DOI:10.13878/j.cnki.jnuist.2019.06.009

张聪聪[1]　何宁[2]

基于关键帧的双流卷积网络的人体动作识别方法

摘要

针对视频序列中人体动作识别存在信息冗余大、准确率低的问题,提出基于关键帧的双流卷积网络的人体动作识别方法.该方法构建了由特征提取、关键帧提取和时空特征融合3个模块构成的网络框架.首先将空间域视频的单帧RGB图像和时间域多帧叠加后的光流图像作为输入,送入VGG16网络模型,提取视频的深度特征;其次提取视频的关键帧,通过不断预测每个视频帧的重要性,选取有足够信息的有用帧并汇聚起来送入神经网络进行训练,选出关键帧并丢弃冗余帧;最后将两个模型的Softmax输出加权融合作为输出结果,得到一个多模型融合的人体动作识别器,实现了对视频的关键帧处理和对动作的时空信息的充分利用.在UCF-101公开数据集上的实验结果表明,与当前人体动作识别的主流方法相比,该方法具有较高的识别率,并且相对降低了网络的复杂度.

关键词

关键帧;双流网络;动作识别;特征提取;特征融合

中图分类号 TP391.41;TP181

文献标志码 A

收稿日期 2019-10-07

资助项目 国家自然科学基金(61872042,61572077);北京市自然科学基金委和北京市教委联合重点项目(KZ201911417048)

作者简介

张聪聪,男,硕士生,研究方向为数字图像处理.1274190198@qq.com

何宁(通信作者),女,博士,教授,研究方向为数字图像处理.xxthening@buu.edu.cn

1　北京联合大学 机器人学院,北京,100101

2　北京联合大学 智慧城市学院,北京,100101

0　引言

随着多媒体技术和网络传输设施的不断发展,对视频数据处理的需求不断增加,视频处理中的一个重要分支是人体行为识别.人体行为识别的目的是分析并理解视频中的人体的动作和行为,研究如何感知目标对象在图像序列中的时空运动变化.视频中的人体行为识别在机器人交互、虚拟现实、视频监控等领域有广泛应用.

目前人体行为识别方法主要分为传统方法和基于深度学习的方法,其中,基于深度学习的方法在处理大数据集上有很大的优势.在人体行为识别中基于双流网络[1]的方法有比较好的成果.双流架构利用视觉帧和相邻帧之间的光流作为网络的两个独立输入,并将其输出融合作为最终预测,许多论文使用并扩展文献[2-4]的双流架构.Zhu等[4]提出了一种新型的卷积神经网络(Convolutional Neural Network,CNN)结构,可以在关联帧中暗中捕捉动作信息,并且能够高效预测光流,端到端地实现人体行为识别.循环神经网络(Recurrent Neural Networks,RNN)能在一定程度上解决视频中的时序处理和预测问题,尤其是对视频序列能够有效建模的长时短期记忆模型(Long Short-Term Memory,LSTM)[5].但是,LSTM的输入是直接从CNN的全连接层中提取的高级特征,而这些特征缺乏时空特征细节.

视频相比图像来说信息更加丰富,但是视频序列里冗余信息太多,针对这种情况本文提出了一种基于关键帧的双流卷积网络的人体动作识别方法来解决视频中连续帧之间的大量冗余的情况,提高识别速度,该方法首先提取出视频序列的深度特征,然后输入到关键帧提取模块,剔除冗余帧,选择出包含足够信息的关键帧,然后结合光流运动特征,对视频动作进行识别.本文借鉴了文献[1,6]方法,在双流网络中加入了关键帧选区模块,去除冗余帧,来提高行为识别的准确率.在数据集UCF101上进行人体行为识别实验,实验结果表明本文中提出的基于关键帧的双流卷积网络的人体动作识别方法具有有效性.

1　相关工作

在人体动作视频序列中,提取的关键帧要能够反映视频序列中要表示的人体动作,因为视频是渐变的,所以帧与帧之间可能存在着冗余,这样会对人体动作识别的识别率产生不良影响.针对现有的人

体动作识别需要输入固定的视频帧,选取的视频帧的帧与帧之间存在信息冗余,采用全部视频序列进行人体动作识别计算量大、效率低,所以选取视频序列中的关键帧,去除冗余信息的视频帧,对于提高动作识别的准确率和实时性非常重要,比较有效的方法是提取视频图像的动作特征,然后基于所提取的特征进行动作识别.

早期针对视频序列中特征提取的算法主要提取视频序列全局特征和局部特征,然后进行字典编码,Bobick 等[7]最早采用轮廓来描述人体的运动信息.Gorelick 等[8]首次从视频序列中的剪影信息得到时空体积(Spatial-Temporal Volume,STV),然后用泊松方程导出局部时空显著点及其方向特征,其全局特征是通过对这些局部特征加权得到的.为了处理不同动作的持续时间不同的问题,Laptev 等[9]使用了局部 HOG(梯度直方图)和 HOF(光流直方图),Klaser 等[10]将 HOG 特征扩展到三维,即形成了 3D-HOG.类似这种将二维特征点检测的算法扩展到三维特征点的工作是将尺度不变特征变换(Scale Invariant Feature Transform,SIFT)算法[11]扩展到三维 SIFT.Scovanner 等[12]在 Wang 等[13]的文章中,比较了各种局部描述算子,并发现在大多数情况下整合了梯度和光流信息的描述算子其效果最好.

在基于深度学习的方法中,Simonyan 等[1]提出了双流深度网络,它结合了空间网络和时间网络,使用 RGB 帧和提取光流相结合来进行动作识别.Ng 等[14]发现了双流网络中的一个缺点,即使用标准图像 CNN 而不是专用网络来训练视频,这导致双流网络无法捕获长期时间信息.后来 LSTM 网络也被用来捕捉视频中的时间动态和序列相关性.在文献[14-15]中提取有序光流用于学习每帧的空间特征,而 LSTM 用于模拟时间演变.本文的关键帧提取模块与基于注意力机制的算法具有相似性,该关键帧提取模块使用评判函数来区别视频序列是否包含足够信息的有用帧,然后将这些有用帧汇集送入分类器,丢弃冗余帧,关键帧模块也使用了递归的思想,不断预测下一帧的重要性来选取有用帧.图 1 显示了关键帧模块的基本示意图.该关键帧提取模块由一个 3 层的多层感知器(Muti-Layer Perceptron,MLP)构成,加入预测函数 $f(\cdot)$ 对视频帧进行预测评判,通过递归的方式使用交叉熵损失函数来更好地选取关键帧,丢弃冗余帧.

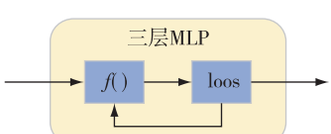

图 1 关键帧提取示意图

Fig. 1 Key frame extraction diagram

2 基于关键帧的双流卷积网络人体动作识别

2.1 整体网络框架

本文提出的网络框架如图 2 所示.该模型主要包含 3 个模块:深度特征提取模块、关键帧提取模块、空间与时间域的特征融合模块.对于深度特征提取模块,通过 VGG16 网络的卷积层和第一个全连接层(FC6)来提取视频图像的深度特征,输入 3×224×224 的图像数据,即一张宽 224、高 224 的视频图像,网络包含 5 个卷积层,5 个池化层,使用 ReLU 激活函数,输出 4 096 个神经元.具体来说,从 1 到 5 的 5 层卷积层的滤波器数目分别是 64、128、256、512、512,1 个全连接层是 4 096 个单元.根据文献[16]对卷积层的不同深度的内核实验研究结果,3×3×3 的核尺寸大小是对所有卷积层来说最佳的选择,因此,在此模块中,所有卷积层均采用 3×3×3 的内核大小,步长为 1×1×1.对于最大池化层,除了第一个最大池化的核大小是 2×2×1,其余 3 个最大池化层的核大小为 2×2×2,特征提取模块主要负责从视频序列中提取视频帧的深度特性 X_t 生成固定维度的特征向量.关键帧提取模块主要负责从深度网络特区的深度特征中选取有效帧.总体而言,本文的网络框架包含了特征提取、关键帧提取和时空特性融合.融合模块是将空间帧和时间帧按照一定的策略进行融合.通过这种关键帧选取方法使得该网络模型对人体动作的识别更加准确.

2.2 特征提取

本文将视频表示为连续帧的表示形式,如式(1)所示,在空间网络表示为 RGB 图像,时间网络表示为相邻帧的光流图像堆栈.首先视频输入后在卷积神经网络中提取视频帧的深度特征 X_t,关键帧模块设置在 VGG-16 网络的第一个全连接层之后,其输入为卷积神经网络提取的深度特征 $a(X_t)$,卷积层是通过多个不同的卷积核对上一层的输入做卷积运算得到多个输出,即多个特征图.卷积公式如式(2)所示.

$$X=[x_1,\cdots,x_T],x_t\in\mathbf{R}^{224\times224\times K},\tag{1}$$

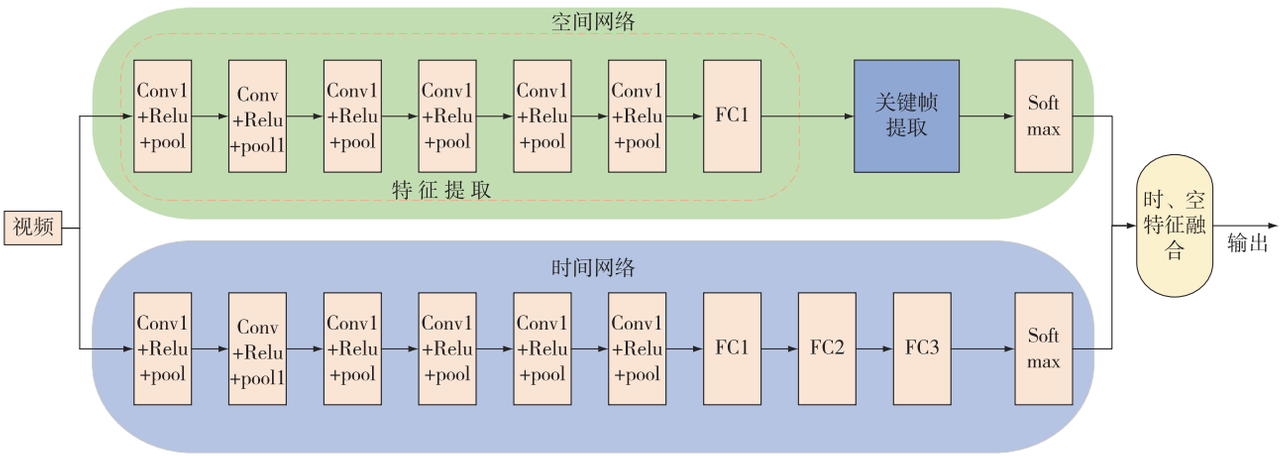

图 2　算法的整体网络框架

Fig. 2　Overall network framework of the proposed algorithm

$$x_j^l = f\Big(\sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_i^l\Big)\,, \tag{2}$$

x_j^l 表示卷积层的输出,即第 l 层的第 j 个特征图,k_{ij}^l 表示第 l 层第 i 个特征图和第 l 层的第 j 个特征图对应的卷积核,b_i^l 表示第 l 层的第 i 个特征图对应的偏置,$\sum_{i\in M_j} x_i^{l-1}$ 表示第 i 个特征图对于的输入特征图的集合,f 表示激活函数.该卷积神经网络使用线性整流函数(ReLU)作为激励函数来协助表达复杂特征,使用 ReLU 激活函数能够有效地避免梯度爆炸和梯度消失问题,同时也简化了计算过程,同时活跃度的分散性使得神经网络整体计算成本下降,其表现形式为式(3).

$$a_{i,j,k}^l = f(x_{i,j,k}^l)\,, \tag{3}$$

$$A_k^l(i,j) = \Big[\sum_{x=1}^{f}\sum_{y=1}^{f} A_k^l(si+x, sj+y)^p\Big]^{\frac{1}{p}}. \tag{4}$$

池化操作是指对池化范围区域内所有像素点求平均值或最大值作为该区域池化后的值,从而实现卷积特征的降维并获得具有空间不变性的特征,本文在卷积层后使用最大池化,即式(4)中 p 趋向无穷时,表示为卷积层后进行最大池化.

2.3　关键帧提取

关键帧的提取通过递归的方式实现,主要分为两步,首先预测视频帧的重要性,通过式 $t(t+1)$ 预测第 $t+1$ 帧的重要性,然后通过汇集当前帧与之前的视频帧,再通过 $f(\cdot)$ 函数判别出当前帧的重要性,给出的重要性系数区间为(0,1).预测帧的重要性如式(5)—(7)所示.其中,式(5)为预测函数,$\alpha(X,t)$ 为卷积神经网络提取深度特征后的 t 个视频帧,$\beta(x_{t+1})$ 为 $t+1$ 帧的深度特征,通过预测函数 $f(\cdot)$ 来预测第 $t+1$ 帧的重要性.式(6)、(7)将预测的第 $t+1$ 帧的重要性与前 t 帧进行加权平均操作.

$$\gamma_{t+1} = f_p(\alpha(X,t),\beta(x_{t+1}))\,, \tag{5}$$

$$\alpha(X,t+1) = \frac{1}{\gamma_{t+1}}(\gamma_t\alpha(X,t) + \gamma_{t+1}\beta(x_{t+1}))\,, \tag{6}$$

$$\gamma_p = \sum_{k=1}^{p}\gamma_k. \tag{7}$$

关键帧模块中的预测函数 f 是基于一个多层感知器(MLP)实现的,f 函数的底层操作只依赖于标准的线性和非线性操作,这样提高了计算速度,能够更好地集成到卷积神经网络中进行端到端的学习,在最终层使用具有 tanh 非线性和 S 型激活的 3 层 MLP.我们将合并向量的初始状态设置为与第一帧的特征相同.本文使用 Glorot 等[17]提出的初始化方法来对关键帧选择模块进行初始化.

为了在关键帧模块能够更好地预测视频帧的重要性和非冗余性,在关键帧模块中加入了当前帧和下一帧之间的相关性差异比较,这样不仅能够丢弃冗余帧,还能更好地选出包含足够信息的有用帧,提高模型的泛化能力.为了能够更好地从一个视频中选择出有用的信息帧,在预测函数后添加了一个基于熵的正则化项,如式(8)所示.这个正则化项加入了峰值分布,能够更好地选择有区别的信息帧,丢弃冗余帧.参数 λ 是一个平衡参数,能够平衡视频帧的选择和更好地减少交叉熵损失函数.如果将 λ 设置为相对较高的值,我们希望选择的帧数少,这会使分类任务变得更加困难,例如每个视频单帧将使其与图像分类相同.同时,如果 λ 的值相对较低,则该模型将选择较大数量的帧,并且可能过度拟合.本文使

用一个标准的交叉损失函数来表示预测的视频帧和真实的视频帧之间的损失函数,如式(9)所示.

$$(X,y) = {}_{CE}(X,y) + \lambda {}_{E}(\ ), \tag{8}$$

$${}_{E}(\ ) = -\sum_{k}\frac{e^{\gamma k}}{N}\log\left(\frac{e^{\gamma k}}{N}\right), \tag{9}$$

其中, ${}_{E}(\ )$ 表示标准的交叉熵损失函数,便于梯度下降反向传播,利于优化多层感知器(MLP),可以更好地选取关键帧,丢弃冗余帧. ${}_{E}(\ )$ 表示添加的基于熵的正则化项,添加正则化项能有效地防止过拟合,减少误差,同时加入平衡参数 λ 进一步提高关键帧的选择.

本文提出的关键帧模块与文献[6]中的自适应扫描池比较相似,自适应扫描池能够识别视频中的信息帧,只对这些对象进行池操作,同时丢弃其他无用帧,以端到端可学习的方式动态地汇集视频帧以进行动作分类,同时产生可解释的中间状态.本文使用评判函数来计算视频帧的相关性,通过递归的思想,不断预测视频帧的重要性,最后丢弃相关性比较小的帧,汇集有用帧来进行行为识别.其中在关键帧模块中预测函数的定义使用评判函数对视频序列的每一帧进行打分,选取分数较高的作为视频序列的有用帧,并使用这些有用帧进行行为识别,以此来提高人体行为识别的识别准确率.同时,本实验使用VGG-16网络模型在一定程度上减少了网络的复杂度.

2.4 时空特征融合

特征融合是将多个基分类器的结果,按照一定的规则融合成一个全局的结果,消除决策本身或决策之间的信息缺陷,提升全局结果的可靠性和稳定性[18].本文使用的基本框架为双流卷积网络,空间网络提取关键帧的表面信息,时间网络输入视频帧后提取帧与帧之间的光流,携带视频帧之间的运动信息,两个深度网络都会输出一个Softmax层,最后通过平均层的决策融合方法,对两个Softmax层的输出结果进行融合.网络结构最后都有时间和空间网络上两个基分类器的识别结果.最终将两个分支的分类结果进行加权融合,以得到关于视频中人体行为类别的最终融合结果.

3 实验结果与分析

3.1 数据集

本文采用UCF101[19]视频动作识别数据集.UCF-101(2012)包含13 320个视频(共27 h),101个人类行为类别,涵盖了较大范围的人体动作,如体育运动、乐器和人物交互等,分辨率为320×240.UCF101在动作的采集上具有非常大的多样性,包括外观变化、姿态变化、物体比例变化、背景变化、光纤变化等.该数据集的大多数视频是在无约束的真实环境下拍摄的,因此视频存在像素低,受到如光照、遮挡等环境因素影响的问题.

3.2 实验结果与分析

在Linux系统搭建Ubuntu16.04的TensorFlow平台下进行实验.由于深度神经网络容易陷入过拟合现象,因此本文将模型中空间网络和时间网络dropout层的丢失率分别设置为0.7和0.8.本文使用Simonyan等[1]的双流网络框架为由VGG16网络组成的双流网络,从每个视频中均匀采样25个帧来作为输入,通过两个方向上的5个相邻帧中堆叠 X 和 Y 方向光流,为时域网络生成20通道光学流输入[1,20].使用Wang等[20]提供的工具来提取光流.采用文献[20]中的TV-L1算法,并通过线性变换离散化了[0,255]范围内的光场来保证和RGB数据同区间.初始化用于训练UCF101的空间网络来自在ImageNet上训练的VGG-16模型[21-22].为了在UCF101上训练时间网络,使用了Wang等[20]提供的16 000次迭代快照初始化其卷积层.后期实验中还使用Resnet网络对空间网络进行训练,通过增加网络的深度,来更好地提取视频序列的深度特征,最后实验结果表明在单只网络上的行为识别准确率比VGG-16的好,但是由于Resnet的网络深度较大,更容易发生过拟合.

将本文的算法与其他人体行为识别方法在UCF101数据集上进行比较,结果如表1所示,可以发现本文提出的算法识别准确率优于其他算法,识别效果较好.本文使用双流卷积神经网络所获取的

表1 不同方法在UCF101数据集上的动作识别准确率

Table 1 Action recognition accuracy of different methods on UCF101 dataset

方法	准确率/%
two-stream convolutional networks[1]	88.0
C3D[23]	85.2
dynamic image networks[24]	89.1
LSTM on long clips[14]	88.6
AdaScan[6]	89.4
two-stream VGG[25]	92.5
本文方法	93.12

特征结合了运动表层特征和时序信息两部分,更好地发掘了视频所包含的信息,同时在关键帧识别模块,我们可以使用更少的视频帧得出更有用的信息,一方面提高了行为识别的准确率,另一方面减少了计算的复杂度.

4 结束语

目前基于深度学习的方法已经广泛应用到模式识别等各个领域的研究中,对于人体动作识别任务,本文提出了基于关键帧的深度神经网络模型,该方法通过不断地预测每个视频帧的重要性,去除冗余帧,选取包含足够信息的有用帧并汇聚起来进行训练,实现了对视频段的关键帧处理和对动作的复杂时空信息的充分利用,从而提高动作的识别率,构建了时空双流深度神经网络架构.将本文模型先在 ImageNet 上进行预训练和微调,然后应用到 UCF101 数据集上,实验结果表明,该方法具有较高的识别率,并且相对降低了网络的复杂度.

参考文献

References

[1] Simonyan K,Zisserman A.Two-stream convolutional networks for action recognition in videos[J].Neural Information Processing Systems,2014,1(2):568-576
[2] Feichtenhofer C,Pinz A,Wildes R P.Spatiotemporal residual networks for video action recognition[J].Neural Information Processing Systems,2016,2(3):3468-3476
[3] Feichtenhofer C,Pinz A,Wildes R P.Spatiotemporal multiplier networks for video action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017,DOI:10.1109/CVPR.2017.787
[4] Zhu Y,Lan Z Z,Newsam S,et al.Hidden two-stream convolutional networks for action recognition [M] // Computer Vision-ACCV 2018. Cham: Springer International Publishing,2019:363-378
[5] Hochreiter S, Schmidhuber J. Long short-term memory [J].Neural Computation,1997,9(8):1735-1780
[6] Kar A,Rai N,Sikka K,et al.AdaScan:adaptive scan pooling in deep convolutional neural networks for human action recognition in videos[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:3376-3385
[7] Bobick A F,Davis J W.The recognition of human movement using temporal templates[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2001,23(3):257-267
[8] Gorelick L,Blank M.Actions as space-time shapes[J].Pattern Analysis and Machine Intelligence, 2007, 29(12):2247-2253
[9] Laptev I,Marszalek M,et al.Learning realistic human actions from movies[J].IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2008,DOI:10.1109/CVPR.2008.4587756
[10] Klaser A, Marszalek M. A spatio-temporal descriptor based on 3D-gradients[C]// British Machine Vision Conference,2008,DOI:10.5244/C.22.99
[11] Mikolajczyk K,Mikolajczyk K.Scale & affine invariant interest point detectors [J]. International Journal of Computer Vision,2004,60(1):63-86
[12] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition[C]// ACM International Conference on Multimedia, 2007:357-360
[13] Wang H,Ullah M M,Klaser A,et al.Evaluation of local spatio-temporal features for action recognition[C]// British Machine Vision Conference,2009,DOI:10.5244/C.23.124
[14] Ng Y H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets:deep networks for video classification[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015:4694-4702
[15] Donahue J,Hendricks L A,Rohrbach M,et al.Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2014,39(4):677-691
[16] Karpathy A,Toderici G,Shetty S,et al.Large-scale video classification with convolutional neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition,2014:1725-1732
[17] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]// International Conference on Artificial Intelligence and Statistics,2010:249-256
[18] 张文宇.基于证据理论的无线传感器网络决策融合算法研究[D].北京:北京交通大学,2016
ZHANG Wenyu.Research on beliffunction based decision fusion for wireless sensor networks [D].Beijing:Beijing Jiaotong University,2016
[19] Soomro K,Zamir A R,Shah M.Ucf101:a dataset of 101 human actions classes from videos in the wild[J].arXiv Preprint,2012,arXiv:1212.0402
[20] Wang L,Xiong Y,Wang Z,et al.Towards good practices for very deep two-stream convnets[J].arXiv Preprint,2015,arXiv:1507.02159
[21] Deng J,Dong W,Socher R,et al.ImageNet:a large-scale hierarchical image database [C] // IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2009, DOI: 10.1109/CVPR.2009.5206848
[22] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition [J]. arXiv Preprint,2014,arXiv:1409.1556
[23] Srivastava N,Mansimov E,Salakhutdinov R.Unsupervised learning of video representations using LSTMs[C]// The 32th International Conference on Machine Learning (ICML),2015:843-852
[24] Bilen H,Fernando B,Gavves E,et al.Dynamic image networks for action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

2016:3034-3042
[25] Feichtenhofer C,Pinz A,Zisserman A.Convolutional two-stream network fusion for video action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:1933-1941

Human motion recognition based on key frame two-stream convolutional network

ZHANG Congcong[1] HE Ning[2]
1 Robotics College,Beijing Union University,Beijing 100101
2 Smart City College,Beijing Union University,Beijing 100101

Abstract Aiming at the problem of large information redundancy and low accuracy in human motion recognition in video sequences,a human motion recognition method is proposed based on key frame two-stream convolutional network. We construct a network framework consisting of three modules:feature extraction,key frame extraction,and spatial-temporal feature fusion.Firstly,the single-frame RGB image of the spatial domain video and the optical flow image superimposed in the time domain multi-frame are sent as input to the VGG16 network model to extract the depth feature of the video;secondly,the importance of each video frame is continuously predicted,then useful frames with sufficient information are pooled and trained by neural network to select key frames and discard redundant frames.Finally,the Softmax outputs of the two models are weighted and combined as the output result to obtain a multi-model fusion.The human body motion recognizer realizes the key frame processing of the video and the full utilization of the spatial-temporal information of the action.The experimental results on the UCF-101 public dataset show that,compared with the mainstream methods of human motion recognition,the proposed method has a higher recognition rate and relatively reduces the complexity of the network.

Key words keyframe;two stream networks;action recognition;feature extraction;feature fusion