基于深度网络模型的视频序列中异常行为的检测方法

吴沛佶*, 梅雪, 何毅, 袁申强
南京工业大学电气工程与控制科学学院, 江苏 南京 211816

摘要 针对视频序列中的几种异常行为,构建训练模型,对其进行识别。使用卷积神经网络(CNN)进行特征提取并采用 Adam 算法(一种基于梯度的优化算法)进行优化。引入自适应池化层,筛选出判别的特征信息,减轻网络的计算量,加快识别视频序列中存在的异常行为。使用 Adam 算法对模型进行优化后,识别率可以达到 87.6%,引入自适应池化层后,识别率可以达到 91.9%。该卷积神经网络对视频序列中基本的异常行为的检测效果比改进的轨迹跟踪(iDT)和双流网络更快更准确;相较于时间分割网络(TSN)和时间关系网络(TRN),识别的准确率稍低,但是识别的速度更快。

关键词 成像系统; 深度学习; 卷积神经网络; 异常行为
中图分类号 TP751 **文献标识码** A **doi**: 10.3788/LOP56.131101

Method of Detecting Abnormal Behavior in Video Sequences Based on Deep Network Models

Wu Peiji*, Mei Xue, He Yi, Yuan Shenqiang
College of Electrical Engineering and Control Science, Nanjing Tech University, Nanjing, Jiangsu 211816, *China*

Abstract In this study, a training model was constructed to identify several abnormal behaviors in video sequences. A convolutional neural network (CNN) was used to extract features, and the features were then optimized using a gradient-based optimization algorithm known as Adam algorithm. The adaptive pooling layer was introduced for feature discrimination to reduce the computational complexity of the network and rapidly identify abnormal behaviors in video sequences. The recognition rate reaches 87.6% after using the Adam algorithm for model optimization. The recognition rate reaches 91.9% when the adaptive pooling layer is introduced. CNN is faster and more accurate than the improved dense trajectories and the two-stream networks in detecting abnormal behaviors in video sequences. Compared with the temporal segment networks and temporal relation networks, the CNN has a lower accuracy but a faster speed.

Key words imaging systems; deep learning; convolutional neural networks; abnormal behavior
OCIS codes 110.0100; 100.3008

1 引 言

异常行为识别在机器视觉和模式识别领域中备受关注,具有广泛的应用价值,如医疗辅助、家居养老等[1]。但是视频中的光线变化及背景的灵活多变、镜头的晃动、待识别对象的遮挡等都会给行为识别带来很大的困难[2]。在深度学习出现之前,有 3 种常见的行为识别方法:参数建模、视频立方体分析和模板匹配[3]。参数建模是建立一个用于表述视频中行为的模型,如隐马尔可夫模型[4]、贝叶斯网络[5]等。视频立方体分析[6]是把待测视频当作一个含有三维时空信息的立方体,并对这个立方体进行一系列的研究。模板匹配[7]是提取视频数据的特征,将之与已经定义的特征模板相匹配后进行识别。上述方法均难以客观地实现行为特征的提取。深度学习在特征提取过程中减少了人为参与,避免了人工选择的主观性和随意性,提高了提取的精确度[8]。

深度学习推动了人工智能的发展,在学术界掀

收稿日期: 2018-12-13; **修回日期**: 2019-01-02; **录用日期**: 2019-01-22
* **E-mail**: 1434519290@qq.com

起了探索和研究的热潮。相较于传统方法，深度学习算法可以有效提高识别精度，基于深度学习的人脸识别可以极大提高识别结果的准确率[9]，实现根据人脸进行性别判断[10]；深度学习也可以实现手写字的辨别[11]以及文本的识别与理解[12-14]，能够有效提取那些因信息量大[15-16]、解析度低[17]而导致特征提取不准确的样本图像的特征。深度学习在行人检测[18]、手势识别[19]、语音识别[20]等领域也有广泛的应用。

在深度学习中，卷积神经网络(CNN)是一种常用的特征提取方法。通过卷积层对视频序列进行特征提取可以避免因人为因素导致特征提取不准确的问题；通过网络的池化层，改善对复杂高维数据的特征提取效果；通过共享卷积核的权值可以直接将未经过预处理的原始视频序列作为 CNN 的输入，有效地提高输入数据特征提取的精确度，提高深度网络的学习性能。基于 CNN 的识别方法大致有以下几种：基于单帧的识别方法——截取视频中的关键帧，然后基于每一帧进行深度学习表达，将截取自视频中的帧输入网络即可获得相应的识别结果；基于 CNN 扩展网络的识别方法——通过在 CNN 框架中找到时间域上的信息来描述局部的动态信息，最终实现整体识别效果的提升；基于双路 CNN 的识别方法——双路 CNN 就是 2 个 CNN，最终的识别结果就是对 2 个 CNN 得到的结果取平均值，其中一个 CNN 是基于单帧的 CNN，另一个则是把一个序列中连续帧的光流叠加后输入；基于长短期记忆网络(LSTM)的识别方法——通过 LSTM 在时间轴上对 CNN 的全连接层进行整合，其好处有 2 个：1) 有充分的时间对 CNN 提取的特征进行融合，2) LSTM可以帮助分辨出视频帧在原始视频中的顺序；三维卷积核(3D CNN)法——将原始视频序列堆砌成一个立方体，扩展卷积核，从相邻帧中得到每一个特征。

本文通过在 CNN 网络中引入自适应池化层，提高了最终的识别结果。

2 引入自适应的卷积神经网络

2.1 CNN 结构

通过卷积层对输入的图片样本进行特征提取，经过自适应池化层对卷积层提取的特征进行特征压缩，降低计算量并提取主要特征，再由全连接层上的 softmax 分类器进行识别，最后输出分类结果。本文使用的网络在网络的池化层中添加了一个特征筛选的依据，将所有特征分为判别的特征和非判别的特征，如图 1 所示。判别的特征即是需要识别的行为正相关的特征，它能帮助池化层更准确地筛选出主要特征，进而提高最终的识别精度。

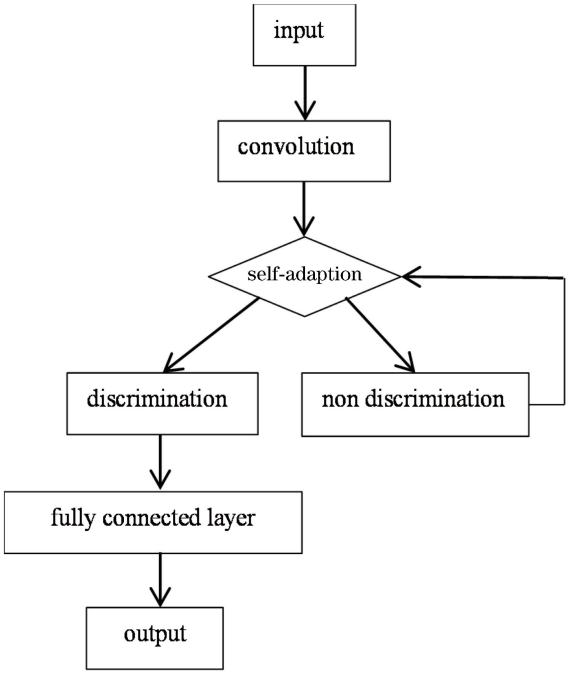

图 1 CNN 流程图

Fig. 1 CNN flow chart

2.2 自适应池化层

自适应池化层是该方法的关键模块，它可以在给定特征向量和已经合并的特征向量的情况下推断当前特征向量的重要性。如果它包含与已定义的正常行为或者异常行为正相关的特征，并且可能与其他行为负相关，那么它就是判别的特征信息；如果是冗余的特征，那么它就是非判别的特征信息。

将自适应合并向量 $\boldsymbol{X}$ 的前 s 个元素表示为 $\psi(\boldsymbol{X},s)$。自适应池化层通过递归计算 2 个操作来实现池化。第 1 个操作，表示为 f_{imp}，预测判别重要性，重要性分数 $\gamma_{s+1}\in[0,1]$，其中第$(s+1)$个元素给出其 CNN 特征，即 $\phi(x_{s+1})$，汇总特征直到第 s 个元素，得到 $\psi(\boldsymbol{X},s)$。将重要性分数表示为实数序列 $\tau=\{\gamma_1,\cdots,\gamma_s\}\in[0,1]$。第 2 个操作是加权平均合并操作，其通过将先前合并的特征与来自当前的特征及其预测的重要性聚合来计算新的合并特征 $\psi(\boldsymbol{X},s+1)$，公式为

$$\gamma_{s+1}=f_{\mathrm{imp}}[\psi(\boldsymbol{X},s),\phi(x_{s+1})], \tag{1}$$

$$\psi(\boldsymbol{X},s+1)=\frac{1}{\hat{\gamma}_{s+1}}[\hat{\gamma}_s\psi(\boldsymbol{X},s)+\gamma_{s+1}\phi(x_{s+1})], \tag{2}$$

$$\hat{\gamma}_p = \sum_{k=1}^{p} \gamma_k , \tag{3}$$

式中：$\hat{\gamma}_p$ 为前 p 个特征的重要性之和；$\hat{\gamma}_s$ 为前 s 个特征的重要性之和；k 为序号。

使用标准交叉熵损失 l_{loss} 来制定损失函数，并添加基于熵的正则化器 l_{c}，如

$$l(\boldsymbol{X}, y) = \lambda l_{\text{c}}(\tau) + l_{\text{loss}}(\boldsymbol{X}, y), \tag{4}$$

$$l_{\text{c}}(\tau) = -\sum_{k} \frac{\exp \gamma_k}{N} \text{lb}\left(\frac{\exp \gamma_k}{N}\right), \tag{5}$$

式中：y 为期望输出；λ 为折中参数；N 为前 s 个特征重要性分数的 e 指数之和，即

$$N = \sum_{s} \exp \gamma_s , \text{ when } \gamma_k \geqslant 0, \lambda \geqslant 0。 \tag{6}$$

正则化器使用 softmax 最小化判别分数的熵，这样有助于选择判别的特征，并丢弃非判别的特征。参数 λ 可以平衡稀疏帧的选择和最小化交叉熵分类损失项。由于期望选择较少数量的特征，若将 λ 设置为相对较高的值，则分类任务很困难；若 λ 的值相对较低，则期望模型有可能过拟合。实验部分展示了不同的 λ 取值带来的影响。

3 仿真与实验

3.1 样本选择

选取法国国家信息与自动化研究所（INRIA）圣诞动作采集序列（IXMAS）作为训练模型的样本。由 11 个人分别表示站立、交叉手臂、挠头、坐下、转身、步行、拳、踢等动作。该数据库从 5 个视角获得，室内 4 个方向和头顶一共安装 5 个摄像头。将拳、踢等行为标记为异常行为，共有 7000 张图片，其中有 6000 张用于训练，1000 张用于测试；其他行为标记为正常行为，共有 12000 张图片，其中有 10000 张用于训练，2000 张用于测试。在 Caffe 框架下使用 CNN 对该数据集进行二分类。图 2 与图 3 是动作序列中的一个人在 2 号摄像头下的部分动作。

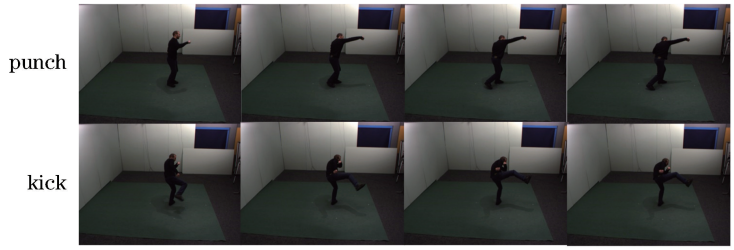

图 2 异常行为

Fig. 2 Abnormal behaviors

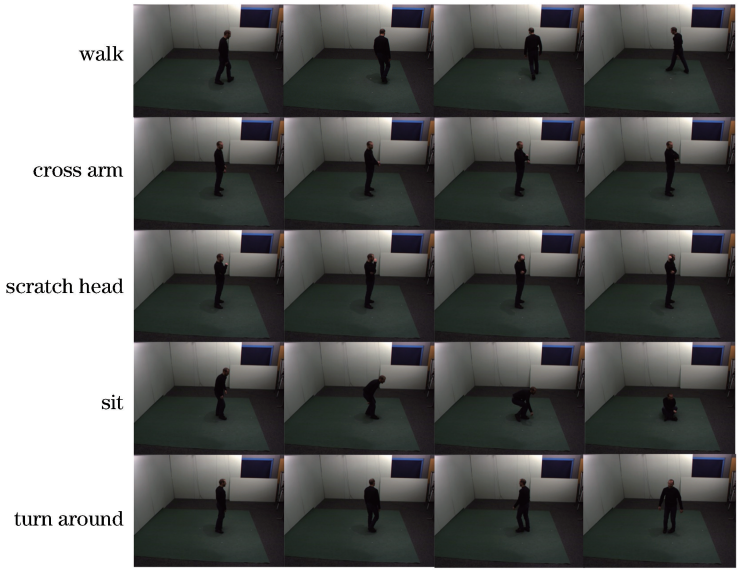

图 3 正常行为

Fig. 3 Normal behaviors

3.2 实验

Caffe 是一种常用的深度学习框架。通过配置 solver 文件，训练模型并对其进行优化。同时还可以选择调用中央处理器(CPU)或者图形处理器(GPU)来训练模型。

3.2.1 网络结构

CNN 共有 1 个输入层，3 个卷积层，3 个池化层，2 个全连接层和 1 个输出层。设定的基本学习率是 0.01，采用的是 step 学习策略，最大迭代次数为 4000。

表 1 卷积神经网络参数

Table 1 Convolutional neural network parameters

Layer	Size /(pixel×pixel)	Number of layers
Input	28×28	1
Convolution kernel	5×5	3
Pooling layer	2×2	3
Fully connected layer	192×1	2
Output	10×1	1

为保障模型的精度，做了 3 组对比实验：

1) 对数据集不做任何处理，按照两类行为的定义分成训练集和测试集并使用 CNN 进行训练和测试，如图 4 所示。

2) 使用 crop(Caffe 深度学习框架里用于图像裁剪的参数)限制待测图片的有效区域，减轻了网络的计算量，最后使用 CNN 进行训练和测试，如图 5 所示。

3) 使用 Caffe(一种深度学习框架)提供的 6 种优化算法对神经网络进行优化。它们分别是随机梯度下降法(SGD)，稳健的学习率方法(AdaDelta)，自适应梯度算法(AdaGrad)，基于梯度的优化方法(Adam 和 RMSProp)，Nesterov 的加速梯度法(NAG)。

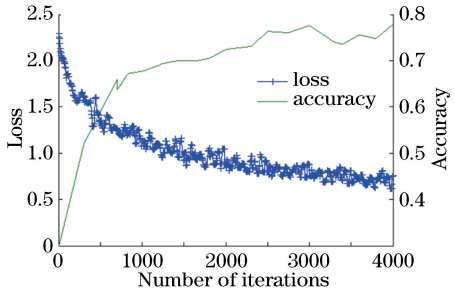

图 4 直接分类的结果

Fig. 4 Direct classification results

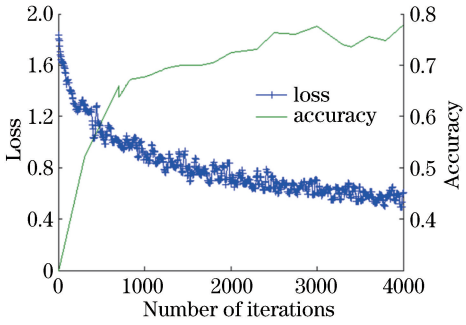

图 5 预处理后的结果

Fig. 5 Processed results

本文采用误识率，即识别错误的结果占识别结果总数的比例，来评价各个算法识别结果的好坏。据表 2 的对比结果，在第 2 组实验中加入 Adam 优化算法，得到的结果如图 6 所示。

表 2 6 种优化算法误识率对比[21]

Table 2 Comparison of misidentification rates of six optimization algorithms[21]

Algorithm	SGD	AdaDelta	NAG	AdaGrad	Adam	RMSProp
False rate /%	16.15	19.56	28.68	18.97	12.33	17.37

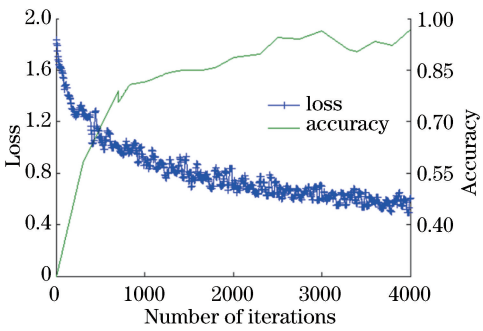

图 6 Adam 算法优化结果

Fig. 6 Adam algorithm optimization results

3.2.2 引入自适应池化层后的结果对比

首先测试 λ 的有效性，确定 λ 可以帮助网络有效减少非判别的特征。如 1.2 节所述，不同的 λ 取值对于特征的过滤能力有很大的影响。如图 7 所示，λ 在 ln 2 到 ln 5 之间时，它的过滤性能保持在 50%不变，说明这是网络自身的过滤能力；而当 λ 超过 ln 6 之后，过滤性能急速下降，当达到 ln 9 的时候只剩下了 30%，这表明 λ 能够基于特征的重要性筛选出判别的特征，提高最终的识别精度。

原始 CNN 与引入自适应池化层的 CNN 在 INRIA 圣诞动作采集序列(IXMAS)上分别迭代

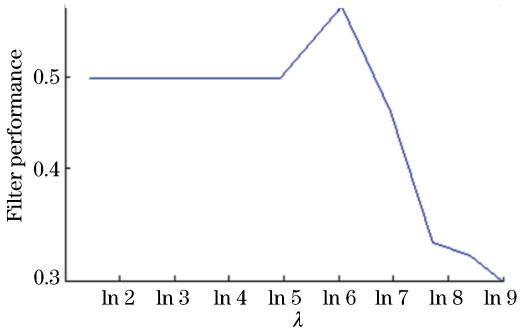

图 7 λ 与过滤性能的关系

Fig. 7 Relationship between λ and filtering performance

4000 次。由图 8 可知，在 4000 左右时基本收敛，且引入自适应池化层的 CNN 收敛效果更好。

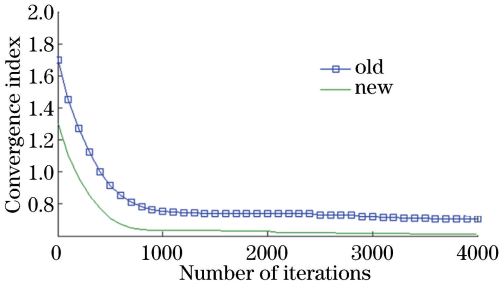

图 8 收敛曲线对比

Fig. 8 Convergence curve comparison

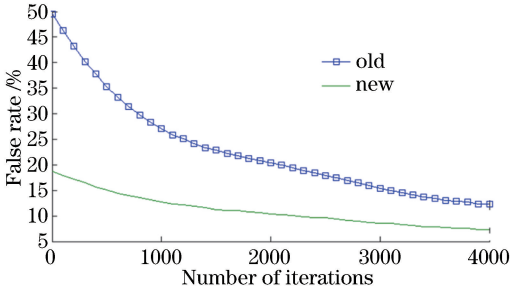

图 9 误识率曲线对比

Fig. 9 Misrecognition rate curve comparison

由表 3 中数据和图 9 可以看出，随着迭代次数的增加，原始 CNN 和引入自适应池化层的 CNN 的误识率都在逐渐降低，但是引入自适应池化层的 CNN 的误识率明显低于原始的 CNN，当迭代收敛时，引入自适应池化层的 CNN 的误识率比原始 CNN 的误识率降低了 35.04%。

本文算法与目前识别效果较好的几种算法作了对比。

1) 改进的轨迹跟踪(iDT)：除深度学习外最好的算法，是改进的 DT 算法，利用相邻两帧的光流消除相机等外在因素的影响，识别结果稳定可靠，但是速度慢。

表 3 误识率对比

Table 3 Misrecognition rate comparison

Number of iterations	Old misrecognition rate /%	New misrecognition rate /%	Reduction rate /%
100	50.44	19.18	61.97
1200	27.15	12.69	53.26
2200	20.26	10.27	49.31
3400	15.34	8.43	45.05
4000	12.33	8.01	35.04

2) 双流融合(Two-Stream Fusion)：对早期的双流网络中加以改进的双流网络，能更好地融合时空信息。

3) 时间分割网络(TSN)：提高了双流网络处理长时间视频序列的能力。

4) 时间关系网络(TRN)：TSN 的改进版本。

由表 4 可以看出，本文提出的 CNN 相较 iDT 和双流网络有更好的识别效果。图 10 给出了表 4 中 4 种深度学习算法的误识率曲线，由表 4 和图 10 可以看出，相较于 TSN 和 TRN，本文算法的识别效果略差。但是对于同一数据集，本文算法完成一次迭代(iter)的平均时间是 1.109 s，TSN 和 TRN 都是 0.953 s，而本文算法在迭代至 4000 次时就已经收敛，TSN 和 TRN 都需要迭代至 5000 次才能收敛，因此本文算法收敛所需要的时间比 TSN 和 TRN 要少。

表 4 不同算法识别效果

Table 4 Recognition effect of different algorithms

Algorithm	Two-stream	TSN	iDT	TRN	This paper
False rate /%	9.37	7.93	8.54	7.26	8.01
Reduction rate /%	14.5	−1.0	6.2	−10.3	—

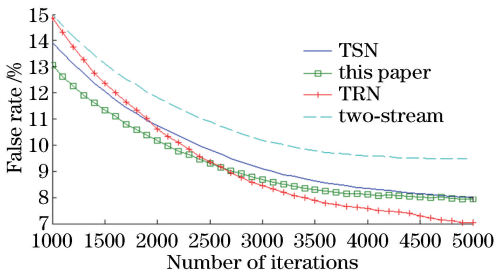

图 10 几种算法的对比

Fig. 10 Comparison of several algorithms

本文还引用了 UT-interaction database 来验证算法的效果。取其中的 punch 和 kick 两种行为共 6000 frame 作为异常行为，其余几种动作共 12000 frame作为正常行为加以测试。经过 10000

次迭代后，所有算法均已收敛，结果如表 5 所示。

表 5 在 UT-interaction database 上的识别效果

Table 5 Recognition effect on UT-interaction database

Algorithm	Two-stream	TSN	iDT	TRN	This paper
False rate /%	12.62	9.49	10.44	8.92	9.53
Reduction rate /%	24.5	−0.4	8.7	−6.8	—

由于 UT-interaction database 是一个交互行为数据库，多人互动带来的干扰对几种模型产生了不同程度的影响。对比表 5 和表 4 的降低率可以看出，本文的模型受到的影响较小，与 Two-Stream 和 iDT 的精度差距进一步扩大，与 TSN 和 TRN 的精度差距则在减小。这也体现了本文提出的自适应池化层能够有效提取判别的特征，提高识别精度。

4 结 论

选用 CNN 对预定义的几种正常与异常行为进行训练并得到识别模型，最后将 Caffe 提供的几种优化算法进行对比。基于 Adam 算法，引入自适应池化层，通过进一步筛选卷积层提取出的特征来提高 CNN 模型的判别能力。与其他几种算法对比，结果表明训练出的模型能够基本满足预期的识别要求。但是本文算法实时性较差，只能用于线下视频的处理，能够以 3 frame/s 的速度处理视频序列。因此，今后研究目标为结合能够实现时序动作检测的 CNN，做到实时识别智能监控系统中的异常行为。

参 考 文 献

[1] Wang L, Ye H, Xia L Z. Discriminative human action recognition using semi-Markov model and large-margin[J]. Journal of Image and Graphics, 2009, 14(11): 2304-2310.
汪力, 叶桦, 夏良正. 基于半马尔可夫和 large-margin 的动作识别[J]. 中国图象图形学报, 2009, 14(11): 2304-2310.

[2] Xu G Y, Cao Y Y. Action recognition and activity understanding: a review[J]. Journal of Image and Graphics, 2009, 14(2): 189-195.
徐光祐, 曹媛媛. 动作识别与行为理解综述[J]. 中国图象图形学报, 2009, 14(2): 189-195.

[3] Hu Q, Qin L, Huang Q M. A survey on visual human action recognition[J]. Chinese Journal of Computers, 2013, 36(12): 2512-2524.
胡琼, 秦磊, 黄庆明. 基于视觉的人体动作识别综述[J]. 计算机学报, 2013, 36(12): 2512-2524.

[4] Fu Y W, Yang S P. Human action recognition by extracting motion trajectories[J]. Proceedings of SPIE, 2015, 9631: 96311H.

[5] Gheisari S, Meybodi M R, Dehghan M, *et al*. BNC-VLA: Bayesian network structure learning using a team of variable-action set learning automata[J]. Applied Intelligence, 2016, 45(1): 135-151.

[6] Chen T T, Ruan Q Q, An G Y. Slow feature extraction algorithm of human actions in video[J]. CAAL Transactions on Intelligent Systems, 2015, 10(3): 381-386.
陈婷婷, 阮秋琦, 安高云. 视频中人体行为的慢特征提取算法[J]. 智能系统学报, 2015, 10(3): 381-386.

[7] Maity S, Bhattacharjee D, Chakrabarti A. A novel approach for human action recognition from silhouette images[J]. IETE Journal of Research, 2017, 63(2): 160-171.

[8] Ma Y J, Li X Y, Song X F. Traffic sign recognition based on improved deep convolution neural network[J]. Laser & Optoelectronics Progress, 2018, 55(12): 121009.
马永杰, 李雪燕, 宋晓凤. 基于改进深度卷积神经网络的交通标志识别[J]. 激光与光电子学进展, 2018, 55(12): 121009.

[9] Long X, Su H S, Liu G H, *et al*. A face recognition algorithm based on angular distance loss function and convolutional neural network[J]. Laser & Optoelectronics Progress, 2018, 55(12): 121505.
龙鑫, 苏寒松, 刘高华, 等. 一种基于角度距离损失函数和卷积神经网络的人脸识别算法[J]. 激光与光电子学进展, 2018, 55(12): 121505.

[10] Wang J M, Lu J F. Face gender recognition based on convolutional neural network[J]. Modern Electronics Technique, 2015, 38(7): 81-84.
汪济民, 陆建峰. 基于卷积神经网络的人脸性别识别[J]. 现代电子技术, 2015, 38(7): 81-84.

[11] Jin L W, Zhong Z Y, Yang Z, *et al*. Applications of deep learning for handwritten Chinese character recognition: a review[J]. Acta Automatica Sinica, 2016, 42(8): 1125-1141.
金连文, 钟卓耀, 杨钊, 等. 深度学习在手写汉字识别中的应用综述[J]. 自动化学报, 2016, 42(8): 1125-1141.

[12] Jaderberg M, Simonyan K, Vedaldi A, *et al*. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 2016, 116(1): 1-20.

[13] Chang L, Deng X M, Zhou M Q, *et al*.

Convolutional neural networks in image understanding[J]. Acta Automatica Sinica, 2016, 42(9): 1300-1312.
常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络[J]. 自动化学报, 2016, 42(9): 1300-1312.

[14] Cai G Y, Xia B B. Multimedia sentiment analysis based on convolutional neural network[J]. Journal of Computer Applications, 2016, 36(2): 428-431, 477.
蔡国永, 夏彬彬. 基于卷积神经网络的图文融合媒体情感预测[J]. 计算机应用, 2016, 36(2): 428-431, 477.

[15] Hou B, Zhang X R, Ye Q, *et al*. A novel method for hyperspectral image classification based on Laplacian eigenmap pixels distribution-flow[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 6(3): 1602-1618.

[16] Najafabadi M M, Villanustre F, Khoshgoftaar T M, *et al*. Deep learning applications and challenges in big data analytics[J]. Journal of Big Data, 2015, 2: 1.

[17] Shi Z T, Wang Z R, Wang R, *et al*. Single image super-resolution based on convolutional neural network[J]. Laser & Optoelectronics Progress, 2018, 55(12): 121001.
史紫腾, 王知人, 王瑞, 等. 基于卷积神经网络的单幅图像超分辨[J]. 激光与光电子学进展, 2018, 55(12): 121001.

[18] Zhang H Y, Wang S N, Hu W B. Improved method for estimating number of people based on convolution neural network[J]. Laser & Optoelectronics Progress, 2018, 55(12): 121503.
张红颖, 王赛男, 胡文博. 改进的基于卷积神经网络的人数估计方法[J]. 激光与光电子学进展, 2018, 55(12): 121503.

[19] Cai J, Cai J Y, Liao X D, *et al*. Preliminary study on hand gesture recognition based on convolutional neural network[J]. Computer Systems & Applications, 2015, 24(4): 113-117.
蔡娟, 蔡坚勇, 廖晓东, 等. 基于卷积神经网络的手势识别初探[J]. 计算机系统应用, 2015, 24(4): 113-117.

[20] Goldberg Y, Hirst G. Neural network methods for natural language processing[M]. Williston, VT: Morgan & Claypool, 2017.

[21] Liu W J, Liang X J, Qu H C. Adaptively enhanced convolutional neural network algorithm for image recognition[J]. Journal of Image and Graphics, 2017, 22(12): 1723-1736.
刘万军, 梁雪剑, 曲海成. 自适应增强卷积神经网络图像识别[J]. 中国图象图形学报, 2017, 22(12): 1723-1736.