这个质点，可以获得其在任意时刻的位置，从而获得高层次的运动信息。

在拉格朗日场算法中，设被跟踪的目标集为 $Z=\left(z_1,z_2,.....\right)$，则 Z 中的质点 z_i 随时间变化的轨迹可以表示为：

$$z_i(t)=(x_i,y_i,t) \tag{1}$$

公式（1）中，如果时间 t 固定，则得到不同质点的位置分布；如果质点 z 固定，则可以得到随时间变化的质点的运动规律。通常称 z、t 为拉格朗日变数。

在暴力行为识别过程中，需要获得质点的运动规律，因此需要在时间轴 t 上追踪质点的位置，如图 1 所示，图中红点为待追踪的质点，可以在视频流上对这个质点进行跟踪，获得一系列离散的点，表示该质点在不同时间的位置。

为获得高层次的运动信息，需利用拉格朗日场对图 1 中的离散点来计算时间尺度 τ 的运动轨迹，方法如下：

1）设质点 z 在一张图像上的位置为 (x,y)，如果知道质点的运动方向和速度，利用公式（2）就可以获得质点在时空维度中的运动轨迹 $z(t)$ 与初始位置 z_0 之间的关系：

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}z(t)\\t\end{bmatrix}=\begin{bmatrix}(z(t),t)\\1\end{bmatrix},\begin{bmatrix}z(t)\\t\end{bmatrix}(0)=\begin{bmatrix}z_0\\t_0\end{bmatrix} \tag{2}$$

2）因为在拉格朗日光流场中，获取轨迹是与时间尺度 τ 相关的，所以在公式（2）的基础上，添加参数 τ，表示这个质点在时间尺度 τ 上的轨迹，如公式（3）。

$$\phi_{t_0}^{\tau}:D\rightarrow D:x_0\rightarrow\phi_{t_0}^{\tau}(z_0)=x(t:t_0,z_0) \tag{3}$$

流函数 $\phi_{t_0}^{\tau}$ 是基于轨迹上离散的点建立的，其作用是通过计算不同类型的场线来导出拉格朗日场，并且其函数与时间尺度 τ 密切相关，不同的 τ 往往得到不同的结果。流函数的最佳选择通常取决于流场的特性、计算开销和分析目标。

(a) 第一帧　(b) 第二帧　(c) 第三帧　(d) 第四帧

图 1　拉格朗日场对质点的跟踪

Fig. 1　Particle tracking in Lagrangian field

2.2 Opt-Lagrange Field构建方法

从上节叙述可知，利用拉格朗日场需要知道质点的运动方向和速度，在视频分析中，可以通过计算相邻图像的光流场得到。光流场是指图像中所有像素点构成的一种二维(2D)瞬时速度场，其中的二维速度矢量是景物中可见点的三维速度矢量在成像表面的投影，所以光流中包含了被观察物体的运动信息。因此，本文基于光流构建拉格朗日场(Optical flow- Lagrange field，简称Opt-lag field)。

光流计算的基本原理如下所述：假定$I(x,y,t)$是时刻t在图像位置(x,y)的灰度值。$u(x,y)$和$v(x,y)$是光流在位置(x,y)的x和y方向的速度分量。如果假设在时刻$t+\delta t$，在位置$(x+\delta x, y+\delta y)$的灰度值保持不变，那么如下等式成立：

$$I(x+u\delta t, y+v\delta t) = I(x,y,t) \tag{4}$$

其中，$\delta x = u\delta t$，$\delta y = v\delta t$，δt代表一个小的时间间隔。接下来，利用泰勒公式展开等式的左侧如下：

$$I(x,y,t) + \delta x\frac{\partial I}{\partial x} + \delta y\frac{\partial I}{\partial y} + \delta t\frac{\partial I}{\partial t} + e = I(x,y,t) \tag{5}$$

这里e表示δx，δy和δt中的二阶和高阶项。消掉左侧和右侧的相同项，并忽略掉e值。对上式分别除δt并整理，有：

$$\frac{\partial I}{\partial x}\frac{\delta x}{\partial t} + \frac{\partial I}{\partial y}\frac{\delta y}{\partial t} + \frac{\partial I}{\partial t}\frac{\delta t}{\partial t} = 0 \tag{6}$$

令$u(x,y) = \frac{\delta x}{\partial t}$，$v(x,y) = \frac{\delta y}{\partial t}$，则有：

$$I_x u(x,y) + I_y v(x,y) + I_t = 0 \tag{7}$$

导数I_x，I_y，I_t可以计算图像灰度差得来，通过引入光流的全局平滑约束条件计算出x和y方向的光流速度分量$(u(x,y), v(x,y))$。本文采用经典的Gunnar Farneback算法求稠密光流[23]。

2.3 基于Opt-Lagrange Field的多级运动特征提取

(1) 基于Opt-Lagrange Field的拉格朗日场

由上述可知，在暴力行为识别过程中，可以通过计算光流场得到基于质点运动信息，然后基于公式(3)计算流函数，从而得到质点的运动规律。在视频监控中，人们关注的是目标的行为变化，利用流函数来为短时运动特征推导出不同时间间隔的拉格朗日场，具体公式如(8)、(9)：

$$\Lambda_X(z,t_0) = \frac{1}{\tau}\int u(\phi(z,t_0,\tau))\partial\tau \tag{8}$$

$$\Lambda_Y(z,t_0) = \frac{1}{\tau}\int v(\phi(z,t_0,\tau))\partial\tau \tag{9}$$

其中，u和v分别是两个方向上的速度函数，可以通过公式(7)计算得到。通过公式(8)、(9)，就可以建立起基于光流的拉格朗日场，场强变化是以时间作为尺度，选择不同的时间尺度，就会得到不同级别的方向场。

传统流函数支持三种形式的拉格朗日场，分别是：有限时间的利物浦指数(Finite-Time Lyapunov Exponent field，FTLE)、速度场和方向场。其中FTLE场获得的脊线提供运动信息的粗略分割，其更侧重描述运动边界。然而，光流场中的FTLE脊线是由各种不同的效应产生的，容易出现复杂运动场的过度分割，不适用与暴力行为识别。而在暴力行为发生过程中，直观的行为表现是质点的运动方向和速度都发生变换，不能仅用单一的速度场和方向场进行计算，因此需要一种新的拉格朗日场表征方法。

(2)基于拉格朗日场的多级运动模块

由于神经网络具有强大的数据表征能力和自学习的特点，本文基于拉格朗日场原理，设计了一种面向暴力行为识别的多级运动(Multilevel-motion)模型，该模型通过自学习的方式得到满足暴力行为需求的流函数，从而实现不同时间尺度下运动特征的提取，具体模型如图2所示。

图2中f_t为第t帧的光流输入，由x方向速度和y方向速度矩阵组成，图中h_t对应质点z在t时刻的坐标位置，应该等于前一刻的质点位置h_{t-1}累加该t帧位置速度乘以帧间隔，由于对于同一视频，帧间隔相同，因此，对于h_t的计算可以简化为：

$$g_t = h_{t-1} + f_{t-1} \tag{10}$$

$$h_t = \mathrm{Conv}_h(g_t) \tag{11}$$

g_t表示质点在x方向和y方向的位移，经过1*1卷积核Conv_h计算出最终t时刻质点的位置坐标。Λ_t对应t级运动特征，C_{t-1}为记忆单元，保存了前$t-1$时刻的轨迹特征，根据拉格朗日场原理，则时间尺度为t的运动特征可以表示为：

$$\Lambda_t = C_{t-1} + \tanh(\mathrm{Conv}_g(g_t)) \tag{12}$$

更新记忆单元：

$$C_t = \Lambda_t \times \mathrm{Sigmoid}(h_t) \tag{13}$$

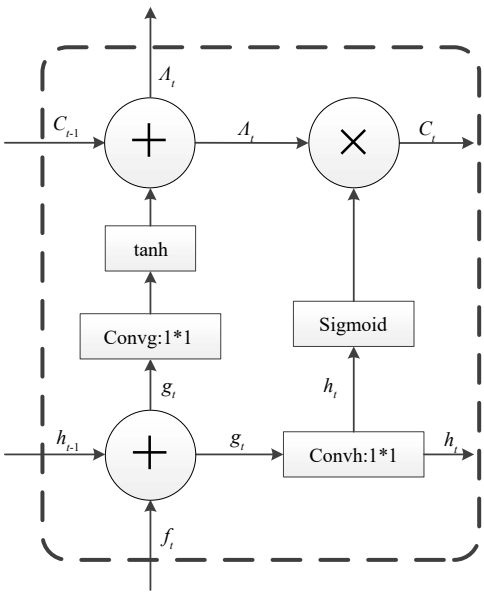


图2　多级运动特征计算模型

Fig. 2　Multilevel-motion feature model

从图2中可以看出，特征的级别受到输入光流帧数的限制，即如果输入τ帧光流图，则可以构建最大包含τ级的运动特征。

由于在拉格朗日场计算过程中，质点的选择非常重要。一般是将视频待跟踪物体作为质点，在时间轴上进行追踪物体位置，从而得到物体运动变化模式。而在暴力视频中，暴力行为发生的位置是随机的，待跟踪物体未知，无法直接选定质点。虽然无法直接选定质点，但是视频图像中的每一个像素都可能是待跟踪物体的组成像素。因此，可以将视频中的每个像素都作为单独质点，利用拉格朗日场计算得到所有像素的运动变化模式，然后利用流量门控制模块来实现对像素运动变化模式的筛选聚焦，从而实现暴力行为的识别。

因此，图2中质点z为图像矩阵中的所有像素点，则在初始t_0时刻h_{t0}如公式(14)所示，其中M和N代表图像矩阵的宽和高。

$$h_{to}=\begin{bmatrix}(0,0) & \cdots & (0,M-1)\\ \vdots & \ddots & \vdots\\ (N-1,0) & \cdots & (N-1,M-1)\end{bmatrix} \tag{14}$$

根据图2，可以将视频序列长度作为时间尺度τ的控制参数，来捕获视频中不同时间跨度的暴力行为特征，为了可视化，将生成的流图值转换为HSV颜色空间，并将生成的颜色投影回原始起始帧(类似于常见的局部光流描绘)。生成的色调值H表示流量图位移的方向，而饱和度S表示位移的大小，V保持不变，具体如图3所示。从图3中可以看出，时间尺度越小，得到物体运动特征表达的越微观，时间尺度越大，得到物体特征越宏观。

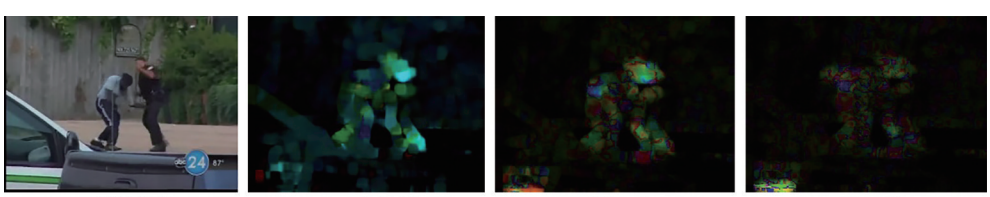
从图3中还可以看出，基于拉格朗日场，可以将相关质点动态运动行为的信息映射到单个帧上，由此产生可以紧凑地表示时间演化的多级运动特征，该特征能够更清晰表达行为规律。

3　基于Opt-Lagrange Field多级运动特征的暴力识别方法

3.1　模型结构

基于图2，本文设计了基于拉格朗日场多级运动特征的暴力行为识别模型。除了运动特征，行为姿态也是判断暴力行为的重要信息，因此在模型设计时本文采用经典的双流网模型架构[24]。具体如图4所示。

从图4中可以看出，该模型有四个模块构成，分别是RGB通道、拉格朗日通道、流量门控制、识别模块全连接层组成。这里RGB通道由4个2d卷积模块组成，主要起到为拉格朗日场计算提供空间信息的补偿作用。拉格朗日通道包含Multilevel-motion模块和3个3d卷积模块，目的提取多级运动特征中隐含的高级行为语义。模型在拉格朗日通道末端采用Relu函数激活，RGB通道末端采用Sigmoid函



(a) 原始图像　(b) 光流图　(c) τ=4　(d) τ=8

图3　不同时间尺度下拉格朗日场特征

Fig. 3　Features of Lagrangian field in different time scales

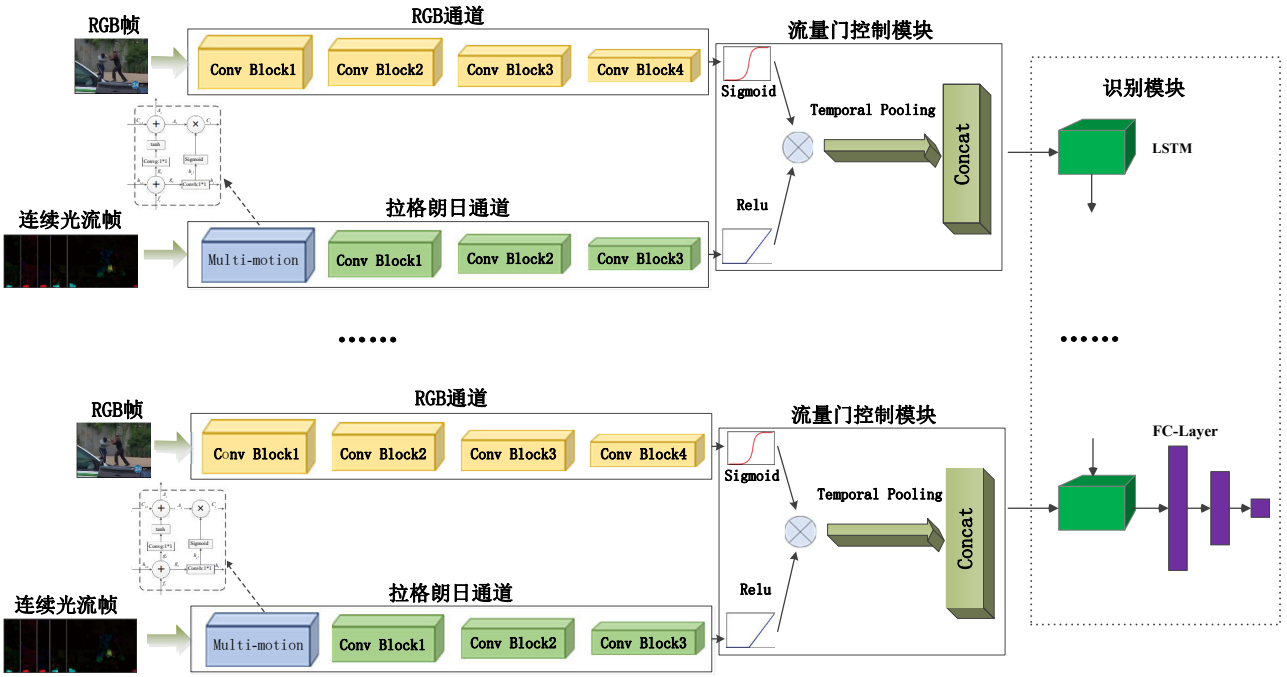


图4 基于Opt-Lagrange filed多级运动特征的暴力识别模型

Fig. 4 Violence recognition model based on Multilevel-motion feature of OPT Lagrange filed

数,由于拉格朗日通道和RGB通道输出是同维度的,因此将来自拉格朗日通道的输出和RGB通道的输出直接相乘,又由于Sigmoid函数的输出介于0和1之间,因此RGB的通道输出可以看作是调整拉格朗日通道输出的权重,具体计算方法如公式(15)所示:

$$y_{\text{flow}} = \text{Sigmoid}(y_{\text{RGB}}) \times \text{Relu}(y_{\text{Lagrange}}) \quad (15)$$

其中y_{RGB}表示RGB通道输出,y_{Lagrange}表示拉格朗日通道输出,y_{flow}表示两个通道的组合输出,并作为时间池Temporal Pooling模块的输入。时间池采用Max Pooling函数处理,由于Max Pooling函数只能保留局部最大值,因此拉格朗日通道乘以趋近1的结果将有更大的保留概率,而乘以趋近0的值将更可能被丢弃,该机制是一种自学习池策略,它利用RGB输出作为流量门对拉格朗日通道输出进行筛选,滤除不必要信息。识别模块包含一个LSTM模型和一个全连接层,用于计算最后的识别结果。此外,为了降低模型参数数量,本文采用MobileNet[25]和Pseudo-3D残差网络[26]中的深度可分离卷积的概念,来修改模型中的3D卷积层,LSTM所有神经元的参数共享,这样可以在不损失性能的情况下,显著降低模型参数。根据图4构建模型,在每个模块内,通过反复试验和优化后,采用如表1所示的模型参数结构。

表1 网络模型参数,i为模型重复次数

Tab. 1 Network model parameters, and i is the number of model repetitions

模块名称	类型	滤波器尺寸	i
Multilevel-motion	Conv3d	1 × 1 × 1@1	2
拉格朗日通道	Conv3d	τ × 3 × 3@8	1
	MaxPool3d	1 × 2 × 2	
	Conv3d	1 × 3 × 3@16	2
	MaxPool3d	1 × 2 × 2	
	Conv3d	1 × 3 × 3@32	
	MaxPool3d	1 × 2 × 2	
RGB通道	Conv2d	3 × 3@16	2
	Conv2d	3 × 3@16	
	MaxPool2d	1 × 2 × 2	
	Conv2d	3 × 3@32	2
	Conv2d	3 × 3@32	
	MaxPool2d	1 × 2 × 2	
流量门控制模块	Multiply	None	1
	MaxPool3d	8 × 1 × 1	
识别模块	FC layer	128	2
	FC layer	128	
	Softmax	2	

3.2　算法设计

模型参数设置好后，对基于 Opt-Lagrange Field 多级运动特征的暴力识别模型进行训练和测试，具体算法见算法 1。

算法 1　基于 Opt-Lagrange Field 多级运动特征的暴力识别算法

Input：包含长度为 T 的视频序列的训练集 Train 和测试集 Test；暴力识别模型 M；运动特征级别控制参数 τ。

Output：Test 中视频序列识别结果；训练好的暴力识别模型 M'。

1. 分别对 Train 和 Test 中视频序列做以下操作：
   (1)将分解视频为长度 T 为图像序列 $I=\{I_0,\cdots I_{T-1}\}$；
   (2)计算光流特征，得到长度为 $T-1$ 的光流序列 $\text{Flow}=\{\text{Flow}_0,\cdots,\text{Flow}_{T-2}\}$；
   (3)计算光流段个数 K：$K=\text{ceil}(\frac{T-1}{\tau})$；
   (4)为每个视频序列计算 RGB 通道和 Lagrange 通道的输入矩阵 $\{\text{RGB}^0,\cdots,\text{RGB}^{K-1}\}$ 和 $\{\text{Lagrange}^0,\cdots,\text{Lagrange}^K\}$，$\forall k\in[0,K]$，有 $\text{RGB}^k=I_k$，$\text{Lagrange}^k=\{\text{Flow}_k,\cdots,\text{Flow}_{k+\tau-1}\}$。
2. 模型初始化参数设置：
   (1)Multy-motion：h_0 初始坐标位置，C_0 初始化为零矩阵；
   (2)LSTM：$\text{LSTM}_{\text{cell}}=k$。
3. 使用 Train 进行训练：
   (1)将 1.4 步得到的矩阵输入暴力识别模型 M；
   (2)计算 t 时刻的误差 E；
   (3)计算 t 时刻各隐藏单元 k 的误差梯度 δ_k；
   (4)更新网络中各单元的权重 W_k；
   (5)满足迭代要求，停止更新，输出训练好的 M'。
4. 使用 Test 进行测试：将 1.4 步得到的矩阵输入暴力识别模型 M'，利用前向算法得到识别结果。

4　实验及结果分析

4.1　模型结构

为验证方法的有效性，本文使用了四个公开的暴力识别数据集，包括电影打架(Movie Fight)[27]和曲棍球打架(Hockey Fight)[28]、密集人群暴乱(Crowd Violence)[29]和 RWF2000 数据集[1]，其中 Movie Fight 数据集有 200 个从短片中提取战斗场景的剪辑；Hockey Fight 数据集中，有 1000 个在全国曲棍球联盟的曲棍球比赛中捕捉到的片段；Crowd Violence 数据集是从 YouTube 下载并剪辑的人群暴力场景，数据集包含 246 个视频；RWF2000，是 2020 年提出的最新的暴力行为识别数据集，该数据集包含 2000 个视频片段，来自于真实监控设备，分为两部分：训练集(80%)和测试集(20%)，一半的视频包含暴力行为，而另一半则属于非暴力活动，作为常见的摄像机设置，视频通常遵循一系列分辨率标准(例如 720P、1080P、2K、4K)。这四个数据集均具有视频级标注，其片段长度、分辨率等细节信息如表 2 所示。

表 2　暴力识别数据集

Tab. 2　Violence dataset

名称	数据规模	片段长度/s	分辨率	标注	场景
Movies Fight	200	1.6~2	720×480	视频级别	电影
Hockey Fight	1000	1.6~1.96	360×288	视频级别	冰球比赛
Crowd Violence	246	1.04~6.52	——	视频级别	自然
RWF-2000	2000	5	——	视频级别	监控视频

4.2　暴力行为识别结果比较

(1)多级运动特征控制参数 τ 的选择

本文利用时间尺度参数 τ 来控制基于 Opt-lagrange Field 构建的多级特征的时间尺度级别，τ 的不同意味着运动特征级别不同。图 5 给出了在 $\tau\in[1,10]$ 时，依据图 4 模型在四个数据集上得到的暴力识别结果，$\tau=1$ 意味着得到的是光流特征。由于 Movies Fight 是来源于电影数据，清晰度高，特征简单，利用光流特征就可以达到 100% 的识别准确率，参考价值较低。本文重点分析剩下的三个数据集。从图 5 中可以看出，与光流特征相比，基于 Opt-Lagrange Field 建立的多级运动特征能够为识别模型提供更多的时空变化信息，总体的识别率较高。但是，模型的识别率与运动级数并不是绝对正相关，在三个数据集上，都出现了识别率先

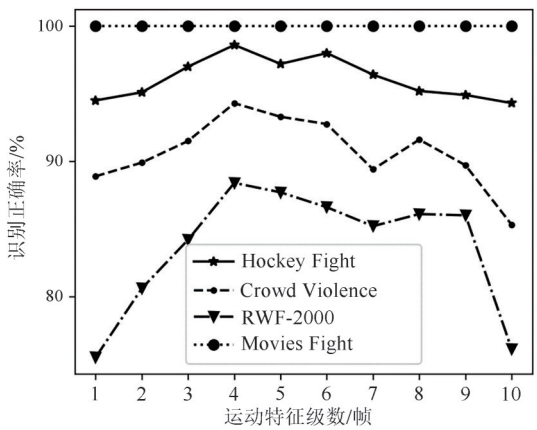


图5 不同τ时模型的暴力识别准确率

Fig. 5 The accuracy of violence recognition based on different τ

升高后下降的情况，这是因为过大的时间尺度会造成模型参数成倍增长，模型容易过拟合，增大运动特征时间尺度提供的信息不足以补偿模型参数爆炸的影响，因此需要折中考虑。根据图5，可以看出在$\tau = 4$时暴力行为识别模型的准确率都达到最大，因此选择$\tau = 4$作为本文Opt-lagrange Field运动特征的级数。在后续试验中均采用$\tau = 4$作为运动特征级数，并以此设定图4模型参数，具体参数见表1。

（2）基于Opt-Lagrange Field的多级运动特征的暴力识别结果分析

为验证拉格朗日多级特征的有效性，文中给出了采用单流通道时，在RWF-2000数据集上暴力行为识别的准确率，如表3所示。从表3中可以看出，拉格朗日特征要比光流特征提供更多有效的信息，其在所有数据上的平均识别准确率为90.51%，整体表现要优于光流特征。

表3 基于单流通道的暴力识别结果

Tab. 3 Violence recognition results based on single stream channel

特征	Movies Fight	Hockey Fight	Crowd Violence	RWF-2000	平均识别率
OPT only	100%	84.10%	84.57%	75.50%	86.04%
Lagrange	100%	92.50%	86.44%	83.12%	90.51%

表4给出经典暴力行为识别模型与本文提出模型在4种暴力识别数据集上的识别结果。从表4中可以看出，本文提出的模型在四种数据集上均有较好的表现，平均识别率达到95.32%，特别对于来源实际监控设备的RWF-2000数据集，可以达到88.40%的识别准确率，这说明，基于Opt-Lagrange Field特征能够提供多级运动信息，对于暴力行为识别来说是非常关键的。

表4 暴力行为识别结果

Tab. 4 Violence identification results

方法	Movies Fight	Hockey Fight	Crowd Violence	RWF-2000	平均识别率
ConvLSTM[17]	100%	97.10%	94.57%	77.00%	89.5%
C3D[15]	100%	96.50%	84.44%	82.75%	87.89%
I3D (RGB only)[16]	100%	98.50%	86.67%	85.75%	90.3%
I3D (OPT only)[16]	100%	84.00%	88.89%	75.50%	82.79%
I3D (Lagrange only)	100%	88.40%	89.10%	83.10%	86.86%
I3D (Fusion)	100%	94.50%	88.95%	84.50%	89.3%
FCN (RGB+OPT)[1]	100%	98.00%	88.87%	87.25%	91.37%
AlexNet+LSTM (RGB only)	100%	97.1%	86.3%	78.8%	90.55%
Ours	100%	98.60%	94.29%	88.40%	95.32%

5 结论

本文提出一种面向暴力行为识别的多级运动特征提取算法，该算法利用了流体运动学中的拉格朗日场来挖掘视频中暴力行为的运动表示。首先设计了基于拉格朗日场的Multilevel-motion模块实现多级运动特征提取；然后设计了基于双流网框架的Opt-Lagrange Field的暴力行为识别模型，该模型中将Multilevel-motion输出的Opt-Lagrange Field多级特征作为时间流，RGB作为空间特征，利用流量门控制模块筛选与运动特征相关的空间特征，实现时空流特征的融合，同时，利用LSTM获取不同时段

的距离依赖信息，解决暴力行为中运动特征中断和时空非局部分布问题；最后，利用全连接模型进行暴力行为判别。通过在 Movies Fight、Hockey Fight、Crowd Violence 和 RWF-2000 四个公开的暴力行为识别数据集进行试验，本文的暴力行为识别率分别为 100%、98.6%、94.29% 和 88.40%，平均识别率达到 95.32%，优于其他传统方法。当然，本文研究中还存在一些不足，如：没有有效利用拉格朗日场的运动预测特性，缺乏对于复杂背景信息以及噪声的考虑；同时对于暴力行为识别的效率问题本文没有提及，尤其是光流计算复杂度较高，应该进行算法优化并提高算法效率，这些都将是下一步我们工作的重点。

参考文献

[1] CHENG Ming, CAI Kunjing, LI Ming. RWF-2000: An open large scale video database for violence detection [C]//2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy. IEEE, 2021: 4183-4190.

[2] DE OLIVEIRA LIMA J P, FIGUEIREDO C M S. Temporal fusion approach for video classification with convolutional and LSTM neural networks applied to violence detection[J]. Inteligencia Artificial, 2021, 24(67): 40-50.

[3] MATKOVIC F, MARČETIC D, RIBARIC S. Abnormal crowd behaviour recognition in surveillance videos[C]// 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). Sorrento, Italy. IEEE, 2019: 428-435.

[4] COLQUE R V H M, CAETANO C, DE ANDRADE M T L, et al. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 673-682.

[5] WANG Tian, SNOUSSI H. Detection of abnormal visual events via global optical flow orientation histogram[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(6): 988-998.

[6] LIU Bin, JI Hao, DAI Yi. Vison-based traffic flow prediction using dynamic texture model and Gaussian process[C]//2017 2nd International Conference on Multimedia and Image Processing (ICMIP). Wuhan, China. IEEE, 2017: 201-205.

[7] FAN Min, HAN Qi, ZHANG Xi, et al. Human action recognition based on dense sampling of motion boundary and histogram of motion gradient[C]//2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). Enshi, China. IEEE, 2018: 1033-1038.

[8] LI Yi, CHENG Jun, FENG Wei, et al. Feature fusion of triaxial acceleration signals and depth maps for human action recognition[C]//2016 IEEE International Conference on Information and Automation (ICIA). Ningbo, China. IEEE, 2016: 1255-1260.

[9] CONG Yang, YUAN Junsong, LIU Ji. Sparse reconstruction cost for abnormal event detection[C]//CVPR 2011. Colorado Springs, CO, USA. IEEE, 2011: 3449-3456.

[10] XIAO Xiang, HU Haifeng, WANG Weixuan. Trajectories-based motion neighborhood feature for human action recognition[C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China. IEEE, 2017: 4147-4151.

[11] 刘强，张文英，陈恩庆. 基于异构多流网络的多模态人体动作识别[J]. 信号处理，2020，36(9)：1422-1428.
LIU Qiang, ZHANG Wenying, CHEN Enqing. Multimodal human action recognition based on heterogeneous multi-stream network[J]. Journal of Signal Processing, 2020, 36(9): 1422-1428. (in Chinese)

[12] LIU Xiao, YANG Xudong. Multi-stream with deep convolutional neural networks for human action recognition in videos[M]//Neural Information Processing. Cham: Springer International Publishing, 2018: 251-262.

[13] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition [C] //Computer Vision-ECCV 2016, 2016: 20-36. DOI:10.1007/978-3-319-46484-8_2.

[14] XU Jiarui, CAO Yue, ZHANG Zheng, et al. Spatial-temporal relation networks for multi-object tracking[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 3987-3997.

[15] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 4489-4497.

[16] WU Qianyu, ZHU Aichun, CUI Ran, et al. Pose-Guided Inflated 3D ConvNet for action recognition in videos[J]. Signal Processing: Image Communication, 2021, 91: 116098.

[17] SUDHAKARAN S, LANZ O. Learning to detect violent videos using convolutional long short-term memory[C]// 2017 14th IEEE International Conference on Advanced

Video and Signal Based Surveillance (AVSS). Lecce, Italy. IEEE, 2017: 1-6.

[18] PEACOCK T, DABIRI J. Introduction to focus issue: Lagrangian coherent structures [J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2010, 20(1): 017501.

[19] KUHN A, SENST T, KELLER I, et al. A Lagrangian framework for video analytics[C]//2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP). Banff, AB, Canada. IEEE, 2012: 387-392.

[20] HALLER G. Lagrangian coherent structures [J]. Annual Review of Fluid Mechanics, 2015, 47(1): 137-162.

[21] ALI S, SHAH M. A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis [C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA. IEEE, 2007: 1-6.

[22] SENST T, EISELEIN V, KUHN A, et al. Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation [J]. IEEE Transactions on Information Forensics and Security, 2017, 12(12): 2945-2956.

[23] BALZ I, ROSENTHAL E, REIMER A, et al. Analysis of the thermo-mechanical mechanism during ultrasonic welding of battery tabs using high-speed image capturing [J]. Welding in the World, 2019, 63(6): 1573-1582.

[24] SIMONYAN K, ZISSEMAN A. Two-stream convolutional networks for action recognition in videos[C]//28th Conference on Neural Information Processing Systems (NIPS). Montreal, Canada, 2014.

[25] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks [C]//CVPR 2018. Salt Lake City, UT, USA. IEEE, 2018: 4510-4520.

[26] QIU Zhaofan, YAO Ting, MEI Tao. Learning spatio-temporal representation with pseudo-3D residual networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy. IEEE, 2017: 5534-5542.

[27] CHEN Lianghua, HSU H W, WANG Liyun, et al. Violence detection in movies[C]//2011 Eighth International Conference Computer Graphics, Imaging and Visualization. Singapore. IEEE, 2011: 119-124.

[28] BERMEJO NIEVAS E, DENIZ SUAREZ O, BUENO GARCIA G, et al. Violence detection in video using computer vision techniques [C]//2011 Fourteenth International Conference Computer Analysis of Images and Patterns(CAIP). Sevilla, Spain. IEEE, 2011: 332-339.

[29] HASSNER T, ITCHER Y, KLIPER-GROSS O. Violent flows: Real-time detection of violent crowd behavior [C]//2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, RI, USA. IEEE, 2012: 1-6.

作者简介

娄 久 男,1977年生,吉林长春人。哈尔滨工业大学高工,硕士研究生,主要研究方向为模式识别、边缘计算。
E-mail: loujiul@hit.edu.cn

左德承 男,1972年生,黑龙江五常人。哈尔滨工业大学教授,博士生导师,主要研究方向为可穿戴计算、计算机系统结构。
E-mail: zuodc@hit.edu.cn

张 展 男,1978年生,黑龙江哈尔滨人。哈尔滨工业大学副教授,硕士生导师,主要研究方向为移动计算技术、可穿戴计算技术、分布式计算技术。
E-mail: zhangzhan@hit.edu.cn

刘宏伟 男,1971年生,黑龙江大庆人。哈尔滨工业大学教授,博士生导师,主要研究方向为容错计算、计算机体系结构、云计算。
E-mail: liuhw@hit.edu.cn