Software Engineering and Applications 软件工程与应用, 2023, 12(2), 198-208
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.122020

Hans汉斯

通过轻量级人体动作识别实现老年人安全实时监控

吴嘉轩*，申　奇

沈阳理工大学，信息科学与工程学院，辽宁 沈阳

收稿日期：2023年2月20日；录用日期：2023年3月24日；发布日期：2023年3月31日

摘　要

本文旨在探讨如何通过轻量级人体动作识别技术实现老年人安全实时监控。老年人安全问题一直是一个备受关注的问题，尤其是在现代化社会，许多老年人居住在独自生活的环境中，缺乏及时的照顾和照料。为了及时监控老年人的安全情况，本文提出了一种面向老年人实时监控系统的轻量级人体动作识别算法。本文使用卷积神经网络与具有长短时记忆的循环神经网络的组合架构，结合了卷积神经网络在图像特征提取过程中的优越性能，以及长短时记忆神经网络对时序数据处理过程的特点，并进行了详细的实验验证，实验结果表明本文提出的轻量级人体动作识别算法具有显著的优势。

关键词

老年人安全监控，轻量级人体动作识别，长短时记忆循环神经网络，动作识别，智能监控

Through Lightweight Human Movement Recognition to Realize the Security of the Elderly Real-Time Monitoring

Jiaxuan Wu*, Qi Shen

School of Information Science and Engineering, Shenyang Ligong University, Shenyang Liaoning

Received: Feb. 20th, 2023; accepted: Mar. 24th, 2023; published: Mar. 31st, 2023

Abstract

The purpose of this paper is to explore how to achieve real-time monitoring of elderly safety

*通讯作者。

文章引用: 吴嘉轩, 申奇. 通过轻量级人体动作识别实现老年人安全实时监控[J]. 软件工程与应用, 2023, 12(2): 198-208.
DOI: 10.12677/sea.2023.122020

through lightweight human motion recognition technology. The safety of the elderly has always been a matter of great concern, especially in modern society where many elderly people live in solitary environments and lack timely care and attention. In order to monitor the safety of the elderly in a timely manner, this paper proposes a lightweight human action recognition algorithm for a real-time monitoring system for the elderly. This paper uses a combined architecture of convolutional neural network and recurrent neural network with long and short term memory, combines the superior performance of convolutional neural network in image feature extraction process and the characteristics of long and short term memory neural network for temporal data processing process, and conducts detailed experimental validation, the experimental results show that the lightweight human action recognition algorithm proposed in this paper has significant advantages.

Keywords

Security Monitoring of the Elderly, Lightweight Human Movement Recognition, Long Short Term Memory Recurrent Neural Network, Action Recognition, Intelligent Monitoring

1. 引言

随着社会的快速发展，老年人口日益增长，老年人的健康和安全问题日益引起社会关注。在现代社会，越来越多的老年人居住在独自生活的环境中，缺乏及时的照顾和照料，这使得老年人的安全问题成为一个备受关注的问题[1]。为了解决老年人的安全问题，许多技术方案被提出，其中最有前途的技术之一是轻量级人体动作识别技术[2]。

从国内社会来看，新中国建立以来，尤其是改革开放之后，我国的老年人人口数量称逐年上升之势。联合国资料显示，我国已于 2000 年步入老龄化社会，且已呈现出“未富先老”的姿态，人口老龄化超前于社会的发展、社会保障制度不健全、家庭模式的改变也导致家庭养老功能弱化等一系列的现实向我们展示，加强对老年人的权益保护已迫在眉睫[3]。但是截至到目前为止，我国尚未在法学领域对老龄化社会的到来带来的为问题做出相应的应对措施，仅一部《老年人权益保障法》(2012 年 12 月 28 日通过，2013 年 7 月 1 日实施)已经不足以全面的保护老年人的合法权益；我国《民法通则》和《民通意见》对于成年人监护制度的规定也只有几条，且只包含了欠缺民事行为能力的精神病人，立法理念落后，对于老年人的监护问题没有涉及[4]。所以必须利用法律手段，将老年人纳入成年人监护体制中，设立老年人监护制度，为老年人尤其是高龄老人的人身健康和财产管理设立保护措施，弥补老年人尤其是高龄老人意思能力或行为能力的缺陷，弥补社会保障体系不健全对老年人的人身和财产权益保护带来的不足，更好的应对人口老龄化给中国带来的挑战。从国际社会上看，近些年来人权保护的呼声日益高涨，对比社会交易安全和第三人的利益价值，人的自由和尊严价值显得更为重要[5]。“维护本人的生活正常化”和“尊重本人的自我决定权”理念为多数国家接受，并在本国的关于成年人监护立法的制定和修改中被纳入，大陆法系的德国、日本，英美法系的英国、美国等，尤其是日本，其修改本国的成年后见制度的背景与我国相识，都为我们以后修改我国的监护制度，建立我国的老年人监护制度提供了有益的借鉴[6]。

本文将介绍如何利用轻量级人体动作识别技术实现老年人的安全监控。首先，本文将介绍人体动作识别技术的基本原理和算法。其次，本文将介绍如何利用轻量级算法实现人体动作识别，以及如何将该算法应用于老年人的安全监控。最后，本文将讨论轻量级人体动作识别技术在老年人安全监控领域的应用前景。

2. 人体动作识别技术的基本原理

人体动作识别算法是一种基于计算机视觉技术的研究方向。Turaga 等人[7]认为，“人体动作”的特点是通常由一个人执行的简单运动模式，而“活动”则更为复杂，涉及少数人之间的协调行动，并研究了识别人类行动和活动的主要方法。Poppe [8]集中讨论了图像表示和动作分类方法，Weinland 等人[9]的主要研究内容也集中在行动表示和分类的方法上。Popoola 和 Wang [10]的研究内容重点是用于监控应用的上下文异常人类行为检测。Ke 等人[11]研究了静态和移动摄像机的人类活动识别方法，涵盖了许多问题，如特征提取、表征算法、人体动作检测和分类。Aggarwal 和 Xia [12]对基于 3D 数据的人类动作识别进行了研究，特别是对使用消费者 3D 传感器如 Kinect [13]传感器获得的 RGB 和深度信息进行了研究。

早期的人体动作识别算法主要基于传统的计算机视觉技术，如特征提取、分类器等。近年来，随着深度学习技术的发展，人体动作识别算法也逐渐向基于深度学习的算法发展。Cheng 等人[7]使用基于方法的分类法研究了人类动作识别的方法，其中所有的方法被分为两类：单阶段识别方法和双阶段识别方法。此外，Vrigkas 等人[14]将人类活动识别方法分为两个主要类别，包括“单模态”和“多模态”。然后，他们分别对这两类方法进行了研究。Subetha 和 Chitrakala [15]的研究内容主要集中在人类活动识别和人－物交互方法上。Presti 等人[16]对基于三维骨架的人类动作识别进行了研究，总结了相关的技术方法，人类动作识别中常见的方法主要是使用手工设计的局部特征，如 HOG/HOF [17] [18]，SIFT [19]，或 SURF [20]。此外，这些方法还在视频处理中进行了扩展，以获得更强的鲁棒性，如 Cuboids [21]和 HOG3D [22]。

人体动作识别算法的基本原理是通过对人体姿态和动作的分析，提取出与动作相关的特征，然后将特征输入分类器进行分类。其主要包括以下几个步骤：

1) 数据采集和预处理：采集人体动作视频数据，并对数据进行预处理，如去除噪声、归一化等；

2) 特征提取：通过对视频数据进行分析，提取与动作相关的特征。传统的特征提取方法主要包括手工设计特征和基于卷积神经网络的特征提取方法。手工设计特征主要是通过对视频中的像素进行处理，提取出与动作相关的信息，如人体的轮廓、关节位置等。而基于卷积神经网络的特征提取方法是通过训练网络，自动学习特征；

3) 分类器：将提取出的特征输入到分类器中进行分类，常用的分类器包括支持向量机、随机森林和深度神经网络等；

4) 动作识别：根据分类器的输出结果，识别出视频中的动作。

3. 轻量级人体动作识别技术

3.1. 循环神经网络算法原理

循环神经网络(Rerrent Neural Network, RNN)在推算视频中各种动作的复杂动态的过程中具有显著优势[23]，因为它的结构允许存储和访问时间序列的长范围背景信息。RNN 和多层感知器的主要区别在于循环连接的存在，如图 1 所示。

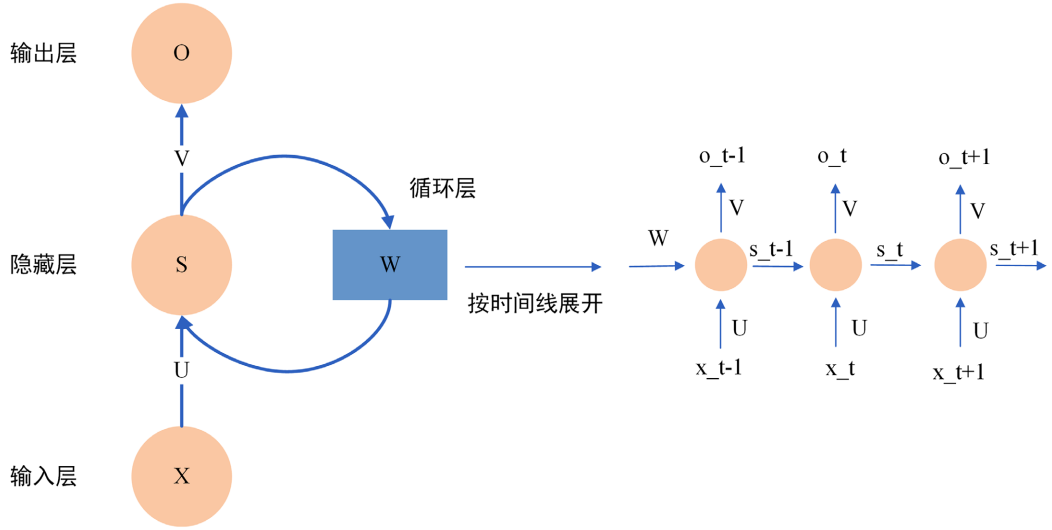

Figure 1. Structure of RNN
图 1. RNN 结构图

在图 1 中，x_t 表示第 t 步的输入，s_t 表示第 t 步隐藏层的状态，是网络的记忆单元，o_t 是第 t 步的输出。由图 1 可知，RNN 可以学习从以前整个历史时刻的输入映射到每个输出节点。然而，由于“梯度弥散问题”，它们的训练非常困难[24]。长短期记忆(Long Short Term Memory, LSTM)方法[25]已被提出来解决这些问题，其结构如图 2 所示。

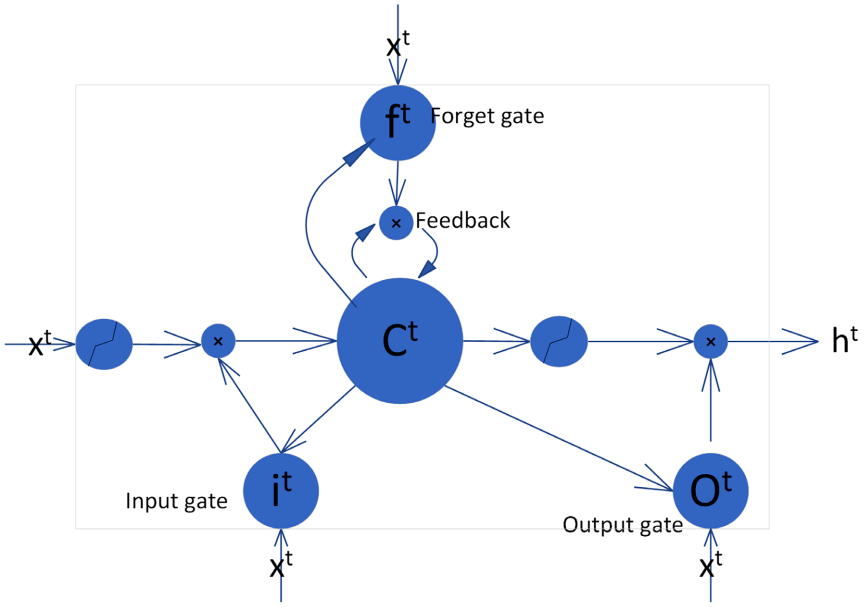

Figure 2. Structure of LSTM
图 2. LSTM 结构图

图 2 描述了 LSTM 的结构和它的信息流，其中包括一个输入门 i^t、一个输出门 o^t、一个遗忘门 f^t、一个输出状态 h^t 和一个存储单元状态 c^t。信息流由下述公式描述。

$$i^t = \sigma\left(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i\right)$$

$$f^t = \sigma\left(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f\right)$$

$$c^t = f^t c^{t-1} + i^t \tanh\left(W_{xc}x^t + \dot{W}_{hc}h^{t-1} + \dot{b}_c\right)$$

$$o^t = \sigma\left(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^t + b_o\right)$$

$$h^t = o^t \tanh\left(c^t\right)$$

上式中，σ 表示 sigmoid 激活函数，x^t 表示网络在 t 时刻的输入，所有矩阵 W 表示单元之间的连接权重，$\odot$表示两个矩阵间对应位置的元素进行乘积。通过用两个独立的隐藏层实现双向处理数据的双向 RNNs 不仅能够利用以前的上下文，而且还能够利用未来的上下文。通过用 LSTM 单元取代双向 RNNs 架构中的非线性单元，我们可以得到双向 LSTM，如图 3 所示。

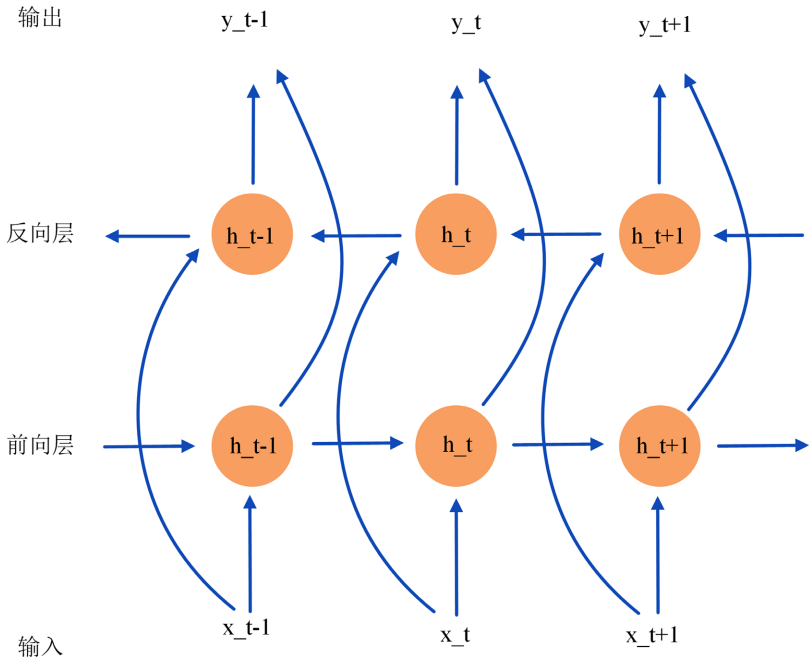

Figure 3. Structure of a Bidirectional-LSTM
图 3. 双向 LSTM 结构图

在图 3 中，每个圆形节点分别代表 LSTM 单元。

3.2. 卷积神经网络与长短时记忆神经网络的组合架构

RNN-LSTM 的主要优势是能够对时间序列的长期背景信息进行建模。这一优势使 RNN-LSTM 成为包括人类动作视觉信息在内的时间序列数据的最佳序列学习器之一[26]。而 CNNs 在图像数据的特征提取性能具有显著的优势，因此，本文采用 CNNs 与 LSTM 结合的组合架构来实现人体动作识别，提高老年人安全实时监控的可靠性与准确性，本文提出的 CNN-LSTM 的结构如图 4 所示。

由图 4 可知，CNN-LSTM 网络的输入数据为 2D 图像，RGB 图像的大小为 256 × 256。经过 CNN 进行特征提取，将提取到的视觉特征信息展开为二维序列，经过 LSTM 对人体行动序列数据的计算后送入注意力机制模块中，由注意力机制模块对人体行动序列数据进行注意力打分，使得模型对重要信息更加关注，同时对冗余信息进行剔除。最后将包含注意力得分的人体行动序列数据送入全连接层，即可得到人体行动的类别识别。其中，各个阶段的网络参数如表 1 所示。

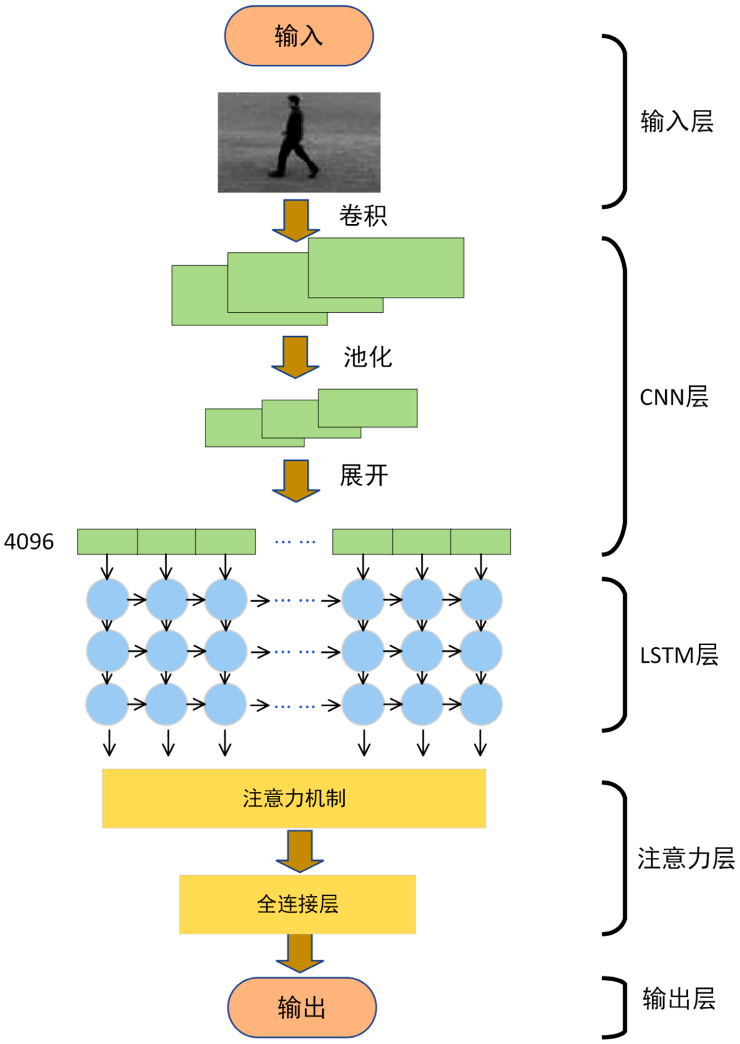

Figure 4. Structure of a Bidirectional-LSTM
图 4. 双向 LSTM 结构图

Table 1. CNN-LSTM network parameters description
表 1. CNN-LSTM 网络参数说明

阶段	层数	参数	输出数据维度
视觉特征提取(卷积)	Conv2d_1	kernel_size = [7, 7, 3, 16]，stride = 2	[126, 126, 16]
	MaxPool_1	kernel_size = [5, 5], stride = 2	[62, 62, 16]
	Conv2d_2	kernel_size = [3, 3, 16, 32]，stride = 1	[60, 60, 32]
	MaxPool_2	kernel_size = [5, 5], stride = 2	[29, 29, 32]
	Conv2d_3	kernel_size = [3, 3, 32, 32]，stride = 1	[27, 27, 32]
	MaxPool_3	kernel_size = [5, 5], stride = 2	[12, 12, 32]
	Conv2d_4	kernel_size = [3, 3, 32, 64], stride = 1	[10, 10, 64]
	MaxPool_4	kernel_size = [5, 5], stride = 2	[4, 4, 64]

Continued

数据展开	Fliping	-	[1, 4096]
LSTM	LSTM_1	kernel_size = [4096, 1024]	[1, 1024]
	LSTM_2	kernel_size = [1024, 256]	[1, 256]
	LSTM_3	kernel_size = [256, 128]	[1, 128]
注意力模块	Attention_1	kernel_size = [128, 16]	[1, 16]
	Attention_2	kernel_size = [16, 128]	[1, 128]
全连接	FC_1	kernel_size = [128, 32]	[1, 32]
	FC_2	kernel_size = [32, 1]	[1, 1]

由表 1 可知，整体网络共包含 11 层，其中池化层及数据展开层因不含可训练的权重参数，不单独算作一个隐藏层。CNN-LSTM 网络的最终输出为人体行动的类别。

4. 应用轻量级人体动作识别技术实现老年人安全实时监控

老年人安全问题一直是一个备受关注的问题，尤其是在现代化社会，许多老年人居住在独自生活的环境中，缺乏及时的照顾和照料。通过使用轻量级人体动作识别技术，可以实现对老年人的安全监控，减少老年人发生意外事件的概率。

老年人安全实时监控系统的设计如图 5 所示：

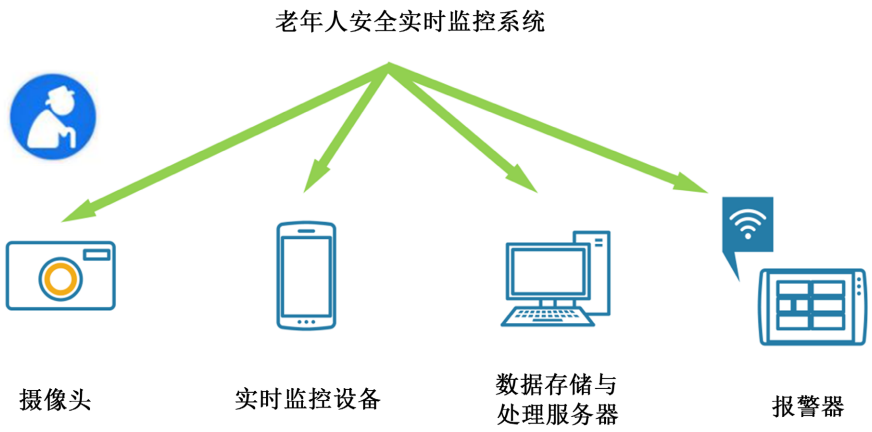

Figure 5. Real-time monitoring system composition for the safety of the elderly
图 5. 老年人安全实时监控系统构成

本文提出的轻量级人体动作识别算法采用 CNN 与 LSTM 级联的组合架构，使用的硬件平台信息如表 2 所示。

Table 2. Hardware platform description
表 2. 硬件平台参数说明

序号	硬件名称	型号	参数说明
1	内存	G.SKILL DDR56000 16GB*2	高性能高稳定性，支持 EXPO 急速超频，传输速度高达 6000 MHz

Continued

2	CPU	Inter Core i9-12900KS	16 核心 24 线程，8 个性能核与 8 个能效核，针对多线程性能进行了优化
3	显卡	NVIDIA RTX 3090	NVIDIA Ampere 架构，同时引入 NVIDIA DLSS(深度学习超级采样)助力性能提升，显存 24GB

本文实验中主要涉及的软件平台及版本说明见表 3。

Table 3. Software platform and version description
表 3. 软件平台与版本说明

序号	软件平台	版本	简单介绍
1	计算机视觉处理库	Opencv-python 4.5.5.64	支持大量与计算机视觉和机器学习相关的算法，对开发人员开源，支持多种编程语言，可在多种平台上开发使用
2	编译语言	Python 3.7.6	作为一款面向对象的解释型语言，其简洁高效且可被移植到不同平台上
3	操作系统	Ubuntu 18.04	作为 Linux 系统中的一员，其完全开源且对开发人员友好
4	CUDA 与 CuDNN	10.0；7.6.5	作为深度神经网络的 GPU 加速库，可加速深度学习训练
5	深度学习框架	Pytorch 1.7.1	深度学习算法的开源框架，集成了卷积、池化等操作，方便快速构建网络模型

本文提出的人体动作识别算法在训练过程中使用的批大小为 16，梯度优化方法使用 Adam 优化器，训练步数为 10,000 步，在不同学习率的情况下，模型的收敛情况如图 6 所示。

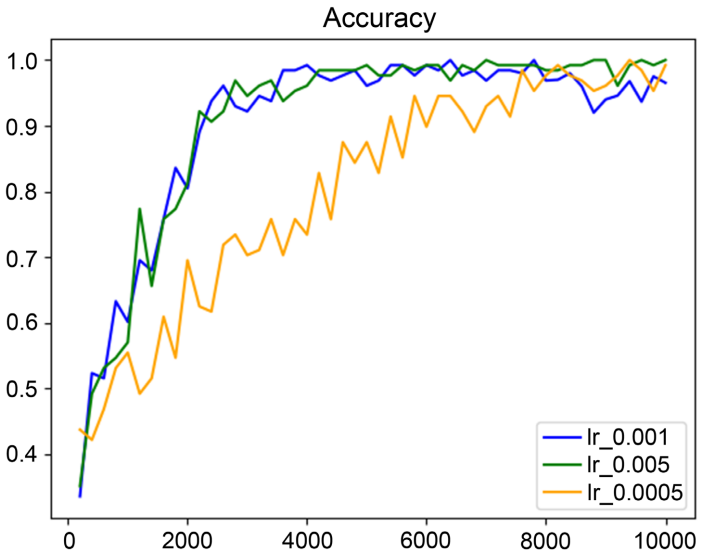

Figure 6. The convergence of the model with different learning rates
图 6. 使用不同的学习率时模型的收敛情况

由图 6 可知，当学习率设置为 0.0005 时模型的收敛速度显著降低，学习率设置为 0.001 与 0.005 时模型的收敛速度相差不大，但学习率设置为 0.005 时效果略佳。为验证本文提出的人体动作识别算法的有效性，与时空注意力时间段网络(Spatial-Temporal Attention Temporal Segment Network, STA-TSN) [27]、

多模态视觉 Transformer (Multi-Modal Video Transformer, MM-ViT) [28]和时空交叉注意力 Transformer (Spatio-Temporal Cross Attention Transformer, STAR-Transformer) [29]算法在公开数据集 UCF101 [30]上进行了对比测试，训练过程中的准确率变化如图 7 所示。

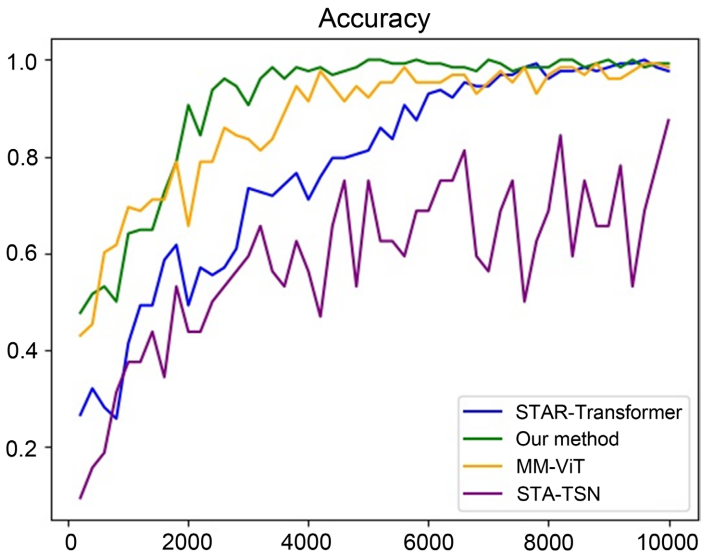

Figure 7. Accuracy curves of different models using the same environment and parameters
图 7. 使用相同的环境及参数时不同模型的准确率变化曲线

由图 7 可知，STA-TSN 算法在训练阶段的检测效果并不理想，本文提出的人体动作识别算法的收敛速度最快，且与 STAR-Transformer、MM-ViT 算法的识别准确率均达到较为理想的效果。四种算法在测试集中的检测结果如表 4 所示。

Table 4. Detection speed and average accuracy of different models on the test set
表 4. 不同模型在测试集上的检测速度及平均准确率

算法名称	每帧检测速度(s)	参数量(M)	测试集中的识别准确率
STA-TSN [27]	0.243178	60	0.89
STAR-Transformer [29]	0.565114	11	0.93
MM-ViT [28]	0.480212	24	0.95
Our method	0.149286	5	0.97

由表 4 可知，本文提出的人体动作检测算法与其它主流算法相比检测速度最快、参数量最少、检测准确率最高，验证了本文提出的人体动作识别算法的可靠性与准确性。

5. 结束语

随着人工智能技术的不断发展，轻量级人体动作识别技术成为实现低功耗、低带宽设备上实时人体动作识别的有效手段。本文采用了 CNN 与 LSTM 级联的组合架构来实现轻量级人体动作识别算法，与其它主流算法相比具有显著优势，验证了本文提出的人体动作识别算法的可靠性与准确性，为老年人安全实时监控系统奠定了基础。

虽然轻量级人体动作识别技术在资源受限的设备上可以实现实时人体动作识别，但是其分类精度仍

需要进一步提高，特别是在复杂场景下的分类精度还有待改进。因此，未来的研究可以重点关注轻量级算法的性能优化和多模态数据的融合，以提高轻量级人体动作识别技术在实际应用中的性能和精度。此外，在老年人安全监控系统的应用中，还需要考虑隐私保护等问题，以避免因技术的应用而侵犯老年人的隐私权。

总之，轻量级人体动作识别技术在老年人安全监控、健康管理、体育训练等领域都有广泛的应用前景。通过对人体动作识别算法的研究和优化，我们可以在低功耗、低带宽等受限资源的设备上实现实时人体动作识别，进一步推动智能化技术的发展，促进智能化生活的实现。

参考文献

[1] 潘泽瀚, 吴连霞, 卓冲, 等. 2010-2020 年中国老年人口健康水平空间格局演变及其影响因素[J]. 地理学报, 2022, 77(12): 3072-3089.

[2] 冉宪宇. 基于图像处理技术的智能化人体行为识别模型研究[J]. 微型电脑应用, 2022, 38(10): 175-178.

[3] 穆光宗, 张团. 我国人口老龄化的发展趋势及其战略应对[J]. 华中师范大学学报(人文社会科学版), 2011, 50(5): 29-36.

[4] 张文范. 我国人口老龄化与战略性选择[J]. 城市规划, 2001, 26(2): 68-72.

[5] 李姝婧, 翟振武. 人口老龄化对中国产业结构演进的影响[J]. 人口学刊, 2022, 44(6): 38-52.

[6] 倪宣明, 贺英洁, 武康平, 等. 人口老龄化, 移民与经济增长[J]. 系统工程理论与实践, 2022, 42(1): 1-12.

[7] Turaga, P., Chellappa, R., Subrahmanian, V.S. and Udrea, O. (2008) Machine Recognition of Human Activities: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, **18**, 1473-1488. https://doi.org/10.1109/TCSVT.2008.2005594

[8] Poppe, R. (2010) A Survey on Vision-Based Human Action Recognition. *Image and Vision Computing*, **28**, 976-990. https://doi.org/10.1016/j.imavis.2009.11.014

[9] Weinland, D., Ronfard, R. and Boyer, E. (2011) A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition. *Computer Vision and Image Understanding*, **115**, 224-241. https://doi.org/10.1016/j.cviu.2010.10.002

[10] Popoola, O.P. and Wang, K. (2012) Video-Based Abnormal Human Behavior Recognition-A Review. *IEEE Transactions on Systems*, *Man*, and *Cybernetics*, *Part C* (*Applications and Reviews*), **42**, 865-878. https://doi.org/10.1109/TSMCC.2011.2178594

[11] Ke, S.R., Thuc, H.L.U., Lee, Y.J., *et al*. (2013) A Review on Video-Based Human Activity Recognition. *Computers*, **2**, 88-131. https://doi.org/10.3390/computers2020088

[12] Aggarwal, J.K. and Xia, L. (2014) Human Activity Recognition from 3D Data: A Review. *Pattern Recognition Letters*, **48**, 70-80. https://doi.org/10.1016/j.patrec.2014.04.011

[13] Zhang, Z. (2012) Microsoft Kinect Sensor and Its Effect. *IEEE Multimedia*, **19**, 4-10. https://doi.org/10.1109/MMUL.2012.24

[14] Vrigkas, M., Nikou, C. and Kakadiaris, I.A. (2015) A Review of Human Activity Recognition Methods. *Frontiers in Robotics and AI*, **2**, Article 28. https://doi.org/10.3389/frobt.2015.00028

[15] Subetha, T. and Chitrakala, S. (2016) A Survey on Human Activity Recognition from Videos. *Proceedings of* 2016 *International Conference on Information Communication and Embedded Systems* (*ICICES*), Chennai, 25-26 February 2016, 1-7. https://doi.org/10.1109/ICICES.2016.7518920

[16] Presti, L.L. and La Cascia, M. (2016) 3D Skeleton-Based Human Action Classification: A Survey. *Pattern Recognition*, **53**, 130-147. https://doi.org/10.1016/j.patcog.2015.11.019

[17] Dalal, N. and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. *Proceedings of* 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*'05), Diego, 20-25 June 2005, 886-893.

[18] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. (2008) Learning Realistic Human Actions from Movies. *Proceedings of* 2008 *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 23-28 June 2008, 1-8. https://doi.org/10.1109/CVPR.2008.4587756

[19] Lowe, D.G. (1999) Object Recognition from Local Scale-Invariant Features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, 20-27 September 1999, 1150-1157.

https://doi.org/10.1109/ICCV.1999.790410

[20] Bay, H., Tuytelaars, T. and Van Gool, L. (2006) Surf: Speeded up Robust Features. In: Leonardis, A., Bischof, H. and Pinz, A., Eds., *ECCV* 2006: *Computer Vision-ECCV* 2006, *Lecture Notes in Computer Science*, Vol. 3951, Springer, Berlin, 404-417. https://doi.org/10.1007/11744023_32

[21] Dollár, P., Rabaud, V., Cottrell, G., et al. (2005) Behavior Recognition via sparse Spatio-Temporal Features. *Proceedings of* 2005 *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, 15-16 October 2005, 65-72.

[22] Klaser, A., Marszałek, M. and Schmid, C. (2008) A Spatio-Temporal Descriptor Based on 3d-Gradients. *Proceedings of BMVC* 2008-19*th British Machine Vision Conference*, Leeds, September 2008, 99.1-99.10. https://doi.org/10.5244/C.22.99

[23] Zaremba, W., Sutskever, I. and Vinyals, O. (2014) Recurrent Neural Network Regularization. (Preprint)

[24] Bengio, Y., Simard, P. and Frasconi, P. (1994) Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks*, **5**, 157-166. https://doi.org/10.1109/72.279181

[25] Graves, A. and Graves, A. (2012) Long Short-Term Memory. In: Graves, A., Ed., *Supervised Sequence Labelling with Recurrent Neural Networks*, *Studies in Computational Intelligence*, Vol. 385, Springer, Berlin, 37-45. https://doi.org/10.1007/978-3-642-24797-2_4

[26] Misgar, M.M., Mushtaq, F., Khurana, S.S. and Kumar, M. (2023) Recognition of Offline Handwritten Urdu Characters Using RNN and LSTM Models. *Multimedia Tools and Applications*, **82**, 2053-2076. https://doi.org/10.1007/s11042-022-13320-1

[27] Yang, G., Yang, Y., Lu, Z., et al. (2022) STA-TSN: Spatial-Temporal Attention Temporal Segment Network for Action Recognition in Video. *PLOS ONE*, **17**, e0265115. https://doi.org/10.1371/journal.pone.0265115

[28] Chen, J. and Ho, C.M. (2022) MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 3-8 January 2022, 1910-1921. https://doi.org/10.1109/WACV56688.2023.00333

[29] Ahn, D., Kim, S., Hong, H., et al. (2023) STAR-Transformer: A Spatio-Temporal Cross Attention Transformer for Human Action Recognition. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 2-7 January 2023, 3319-3328. https://doi.org/10.1109/WACV56688.2023.00333

[30] Soomro, K., Zamir, A.R. and Shah, M. (2012) UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. (Preprint)