❖特约专栏❖

时空特征融合深度学习网络人体行为识别方法

裴晓敏[1,2],范慧杰[2],唐延东[2]

(1. 辽宁石油化工大学 信息与控制工程学院,辽宁 抚顺 113001;
2. 中国科学院沈阳自动化研究所 机器人学国家重点实验室,辽宁 沈阳 110016)

摘 要:基于自然场景图像的人体行为识别方法中遮挡、背景干扰、光照不均匀等因素影响识别结果,利用人体三维骨架序列的行为识别方法可以克服上述缺点。首先,考虑人体行为的时空特性,提出一种时空特征融合深度学习网络人体骨架行为识别方法;其次,根据骨架几何特征建立视角不变性特征表示,CNN(Convolutional Neural Network)网络学习骨架的局部空域特征,作用于空域的 LSTM (Long Short Term Memory)网络学习骨架空域节点之间的相关性特征,作用于时域的 LSTM 网络学习骨架序列时空关联性特征;最后,利用 NTU RGB+D 数据库验证文中算法。实验结果表明:算法识别精度有所提高,对于多视角骨架具有较强的鲁棒性。

关键词:时空特征; 融合; 骨架; 视角不变

中图分类号:TP183 **文献标志码**:A **DOI**:10.3788/IRLA201847.0203007

Action recognition method of spatio-temporal feature fusion deep learning network

Pei Xiaomin[1,2], Fan Huijie[2], Tang Yandong[2]

(1. School of Information and Control Engineering, Liaoning Shihua University, Fushun 113001, China;
2. State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China)

Abstract: Action recognition from natural scene was affected by complex illumination conditions and cluttered backgrounds. There was a growing interest in solving these problems by using 3D skeleton data. Firstly, considering the spatio-temporal features of human actions, a spatio-temporal fusion deep learning network for action recognition was proposed; Secondly, view angle invariant character was constructed based on geometric features of the skeletons. Local spatial character was extracted by short-time CNN networks. A spatio-LSTM network was used to learn the relation between joints of a skeleton frame. Temporal LSTM was used to learn spatio-temporal relation between skeleton sequences. Lastly, NTU RGB+D datasets were used to evaluate this network, the network proposed achieved the state-of-the-art performance for 3D human action analysis. Experimental results show that this network has strong robustness for view invariant sequences.

Key words: spatio-temporal feature; fusion; skeleton; view invariant

---

收稿日期:2017-08-10; **修订日期**:2017-10-28

基金项目:国家自然科学基金(61401455)

作者简介:裴晓敏(1981-),女,讲师,博士后,主要从事机器视觉方面的研究。Email:pxm_neu@126.com

导师简介:唐延东(1962-),男,博士生导师,博士,主要从事图像处理与模式识别、机器视觉方面的研究。Email:ytang@sia.cn

0 引 言

近年来，行为识别技术成为机器视觉领域的研究热点之一。人体行为识别技术可广泛应用于智能视频监控、病人看护、机器人、人机交互等领域。传统行为方法采用自然场景图像序列识别行为，容易受背景运动、光照不均、遮挡等环境因素影响。

随着深度相机的普及，实时获取行为人骨架成为可能。基于三维骨架的行为识别方法因其具有不受遮挡、背景干扰等优点受到业界广泛关注[1-4]。

深度学习 RNN(Recurrent Neural Networks)具有记忆功能，在序列行为识别、预测中取得了较好的效果。典型的骨架行为识别方法多采用 RNN 或其改进模型，主要有 Yong Du 等提出基于分层 RNN 骨架识别方法，根据人体结构先验知识将骨架分组后逐层融合输入到 RNN[5]。

Veeriah 等提出基于差分 RNN 的行为识别方法，通过 RNN 学习连续帧间的骨架节点变化[6]。Wentao Zhu 等提出基于共生性特征学习的正则化深度 LSTM(Long Short Term Memory)网络骨架行为识别，利用全连接网络学习骨架的共生性[7]。Amir Shahroudy 等提出 Part-Aware LSTM 方法，将人体骨架分成五部分输入到网络，通过 LSTM 网络学习骨架的长时组合特征表示 [8]。Liu 等提出带有 Trust Gates 的 LSTM 模型学习骨架序列的可靠性 [9]。Liu 等提出基于全局内容显著性的 LSTM 网络行为识别方法[10]，上述识别方法对于典型、固定视角骨架库均取得了较好的识别效果。然而，以上方法对于多视角变换骨架并未深入讨论，而实际应用中人体行为往往为多角度变化骨架。

考虑到人体行为序列的时空特性，文中提出时空特征融合深度学习网络行为识别模型。首先建立骨架的视角不变性时空描述，然后采用 CNN (Convolutional Neural Network) 提取骨架局部空域特征，LSTM 网络学习骨架节点的空间关联性特征，最后利用 LSTM 网络学习骨架时空融合特征。实验结果表明，融合时空特征的深度学习网络较于前述网络识别效果有明显提高，并且具有视角不变性特征。

1 骨架视角不变性特征提取

骨架生成过程中因成像条件不同，如摄像头相对距离、角度、相对运动等，造成较大差异。首先将骨架序列规整化；然后采用视角不变性变换处理；最后生成运动特征图。

骨架以三维点序列的形式保存，人体骨架是 n 个骨架节点的三维坐标(x,y,z)，为消除骨架拍摄视角对识别结果的影响，文中采用骨架距离图和骨架角度图描述骨架的空间特征(如图 1 所示)。为使骨架具有视角不变性，以人体骨架脊柱点 2 为中心点，脊柱根节点 1 到中心点连线为中心线 S_{21}，计算骨架的距离运动图和角度运动图。公式(1)计算 t 时刻骨架各节点到中心点距离 $D_{\text{dist}}(n,t)$，生成骨架距离运动图。根据公式(2)计算骨架上(除中心点外)各点到中心点连线 S_{n2} 与中心线 S_{21} 的夹角 $D_{\text{angle}}(n,t)$，得到骨架角度运动图。并在整个序列内对 $D_{\text{dist}}(n,t)$、$D_{\text{angle}}(n,t)$归一化处理。

$$D_{\text{dist}}(n,t)=\left|\left|(x,y,z)_{n,t}-(x,y,z)_{2,t}\right|\right|^2 \tag{1}$$

$$D_{\text{angle}}(n,t)=\frac{S_{n2}S_{21}}{\left|S_{n2}S_{21}\right|} \tag{2}$$

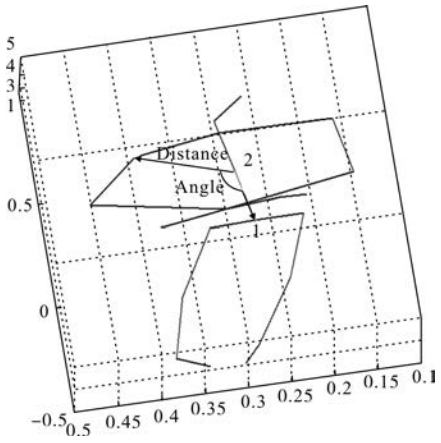

图 1 人体骨架图

Fig.1 Human skeleton

2 时空特征融合深度学习网络

2.1 CNN 和LSTM 网络

CNN 由卷积层 (Convolutional Layer)、池化层(Pooling Layer)构成。卷积层输出特征面的每个神经元与其输入局部连接，通过对应的连接权值与局部输入进行加权求和再加上偏置值，得到该神经元输出[12]。卷积层根据公式(3)选取不同卷积核 W^k 提取

输入的不同特征 h_{ij}^{k} 可表示为：

$$h_{ij}^{k}=\tanh((W^{k}\times x)_{ij}+b_{k}) \tag{3}$$

池化层在卷积层后，完成二次特征提取。池化层的每个神经元对局部接受域聚合，提取概要特征。池化处理不仅降低特征维度，还会防止过拟合。

RNN 用来处理序列数据。网络对前面的信息进行记忆，并应用于当前输出的计算中，即隐藏层的节点之间有连接。隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。LSTM 网络是目前较常用的 RNN，LSTM 网络能够学习长期依赖关系。

LSTM 网络模型由 LSTM 单元组成，定义一个使用 Sigmoid 神经网络和一个按位做乘法的操作 σ 为门。每个 LSTM 单元包含输入门、输出门、遗忘门和记忆门。遗忘门 f_t 计算从当前 LSTM 单元中舍去的信息，如公式(4)所示：

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{4}$$

输入门 i_t 计算当前 LSTM 单元输入信息，见公式(5)：

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{5}$$

$\widetilde{C}_t$ 备选当前单元需要更新内容，见公式(6)：

$$\widetilde{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{6}$$

记忆门 C_t 计算当前 LSTM 单元更新内容，见公式(7)：

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\widetilde{C}_t \tag{7}$$

输出门 o_t 确定 LSTM 单元的哪个部分将输出，见公式(8)：

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{8}$$

对 C_t 通过 tanh 进行处理，得到的输出与 o_t 相乘，公式(9)确定最终输出 h_t。

$$h_t=o_t\cdot\tanh(C_t) \tag{9}$$

2.2 时空特征融合深度学习网络

行为识别属于时空序列处理问题，既需要提取序列的空间特征，又需要考虑序列的时间特征。为此，文中采用时空特征融合深度学习网络学习骨架序列的融合特征。

CNN 卷积层与输入骨架局部连接，选择多组卷积核可提取骨架的空域局部信息，经过 Maxpooling 处理之后，得到骨架空域局部特征极值。LSTM 网络输出不但与当前时刻输入有关，还与之前时刻的输入有关，具有很好的记忆性；利用作用于空域的 LSTM 网络，学习骨架多组空域局部特征之间的关联信息，得到空域的关联性特征表示；利用作用于时域的 LSTM 网络学习序列骨架空间特征的时间关联性，最终得到骨架序列的时空融合特征。

首先，提取骨架的视角不变性特征，分别计算 D_{dist}、D_{angle}，并将 $D_{\text{dist}}(n,t)$、$D_{\text{angle}}(n,t)$ 规则化处理，调整为 $n\times t$，n 为骨架节点数，t 为时间序列长度；其次，保持整个序列的时间先后顺序，利用 CNN 提取局部、短时空间特征，获得空间局部最大值，滤除噪声；然后，空域 LSTM 网络学习各节点之间的空域相关性；最后，时域 LSTM 网络学习序列的时间相关性。网络结构如图 2 所示。

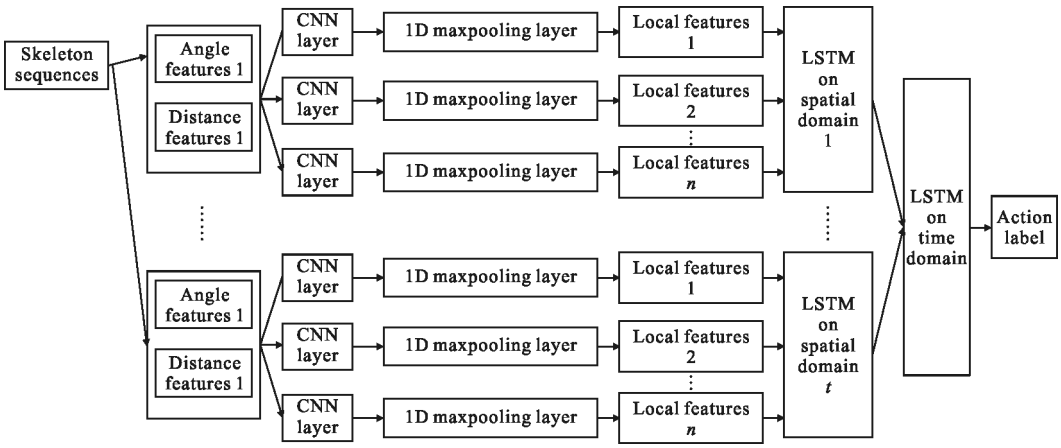

图 2 网络结构图

Fig.2 Net framework

2D CNN 提取局部短时空间特征，卷积核作用于运动特征图上，时间轴尺度足够小，空间轴尺度在 3~5 之间。1D maxpooling 层提取局部空间最大值，获得多维时空特征图。LSTM 作用于空域，学习单幅

骨架各节点之间的空域相关性；得到序列骨架的空间相关性表述。LSTM 网络作用于序列的时间维度，学习序列骨架沿时间变化的特征。最后得到骨架的时空表述。 全连接层、softmax 层作用于时空表述特征之后，学习到序列的行为标签。整个网络特征变换过程如图 3 所示。

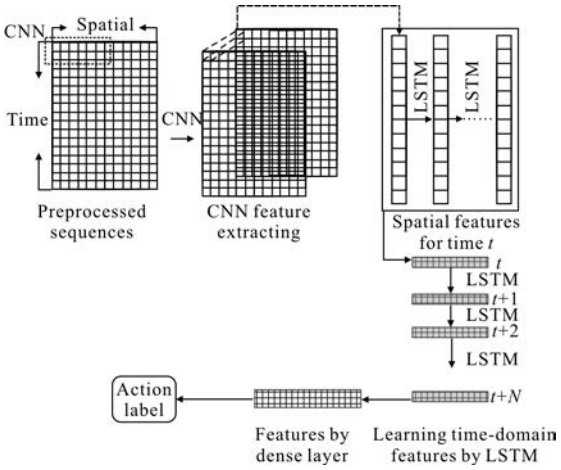

图 3 特征变换图

Fig.3 Feature transformation

3 实验结果及分析

实验采用 NTB-dataset 数据库的骨架库，通过补齐，截短等方式固定骨架序列长度为 50 帧。在每种行为中选取 1 000 组骨架序列作为训练样本，100 组骨架序列作为测试样本。具体网络结构设置如表 1。

表 1 网络结构表

Tab.1 Net framework

Net layer	Name	Size
1	Input layer	
2	Conv 2D CNN	[2,5]
3	Max-pooling	[1,2]
4	Spatial LSTM	64
5	Time domain LSTM	128
6	Full connect layer	64
7	Full connect layer	10
8	Softmax layer(Sigmoid)	-

经过训练后，行为识别网络对于交叉目标分类准确率可达到 83.2%，对数据库中 10 种行为的混淆矩阵(Confusion Matrix)如图 4 所示，对角线元素表示正确识别率。其中站立行为的正确识别率最高，吃饭被错误判为刷牙错误识别率最高。

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
A1	0.76	0	0	0	0	0.02	0	0.16	0	0
A2	0	0.76	0.10	0.14	0	0	0	0	0	0
A3	0.12	006	0.74	0	0	0.04	0	0.02	0	0
A4	0.02	0.14	0.06	0.08	1.10	0	0	0	0.02	0
A5	0.02	0	0.02	0.06	0.82	0	0	0.02	0	0
A6	0.08	0.02	0.08	0	0.04	0.84	0	0.04	0	0
A7	0	0	0	0	0	0.02	0.94	0	0.04	0
A8	0	0	0	0	0.04	0.08	0	0.76	0	0
A9	0	0	0	0	0	0	0	0	0.90	0
A10	0	0.02	0	0	0	0	0.06	0	0.04	1.00

图 4 10 种行为混淆矩阵

Fig.4 Confusion-matrix of 10 action

文中算法与典型算法对比如表 2，典型算法性能数据来源于参考文献[8-11]。由表 2 可见，文中方法在交叉目标和交叉视角测试中都能取得较好的效果，说明文中基于时空特征融合的深度学习网络在行为识别精度上有所改善。

表 2 典型算法效果对比

Tab.2 Comparsion of typical algorithm

Method	Cross subject accuracy	Cross view accuracy
Part-aware LSTM network[8]	62.93%	70.27%
ST-LSTM trust gate[9]	69.2%	77.7%
Hierarchical RNN[5]	59.1%	64%
Clips+CNN+MTLN[10]	79.57%	84.83%
Lie group[11]	50.1%	52.8%
Proposed algorithm	83.2%	85.2%

文中采用归一化之后的视角不变性特征描述骨架，Angle、Distance 视角不变性特征和直角坐标系下 x、y、z 特征对比见图 5。图 5 中表示同一个人在不同位置，不同角度相同行为的节点轨迹(共 10 组)。利用 10 组曲线的峰值个数方差及波动幅度均值参数来衡量以上五组特征，见表 3。由表 3 及图 5 可见，由文中提出的视角不变性特征描述骨架，相同行为的骨架相似性更高，幅度变化更小。说明文中所提出的骨架特征描述方法对于视角变化、位置变化鲁棒性更强。

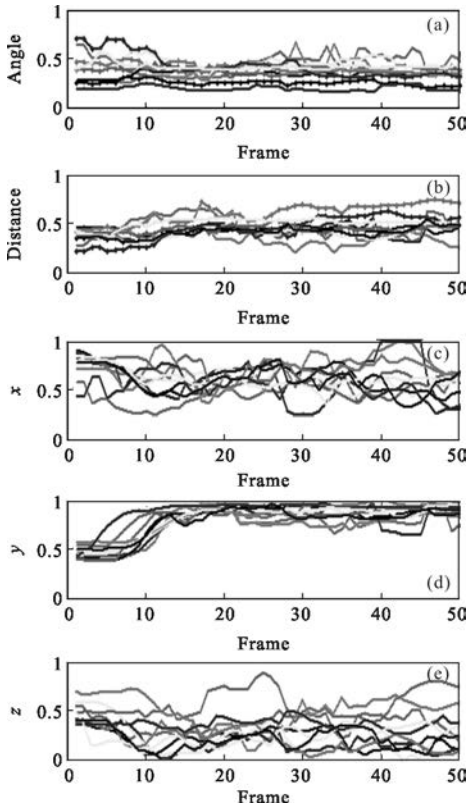

图 5 节点轨迹图

Fig.5 Joint trajectory

表 3 节点轨迹特征对比

Tab.3 Joint trajectory characteristic comparsion

Characteristic	Peak deviation	Range
Angle	3.12	0.238 8
Distance	5.11	0.290 9
x	7.28	0.543 6
y	6.05	0.582 1
z	6.01	0.415 8

4 结 论

提出一种基于深度融合网络的人体骨架行为识别方法,在视角不变性特征提取、时空特征融合网络模型创建方面做了改进和创新，实现具有多视角适应性和较高准确性的行为识别方法，比之目前基于骨架行为识别领域典型算法,文中识别准确率更高,考虑成像过程中的视角变化。

该方法在多视角、多参与人、多位置数据库上取得了一定的研究成果,对于不同人、不同位置、不同角度的同一行为识别准确率更高,对于不同人、不同位置、不同角度的不同行为区分性更强。

文中还存在如下缺点，针对于易混淆行为如(吃饭、喝水等)识别率还有待提高,下一步将继续调整网络结构,使其在空间,时间上能更准确地提取有代表性的差异性特征,提高易混行为的识别准确率。

参考文献:

[1] Wang Jiang, Liu Zicheng. Mining actionlet ensemble for action recognition with depth cameras [C]//IEEE Conference on Computer Vision and Pattern Recognition,2012: 1290–1297.

[2] Luvizon D C, Tabia H. Learning features combination for human action recognition from skeleton sequences[J]. ***Pattern Recognition Letters***, 2017, 99(11): 13–20.

[3] Ji XiaoPeng, Cheng Jun. The spatial laplacian and temporal energy pyramid representation for human action recognition using depth sequence [J]. ***Knowledge–Based System***, 2017, 122: 64–74.

[4] Zhang Pengfei, Lan Cuiling. View adaptive recurrent neural networks for high performance human action recognition from skeleton data [C]//ICCV 2017. International Conference on Computer Vision, 2017: 2136–2145.

[5] Du Yong, Wang Wei, Wang Liang. Hierarchical recurrent neural network for skeleton based action recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1110–1118.

[6] Vivek Veeriah, Naifan Zhuang. Differential recurrent neural networks for action recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4041–4049.

[7] Zhu Wentao, Lan Cuiling. Co–occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]//AAAI, 2016: 3697–3704.

[8] Amir Shahroudy, Liu Jun. NTU RGB+D: A large scale dataset for 3D human activity analysis [C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010–1019.

[9] Liu Jun, Amir Shahroudy. Spatio–temporal LSTM with trust gates for 3D human action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[10] Liu Jun, Wang Gang. Skeleton based human action recognition with glo bal context–aware attention LSTM networks [C]//IEEE Conference on Computer Vision and

Pattern Recognition, 2017: 3671-3680.

[11] Huang Zhiwu, Wan Chengde. Deep learning on Lie groups for skeleton-based action recognitio[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1243-1252.

[12] Zhou Feiyan, Jin Linpeng, Dong Jun. Review of convolutional neural networks [J]. ***Chinese Journal of Computers***, 2017, 40(6): 1229-1250. (in Chinese)
周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1250.

[13] Luo Haibo, Xu Lingyun, Hui Bin, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)
罗海波, 许凌云, 惠斌, 等. 基于深度学习的目标跟踪方法研究现状与展望 [J]. 红外与激光工程, 2017, 46(5): 0502002.

[14] Shao Chunyan, Ding Qinghai, Luo Haibo, et al. Target tracking using high-dimension data clustering [J]. ***Infrared and Laser Engineering***, 2016, 45(4): 0428002. (in Chinese)
绍春艳, 丁庆海, 罗海波, 等. 采用高维数据聚类的目标跟踪[J]. 红外与激光工程, 2016, 45(4): 0428002.