基于卷积神经网络与长短期记忆神经网络的多特征融合人体行为识别算法

黄友文，万超伦*，冯恒
江西理工大学信息工程学院，江西 赣州 341000

摘要 提出了一种基于卷积神经网络和长短期记忆(LSTM)神经网络的深度学习网络结构。采用特征融合的方法，通过卷积网络提取出浅层特征与深层特征并进行联接，对特征通过卷积进行融合，将获得的矢量信息输入 LSTM 单元。分别使用数据光流信息与红绿蓝信息训练网络，将各网络的结果进行加权融合。实验结果表明，所提模型有效地提高了行为识别精度。

关键词 机器视觉；深度学习；行为识别；卷积神经网络；长短期记忆神经网络
中图分类号 TP183 **文献标识码** A **doi**：10.3788/LOP56.071505

Multi-Feature Fusion Human Behavior Recognition Algorithm Based on Convolutional Neural Network and Long Short Term Memory Neural Network

Huang Youwen, Wan Chaolun*, Feng Heng
School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, *China*

Abstract A deep learning network structure based on the convolutional neural network and long short term memory (LSTM) neural network is proposed. The feature fusion is used to extract the shallow features and deep features through the convolutional network, and the features are fused by convolution, and the the obtained vector information is input into the LSTM unit. Networks are trained separately using the optical flow images and the red green blue information, and the results from each network are fused with weights. The experimental results show that the proposed model effectively improves the accuracy of behavior recognition.

Key words machine vision; deep learning; behavior recognition; convolutional neural network; long short term memory; neural network
OCIS codes 150.1135; 150.0155; 150.4620

1 引 言

人机交互、行为检测等技术的广泛运用使得人体行为识别成为了热点领域。人体行为识别的关键在于人类行为特征的提取，早期提取特征的方式通常是对手动设计特定的特征进行提取描述，典型的有方向梯度直方图(HOG)、光流方向信息直方图(HOF)[1]等方法。文献[2]将深度图序列转换为三维(3D)点云序列并获得方向特征，再进行多层时域重叠分割，获得时空方向主成分直方图(HSTOPC)特征描述子，最后采用支持向量机进行训练。文献[3]提出的一种将稠密光流轨迹与稀疏编码框架融合的无监督行为特征提取方法，将处理得到的行为特征送入支持向量机中得到模型进行分类；文献[4]提出了基于局部和全局特征视觉单词的方法，提高人物区域内兴趣点，再结合局部 3D 图像配准算法(3D-SIFT)与全局光流方向直方图(HOOF)特征去描述人物行为，最后通过谱聚类生成视觉单词输入

收稿日期：2018-09-21；**修回日期**：2018-10-22；**录用日期**：2018-10-30
基金项目：江西省教育厅科技项目(GJJ150683)、江西理工大学校级重点课题(NSFJ2014-K18)
* **E-mail**：353382420@qq.com

Topic Model by Belief Propagation(TMBP)模型进行人体行为识别。然而,传统方法得到的特征在面对复杂光照、复杂背景的现实情况下往往难以获得好的识别效果[5]。

近几年,随着深度学习模型的发展,深度学习模型提取特征的过程取代了原本的人工设计特征的过程,这也消除了人工设计过程中的盲目性和差异性,实现了特征的自动提取。深度学习模型之一,即卷积神经网络(CNN),通过对输入数据的卷积与池化操作,从浅层特征开始,逐层提取出深层特征,其在图像识别领域已经取得了优良的效果。2012 年的 AlexNet 网络[6]取得了 ImageNet Large Scale Visual Recognition Competition(ILSVRC)比赛的冠军,其将 ImageNet 数据集上的 Top-5 错误率降低到 16.4%; 2016 年,以通道分离式卷积为基本思想的 Xception 网络[7]将 Top-1 准确率提高到了 79.0%;次年的 MobileNets 网络[8]将关注重点放到了在压缩模型的基础上同时保证精度不降。

针对人体行为识别问题,视频中连续帧之间具有时间关联性,深度学习中通常采用递归神经网络(RNN)模型来处理此类问题。RNN 是一个包含循环的网络,可以被看作对同一神经网络的多次赋值,允许了信息的持久化。但 RNN 存在着梯度消失的问题,为了解决这个问题,Hochreiter 等[9-10]在传统 RNN 的基础上引入存储单元,改进为长短期记忆(LSTM)递归神经网络,通过刻意的设计避免了长期依赖问题的发生。Donahue 等[11]初次将 CNN 与 LSTM 进行结合,提出了长效递归卷积神经网络(LRCNN),运用在了视频识别、视频描述等领域。除此之外,文献[12]设计了一种时间段网络,将视频数据提取出的稠密光流与红绿蓝(RGB)帧数据分别作为 CNN 的输入得到模型,进行结果融合;文献[13]将数据通过一组硬连接内核将 7 帧的数据处理得到 5 个信道后,利用 3D 卷积网络进行训练提取时域信息;文献[14]将 3D 卷积处理得到的含有时间信息的特征送入 LSTM 进行处理,再用于人体行为识别。

本文提出了一种 CNN 与 LSTM 结合的网络结构,在 CNN 中利用 Concat 层将浅层特征与深层特征进行融合,再利用卷积层完成对图像序列特征的提取与矢量化,获得相应维度的特征序列后输入双层的 LSTM 神经网络中进行处理。视频数据提取出的光流与 RGB 帧数据分别被输入网络中进行训练,最后将各自训练得到的分类结果进行加权融合,得到最终的分类结果,用于人体行为识别。

2 模型架构

2.1 总体框架

提取出视频数据的光流与 RGB 帧数据之后,将数据集分为训练集与测试集,对训练集数据采用镜像、裁剪、尺度抖动等操作进行数据增强预处理,之后训练集数据被用于模型的构建与参数训练调整,训练完毕后,用测试集数据去验证模型的性能。系统框图如图 1 所示。

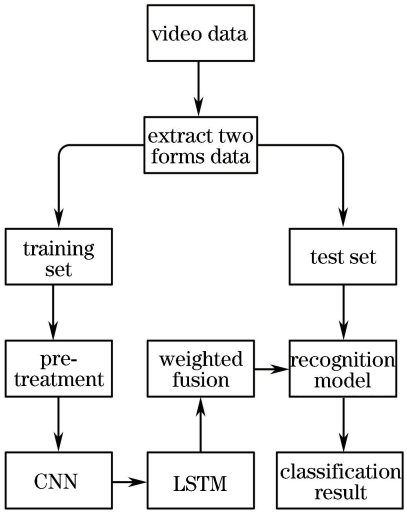

图 1 系统框图

Fig. 1 System block diagram

2.2 CNN

本文 CNN 由卷积层、池化层、局部响应归一化(LRN)层和 Concat 层堆叠而成。卷积层利用多个卷积核提取图片特征,生成特征图组;池化层进行下采样,合并特征缩减维度;LRN 层被用来对局部神经元的活动创建竞争机制,放大响应较大的值,抑制反馈较小的神经元,增强泛化能力[6];Concat 层的加入可有效降低过拟合的风险[15],相较直接依赖网络最后一层复杂度最高的特征,本文网络将浅层特征与深层特征进行 Concat 融合,使其更容易得到一个光滑的、具有更好泛化能力的结果。Concat 层的输出为

$$x_l = H_l[(x_0, x_1, \ldots, x_{l-1})], \tag{1}$$

式中:H_l 表示一个非线性变换,l 表示第 l 个卷积层;$(x_0, x_1, \ldots, x_{l-1})$表示多个特征图集合。

2.3 长短期记忆神经网络

RNN 被用来处理具有时间相关性的视频数据,时间序列的历史信息被存储在网络隐藏层之中[11],前向公式可表示为

$$h_t=\mathrm{sigmoid}(W_{\mathrm{xh}}x_t+W_{\mathrm{hh}}h_{t-1}+b_{\mathrm{h}}),\quad(2)$$

$$z_t=\mathrm{sigmoid}(W_{\mathrm{hz}}h_t+b_{\mathrm{z}}),\quad(3)$$

式中：sigmoid 表示 S 型函数，即神经网路中的阈值函数；t 代表某一确定时刻点；h_t 代表 RNN 隐藏层的输出；x_t 为当前时刻的输入；h_{t-1} 为上一隐藏层的输出；z_t 为 RNN 网络当前时刻的输出；W_{xh}、W_{hh}、W_{hz} 与 b_{h}、b_{z} 分别表示加权项与偏置项。

在 RNN 的基础上，LSTM 被设计，信息传入 LSTM 之后的第一步会经过遗忘门的结构，遗忘门公式为

$$f_t=\mathrm{sigmoid}[W_{\mathrm{f}}\cdot(h_{t-1},x_t)+b_{\mathrm{f}}],\quad(4)$$

式中：h_{t-1} 代表前一单元的输出；x_t 表示当前时刻单元的输入；f_t 代表遗忘层的输出；W_{f} 与 b_{f} 分别表示加权项与偏置项。

第二步，确认更新的信息，即

$$i_t=\mathrm{sigmoid}[W_{\mathrm{i}}\cdot(h_{t-1},x_t)+b_{\mathrm{i}}],\quad(5)$$

$$\tilde{C}_t=\tanh[W_{\mathrm{c}}\cdot(h_{t-1},x_t)+b_{\mathrm{c}}],\quad(6)$$

式中：i_t 与 $\tilde{C}_t$ 被用来确认更新状态并加入到更新单元中去；C_{t-1} 为更新前的单元；W_{i}、W_{c} 与 b_{i}、b_{c} 分别表示加权项与偏置项。

得到更新后的单元状态为

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tilde{C}_t,\quad(7)$$

式中 C_t 即为更新后的单元。

最后，决定给下一个隐藏层的输出内容，即

$$o_t=\mathrm{sigmoid}[W_{\mathrm{o}}(h_{t-1},x_t)+b_{\mathrm{o}}],\quad(8)$$

$$h_t=o_t\cdot\tanh(C_t),\quad(9)$$

式中：o_t 作为中间项被用来与 C_t 得到输出项 h_t；W_{o} 与 b_{o} 分别表示加权项与偏置项。

2.4 融合模型

所采用的 CNN 结构参考 Alex 等提出的经典 CNN 框架 AlexNet，改变了卷积核数量与大小，加入了 Concat 层进行特征融合，同时删去了最后的三个全连接层，转而加入一个卷积层以进行特征提取与特征图的矢量化。

将提取后的长段视频实验数据随机剪辑为 25 帧的视频序列，对单帧图片，将尺寸扩充为 227 pixel×227 pixel，以加载预先在 ImageNet 数据集下训练好的预训练模型参数。输入卷积网络的数据规模为 25×227 pixel×227 pixel×3，其中 25 为输入视频序列长度，227 pixel×227 pixel 为单张图片大小，3 为图片的三个通道。融合模型的 CNN 部分具体参数如图 2 所示。

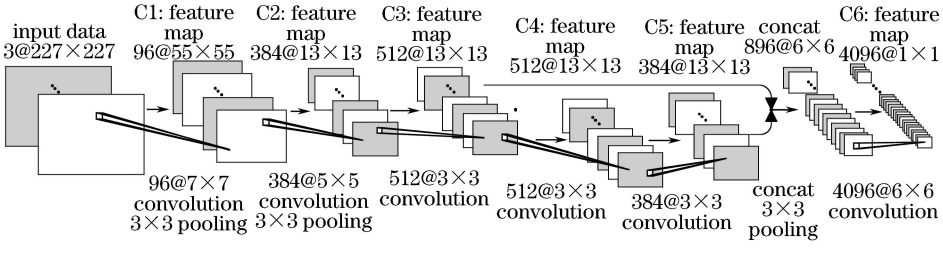

图 2 CNN 及各层参数

Fig. 2 CNN and parameters of each layer

图 2 中连接线下方表示神经网络操作层的维度与功能，特征图上方表示经过操作后得到的特征图的维度大小。人体行为识别 CNN 部分的模型一共有 6 个卷积层，3 个池化层。第 1 个卷积层的尺寸为 7×7，步长为 2；第 2 个卷积层的尺寸为 5×5，步长为 2；第 3～5 个卷积层的尺寸为 3×3，步长为 1；第 6 个卷积层的尺寸为 6×6，步长为 1；池化层的尺寸都为 3×3，步长为 2。为了保证模型的非线性，每个卷积层之后都添加一个非线性激活函数 Relu，同时为了提高模型的泛化能力并加速训练，前两个池化层后也都添加了一个 LRN 层。C3 浅层特征图与 C5 深层特征图将会通过 Concat 层进行特征融合操作，再输入最大池化层进行降维。如图 3 所示，在模型的最后，由于本文删去了全连接层，因此得到

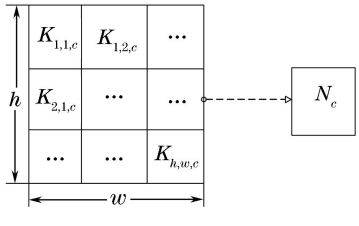

图 3 CNN 隐藏层维度与矢量化

Fig. 3 CNN hidden layer dimension and vectorization

的 CNN 模型的隐藏层输出特征维度是长(H)×宽(W)×特征图数量(C)，是一个立方体特征，需要将其化为矢量特征后才能输入 LSTM 网络中。

图 3 中 K 表示像素点，c 表示特征图数量，h 与 w 分别表示特征图的长与宽。

直接将 $H\times W$ 个 C 维向量展开成为矢量特征，

特征维度过高，会影响分类的效率，也容易过拟合。因此，本文采用卷积的方法来进行特征的矢量化，利用 CNN 的第 6 个卷积层，具有 $H\times W$ 维度大小的卷积核、步长为 1，使特征的维度变为 $1\times1\times C$。在进行特征矢量化的同时，也对之前融合得到的特征做了进一步的特征提取。

之后，数据被调整规模后输入 LSTM 中，LSTM 按时序做递归运算，每次递归运算的结果是前面所有特征和当前特征的融合。图 4 为本文 LSTM 模型，包含双层 LSTM 结构。

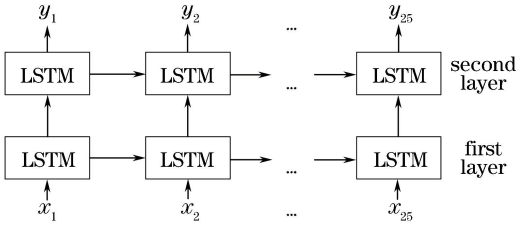

图 4 所提 LSTM 模型示意图

Fig. 4 Schematic of proposed LSTM model

融合后的模型网络结构如图 5 所示，采用文献[16]提出的光流提取方法将水平方向光流值与垂直方向的光流值尺度缩放至[−128，+128]，再利用水平方向光流值、垂直方向光流值与光流模值构建三通道的光流作为光流数据，与 RGB 帧数据共同进行实验。将提取出的长段视频图片数据经扩充等预处理后，随机剪辑为多帧的序列长度，再将剪辑好的两种数据流数据都以三通道的形式输入各自的 CNN 中，得到的矢量化特征输入双层的 LSTM 神经网络中进行处理，通过各自的 Softmax 分类器将多帧的各帧训练得到的预测标签进行概率平均处理，得到 RGB 帧数据与光流数据的视频段在各自网络结构中训练得出的分类结果。最后将两种数据流在各自网络中得到的两个分类结果进行加权融合，用于最终的人体行为识别任务。

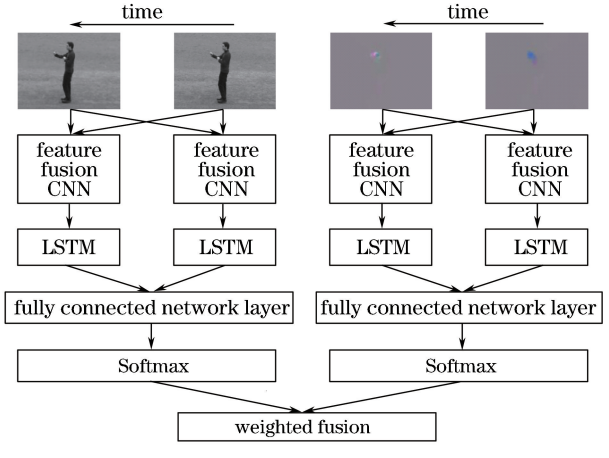

图 5 模型的网络结构

Fig. 5 Network structure of model

3 实验过程

3.1 数据集

使用 Kungl tekniska högskolan (KTH)数据集与 University of Central Florida Sports (UCF Sports)数据集作为实验的测试数据。KTH 数据集由固定的摄像机拍摄采集的 600 个动作视频组成。视频中每帧图片的像素大小为 160 pixel × 120 pixel，每个视频大约 360 帧，帧率为 25 frame /s。共 25 个实验对象，包含 4 个场景，分别为户外、户外(远近尺度变化)、户外(衣着变化)、室内；6 种行为，分别为散步、慢跑、奔跑、挥手、拍手、拳击。UCF Sports 数据集包含 10 类动作，分别为跳水、高尔夫秋千、踢、吊装、骑马、跑步、滑板、秋千长椅、摆动侧、步行。

本文将两个数据集分别以动作进行划分，KTH 数据集每个动作的 80%作为训练集，20%作为测试集；UCF Sports 数据集每个动作的 70%作为训练集，30%作为测试集。本节以 KTH 数据集为例，视频数据的 RGB 帧与光流被分别提取出，经预处理后，再将长段的视频数据随机剪辑为 25 帧的较短视频段输入进行训练，用以数据增强。每张图片的大小被扩充成了 227 pixel ×227 pixel，在保持特征不损失的情况下，去加载适配预训练模型，防止过拟合[17]。

3.2 训练设置

实验利用具有 python 接口的深度学习库 caffe 在图形处理器(GPU)加速环境下进行实验。实验环境如下：Intel i3-7100 CPU 3.90 GHz；NVIDIA GeForce GTX 1070 (1920 个 CUDA 处理核心)；8 GB内存；Ubuntu 16.04 64 位。

在训练中，为增强模型的稳健性，加载了网络预先在 ImageNet 数据集下训练 30 万次得到的预训练模型参数。在实验开始前，预先切断了前 5 个卷积层的反向传播计算，单独训练融合模型中 CNN 添加的层与双层的 LSTM 网络参数。

之后，再打开之前切断反向传播计算的网络层，以 10^{-6} 的学习率进行微调，优化参数。图 6 为训练过程中，随着训练次数的增加，光流数据和 RGB 帧数据对数据集识别率的变化情况。

3.3 融合权重选择

得到光流与 RGB 帧数据分别训练出的分类模型

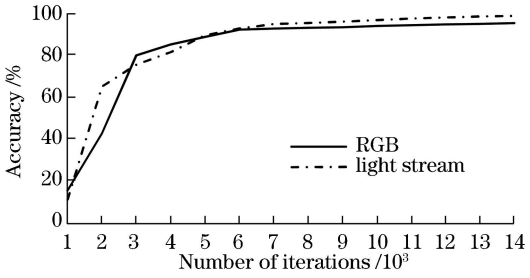

图 6 KTH 数据训练准确率

Fig. 6 KTH data training accuracy

之后,将两个模型的分类结果进行加权融合。如图 7 所示,横轴代表 RGB 帧数据模型占比,纵轴代表对数据集视频的识别率。这里 RGB 帧数据占比以 0.05 的步长进行递增取点,并逐步提高 RGB 帧数据模型的权重占比。可以看出,当 RGB 帧与光流的权重之比为 0.35∶0.65 时,整个模型达到最好的识别效果。

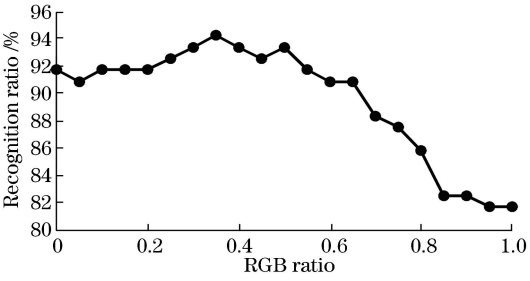

图 7 不同权重比下的识别率

Fig. 7 Recognition ratio of different weight ratios

4 实验结果对比与分析

如表 1 所示,经过训练后,选取得到的识别率最高的权重比例(RGB 帧与光流的权重之比为0.35∶0.65)的算法模型在 KTH 数据集测试集上与已有的算法模型和未采用特征融合的网络进行了对比。

从表 1 可以看出,在 KTH 数据集中,所提基于 LSTM 与 CNN 结合的多特征融合网络结构模型在以 RGB 帧与光流的权重比为 0.35∶0.65 进行加权融合之后,可以得到优于文献[4]与[3]方法的结果;相比文献[13]与[14]的基于深度学习的方法,本文模型在测试集上可以得到更好的效果。同时,对去除特征融合的网络结果进行了实验,识别精度为 93.3%,实验结果显示,采用特征融合的方法将精确度提升了 0.9%。

表 1 不同算法在 KTH 数据集上的准确率对比

Table 1 Comparison of accuracy of different algorithms on KTH dataset

Algorithm	Accuracy /%
Method in Ref. [4]	93.7
Method in Ref. [3]	90.2
Method in Ref. [13]	90.2
Method in Ref. [14]	93.7
Model without concat	93.3
Proposed method	94.2

如表 2 所示,依然选取了 RGB 帧与光流的权重比为 0.35∶0.65 的算法模型在 UCF Sports 数据集上与已有的算法模型进行了对比。

表 2 不同算法在 UCF Sports 数据集上的准确率对比

Table 2 Comparison of accuracy of different algorithms on UCF Sports dataset

Algorithm	Accuracy /%
Method in Ref. [18]	88.0
Method in Ref. [19]	88.4
Proposed method	88.89

由表 2 可以看出,在 UCF Sports 数据集上,本文算法准确率依然有所提升,证明了其可行性。

表 3 通过混淆矩阵,将模型对 KTH 测试集中 6 种行为的识别结果进行可视化,用以观测模型的表现,对角线元素表示正确识别率。可以看出,就 KTH 数据集而言,慢跑和奔跑行为的识别错误情况较多,而拍手、挥手、拳击、散步的识别率较高。通过观察原始视频可以看出,慢跑和奔跑的区分度并不大,数据本身有着很高的相似性。因此,模型依旧具有良好的泛化能力与稳健性。

5 结 论

提出了一种基于多特征融合的人体行为识别深

表 3 基于 KTH 数据集的行为识别混淆矩阵

Table 3 Behavior recognition confusion matrix based on KTH dataset

Actual classification	Model prediction					
	Walking	Jogging	Running	Boxing	Hand waving	Hand clapping
Walking	0.95	0.05	0.00	0.00	0.00	0.00
Jogging	0.10	0.85	0.05	0.00	0.00	0.00
Running	0.00	0.10	0.90	0.00	0.00	0.00
Boxing	0.00	0.00	0.00	1.00	0.00	0.00
Hand waving	0.00	0.00	0.00	0.00	1.00	0.00
Hand clapping	0.00	0.00	0.00	0.00	0.05	0.95

度学习模型，采用 CNN 和 LSTM 神经网络相结合的结构。在 CNN 部分，将浅层特征与深层特征进行融合，再将矢量化后的融合特征送入双层的 LSTM 神经网络进行训练，最后将光流与 RGB 帧数据分别训练得到的结果进行加权融合。所提模型在 KTH 数据集的测试集上识别率达到了 94.2%，在 UCF Sports 数据集测试集上的准确率达到了 88.89%。相比于已有方法，所提模型能更好地提取视频每帧的特征和帧与帧之间的时序特征，识别率较好。整个模型基于 CNN，无需先验经验，因此具有良好的泛化性。同时该算法依旧存在着一些不足，如未采用更新颖的方式去提取视频数据的光流形式等，将在以后的工作中进行改进。

参 考 文 献

[1] Laptev I, Marszalek M, Schmid C, *et al*. Learning realistic human actions from movies [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[2] Xu H Y, Kong J, Jiang M, *et al*. Action recognition based on histogram of spatio-temporal oriented principal components [J]. Laser & Optoelectronics Progress, 2018, 55(6): 061009.
徐海洋, 孔军, 蒋敏, 等. 基于时空方向主成分直方图的人体行为识别[J]. 激光与光电子学进展, 2018, 55(6): 061009.

[3] Zhao X J, Zeng X Q. Action recognition method based on dense optical flow trajectory and sparse coding algorithm [J]. Journal of Computer Applications, 2016, 36(1): 181-187.
赵晓健, 曾晓勤. 基于稠密光流轨迹和稀疏编码算法的行为识别方法[J]. 计算机应用, 2016, 36(1): 181-187.

[4] Xie F, Gong S R, Liu C P, *et al*. Human action recognition by visual word based on local and global features[J]. Computer Science, 2015, 42(11): 293-298.
谢飞, 龚声蓉, 刘纯平, 等. 基于局部和全局特征视觉单词的人物行为识别[J]. 计算机科学, 2015, 42(11): 293-298.

[5] Luo H L, Wang C J, Lu F. Survey of video behavior recognition[J]. Journal on Communications, 2018, 39(6): 169-180.
罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169-180.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[7] Chollet F. Xception: deep learning with depthwise separable convolutions [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1800-1807.

[8] Howard A G, Zhu M, Chen B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications [J]. arXiv preprint arXiv: 1704.04861, 2017.

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

[10] Cai M, Liu J. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition[J]. Speech Communication, 2016, 77: 53-64.

[11] Donahue J, Hendricks L A, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2625-2634.

[12] Wang L M, Xiong Y J, Wang Z, *et al*. Temporal segment networks: towards good practices for deep action recognition[M]. Cham: Springer International Publishing, 2016: 20-36.

[13] Ji S W, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[14] Qin Y, Mo L F, Guo W K, *et al*. Combination of 3D CNNs and LSTMs and its application in activity recognition [J]. Measurement & Control Technology, 2017, 36(2): 28-32.
秦阳, 莫凌飞, 郭文科, 等. 3D CNNs 与 LSTMs 在行为识别中的组合及其应用[J]. 测控技术, 2017, 36(2): 28-32.

[15] Huang G, Liu Z, Maaten L V D, *et al*. Densely connected convolutional networks [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2261-2269.

[16] Brox T, Bruhn A, Papenberg N, *et al*. High accuracy optical flow estimation based on a theory for warping [C] // European Conference on Computer Vision, 2004: 25-36.

[17] Qu L, Wang K R, Chen L L, *et al*. Fast road detection based on RGBD images and convolutional neural network [J]. Acta Optica Sinica, 2017, 37(10): 101003.

曲磊, 王康如, 陈利利, 等. 基于RGBD图像和卷积神经网络的快速道路检测[J]. 光学学报, 2017, 37(10): 101003.

[18] Yang Y, Saleemi I, Shah M. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(7): 1635-1648.

[19] Lu T R, Yu F Q, Yang H Z, *et al*. Human action recognition based on dense trajectories with saliency detection [J]. Computer Engineering and Applications, 2018, 54(14): 163-167.
鹿天然, 于凤芹, 杨慧中, 等. 基于显著性检测和稠密轨迹的人体行为识别[J]. 计算机工程与应用, 2018, 54(14): 163-167.