文章编号：2096-1618(2023)04-0387-05

基于双流网络和迁移学习的红外人体行为识别

程 睿[1]， 陈 冲[1,2]， 黄瑞丰[1,2]， 郭 坤[3]

(1. 安徽建筑大学电子与信息工程学院，安徽 合肥 230601；2. 安徽省古建筑智能感知与高维建模国际联合研究中心，安徽 合肥 230601；3. 江苏省句容市边城镇徐家边技术维护室，江苏 镇江 212416)

摘要：针对传统预训练模型无法充分利用红外人体行为数据时间信息的问题，提出一种基于双流网络和迁移学习的红外人体行为识别方法。首先对原始视频提取运动历史图和光流图，利用滑动窗口的思想进行堆叠处理；其次根据可见光行为数据和红外行为数据之间的相似性，设计双流预训练网络，并通过迁移学习，将可见光行为数据预训练的网络模型参数共享给红外人体行为识别网络模型，以此提取红外行为数据的特征；然后，将提取的特征输入至双流网络中，进一步提取红外人体行为信息，在特征融合处采用并联特征融合方式替换 Softmax 融合方式；最后，使用支持向量机对融合的特征进行人体行为分类。实验结果表明，所提方法在 NTU RGB+D 数据集上达到 78.52% 的准确率，具有较好的分类效果。

关 键 词：人体行为识别；迁移学习；双流网络；红外；可见光

中图分类号：TP391.4 **文献标志码**：A

doi：10.16836/j.cnki.jcuit.2023.04.002

0 引言

人体行为识别是计算机视觉重要研究领域之一，一直受到国内外学者的广泛研究和关注，其任务是对数据中的人体行为进行分类。当前，许多智能服务都以高精度的人体行为识别为基础，如视频监控、智能家居、虚拟现实和人机交互等[1]。因此，人体行为识别研究具有重要的研究意义和广泛的应用价值。

双流网络[2](two-Stream convolutional neural network，TCN)在人体行为识别中一直都表现优异，原因在于不仅充分利用了人体行为数据的空间信息，也充分利用了时间信息。为提高双流网络的检测效果，国内外学者针对双流网络的改进方法开展了大量研究，如时域分割网络[3](temporal segment networks，TSN)、双流网络+长短时记忆网络[4](two-stream +long short term memory，TCN+LSTM)、三流网络[5]等。双流网络在红外人体行为识别中同样表现突出。Gao 等[6]采用双流网络验证其自建红外数据集：InfAR 数据集，经过实验分析，得出在红外人体行为识别中时间信息优于空间信息的结论。为更充分利用时间信息，Liu 等[7]提出基于全局时域的三流网络，通过综合考虑局部时域信息、全局时域信息和空间信息后，从红外动作数据中提取鲁棒的判别特征。

在可见光人体行为识别的过程中，国内外学者将行为数据先输入预训练网络中，初步去除数据中的冗余信息，减小训练复杂度。Romaissa 等[8]将预训练模型加在双流网络之前，以便提取表征能力更强的行为特征。由于目前的预训练模型均在 ImageNet 数据集上进行预训练，而 ImageNet 数据集中仅包含空间信息，导致传统预训练模型无法充分提取时域信息。为此，本文将结合人体行为的特点，设计和训练新的预训练模型，使其针对时域信息更新参数，从源头减少预训练模型对分类结果造成的影响。

双流网络可以较好地提高人体行为识别的准确率，迁移学习能够将不同领域的信息进行共享。因此，本文通过结合双流网络和迁移学习的优点，提出一种基于双流网络和迁移学习的红外人体行为识别方法。为解决传统预训练模型在红外人体行为识别中的问题，使用可见光数据预训练双流结构模型，使其作为预训练模型，可以充分利用红外人体行为数据的时间信息。

1 方法

1.1 总体架构设计

本文所提出方法的总体框架如图 1 所示。框架分为 4 个部分：堆叠运动历史图(motion history image，MHI)和堆叠光流图(optical flow，OF)生成、MHI 和 OF 特征提取、MHI 和 OF 融合特征提取、支持向量机(support vector machine，SVM)分类。

收稿日期：2022-10-19

基金项目：国家自然科学基金资助项目(62001004)；安徽省高校省级自然科学研究项目(KJ2019A0768)

通信作者：陈冲. E-mail：shchshch@ustc.edu.cn

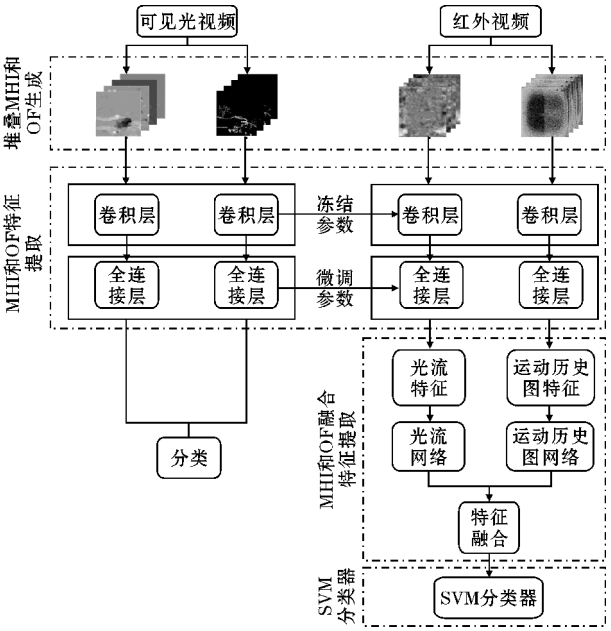

图 1 网络整体结构图

首先,对原始输入数据提取相应的运动历史图和光流图,并对其进行堆叠处理。从一段视频中提取的图片数量为 K 张,表示为 $\{S_1,S_2,S_3,\cdots,S_k\}$。根据滑动窗口的思想,步长设置为 1,形成 $(K-3)$ 组堆叠数据,表示为

$\{\{S_1,S_2,S_3,S_4\},\{S_2,S_3,S_4,S_5\},\cdots,\{S_{k-3},S_{k-2},S_{k-1},S_k\}\}$

其次,使用可见光数据训练双流预训练网络,并迁移至红外数据。以 UCF101 数据集预训练的双流网络模型为基础,冻结所有卷积层的参数,并微调全连接层,达到迁移学习的目标,以此提取出 MHI 特征和 OF 特征。

然后,将提取的特征分别输入至双流网络。卷积层进一步充分提取红外数据中的行为特征,通过全连接层进行并联特征融合操作,形成融合特征。

最后,将 MHI 和 OF 融合特征输入到 SVM 分类器中,对红外人体行为进行分类。

1.2 双流预训练网络模型

迁移学习是利用数据、任务或者模型之间的相似性,将旧领域学习过的模型和知识应用到新的领域[9]。可见光和红外人体行为数据包含的信息大致相似,但是由于可见光成像原理的问题,包含更多的冗余信息,导致网络的泛化性能受到影响。红外数据虽然可以有效地解决信息冗余的问题,但是红外人体行为数据量少、获取难度大,亦难获取泛化性能强的网络模型。为利用可见光行为数据和红外行为数据在表现形式和任务的相似性,同时弥补红外数据的相关缺点,本文参考双流网络的结构,设计双流预训练网络模型。具体结构如图 2 所示。

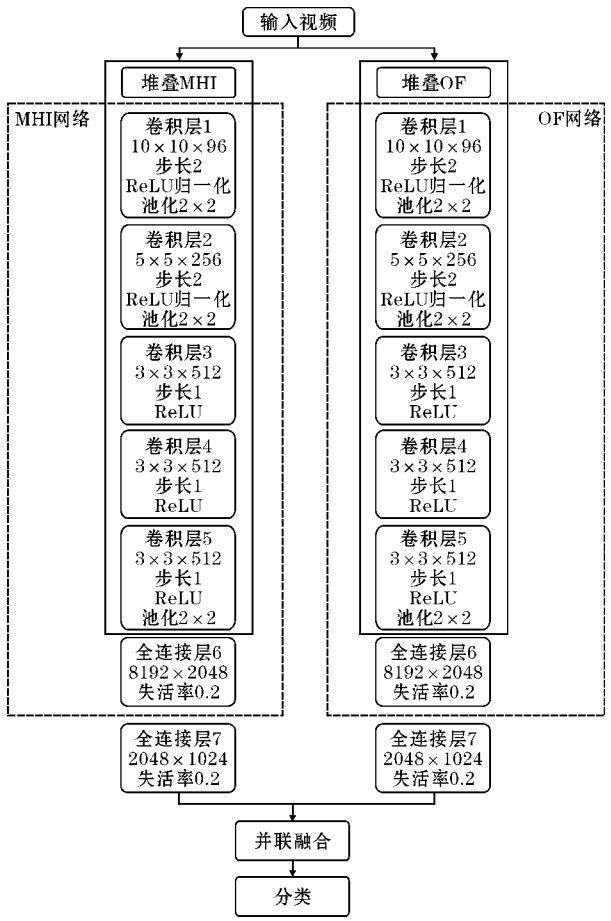

图 2 双流预训练网络模型结构

在可见光的人体行为识别中,传统双流网络的输入分别采用单帧 RGB 图像和叠加的光流图像,在空间信息和时间信息中达到平衡。红外行为视频的时间信息更重要,故双流预训练网络的输入分别采用运动历史图和光流图。为增加时域信息的长度,有效利用红外视频中的时间信息,对运动历史图和光流图分别进行堆叠,形成堆叠运动历史图和堆叠光流图输入双流预训练网络中。

1.3 MHI 和 OF 特征融合

红外人体行为数据相对于可见光人体行为数据缺少相应的色彩信息和细节纹理信息[10],这些信息对于动作识别属于非必要信息。色彩信息在数据处理过程中通过灰度处理解决,细节纹理信息通过预训练网络仍然得到保留,故使用双流网络剔除相应的冗余信息,进一步提高准确率。

一般情况下,单个网络提取的特征表征能力不足,训练的模型泛化能力不高。文献[11]表明多个特征可以有效克服单个网络表征能力不足的缺点,特征融

合可以充分提高模型的泛化能力和分类准确率。目前常用的特征融合方法有 3 种: Softmax 融合方式、串联特征融合和并联特征融合[12-13]。本文采用并联融合方式替代传统双流网络的 Softmax 融合方式。具体融合策略表示如下:

$$f_{\text{Fusion}} = \lambda F_{\text{OF}} + (1-\lambda) f_{\text{MHI}}$$

其中,f_{OF}和f_{MHI}分别表示堆叠 OF 输入的分支和堆叠 MHI 输入的分支中的特征图表示。λ 是人为输入的系数,表示不同分支网络的重要性程度。

2　实验结果和分析

2.1　数据集

采用两个数据集,分别为 UCF101 数据集和 NTU RGB+D 数据集。

UCF101 数据集不论类别还是数量都是人体行为识别领域最大的数据集,其视频源于真实场景,包含 101 个分类,13320 个视频。经过数据预处理后有 160 万余条数据,用此数据集训练的预训练模型可以充分保证泛化性能。

NTU RGB+D 数据集包含可见光、红外、骨骼和深度图数据,拥有 60 个分类,共 56880 个视频,本文从中选取 10 个行为进行实验,每个行为选择 50 个视频,共 500 个视频。

2.2　基本参数设置

本文实验均在 Ubuntu 16.04 操作系统下进行,基于 Pytorch 框架实现。双流预训练网络和双流网络的初始学习率均设置为0.001,随着训练不断衰减,衰减率为 0.25。网络使用的优化器为自适应矩估计(adaptive moment estimation, Adam)优化器,损失函数为交叉熵损失(cross entropy loss)。在训练过程中批尺寸表示一次送入网络训练的样本数量,批尺寸越大,网络收敛得越快,本文的批尺寸设置为 16。为防止过拟合,训练时采用早停机制和随机丢失部分神经元的操作,丢失率(dropout)设置为0.2。在网络训练过程中,本文根据硬件条件在合理的范围内选择批尺寸值为 16。分类器采用线性函数作为核函数,惩罚系数 C = 0.25。

2.3　消融实验

本文消融实验主要分为两个部分,分别为预训练网络对比和融合策略的对比。

预训练网络的对比结果如表 1 所示。ResNet、AlexNet 和 VGG16 预训练网络模型均是在 ImageNet 数据集上训练,网络参数针对空间信息更新,本文设计的双流预训练网络参数针对时域信息更新,故效果优于 AlexNet 和 VGG16。与无预训练网络的结果对比,可以得到双流预训练网络能够有效利用可见光行为数据和红外行为数据的相似性,提高红外人体行为识别的准确率。

表 1　不同预训练网络对比的结果

预训练网络	准确率/%
无预训练网络	66.73
ResNet	74.85
AlexNet	71.25
VGG16	73.89
双流预训练网络	78.52

融合策略的对比如表 2 所示。表中先后显示了单独 OF 输入、单独 MHI 输入、Softmax 融合方式、串联特征融合方式和并联特征融合方式的准确率。由表 2 可知,并联特征融合效果优于其他融合方式。

表 2　不同融合方式对比结果

融合方式	准确率/%
单独 OF	65.16
单独 MHI	68.13
Softmax 融合	66.28
串联特征融合	69.25
并联特征融合	78.52

2.4　对比分析

本文选取具有代表性的人体行为识别方法与本文方法进行比较,结果如表 3 所示。与 TCN、TSN 和 TSTDDs 等多分支结构的人体行为识别方法相比,本文所使用的预训练网络和并联特征融合方式可以提高红外人体行为识别的准确率;与 C3D 等方法比较,说明本文提出的方法可以提取表征能力更强的行为特征,优于现有先进的红外人体行为识别方法。

表 3　NTU RGB+D 数据集上不同方法的识别结果

方法	准确率/%
TSTDDs[7]	69.29
TCN[14]	66.28
TSN[3]	72.46
TCN+LSTM[4]	68.27
C3D[15]	70.56
RGB+Depth+3D skeletons[8]	75.50
本文方法	78.52

分类器的混淆矩阵可以表明每个类别正确和错误的分类情况,并直观地显示模型的性能。本文方法在 NTU RGB+D 数据集 10 个分类的混淆矩阵如图 3 所示,可以看出,本文方法不仅可以分类差别较大的动作,而且可以区分差距较小的动作。

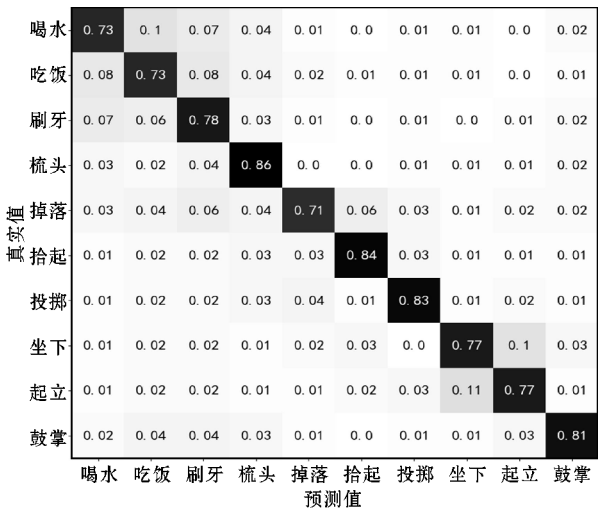

图 3 NTU RGB+D 数据集前 10 分类的混淆矩阵

5 结论

本文结合双流网络和迁移学习的优势,提出一种基于双流网络和迁移学习的红外人体行为识别方法。针对传统预训练网络无法充分提取人体行为中的时间信息,设计并预训练新的预训练网络模型。消融实验结果表明,本文所使用的预训练网络和特征融合方法至少提升 4.6% 红外人体行为识别准确率;通过于现有先进方法的比较,效果至少提升 3.9%。

致谢:感谢安徽建筑大学引进人才科研启动项目(2020QDZ24)对本文的资助

参考文献:

[1] Li J J, Han Y, Zhang M, et al. Multi-scale residual network model combined with Global Average Pooling for action recognition [J]. Multimedia Tools and Applications, 2022, 81(1): 1375-1393.

[2] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos [M]. Ghahramani Z, Welling M, Cortes C, et al. Advances in Neural Information Processing Systems 27 Cambridge, USA: The MIT Press, 2014: 568-576.

[3] Wang L M, Xiong Y J, Wang Z, et al. Temporal Segment Networks for Action Recognition in Videos [J]. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(11): 2740-2755.

[4] Ye W, Cheng J, Yang F, et al. Two-Stream Convolutional Network for Improving Activity Recognition Using Convolutional Long Short-Term Memory Networks [J]. Ieee Access, 2019, 7: 67772-67780.

[5] Huang R, Chen C, Cheng R, et al. Human Action Recognition Based on Three-Stream Network with Frame Sequence Features [J]. 2022 7th International Conference on Image, Vision and Computing (ICIVC), 2022: 37-44.

[6] Gao C, Du Y, Liu J, et al. Infar dataset: Infrared action recognition at different times [J]. Neurocomputing, 2016, 212: 36-47.

[7] Liu Y, Lu Z, Li J, et al. Global Temporal Representation Based CNNs for Infrared Action Recognition [J]. IEEE Signal Processing Letters, 2018, 25(6): 848-852.

[8] Romaissa B D, Mourad O, Brahim N. Vision-Based Multi-Modal Framework for Action Recognition [C]. 2020 25th International Conference on Pattern Recognition (ICPR), 2021: 5859-5866.

[9] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis [J]. IEEE transactions on neural networks, 2010, 22(2): 199-210.

[10] 邓茜文,冯子亮,邱晨鹏. 基于近红外与可见光双目视觉的活体人脸检测方法[J]. 计算机应用,2020,40(7):2096-2103.

[11] Wu J, An Y Y, Shi Q W, et al. Behavior Recognition Algorithm Based on the Fusion of SE-R3D and LSTM Network [J]. Ieee Access, 2021, 9: 141002-141012.

[12] Hall D L, Llinas J. An introduction to multisensor data fusion [J]. Proceedings of the IEEE, 1997, 85(1): 6-23.

[13] Yang J, Yang J Y, Zhang D, et al. Feature fusion: parallel strategy vs. serial strategy [J]. Pattern recognition, 2003, 36(6): 1369-1381.

[14] Gu Y, Ye X, Sheng W, et al. Multiple stream deep learning model for human action recognition [J]. Image and Vision Computing, 2020, 93: 10381-1038.

[15] Arif S, Wang J, Siddiqul A A, et al. Bidirectional LSTM with saliency-aware 3D-CNN features for human action recognition [J]. Journal of Engineering Research, 2021, 9(3A): 115-133.

Infrared Human Action Recognition based on Two-Stream Network and Transfer Learning

CHENG Rui[1], CHEN Chong[1,2], HUANG Ruifeng[1,2], GUO Kun[3]

(1. School of Electronics and Information Engineering, Anhui Jianzhu University, Hefei 231600, China; 2. Anhui International Joint Research Center for Ancient Architecture Intellisencing and Multi-Dimensional Modeling, Hefei 231600, China; 3. Border Town Xujiabian Technical Maintenance Room of Jurong City, Zhenjiang Jiangsu Province, Zhenjiang 212416, China)

Abstract: Aiming at the problem that traditional pre-training models cannot make full use of the temporal information of infrared human action data, this paper proposes an infrared human action recognition method based on two-stream network and transfer learning. In this paper, firstly, the images of motion history and optical flow are extracted from the original video and stacked using asliding window; secondly, based on the similarity between visible action data and infrared action data, a two-stream pre-training network is designed and the parameters of the network model pre-trained with visible action data are shared with the infrared human action recognition network model through transfer learning, so as to extract the features of infrared action data. Then, the extracted features, which is the input of the two-stream network, are used to further extract infrared human action information, and Softmax fusion is replaced by parallel feature fusion in the feature fusion. The experimental results show that the proposed method achieves an accuracy of 78.52% on the NTU RGB+D dataset, which has a good classification effect.

Keywords: human action recognition; transfer learning; two-stream network; infra-red; visible light