基于深度残差双单向 DLSTM 的时空一致视频事件识别

李永刚[1),2)] 王朝晖[1)] 万晓依[1)] 董虎胜[1)] 龚声蓉[1),3),4)]
刘纯平[1),5)] 季 怡[1)] 朱 蓉[2)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)
[2)](嘉兴学院数理与信息工程学院 浙江 嘉兴 314001)
[3)](常熟理工学院计算机科学与工程学院 江苏 常熟 215500)
[4)](北京交通大学计算机与信息技术学院 北京 100044)
[5)](吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

摘 要 监控视频下的事件识别是近期计算机视觉领域的研究热点之一.然而,自然场景下监控视频往往具有背景复杂、事件区域内对象遮挡严重等特点,使得事件类内差异大、类间差异小,给识别带来了很大的困难.为解决复杂背景下事件识别问题,提出了一种基于深度残差双单向 DLSTM(DRDU-DLSTM)的时空一致视频事件识别方法.该方法首先从训练好的时间 CNN 网络和空间 CNN 网络获取视频的时空深度特征,经 LSTM 同步解析后形成时空特征数据联接单元 DLSTM,并作为残差网络的输入.双单向传递的 DLSTM 联接后构成 DU-DLSTM 层;多个 DU-DLSTM 层再加一个恒等映射形成残差模块;在此基础上,多层的残差模块堆叠构成了深度残差网络架构.为了进一步优化识别结果,设计了基于双中心 Loss 的 2C-softmax 目标函数,在最大化类间距离的同时最小化类内间隔距离.在监控视频数据集 VIRAT 1.0 和 VIRAT 2.0 上的实验表明,该文提出的事件识别方法有很好的性能表现和稳定性,识别准确率分别提高了 5.1%和 7.3%.

关键词 事件识别;时空一致;残差网络;LSTM;双单向;DLSTM;深度特征;监控视频
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2018.02852

Deep Residual Dual Unidirectional DLSTM for Video Event Recognition with Spatial-Temporal Consistency

LI Yong-Gang[1),2)] WANG Zhao-Hui[1)] WAN Xiao-Yi[1)] DONG Hu-Sheng[1)] GONG Sheng-Rong[1),3),4)]
LIU Chun-Ping[1),5)] JI Yi[1)] ZHU Rong[2)]

[1)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)
[2)](*College of Mathematics Physics and Information Engineering, Jiaxing University, Jiaxing, Zhejiang* 314001)
[3)](*School of Computer Science and Engineering, Changshu Institute of Science and Technology, Changshu, Jiangsu* 215500)
[4)](*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)
[5)](*Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun* 130012)

Abstract Event recognition in surveillance video is attracting growing interest in recent years. Nevertheless, event recognition in real-world surveillance video still faces great challenges due to

收稿日期:2017-08-10;在线出版日期:2018-03-24.本课题得到国家自然科学基金(61773272,61170124,61272258,61301299)、教育部科技发展中心"云数融合科教创新"基金(2017B03112)、江苏省自然科学基金(BK20151260,BK20151254)、江苏省"六大人才高峰"项目(DZXX-027)、吉林大学符号计算与知识工程教育部重点实验室基金项目(93K172016K08)、浙江省自然科学基金(LY15F020039)、江苏省研究生科研与实践创新计划项目(KYCX17_2006)资助.**李永刚**,男,1979 年生,博士研究生,讲师,中国计算机学会(CCF)会员,主要研究方向为计算机视觉、图像视频处理和模式识别.E-mail:lyg_gang@163.com.**王朝晖**,女,1967 年生,硕士,副教授,主要研究方向为模式识别与图像处理.**万晓依**,女,1993 年生,硕士研究生,主要研究方向为 3D 行为识别.**董虎胜**,男,1981 年生,博士研究生,讲师,主要研究方向为计算机视觉与机器学习.**龚声蓉**(通信作者),男,1966 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为图像视频处理、模式识别和计算机视觉.E-mail:shrgong@suda.edu.cn.**刘纯平**(通信作者),女,1971 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为视觉显著性检测、对象检测与识别和场景理解.E-mail:cpliu@suda.edu.cn.**季 怡**,女,1973 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为 3D 行为识别和复杂场景理解.**朱 蓉**,女,1973 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为智能信息处理、机器学习和数据挖掘.

various facets such as cluttered background, severe occlusion in event bounding box, tremendous intra-class variations while small inter-class variations, etc. A pronounced tendency is that more researches focus on learning deep features from raw data. Two-stream CNNs (Convolutional Neural Networks) architecture becomes a very successful model in video analysis field, in which appearance features and short-term motion features are utilized. In contrast, Long Short-Term Memory (LSTM) network can learn long-term motion features from the input sequence, which is widely used to process those tasks with quintessential time series. In order to combine the advantages of the two types of networks, in this paper, we propose a deep residual dual unidirectional double LSTM (DRDU-DLSTM) for video event recognition in surveillance video with complex scenes. In the first place, deep features are extracted from the fine-tuned temporal CNN and spatial CNN. Since fully connected layers (FC) takes more semantic information than convolutional layers, which are more suitable as the inputs of LSTM network, we extract FC6 feature of spatial CNN and FC7 feature of temporal CNN respectively. Secondly, to reinforce spatial-temporal consistency, the deep features are transformed by spatial LSTM (SLSTM) and temporal LSTM (TLSTM) respectively, and conjugated as a unit called double-LSTM (DLSTM), which forms the input of the residual network. DLSTM cells increase the number of hidden nodes of LSTM cells, and expand the width of the networks. The input features of spatial CNN and temporal CNN are deeply intertwined by DLSTM cells. At the same time, the features will be transmitted and evolved simultaneously, which will increase the consistency of spatial and temporal features. Furthermore, dual unidirectional DLSTMs are concatenated as DU-DLSTM layer. Compared with the shallow bidirectional recurrent network, the deep dual unidirectional network captures the global information better. The architecture of DU-DLSTM can further ramp up the capacity of hidden nodes in networks. The wider networks augment the optional range of features and enhance the coupling capacity of the feature. One or more DU-DLSTM layers are added to an identity mapping to form a residual block, in which the identity shortcut is a good solution to the problem of deep network vanishing gradient. Stacked residual blocks construct the deep residual architecture. The LSTM network with residual structure can reach up to 10 layers, which deepens the depth of the recurrent network. What's more, the network's optimization ability will be greatly enhanced. At last, to further optimize the recognition results, we design 2C-softmax objective function based on two-center Loss, which computes the center of spatial feature C_S and the center of temporal feature C_T separately. C_S and C_T will be fused as one center of mass according to the set weight coefficient. 2C-softmax objective function can minimize the intra-class variations while keep the features of different classes separable. Experiments on VIRAT 1.0 Ground Dataset and VIRAT 2.0 Ground Dataset demonstrate that the proposed method has good performance and stability, which can achieve superior performance by 5.1% and 7.3% respectively compared with the state-of-the-art methods.

Keywords　even recognition; spatial-temporal consistency; residual network; long short-term memory; dual unidirectional; double long short-term memory; deep feature; surveillance video

1　引　言

视频事件识别是指从视频中识别出事件的时空视觉模式[1]. 随着视频监控在现实生活中的广泛应用，监控视频事件识别受到了广泛关注，并取得了一系列的研究成果[2-3]. 然而监控视频的事件识别仍然面临着巨大的挑战和困难，比如自然场景下监控视频背景复杂、事件区域对象遮挡严重、摄像头视角变化等因素，导致事件类间距离小、类内距离大. 图 1

展示部分从监控视频中截取的视频帧,其中(a)中不同的图形框圈出了正在发生的事件,包含开车门、关车门和装载货物三类事件,(b)中的椭圆框圈出了多个关车门事件.从图 1 可以看出,即便人工准确识别监控视频中的事件也较为困难.

(a) 事件帧1 (b)事件帧2

图 1 监控视频事件帧图

为了解决监控视频事件识别困难的问题,研究者们提出了很多解决方案,其中基于视觉词袋模型(BOW)的方法[4]和基于运动轨迹的方法[2]是佼佼者.然而,传统的手工特征识别方法遇到了精度难以进一步提高的瓶颈.近年来,深度学习成为人工智能领域研究的热点,并开始应用于视频监控中的事件检测、行为识别等领域.Xu 等人[5]将深度特征用于视频事件检测,对卷积神经网络(Convolutional Neural Network,CNN)特征作费舍尔向量(Fisher Vector,FV)编码并利用潜在概念描述符生成更好的特征描述符,但其并没有有效利用视频的动态特征.Simonyan 等人[6]提出双流 CNN 网络用于行为识别,时间 CNN 网络利用视频的静态帧信息,空间 CNN 网络利用视频的光流信息.虽然 CNN 在视频处理任务上取得了较大的成功,但是以双流 CNN 网络为代表的方法仅仅利用了视频的短时动态特征,并没有有效利用视频的长时动态特征.

长短时记忆(LSTM)网络可以从输入序列中递归学习长时动态特征,因此在处理具有典型时间序列的任务上取得了很大成功,比如语音识别[7]、行为识别[8]等.Donahue 等人[8]提出长时递归卷积网络(Long-term Recurrent Convolutional Networks,LRCN),LRCN 利用 CNN 网络提取特征,然后送入 LSTM 网络获得识别结果.深层架构提高了 CNN 和 LSTM 网络的识别能力.但是,无论是 CNN 还是 LSTM,随着网络深度的增加,都会遇到梯度消失问题,很难训练深度非常深的网络.

深度残差网络(ResNet)[9]带来了卷积神经网络深度的革命.ResNet 引入了残差块(residual block),使得训练深达数百甚至超过千层的网络成为可能,而且性能依然优异.残差连接很好地解决了深度网络梯度消失的问题,且由于其表征能力强,ResNet 在图像分类之外的很多计算机视觉应用上也取得了巨大的性能提升.视频事件识别最大的难度在于事件类间距离小,而类内距离大,中心 Loss 是个好的解决方案[10].中心 Loss 对每一个类别学习一个中心,并根据样本特征与类中心的距离进行惩罚,进而缩少类内的距离,使学习到的特征具有更好的泛化能力和辨别能力.

基于以上分析,我们在吸收多种网络结构优点的基础上,提出深度残差双单向-双单元长短时记忆网络(DRDU-DLSTM)架构来解决复杂场景下监控视频事件识别问题.在双流 CNN 网络的基础上,通过 LSTM 递归网络进一步利用了视频的长时动态特征,并利用残差架构训练更深的网络提升特征的表征能力,其总体框架如图 2 所示.DRDU-DLSTM 首先从时空双流 CNN 网络中提取深度特征,空间、时间特征数据经 LSTM 解析后分别得到 SLSTM (Spatial LSTM)和 TLSTM(Temporal LSTM),SLSTM 和 TLSTM 联接形成时空特征数据联接单元 DLSTM,并作为残差网络的输入.数据经残差层处理后采用 2C-softmax 目标函数对网络进行了优化.

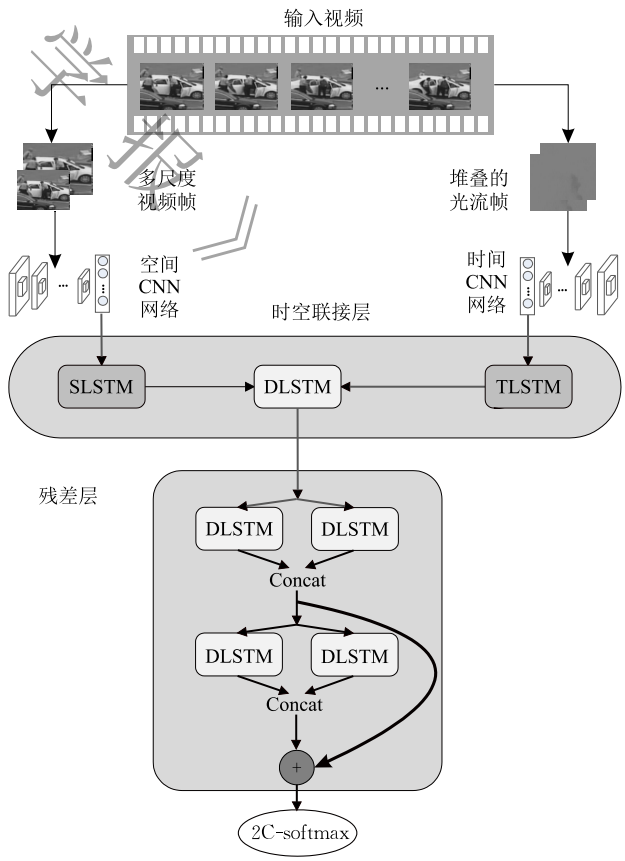


图 2 系统总体框架图

本文的创新点主要有：(1) 设计了时空特征数据联接层，时空特征数据经 LSTM 同步解析后形成时空特征数据联接单元 DLSTM，突出时空信息一致性；(2) 设计了双单向结构 DU-DLSTM，拓宽了网络的宽度，增加了特征选择范围；设计了残差模块 RDU-DLSTM，解决更深层次的网络梯度消失问题. 更深更宽的网络增强了特征的表征能力；(3) 设计了 2C-softmax 目标函数，扩大类间距离的同时兼顾缩小类内距离.

本文第 2 节介绍相关工作；第 3 节描述时空深度特征提取方法；第 4 节详细说明 DRDU-DLSTM 网络模型的设计方法；第 5 节描述网络优化方案；第 6 节通过实验验证网络架构的合理性并对识别结果作深入分析；最后是本文的总结.

2　相关工作

视频监控在现实生活中应用广泛，通过视频分析技术实现视频事件的自动识别，已成为计算机视觉领域的研究热点. Xian 等人[3]采用 FV 表征视频片段的低层特征，利用随机森林作为学习模型实现监控视频事件检测. Wang 等人[11]提出基于层次化上下文的贝叶斯网络模型(BN)，将上下文信息分为特征级上下文、语义级上下文和先验级上下文，并设计推理算法解决多层次特征集成问题，从而实现监控视频事件识别. Coşar 等人[12]根据对象运动轨迹的速度和方向检测监控视频中的异常事件. Zhu 等人[13]提出一种结构化模型，该模型从每个视频类别中学习上下文特征、运动特征及时空关系等. Wang 等人[1]提出一种深度层次化上下文模型，该模型首先从事件边界的邻居区域提取外观上下文和交互上下文等低层特征，然后利用深度玻尔兹曼机学习中层语义表征并联合低层的特征层、中层的语义层和高层的先验层实现监控视频事件识别. Zhu 等人[14]提出一种层次化条件随机场(CRF)模型，该模型将运动信息和上下文特征在不同层次上集成，并将分类问题转化为 CRF 模型解决. 上述方法主要是通过提取视频的上下文信息或者低层特征实现事件识别. Xu 等人[2]提出一种室内监控视频多人事件检测方法，该方法基于直方图的特征描述子可以捕获轨迹间的角度，进而可以捕获多人事件的运动模式. Zaidenberg 等人[15]根据视频中的突然运动识别群体事件. 这些方法主要是通过捕获视频的运动轨迹识别突发的群体事件.

近年来，深度卷积神经网络在计算机视频领域产生了深远的影响，Alexnet[16]、GoogLeNet[17]、VGGNet[18]等经典 CNN 架构不仅在图像处理任务上取得突破性的成绩，在视频处理任务上也得到了广泛的应用. 何克磊等人[19]提出一种多示例深度卷积网络模型，通过引入原型学习层实现示例特征至包特征的映射，解决弱标记环境下的多示例学习问题. Simonyan 等人[6]提出一个双流 CNN 网络架构，空间 CNN 网络和时间 CNN 网络分别利用视频帧和堆叠光流帧训练，开创了时空双模态视频处理的新模式. 在双流网络的基础上，Wang 等人[20]结合了 VGGNet 的 16 层模型架构，重新训练了双流网络，称之为非常深双流 CNN 网络，取得了比双流网络更好的性能. Feichtenhofer 等人[21]提出双流网络融合方法，利用 3D 卷积和 3D 池化融合方法形成时空数据流. Xu 等人[5]对 CNN 特征作 FV 编码并利用潜在概念描述符生成更好的特征描述符实现事件识别. Hou 等人[22]首先利用 VGG 获得全连接层的深度特征，并分别送入捕获运动特征的递归神经网络(Recurrent Neural Networks，RNN)和捕获外观特征的空间网络，融合后得到事件检测结果. Gan 等人[23]设计了深度事件网络(DevNet)，利用 CNN 生成视频关键帧的时空显著图检测视频事件. 王梦来等人[24]引入级联 CNN 网络和轨迹分析方法在复杂场景中检测监控视频事件. 更深的网络往往意味着特征的抽象能力更强、语义信息更丰富. 然而在训练非常深的 CNN 网络模型时，会遇到梯度消失的问题，CVPR 2016 最佳论文 ResNet[9]通过残差连接提供了非常理想的解决方案，成为近两年来最为火热的 CNN 架构之一. Feichtenhofer 等人[25]设计了时空残差网络(ST-ResNet)，在时间残差网络和空间残差网络之间建立残差连接，该方法可以进一步提升视频行为识别的准确率. Feichtenhofer 等人[26]提出一种应用于动态场景识别的时间残差网络(T-ResNet)，T-ResNet 将空间 ResNet 架构转化为时空架构，时空残差单元通过时间滤波器将时间信息层次化地投射到残差块中. CVPR 2017 最佳论文 DenseNet[27]在 ResNet 的基础上进行了拓展，在网络的各层建立稠密连接，即每层以之前层的输出作为该层的输入，第 L 层一共有 L 个连接，对于 DenseNet，则有 $L(L+1)/2$ 个连接. DenseNet 在大型图像数据集上进一步地提升了性能. 可以预见，基于残差网络的 CNN 模型架构将是在未来一段时间内的研究热点之一.

与卷积神经网络自动学习输入数据的全局特征相比,RNN 可以从输入序列中递归学习复杂的时间动态特征,适合处理具有典型时间序列的任务. Adi 等人[28]利用 RNN 分割声音序列实现语音识别. Woo 等人[29]基于 RNN 实现多人环境下的行为识别. 为了解决 RNN 可能会遇到的梯度弥散问题,Hochreiter 等人[30]提出用 LSTM 网络替代传统的 RNN. Gammulle 等人[31]设计双流融合 LSTM 架构用于行人识别. Zhao 等人[32]提出一种长时残差递归网络(LRRN),实现监控视频双人交互行为识别. 这些方法主要从行为的时序入手识别简单场景下的行为.

综上所述,目前的事件识别方法主要从增大类间距离的目标入手,在处理复杂场景下的事件识别上效果不理想. 本文提出的 CNN+深度残差 LSTM 网络架构,获取的特征更具区分能力,更大程度上增加了事件类间距离,同时双中心 Loss 缩小了类内距离,更适合复杂背景下的事件识别.

3 时空深度特征提取

在双流卷积神经网络的框架下,我们微调了 Wang 等人设计的非常深双流模型(Very deep two-stream CNNs)[20],如图 3 所示.

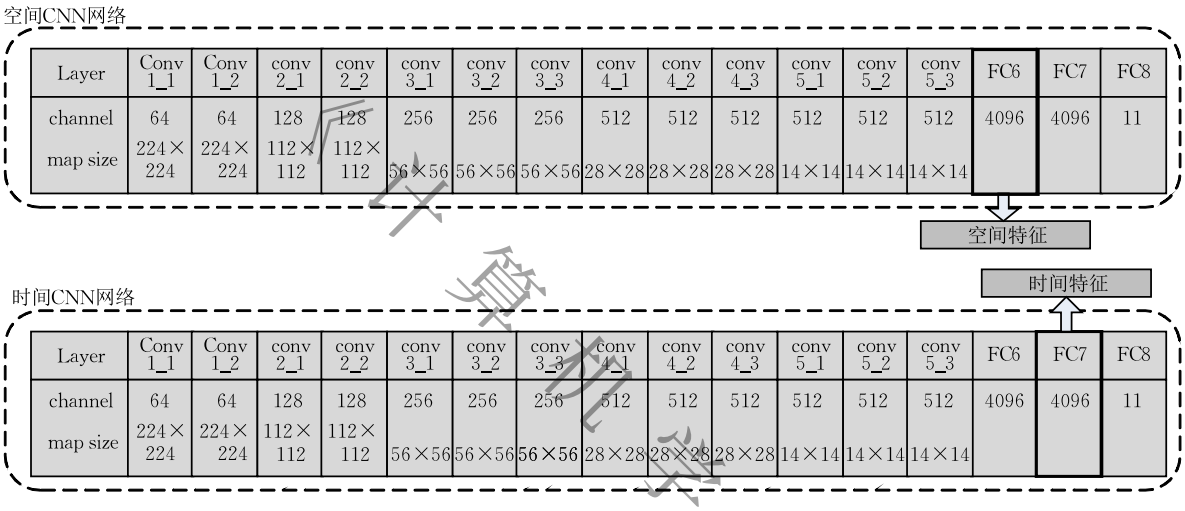


图 3 基于 VGG 16 层架构的非常深双流模型

该模型结合了双流网络和 VGG 模型的优点,具有比双流网络更大的优势. 对于空间网络,从视频中提取视频帧作为输入. 对于时间网络,首先采用 Wang 提供的稠密光流帧提取方法①,利用 OpenCV 从视频中获取水平方向和垂直方向的光流帧,然后将 20 个光流帧构成一个光流组(10 对 flow_x 和 flow_y)作为时间 CNN 网络的输入. 相比于卷积层特征,全连接层特征具有更高层次的抽象和更好的语义信息,更适合作为 LSTM 网络的输入. 在实验部分,我们进一步证明了空间 CNN 网络的 FC6 层特征和时间 CNN 网络的 FC7 层特征具有更好的效果,因此我们分别抽取空间 CNN 网络的 FC6 层特征和时间 CNN 网络的 FC7 层特征作为 LSTM 的输入.

4 DRDU-DLSTM 网络模型

本节主要阐述了 DRDU-DLSTM 模型的网络架构及设计方案,其中 4.1 节设计了模型架构,4.2 节介绍了 LSTM 工作原理及双单向 LSTM 设计方法,4.3 节描述了时空特征数据联接单元 DLSTM,4.4 节阐述了残差双单向 DLSTM 设计方案.

4.1 模型架构

DRDU-DLSTM 网络模型主要由时空特征数据联接层和堆叠的残差层组成,如图 4 所示. 图 4 描述了网络在 1、2 以及 τ 时刻的状态,每个状态描述了各层之间的关系.

从图 4 可以看到,时空特征数据联接层接收从深度 CNN 网络提取的时间特征 $\boldsymbol{x}^{\mathrm{T}}$ 和空间特征 $\boldsymbol{x}^{\mathrm{S}}$,时空特征数据经 LSTM 同步解析后形成时空特征数据联接单元 DLSTM. 双单向的 DLSTM 联接后构成 DU-DLSTM 层,多个 DU-DLSTM 层再加一个恒等映射形成一个残差模块. 数据经过深层的残差网络计算后反向传播,误差逐渐得到优化.

① Wang L. Opencv implementation of different optical flow algorithms. https://github.com/wanglimin/dense_flow

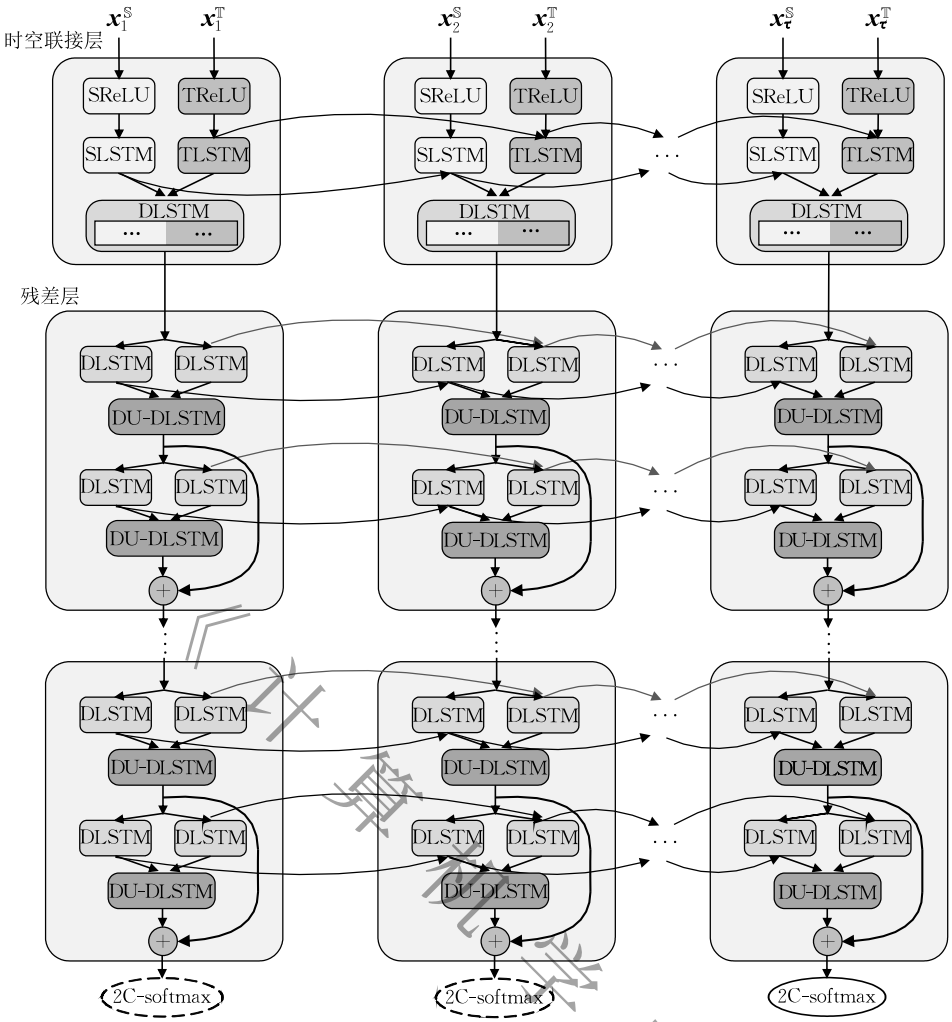


图 4　DRDU-DLSTM 网络模型

4.2　LSTM

视频事件识别的目标是从视频中识别出事件的时空视觉模式，而事件的时序是比较复杂的，给识别任务带来了困难. LSTM 可以从输入序列中递归学习长时动态特征，因此在处理复杂时序任务时能够取得很大的成功.

传统的递归神经网络能够将输入序列$(x_1,\cdots,x_t)$映射为隐结点序列$(h_1,\cdots,h_t)$，从而可以从输入序列中递归学习复杂的时间动态特征. 然而，当时序信息经过多个时间步传递后，RNN 可能会遇到梯度弥散或梯度爆炸问题，网络参数更新受到阻碍，从而很难学到长时动态信息. LSTM 提供了解决上述问题的方案. LSTM 提出几种类型的门结构和记忆单元，如图 5 所示，可以使得梯度在时间轴上长时传播，进而能够学到长时的动态特征.

LSTM 的门结构通过非线性激活函数控制. 设$\sigma(x)=(1+e^{-x})^{-1}$表示 sigmoid 非线性函数，其取值范围为$[0,1]$，$\psi(x)=\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$表示双曲正切非线性函数，其取值范围为$[-1,1]$. LSTM 的门结构可由式(1)表示：

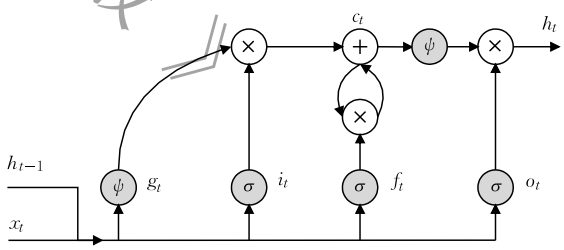


图 5　LSTM 结构

$$\begin{aligned} i_t&=\sigma(U_{xi}x_t+W_{hi}h_{t-1}+b_i),\\ f_t&=\sigma(U_{xf}x_t+W_{hf}h_{t-1}+b_f),\\ o_t&=\sigma(U_{xo}x_t+W_{ho}h_{t-1}+b_o),\\ g_t&=\psi(U_{xc}x_t+W_{hc}h_{t-1}+b_c) \end{aligned} \tag{1}$$

其中 i_t、f_t、o_t和 g_t分别表示输入门、遗忘门、输出门和输入调制门，$\boldsymbol{U}$、$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示各个门的输入权重、递归权重和偏置项，x_t表示 t 时刻的输入，h_{t-1}表示 $t-1$ 时刻的输出. i_t决定更新什么信息，f_t决定丢

弃什么信息，o_t确定记忆单元的哪个部分将输出，g_t是由双曲正切函数ψ创建的一个新的候选值向量.

记忆单元c_t是LSTM的核心部分，决定什么样的信息被保存，如式(2)所示. c_t由两部分构成，一部分由上一时刻记忆单元c_{t-1}与遗忘门f_t相乘获得，另一部分由输入门i_t和输入调制门g_t相乘得到.

$$c_t = f_t \times c_{t-1} + i_t \times g_t \tag{2}$$

LSTM的输出h_t可由输出门o_t控制是否激活记忆单元c_t，c_t通过ψ处理(取值为[-1,1])后并将它和输出门o_t相乘，如式(3)所示：

$$h_t = o_t \times \psi(c_t) \tag{3}$$

4.3 时空特征数据联接单元 DLSTM 设计

采用双流CNN网络分别提取视频的时间特征和空间特征，然后采用后融合方法得到分类结果，这是常用的视频识别方法[6,20,33]. 但是上述方法仅从结果上对时空特征融合，没有考虑时空信息在迁移过程中的一致性. 为了充分利用视频长时特征和视频时空一致性提高事件识别准确率，我们设计了时空特征数据联接层，如图4所示. 时空特征数据联接层采用两个LSTM单元，分别记为SLSTM和TLSTM. SLSTM接收来自空间CNN网络的特征，TLSTM接收来自时间CNN网络的特征. LSTM单元在接收输入前，需采用非线性激活函数对输入数据处理，本文采用ReLU激活函数. ReLU激活函数既可以解决梯度消失问题，又因其神经元的稀疏激活性而能够数倍地提高网络的速度[16]，近来被广泛使用. SLSTM和TLSTM经联接操作形成一个新的单元DLSTM，DLSTM的结构如式(4)所示.

$$h_{DL} = \chi(\delta(W_S h_{SL} + b_S), \delta(W_T h_{TL} + b_T)) \tag{4}$$

其中δ表示ReLU激活函数，χ表示联接操作，h_{SL}和h_{TL}分别表示SLSTM和TLSTM单元的输入，h_{DL}为DLSTM的输出，$\boldsymbol{W}$和$\boldsymbol{b}$分别表示权重和偏置项. DLSTM单元增加了LSTM单元隐结点的数量，拓展了网络的宽度，如图4中的DLSTM单元所示. 同时，DLSTM单元将时空输入特征紧密联系在一起，时空特征在LSTM网络中同时传递和进化，增强了时空特征的一致性.

4.4 双单向结构 DU-DLSTM 设计

尽管LSTM能够捕获长时信息，但其有较强的先后关系，即t时刻的状态只能捕获$x_1,\cdots,x_t$的输入信息. 在有些应用中，希望t时刻的输出依赖于全部输入信息. Schuster等人[34]提出了双向RNN，并在自然语言处理等领域[35]获得成功. Chevalier等利用双向LSTM实现了行为识别①. 双向LSTM结构如图6所示.

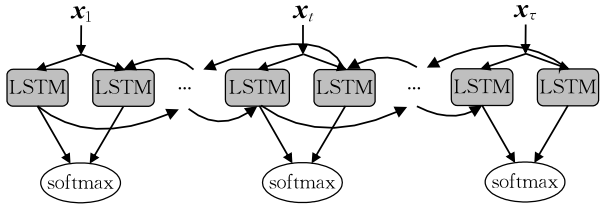


图6 双向结构 Bi-LSTM

然而，深层的双向LSTM网络容易出现优化瓶颈，从而不能有效提高预测结果. 为此，我们设计了双单向DLSTM(DU-DLSTM)，其结构如图7所示，其中每个DLSTM单元包含了来自时间CNN网络和空间CNN网络的输入，两个单向传递的DLSTM联接后构成DU-DLSTM单元. DU-DLSTM单元可描述为

$$h_{DU} = \chi(\delta(W_1 h_{DL_1} + b_1), \delta(W_2 h_{DL_2} + b_2)) \tag{5}$$

其中h_{DL_1}和h_{DL_2}表示两个相同传递方向DLSTM单元的输入，$\boldsymbol{W}$和$\boldsymbol{b}$分别表示权重和偏置项，h_{DU}为DU-DLSTM的输出. 设计DU-DLSTM结构的原因在于：随着LSTM网络深度的增加，信息也在不断地进化，视频帧的时序信息经过深层的传递后，在全局范围内得到融合，相比浅层的双向递归网络，深层的双单向网络更好地捕获了全局信息，也就能获得更好的结果. 而且，我们设计的DLSTM单元及DU-DLSTM结构增加了网络中隐结点的数量，从而增加了网络的宽度，这一点与GoogLeNet的Inception结构相似. 更宽的网络增加了特征选择的范围，增强了特征的耦合能力.

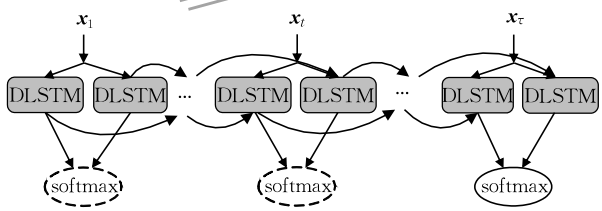


图7 双单向结构 DU-DLSTM

4.5 残差结构设计

4.5.1 残差网络

无论是卷积神经网络还是递归神经网络，网络的层数越多，往往意味着提取到的层特征越丰富，也就能提取到更多的语义信息. 但是，简单地增加网络的深度会导致梯度消失问题，反而降低识别率. 深度残差网络可以解决网络深度增加导致的梯度消失

① Chevalier G. Lstms for human activity recognition. https://github.com/guillaume-chevalier/LSTM-Human-Activity-Recognition, 2016

问题[9]. 基于残差思想构造更深的网络架构，也就能获得更鲁棒的事件特征表达，从而可以提高事件识别率.

深度网络的误差可以表示为

$$Loss=F_N(X_{L_N},W_{L_N},b_{L_N}) \tag{6}$$

$$X_{L_N}=F_{N-1}(X_{L_{N-1}},W_{L_{N-1}},b_{L_{N-1}})$$

$$\cdots \tag{7}$$

$$X_{L_2}=F_1(X_{L_1},W_{L_1},b_{L_1})$$

当梯度反向传播到第一层时，其偏导如式(8)所示：

$$\begin{aligned}\frac{\partial Loss}{\partial X_{L_1}}&=\frac{\partial Loss}{\partial X_{L_N}}*\frac{\partial X_{L_N}}{\partial X_{L_{N-1}}}*\cdots*\frac{\partial X_{L_2}}{\partial X_{L_1}}\\&=\frac{\partial F_N(X_{L_N},W_{L_N},b_{L_N})}{\partial X_{L_N}}*\\&\quad\frac{\partial F_{N-1}(X_{L_{N-1}},W_{L_{N-1}},b_{L_{N-1}})}{\partial X_{L_{N-1}}}*\cdots*\\&\quad\frac{\partial F_1(X_{L_1},W_{L_1},b_{L_1})}{\partial X_{L_1}}\end{aligned} \tag{8}$$

从式(8)可以看到，梯度传播到前几层的时候就会越来越小，就会产生梯度消失的问题. 深度残差网络通过添加残差连接解决因网络深度增加导致的梯度消失问题. 深度残差网络设计为 $H(x)=F(x)+x$，如图 8 所示. $F(x)$是一个关于恒等 x 的残差映射，$H(x)$是任意一种理想的映射. 通过求偏导可以发现：

$$\begin{aligned}\frac{\partial X_L}{\partial X_l}&=\frac{\partial X_l+\partial F(X_l,W_l,b_l)}{\partial X_l}\\&=1+\frac{\partial F(X_l,W_l,b_l)}{\partial X_l}\end{aligned} \tag{9}$$

这样就算深度很深，梯度也不会消失了.

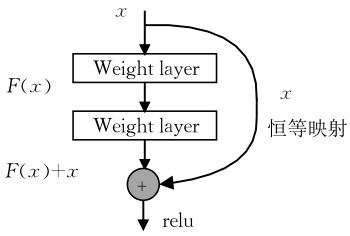


图 8　经典残差模块

残差学习结构可以通过前向神经网络加快捷连接实现[9]. 快捷连接块定义为

$$y=F(x,\{W_i\})+x \tag{10}$$

其中 x 和 y 分别表示网络层的输入和输出，函数 $F(x,\{W_i\})$表示待学习的残差映射. 快捷连接相当于简单执行了恒等映射，不会增加计算复杂度，而且网络也可以通过端到端反向传播训练.

4.5.2　RDU-DLSTM 模块设计

我们的残差模块参考了文献[9]的设计，将 DU-DLSTM 结构作为一个网络层，第一个 DU-DLSTM 结构的输出 h_{DU} 作为 x，如图 4 所示，快捷连接对 h_{DU} 做一个线性变换 W_i，残差双单向 DLSTM(RDU-DLSTM)模块的输出如式(11)所示：

$$h=F(h_{DU},\{W_i\})+h_{DU} \tag{11}$$

通常情况下，LSTM 网络可以堆叠 3～5 层，更深层次的 LSTM 网络也会遇到梯度消失问题，使得网络的精确度下降. 而采用残差结构的 LSTM 网络可以达到 10 层以上，加深了网络的深度，网络的优化能力也会得到加强.

5　网络优化

网络反向传播通过计算损失函数实现，通常情况下可以用 softmax 的 Loss：

$$L_J=-\sum_{i=1}^{m}\log\frac{e^{\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+b_{y_i}}}{\sum_{j=1}^{n}e^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i+b_j}} \tag{12}$$

其中 $\boldsymbol{x}_i$表示第 i 个特征向量，y_i表示类别标签，n 为类别数，m 表示小批量(mini-batch)的大小，$\boldsymbol{W}$ 为权重，$\boldsymbol{b}$ 是偏置项.

为了防止过拟合，可以给 softmax 的 Loss 加上正则项. DLSTM 单元对网络具有重要的影响，因此可以加入 DLSTM 单元权重的二范数作为正则项：

$$L_{DL}=\frac{1}{2}\sum_{i=1}^{m}\|W_{DL_i}\|_2^2 \tag{13}$$

$$L=L_J+\alpha\sum_{k=1}^{D}L_{DL_k} \tag{14}$$

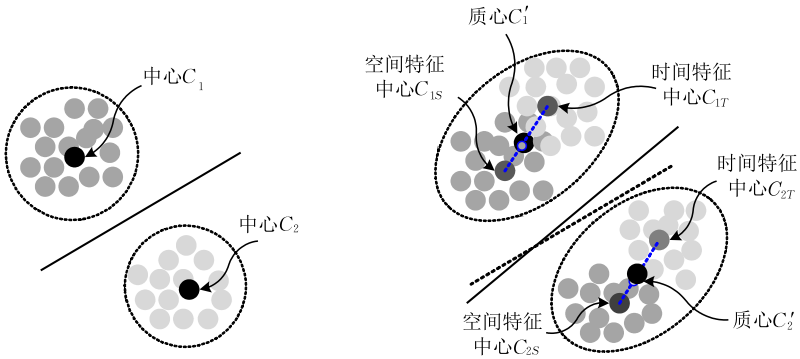
其中 m 表示小批量的大小，W_{DL_i} 表示第 i 个样本的权重，D 表示 DLSTM 单元的个数，α 为正则项系数.

Wen 等人[10]设计了中心 Loss 函数，对每个类别在特征空间都维护一个类中心 C，如图 9(a)所示，此后如果新增样本的特征距离类中心的特征太远就要惩罚，从而兼顾了缩小类内距离与扩大类间距离. 同样是作为训练阶段的辅助 Loss，中心 Loss 和 Contrastive Loss[36] 和 Triplet Loss[37] 相比，其优点在于省去了复杂并且含糊的样本对构造过程. Center Loss 的计算公式如式(15)所述：

$$L_C=\frac{1}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-c_{y_i}\|_2^2 \tag{15}$$

其中 $\boldsymbol{x}_i$ 表示第 i 个样本的特征向量，c_{y_i} 表示该样本所属类别的特征值中心.

在我们的事件识别算法中，DRDU-DLSTM 网络的输入来自于时间 CNN 网络和空间 CNN 网络的两类特征，因此，我们设计一个双中心 Loss，如图 9(b)所示.



(a) 中心Loss (b) 双中心Loss

图 9 样本中心 Loss 的特征值划分

双中心 Loss 分别维护空间特征中心 C_S 和时间特征中心 C_T，C_S 和 C_T 按一定权重系数 β_S 和 β_T 融合形成质心 C'. 为了让质心 C' 在 C_S 和 C_T 的连线上，此时 C' 同时离 C_S 和 C_T 最近，因此采用线性加权方式确定 β_S 和 β_T 权重系数. 加入双中心 Loss 可用式(16)描述：

$$L'=L_J+\beta_S L_{C_S}+\beta_T L_{C_T} \tag{16}$$

进一步地，我们加入 DLSTM 单元的正则项，以防止目标函数过拟合，如式(17)所示：

$$L''=L_J+\alpha\sum_{k=1}^{D}L_{DL_k}+\beta_S L_{C_S}+\beta_T L_{C_T} \tag{17}$$

将式(12)、(13)、(15)代入式(17)得目标函数：

$$L''=-\sum_{i=1}^{m}\log\frac{e^{\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+b_{y_i}}}{\sum_{j=1}^{n}e^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i+b_j}}+\frac{\alpha}{2}\sum_{k=1}^{D}\sum_{i=1}^{m}\|W_{DL_{ki}}\|_2^2+$$

$$\frac{\beta_S}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-c_{S_i}\|_2^2+\frac{\beta_T}{2}\sum_{i=1}^{m}\|\boldsymbol{x}_i-c_{T_i}\|_2^2 \tag{18}$$

式(18)称之为 2C-softmax，softmax 的 Loss 保证类间距离尽量大，双中心 Loss 可以缩小类内距离，DLSTM 正则项防止目标函数过拟合.

6 实验及结果分析

本节首先介绍了实验平台搭建情况，然后在 VIRAT 2.0 数据集上对数据输入方式、LSTM 传递方向、残差单元和堆叠层数、LSTM 网络传播误差等方面进行了详细分析，结合案例对识别结果作了分析，并在 VIRAT 1.0 和 VIRAT 2.0 两个数据集上和其他方法作了对比.

6.1 实验平台搭建

6.1.1 数据集

VIRAT 1.0 数据集[38]包含了约 3 小时的监控视频，180 多个事件案例. 视频由安装在校园停车场的固定高清摄像机拍摄，分辨率为 1280×720 像素或 1920×1080 像素. VIRAT 1.0 数据集的事件类型包括 6 类人车交互事件：(1)装载货物(loading)，(2)卸载货物(unloading)，(3)打开车门(opening)，(4)关闭车门(closing)，(5)进入车辆(into vehicle)，(6)走出车辆(out vehicle).

VIRAT 2.0 数据集[38]包含了 8.5 小时的监控视频，11 类事件，1500 多个事件案例. 视频由安装在校园停车场、商场入口、建筑工地等场所的固定高清摄像机拍摄，分辨率为 1280×720 像素或 1920×1080 像素. VIRAT 2.0 数据集扩展自 VIRAT 1.0 数据集，事件类别由 6 类扩展为 11 类，原有的 6 类事件增加了部分事件案例，新增的事件类别涉及人与建筑物、人与物以及人体行为等，新增的事件类型有：(1)进入商场(entering facility)，(2)走出商场(exiting facility)，(3)打手势(gesturing)，(4)搬运物体(carrying)，(5)跑步(running).

6.1.2 实验参数设置

对于微调的非常深双流 CNN 模型，空间 CNN 网络的输入是从视频事件片段中提取的视频帧(224×224×3)，时间 CNN 网络的输入是由 10 对(x、y 方向)光流帧堆叠而成的光流组(224×224×20). 空间 CNN 网络和时间 CNN 网络均采用随机梯度下降法(Stochastic Gradient Decent，SGD)每次输入一个微型集(mini-batch)对网络进行训练，动量

均为 0.9. 空间 CNN 网络的学习率为 0.01，权重衰减系数为 0.0005，训练过程迭代次数为 10 K. 时间 CNN 网络的学习率为 0.005，权重衰减系数为 0.0005，训练过程迭代次数为 30 K. 视频事件片段通过事件邻域(event neighborhood)[1]的方式从原始视频中获得，其邻域参数 λ 设置为 0.35. CNN 网络采用 Caffe 工具箱实现. 实验在 GPU 服务器上完成，操作系统为 Centos 7，使用了 2 个 K20 加速卡. VIRAT 1.0 数据集中的 180 多个事件案例视频、VIRAT 2.0 数据集中的 1500 多个事件案例视频分别提取了空间 CNN 网络的全连接层 FC6 的特征和时间 CNN 网络的全连接层 FC7 特征，生成视频特征数据文件. 视频特征数据文件按文件名随机置乱后，选取其中的 70% 为 DRDU-DLSTM 网络的训练数据，剩余的 30% 数据作为测试数据，并作为 DRDU-DLSTM 网络时空特征数据联接层的输入.

DRDU-DLSTM 网络的学习率设为 0.001，梯度阈值设为 15.0，dropout 设为 0.85，批次大小为 100、250 个训练周期，LSTM 单元的隐结点数设为 28. 误差正则化系数 $\alpha=0.0001$，$\beta_S=0.000\,06$，$\beta_T=0.000\,04$.

6.1.3　对比方法

为了验证本文方法的有效性，本文与以下方法作了对比：(1) Jiang 等人[4]提出的基于词袋(BOW)的方法，该方法采用软加权评估视觉单词的重要性，BOW 是经典的视频分类方法；(2) Amer 等人[39]设计的 SPN 网络(Sum-Product Network)架构，网络由 BOW 构成的结点组成，加和结点对所有的动作编码，乘积结点则对特定动作编码；(3) Gaur 等人[40]设计的 SFG(String of Feature Graphs)模型，构建局部特征图描述时空特征之间的联系；(4) Zhu 等人[13]提出的结构化模型(Structural Model)；(5) Zhu 等人[14]提出的层次化 CRF 模型(Hierarchical-CRF)；(6) Wang 等人[11]提出基于层次化上下文的贝叶斯网络模型(BN)；(7) Wang 等人[41]提出的深度层次化上下文模型(DHCM)，该模型采用深度玻尔兹曼机实现. (1)～(3)是传统的视觉模型，(4)～(6)是结构化模型，(7)采用了深度模型.

6.2　实验结果分析

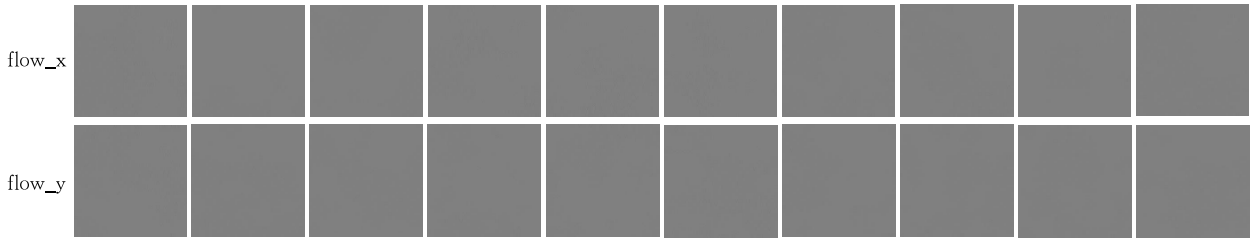
实验首先比较了 VIRAT 2.0 数据集上数据输入模态对 DRDU-DLSTM 网络的影响，以验证时空特征数据联接单元的有效性，如表 1 所示. DRDU-DLSTM 网络的结构为 1 个残差单元、5 个堆叠层，Loss 为式(14)所述的 L.

表 1　输入方式对网络的影响

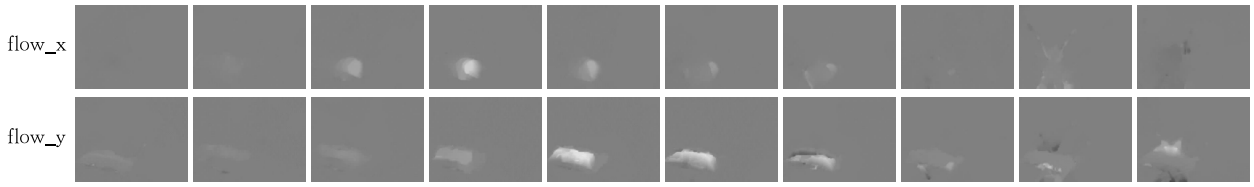
数据输入模态	数据层	准确率/%
空间数据流	S_FC6	80.67
	S_FC7	79.71
时间数据流	T_FC6	68.49
	T_FC7	69.21
双流独立输入后融合	S_FC6&T_FC7	80.67
双流联接输入	S_FC6&T_FC7	**82.58**

从表 1 可以看出，无论是时、空数据流分别作为独立输入，还是取双流独立输入后融合的结果，并不能提高识别准确率. 我们分析发现，在时间数据流检测正确而空间数据流检测错误的案例中，由于时间数据流仅有微弱的概率优势，并没有做到和空间数据流互补. 而我们设计的双流联接输入模式，准确率可以提高 2% 左右，其原因主要在于深层的残差 DLSTM 结构在传递过程中，时空双流连接输入单元 DLSTM 加深了时空信息的融合，时空信息做到了最大程度上的互补.

从表 1 可以发现，在两种数据流的全连接层特征中，空间数据流的 S_FC6 层特征比 S_FC7 层特征好，时间数据流的 T_FC7 层特征比 T_FC6 层特征好，而空间数据流明显高于时间数据流的识别结果. 文献[6]所述的双流网络，在行为识别数据集 UCF101 上，其时间网络的准确率要优于空间网络，文献[20]所讲练的非常深双流 CNN 模型在 UCF101 上也印证了这一结果. 对比 UCF101 和 VIRAT 2.0 两个数据集可以发现，UCF101 数据集中的视频大多为近距离拍摄，行为动作明显，光流信息有很强的区分度，以堆叠光流作为输入的空间 CNN 网络能够发挥非常突出的作用. 而 VIRAT 2.0 数据集中的视频为高空拍摄的远景监控视频，从整个监控画面定位、截取的事件视频动作隐蔽且动作幅度小，从事件视频提取的光流帧所含的信息量偏小，容易导致时间 CNN 网络识别效果差. 图 10(a)为一个检测失败的案例，从图 10(a)可以看出，由 10 对光流帧构成的一个光流组，其包含的信息量非常低. 而光流信息丰富的光流组，如图 10(b)所示，则能够很好地描绘运动信息，因而能够正确检测.



(a) 失效案例的光流组



(b) 有效案例的光流组

图 10 堆叠的光流组

表 2 给出了 VIRAT 2.0 上不同的传递方向对网络的影响. 实验采用了双流联接输入方式. 从表 2 可以看出,在语音识别等领域获得成功的双向传递方式识别准确率不高,甚至低于单向传递方式,说明在事件识别中,后续帧对前帧的正向影响不大,时序的先后关系更重要;而本文设计的双单向传递方式获得了最好的结果,表明双单向传递的 DLSTM 单元拓宽网络的宽度,增加了特征选择的范围,增强了特征的耦合能力.

表 2 DLSTM 的传递方向对网络的影响

DLSTM 的传递方向	准确率/%	$F1$/%
单向传递	81.38	81.67
双向传递	79.95	79.79
双单向传递	**82.58**	**82.48**

为了说明网络层次结构对识别结果的影响,表 3 对比了 VIRAT 2.0 上残差单元数量和堆叠深度对网络的影响. 实验采用双流联接输入方式,Loss 为式(18)所述的 L''. 表 3 给出了网络在 1～2 个残差单元、堆叠深度为 2～6 层时的准确率和 $F1$ 值. 其中 $F1$ 的计算公式如式(19)所示.

$$F1=\frac{2PR}{P+R} \tag{19}$$

其中 P 表示精准率,R 表示召回率.

实验结果表明,不同的层次结构对网络有一定的影响,残差单元和堆叠深度取值应适中,更多的残差单元或更深的堆叠深度并不能提高准确率和 $F1$ 值. 本文中残差单元为 1、堆叠深度为 5 时取得了最好的结果.

表 3 残差单元和堆叠层数对网络的影响

残差单元	堆叠深度	准确率/%	$F1$/%
1	2	83.53	83.22
1	3	83.77	83.98
1	4	83.53	83.54
1	5	**84.73**	**84.75**
1	6	81.14	80.93
2	2	80.19	79.95
2	3	83.77	83.80
2	4	81.62	81.71
2	5	80.42	80.28
2	6	81.62	81.58

另外,为了说明优化后的 Loss 在网络的作用,表 4 对比了 VIRAT 2.0 上不同的 Loss 设计方案对网络的影响. 实验表明,仅用双中心 Loss 并不能有效提高识别结果,双中心 Loss 与 DLSTM 单元正则项结合可以取得更好的效果.

表 4 Loss 对网络的影响

Loss	准确率/%	$F1$/%
$L_J+\alpha\sum_{k=1}^{D}L_{DL_k}$	82.58	82.48
$L_J+\beta_S L_{C_S}+\beta_T L_{C_T}$	81.62	81.19
$L_J+\alpha\sum_{k=1}^{D}L_{DL_k}+\beta_S L_{C_S}+\beta_T L_{C_T}$	**84.73**	**84.75**

本文算法在 VIRAT 2.0 上的混淆矩阵展示了容易错误识别事件间的关系,如图 11 所示. 从图 11 中可以看出,进入商场事件和走出商场事件之间的混淆度较高,均达到 25%以上,而这两类事件与其它事件的混淆度则接近于 0. 而另外两类识别准确率较低的事件,进入车辆事件和走出车辆事件间的混淆度也达到 20%以上. 而同样具有易混动作的事

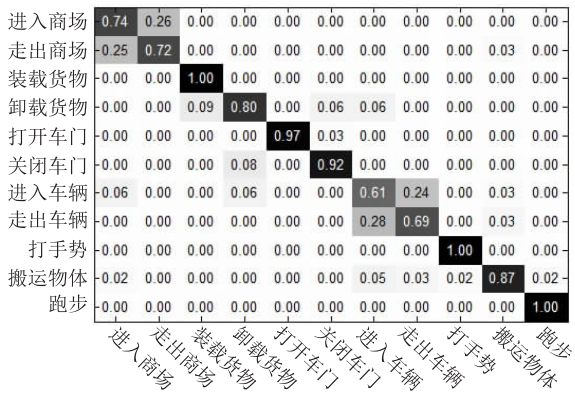


图 11　VIRAT 2.0 的混淆矩阵

件，比如打开车门事件和关闭车门事件间的混淆度则较小. 分析混淆矩阵我们发现，本文算法较好解决了对象遮挡对事件识别带来的不利影响，同时缩小类内间隔、扩大类间间隔，有效地提高了识别准确率，但也存在部分事件的运动信息没有捕获到导致错误识别的情况.

我们对实验中的部分检测案例进行了分析. 从表 5 中可以看出，大部分案例均能正确识别，在这些案例中，虽然识别难度也很大，例如，A1、B1 背景复杂、图像不够清晰，C2、C3、E1、F1 事件区域内的对象遮挡较为严重，我们的算法很好地解决了这些问题. 而在错误识别的案例中，A2 同时有人出入，事件标签本身就有二义性；在 D4 从车辆上卸载货物事件里，除了卸载的物体被遮挡严重外，其事件里附加有开车门、关车门等动作，容易和开车门、关车门事件混淆；E2、F3 等事件，存在空间特征和运动特征区分度都很小的情况，相似事件容易混淆.

表 5　事件识别案例分析

序号	A. 进入商场	B. 走出商场	C. 装载货物	D. 卸载货物	E. 进入车辆	F. 走出车辆
1						
	✓	✓	✓	✓	✓	✓
2						
	×	✓	✓	×	×	✓
3						
	✓	✓	✓	✓	✓	×

为了和文献[11]、文献[41]的 6 类事件识别准确率作对比，我们选取了相同的 6 类事件，如表 6 所示. 从表 6 中可以看出，我们的算法在其中的 5 类中都取得了更好的识别准确率. 文献[41]提取了视频三个层次的上下文特征，并用受限玻尔兹曼机学习. 相比于文献[41]，我们的视频特征采用了 VGG16 层卷积网络架构，分别提取了视频的空间特征和时间特征，特征描述更有表达能力和分辨能力，更关键的是，我们设计的时空特征数据联接单元更好地利用了空间特征和时间特征的互补性，深度残差 LSTM 结构有效利用了视频的长时特性，因而能够取得更好的识别结果.

表 6　VIRAT 2.0 中 6 类事件的识别准确率

事件类别	BN[11]/%	SVM-Context[41]/%	Model-BM[41]/%	SVM-TIP[41]/%	DHCM[41]/%	DRDU-DLSTM/%
装载货物	77.78	66.67	66.67	44.44	66.67	**100.00**
卸载货物	58.62	62.07	68.97	51.72	68.97	**80.00**
打开车门	35.00	15.00	25.00	10.00	45.00	**97.37**
关闭车门	63.16	63.16	84.21	52.63	89.47	**91.89**
进入车辆	68.75	64.58	52.08	58.33	**70.83**	60.61
走出车辆	48.89	40.00	55.56	33.33	57.78	**68.97**
平均值	58.70	51.91	58.75	47.74	66.45	**83.14**

我们在 VIRAT 1.0 和 VIRAT 2.0 两个数据集上和更多的算法作了进一步的对比，如表 7 所示. BOW[4] 虽然在视频检索、行为识别等视频处理任务上取得了广泛的应用和不错的业绩，但其在更具有挑战性的监控视频事件识别任务上效果一般. SPN[39] 在 BOW 的基础上更好地结合了视频的全

局特征和局部特征，取得了比 BOW 更好的效果. SFG[40]考虑了类内的时空联系，但是没有考虑类间的关系. Structural Model[13]、Hierarchical-CRF[14]和 BN[11]均利用了视频的上下文信息和时空特征，虽然也取得了很好的结果，但手工特征的选择禁锢了算法识别的上限.

表 7 VIRAT 1.0 和 VIRAT 2.0 上的对比实验

算法	准确率/%	
	VIRAT 1.0	VIRAT 2.0
BOW+SVM[4,13]	45.80	55.40
SPN[39]	/	70.00
SFG[40,13]	57.60	/
Structural Model[13]	62.90	73.50
Hierarchical-CRF[14]	66.20	75.10
BN[11]	65.80	77.42
DHCM[41]	69.88	77.47
DRDU-DLSTM	**75.00**	**84.73**

相比于文献[41]所用的深度受限玻尔兹曼机，我们采用卷积网络能够从视频中获得更加鲁棒的自动特征，结合残差结构的 LSTM 递归网络，我们的算法能够最大程度上利用视频的空间信息、短时信息、长时信息和时空融合信息，识别准确率也得到了较大幅度的提高. VIRAT 1.0 数据集涉及人车交互事件，VIRAT 2.0 数据集涵盖了人车交互事件、人与建筑物交互事件、人与物体交互事件以及人体自身的行为等多种场景，DRDU-DLSTM 在两个数据集上均取得了最好的结果，说明本文提出的算法具有较好的推广能力.

7 结束语

本文设计了一个深度残差双单向递归网络模型 DRDU-DLSTM，并用于监控视频事件识别任务. 该模型从双流 CNN 网络获得深度特征输入，从输入序列中递归学习长时动态特征，残差模块较好地解决了深度堆叠的 LSTM 梯度消失的问题，2C-softmax 目标函数提高了模型的辨识能力. DRDU-DLSTM 综合利用了 CNN 网络、LSTM 网络和深度残差网络的优点，因而能够学习到表征能力更强的特征. 监控视频数据集上的实验验证了模型的有效性. 从实验结果可以发现：(1) 深度更深的 CNN 网络，其特征的表征能力更强，但不同层上的特征所包含的语义不同，若对深度特征进一步编码，需要选择合适层上的特征；(2) 与两种数据流分别输入到两个网络处理最后结果后融合的方式相比，时空特征数据联接单元的设计更好地体现了时空信息的一致性；(3) 双向 LSTM 网络在有些应用上不仅没有优势，甚至效果不如单向 LSTM，而双单向 LSTM 拓宽了网络的宽度，是一种有益的探索；(4) 残差理论在 LSTM 网络中也能起到很好的效果，残差模块的大小要合适，2～3 层的残差模块比较理想；(5) 合适的目标函数对识别结果影响较大，中心 Loss 虽然性能不错，但要与合适的正则项联合使用.

监控视频事件识别的准确率还有较大的提升空间，深度模型还有很多未知因素可以探索，比如深度模型在每一层的运算中注意力重点集中哪些环节，如何设计更合理的残差快捷连接等，这将是我们未来研究的方向.

参 考 文 献

[1] Wang X, Ji Q. Video event recognition with deep hierarchical context model//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4418-4427

[2] Xu J, Denman S, Sridharan S, et al. An efficient and robust system for multiperson event detection in real-world indoor surveillance scenes. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(6): 1063-1076

[3] Xian Y, Rong X, Yang X, et al. Evaluation of low-level features for real-world surveillance event detection. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 624-634

[4] Jiang Y G, Ngo C W, Yang J. Towards optimal bag-of-features for object categorization and semantic video retrieval//Proceedings of the ACM International Conference on Image and Video Retrieval. Amsterdam, Netherlands, 2007: 494-501

[5] Xu Z, Yang Y, Hauptmann A G. A discriminative CNN video representation for event detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1798-1807

[6] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 568-576

[7] Graves A, Jaitly N, Mohamed A R. Hybrid speech recognition with deep bidirectional LSTM//Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic, 2013: 273-278

[8] Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2625-2634

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[10] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 499-515

[11] Wang X, Ji Q. A hierarchical context model for event recognition in surveillance video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 2561-2568

[12] Coşar S, Donatiello G, Bogorny V, et al. Toward abnormal trajectory and event detection in video surveillance. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 683-695

[13] Zhu Y, Nayak N M, Roy-Chowdhury A K. Context-aware modeling and recognition of activities in video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2491-2498

[14] Zhu Y, Nayak N M, Roychowdhury A K. Context-aware activity modeling using hierarchical conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(7): 1360-1372

[15] Zaidenberg S, Bilinski P, Brémond F. Towards unsupervised sudden group movement discovery for video surveillance//Proceedings of the International Conference on Computer Vision Theory and Applications. Lisbon, Portugal, 2014: 388-395

[16] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

[17] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1-9

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014

[19] He Ke-Lei, Shi Ying-Huan, Gao Yang, et al. A prototype learning based multi-instance convolutional neural network. Chinese Journal of Computers, 2017, 40(6): 1265-1274(in Chinese)
(何克磊，史颖欢，高阳等. 一种基于原型学习的多示例卷积神经网络. 计算机学报，2017，40(6)：1265-1274)

[20] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets. arXiv preprint arXiv: 1507.02159, 2015

[21] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1933-1941

[22] Hou J, Wu X, Yu F, et al. Multimedia event detection via deep spatial-temporal neural networks//Proceedings of the IEEE International Conference on Multimedia and Expo. Seattle, USA, 2016: 1-6

[23] Gan C, Wang N, Yang Y, et al. DevNet: A deep event network for multimedia event detection and evidence recounting//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2568-2577

[24] Wang Meng-Lai, Li Xiang, Chen Qi, et al. Surveillance event detection based on CNN. Acta Automatica Sinica, 2016, 42(6): 892-903(in Chinese)
(王梦来，李想，陈奇等. 基于 CNN 的监控视频事件检测. 自动化学报，2016，42(6)：892-903)

[25] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition//Proceedings of the International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 3468-3476

[26] Feichtenhofer C, Pinz A, Wildes R P. Temporal residual networks for dynamic scene recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7435-7444

[27] Huang G, Liu Z, Weinberger K Q. Densely connected convolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2261-2269

[28] Adi Y, Keshet J, Cibelli E, et al. Sequence segmentation using joint RNN and structured prediction models//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA, 2017: 2422-2426

[29] Woo S, Byun J, Kim S, et al. RNN-based personalized activity recognition in multi-person environment using RFID//Proceedings of the IEEE International Conference on Computer and Information Technology. Helsinki, Finland, 2017: 708-715

[30] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[31] Gammulle H, Denman S, Sridharan S, et al. Two stream LSTM: A deep fusion framework for human action recognition //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Santa Rosa, USA, 2017: 177-186

[32] Zhao Y, Sun T, Jiang X, et al. Long-term residual recurrent network for human interaction recognition in videos//Proceedings of the International Congress on Image and Signal Processing, Biomedical Engineering and Informatics. Datong, China, 2016: 78-83

[33] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition //Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 20-36

[34] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681

[35] Graves A. Generating sequences with recurrent neural networks. arXiv preprint arXiv: 1308.0850, 2013

[36] Sun Y, Wang X, Tang X. Deep learning face representation by joint identification-verification//Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 1988-1996

[37] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 815-823

[38] Oh S, Hoogs A, Perera A, et al. A large-scale benchmark dataset for event recognition in surveillance video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3153-3160

[39] Amer M R, Todorovic S. Sum-product networks for modeling activities with stochastic structure//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1314-1321

[40] Gaur U, Zhu Y, Song B, et al. A "string of feature graphs" model for recognition of complex activities in natural videos//Proceedings of the International Conference on Computer Vision. Barcelona, Spain, 2011: 2595-2602

[41] Wang X, Ji Q. Hierarchical context modeling for video event recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1770-1782

LI Yong-Gang, born in 1979, Ph. D. candidate, lecturer. His research interests include computer vision, image and video processing and pattern recognition.

WANG Zhao-Hui, born in 1967, M. S., associate professor. Her research interests include pattern recognition and image processing.

WAN Xiao-Yi, born in 1993, M. S. candidate. Her research interest is 3D action recognition.

DONG Hu-Sheng, born in 1981, Ph. D. candidate, lecturer. His research interests include computer vision and machine learning.

GONG Sheng-Rong, born in 1966, Ph. D., professor, Ph. D. supervisor. His research interests include image and video processing, pattern recognition and computer vision.

LIU Chun-Ping, born in 1971, Ph. D., professor, Ph. D. supervisor. Her research interests include computer vision, image analysis and recognition, in particular in the domains of visual saliency detection, object detection and recognition and scene understanding.

JI Yi, born in 1973, Ph. D., associate professor. Her research areas are 3D action recognition and complex scene understanding.

ZHU Rong, born in 1973, Ph. D., professor. Her research interests include intelligent information processing, machine learning and data mining.

Background

Video monitoring is widely used in our lives. Nonetheless, important information extracted from the surveillance videos is mainly by hand-crafted analysis, which leads to high cost and low efficiency. Event recognition in surveillance video is attracting growing interest in recent years. Nevertheless, event recognition in real-world surveillance video still faces great challenges due to various facets such as cluttered background, severe occlusion, etc.

Recently, a pronounced tendency is that more researches focus on learning deep features from raw data. Two-stream CNNs architecture becomes a very successful model in video analysis field, in which appearance features and short-term motion features are utilized. LSTM network can learn long-term motion features from the input sequence, which is widely used to process those tasks with quintessential time series. Whatever CNN or LSTM, they will meet vanishing gradient problem with the depth's increasing. Accordingly, training a very deep network is very difficult. Deep residual network provides a good solution by using residual blocks, which makes it possible to train network with hundreds even more than a thousand of layers, and the performance is still excellent. The design of shortcut ingeniously solves the vanishing gradient problem. We combine the merits of several networks aforementioned, and propose a deep residual dual unidirectional DLSTM for video event recognition in surveillance video.

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61773272, 61170124, 61272258, 61301299), the Integration of Cloud Computing and Big Data, Innovation of Science and Education (Grant No. 2017B03112), the Provincial Natural Science Foundation of Jiangsu, China (Grant Nos. BK20151260, BK20151254), the Six Talent Peaks Project in Jiangsu Province, China (Grant No. DZXX-027), the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (Grant No. 93K172016K08), the Provincial Natural Science Foundation of Zhejiang, China (Grant No. LY15F020039), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX17_2006).