基于神经网络的学生行为检测算法研究

苏寒松，刘腾腾，刘高华*，田曦初
天津大学电气自动化与信息工程学院，天津 300072

摘要 关于行为检测的算法有很多，但针对教室场景下的学生行为检测算法却略显缺乏。为了使学生行为检测算法具有较好的准确率和实时性，在 MTCNN 的基础上改进了网络结构，并提出了一种新的激活函数和损失函数以检测学生图像和关键点。同时，提出了通过图像分类网络和关键点分类网络对学生行为进行联合分类的策略。实验结果表明，所提出的改进措施均有效提升了学生行为检测的准确率，最终模型的检测准确率为 78.6%。在嵌入式开发板 Jetson TX2 上，所提算法的实时检测准确率和速度优于 YOLOv3 和 SSD 等算法。

关键词 图像处理；机器视觉；神经网络；行为检测；教室场景
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.221016

Algorithm for Student Behavior Detection Based on Neural Network

Su Hansong, Liu Tengteng, Liu Gaohua*, Tian Xichu
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract There are many algorithms for behavior detection in different datasets, but there is a little lack of algorithms for student behavior detection in classroom. In order to achieve better accuracy and real-time of student behavior detection, this paper improves the network structure based on MTCNN, and proposes a new activation function and a loss function to detect student images and landmark localization. Meanwhile, this paper proposes the strategy of joint classification of student behaviors through the image classification network and the landmark localization classification network. The experimental results show that the proposed improvement actions effectively improve the accuracy of student behavior detection and the final detection accuracy of the model is 78.6%. On the embedded development board of Jetson TX2, the proposed algorithm has the real-time detection accuracy and speed superior to those of the other algorithms such as YOLOv3 and SSD.

Key words image processing; machine vision; neural network; behavior detection; classroom scene
OCIS codes 100.2000; 150.1135; 100.3005

1 引 言

随着神经网络和深度学习的快速发展，计算机视觉在不同领域都有了长足的进步。行为检测作为计算机视觉的重要领域，也出现了很多经典算法。文献[1]提出了时空双流网络结构(Two Stream Network)并进行了视频的行为检测，该网络使用两个相同的卷积神经网络(Convolutional Neural Network，CNN)来分别训练视频图像和光流，最后进行融合分类。文献[2]在原始双流网络的基础上，使用 LSTM[3]方法进行空间流和时间流的融合，进一步提升了检测效果。为了解决双流网络对长视频检测效果差的问题，文献[4]提出了 TSN 网络，其主要思想是对长视频进行分段处理。基于双流网络的不同算法在数据集上的表现良好，但由于网络复杂，检测速度较慢。文献[5]使用 3D 卷积构建网络，提出了 C3D 结构并证明了其在时空特征提取上的有效性，该算法提升了检测速度但准确率较低。在具体的应用中，文献[6]通过结合灰度值与光流场的分布来提取运动区域，并进行了人群行为的异常检测，

收稿日期：2020-03-19；**修回日期**：2020-04-08；**录用日期**：2020-04-20
基金项目：广州市科技计划(201802020008)
* **E-mail**：suppig@126.com

该方法抗干扰能力较强。Hu 等[7]提出了基于语义的人体行为识别方法，该方法在室内人体行为识别上具有较大优势。Hu 等[8]根据人体动作与预先建立的动作模型的相似度来进行人体动作识别。

尽管许多算法在数据集或其他场景中表现优异，但针对于教室场景下学生行为检测的算法却难以直接套用现有算法。其难点在于：1）针对教室场景下的学生行为检测，需要特定的数据集，且训练神经网络的数据集较大。2）深度学习的模型一般较大，应用时将面临模型承载设备（嵌入式开发板等）计算能力不足等问题。3）对学生在教室中的行为的判定存在一定的歧义，难以实现行为的分类。为了解决这些问题并设计出可以应用在教室场景下的算法，本文首先拍摄并制作了学生教室行为的数据集，包括教室情景下最常出现的五类行为。其次，本文基于目标检测和图像分类进行了行为检测，并加入了关键点检测和分类，提高了行为检测的准确率。利用改进的 MTCNN[9] 网络结构进行了人体框预测和关键点预测，通过图像和关键点联合分类对学生行为进行了识别。最后在 Jetson TX2 上构建了以本文算法为核心的行为检测系统，实现了算法和应用的对接。

2 原理和方法

2.1 网络架构

本文采用一种自上而下的行为检测的网络结构，即先用目标检测网络进行人体的定位检测，然后将检测到的人体图像和关键点送至分类网络并进行识别，最后输出行为检测结果。MTCNN 网络包括 P-Net，R-Net 和 O-Net 三个网络，在人脸检测任务中有很好的表现，但在人体检测时却不够精准且无法进行图像的分类。因此，本文改进了 MTCNN 网络并通过图像分类网络和关键点分类网络对学生行为进行了联合分类。

本文对 MTCNN 网络的改进集中于 O-Net 模块，如图 1 所示。采用 48×48 的输入尺寸，在第二个卷积层后加入一个尺寸为 3×3 的卷积层和两个尺寸为 1×1 的卷积层以扩大网络的感受野，如图 1 中左侧虚线框所示。之后参考 inception 模块[10]，利用 3×3 和 1×1 大小的卷积核分别对前面一层的输出进行卷积操作并将两个输出进行拼接，得到更多的特征，如图 1 中右侧虚线框所示，其中 landmark localization 表示关键点坐标。改进的 O-Net 网络最终输出人体预测框和 7 个关键点坐标。

图像分类网络结构如表 1 所示，其中 Conv 表示卷积层，Output size 表示输出的特征图大小，Ksize 表示卷积核大小，Padding 表示填充操作，Stride 表示卷积步长，Fc 表示全连接层，Same 和 Valid 是填充操作中常用的两种方式，Same 方式表示填充后经卷积操作得到的特征图与输入的特征图大小相同，Valid 方式表示填充后经卷积操作得到的特征图会变小。根据 O-Net 输出的人体框坐标，将原图上裁剪下来的图像尺寸转换成 100×100 大小的输入图像分类网络，经过三个卷积层后对其进行一次最大池化（Max pool）处理，之后再通过三个卷积层提取特征及全连接层分类后，最终输出学生各个行为的概率。

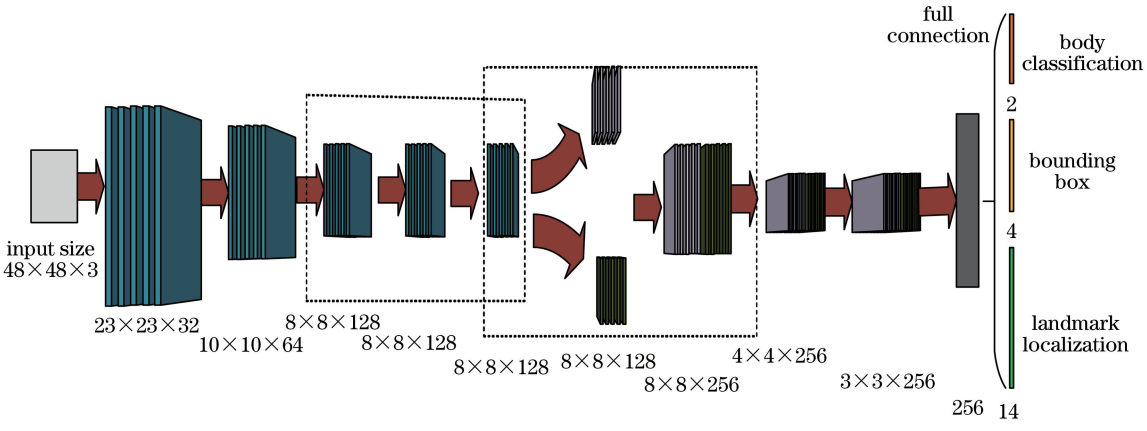

图 1 改进的 O-Net 网络

Fig. 1 Improved O-Net network

表 1　图像分类网络结构

Table 1　Network structure for image classification

Layer	Output size	Ksize	Padding	Stride
Conv 1	100×100	5×5	Same	1
Conv 2	50×50	2×2	Valid	2
Conv 3	50×50	5×5	Same	1
Max pool	25×25	2×2	Valid	2
Conv 4	25×25	3×3	Same	1
Conv 5	12×12	3×3	Valid	2
Conv 6	12×12	3×3	Same	1
Conv 7	6×6	2×2	Valid	2
Fc 1	1152			
Fc 2	192			
Fc 3	5			

关键点分类网络是一个纯全连接网络，以改进的 MTCNN 网络输出的 7 个关键点坐标作为关键点分类网络的输入，中间有两个隐含层（神经元数分别为 512 和 126），输出为五种行为类别的概率。关键点的选取主要以人体上半身关节为主，分别为双肩、双肘和双手六个关节，再加上额头共 7 个关键点。

改进的 MTCNN 网络、图像分类网络及关键点分类网络组成了本文的核心网络，整体结构如图 2 所示。改进的 MTCNN 网络负责检测人体图像和关键点并将其分别作为图像分类网络和关键点分类网络的输入。学生行为的判定结果以图像分类网络的分类结果作为主要分类依据，以关键点分类网络的分类结果作为辅助判定依据。

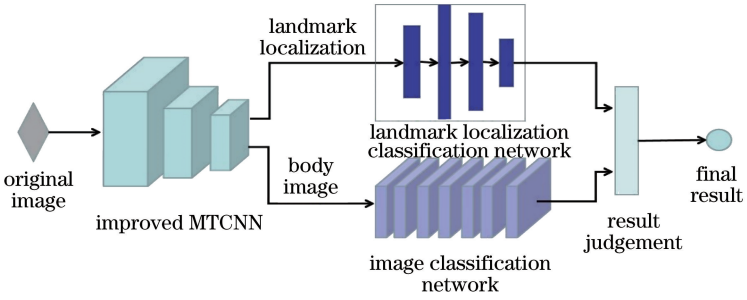

图 2　整体网络结构

Fig. 2　Overall network structure

2.2　激活函数

激活函数是神经网络的重要组成部分，激活函数的选择和好坏影响着神经网络的效果。Swish[11] 是 Google 提出的一种激活函数，有着不错的效果，在许多数据集上的表现都超过其他激活函数。Swish 函数的曲线如图 3 所示。虽然 Swish 函数使得神经网络的精度提升，但其计算量较大。文献[12]在此基础上提出了 hard Swish（h-Swish，h_{hSwish}）：

$$h_{\mathrm{hSwish}}(x)=x\frac{r_{\mathrm{relu6}}(x-3)}{6},\tag{1}$$

式中：x 为激活函数的输入；r_{relu6} 是指 ReLUctant[13] 函数输出大于 6 的部分全部取 6。h-Swish 函数的曲线如图 3 所示。h-Swish 的思想是通过损失精度（使函数变“hard”）来减少计算量。具体方法是利用 ReLUctant 函数的平移和缩放等变换来拟合 sigmoid 函数，进而乘以 x 以拟合 Swish。在 h-Swish 中，对于 $x<0$，只有在 $-3\leqslant x<0$ 的区间内，函数的输出才取到非零值。为了扩大激活函数在 $x<0$ 时输出不为零的取值范围，本文提出了比 h-Swish 非零输出范围更宽的激活函数 broader h-Swish（bh-Swish，b_{hSwish}）：

$$b_{\mathrm{hSwish}}(x)=\begin{cases}0, & x<-6\\ x(0.025x+0.15), & -6\leqslant x<-2\\ x(0.2x+0.5), & -2\leqslant x<0\\ x(0.167x+0.5), & 0\leqslant x<3\\ x, & x\geqslant 3\end{cases},\tag{2}$$

式中：$x\geqslant 0$ 的部分沿用 h-Swish 的计算方法；$x<0$ 的部分使用三个函数组合。从图 3 可以看出，bh-Swish 函数在负值区间（$x<0$）得到非零输出对应的 x 范围比 h-Swish 大了一倍，且其在 $-6\leqslant x\leqslant -3$ 区间内输出的绝对值大于 Swish。bh-Swish 理论上增大了 h-Swish 函数的精度但并未增加其计算量。

2.3　相对陡峭损失

利用损失函数来计算神经网络模型的预测值和准确值之间的差异幅度。坐标点损失函数的选择对人体检测网络预测人体框和关键点的准确率有着至关重要的影响。MTCNN 网络使用平方欧氏距离（SEUCLID）作为坐标点的损失函数。

图 4 所示为损失函数曲线。从图 4 可以看出，当预测值和真实值的距离小于 1 时，平方欧氏距离

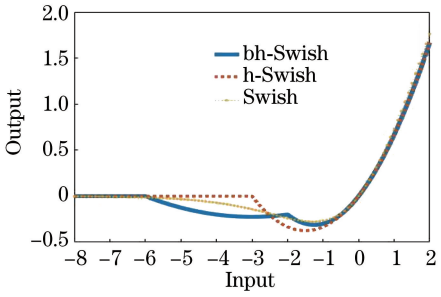

图 3 激活函数曲线

Fig. 3 Activation functional curves

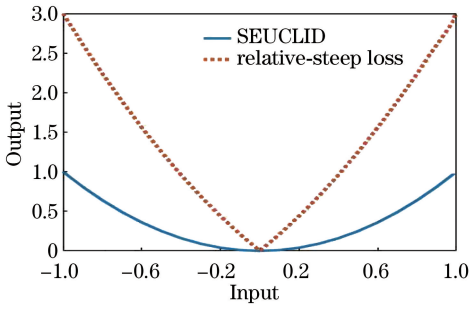

图 4 损失函数曲线

Fig. 4 Loss functional curves

输出的变化幅度很小，平方欧氏距离输出越接近零，损失变化越小。这使得神经网络模型在进行多次训练迭代之后效果变差。为了改善这一问题，本文提出了相对陡峭损失函数(Relative steep loss)，其函数曲线如图 4 所示，其在所有取值范围内的输出均大于平方欧氏距离，且当预测值和真实值的距离小于 1 时也保持输出的幅度基本不变。相对陡峭损失函数为

$$l=(|P-T|+1)^2-1, \tag{3}$$

式中：l 表示坐标点预测的损失；P 表示神经网络对坐标点的预测值；T 表示目标值。由于输入图像经过处理后，$|P-T|$ 的值总是小于1，而小于1的数的平方只会更小。为了扩大损失值以便于后期神经网络参数的优化，本文尝试 $|P-T|$ 加上一个常数 α。这个常数最终选择为 $\alpha=1$，1 是保证 $|P-T|+\alpha>1$ 的最小数，这使得在增大损失时不会出现梯度爆炸现象。最后，在平方项之后再减去 1，使预测值和真实值相等时损失为 0。

2.4 训练模型及在 Jetson TX2 上的实现

本文收集了课堂中普遍存在的五种行为，包括听讲、写字、站立(回答)、举手和睡觉在内的 10000 张图像，利用这些图像标注的行为框和人体上半身的 7 个关键点制作成数据集，并拍摄制作了教室场景的视频以用作检测。数据集如图 5 所示。

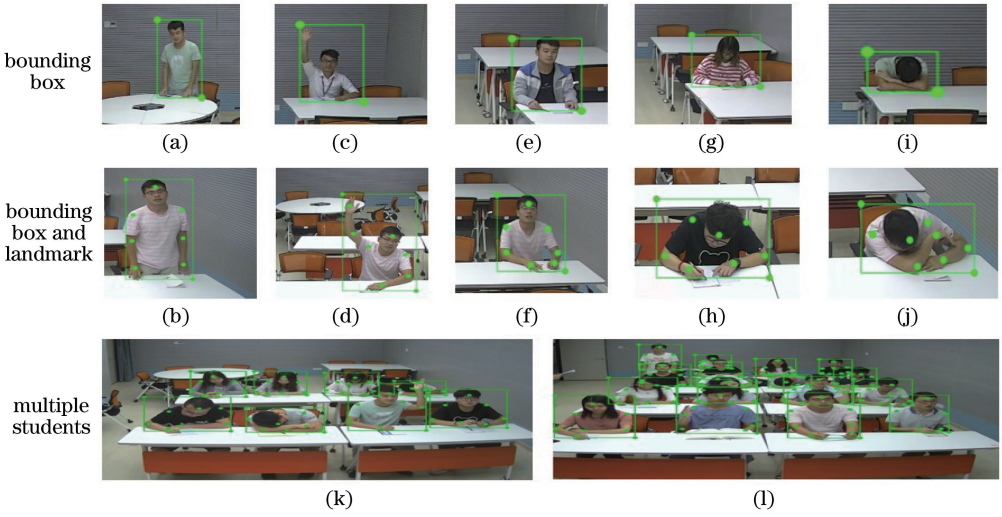

图 5 数据集示例。(a)(b)站立；(c)(d)举手；(e)(f)听讲；(g)(h)写字；(i)(j)睡觉；(k)(l)多人图像

Fig. 5 Dataset examples. (a)(b)Stand; (c)(d) hand up; (e)(f) listen; (g)(h) write; (i)(j) sleep; (k)(l) multiple students

使用图 5 所示数据集训练不同激活函数或损失函数下的 MTCNN 网络和改进的 MTCNN 网络。训练完成后在数据集上检测不同改进措施对模型准确率的影响。利用同样的数据集训练模型 SSD[14]、YOLOv3[15]、轻量化模型 tiny-YOLOv3 及基于 MobileNetV2[16] 和 MobileNetV3[12] 的 SSD。使用数据集和视频测算不同算法的准确率，其中测算视频检测准确率的方法为：只计算各算法输出的正确行为分类数量占总输出分类数量的比例，而不计算各算法的漏检情况。

模型训练时的环境为 Ubuntu16.04 系统和 tensorflow 框架。本实验使用 GPU(显卡型号为

NVIDIA GeForce GTX 1080Ti)进行训练,批处理(batch size)大小为 24,总计训练 20000 次。使用随机梯度下降(SGD)训练网络,初始学习率(learning rate)设置为 0.1,动量值(momentum)为 0.9。

Jetson TX2 是 NVIDIA 推出的一款嵌入式开发平台,采用 NVIDIA Pascal™ 架构。很多神经网络模型以及功能在 Jetson TX2 上得以实现[17-21]。基于 Jetson TX2,以本文提出的算法为核心构建课堂行为检测系统。利用网络摄像头获取教室图像并发送至 Jetson TX2,系统每隔 12 frame 输入一次图像并进行质量检测。设定阈值 a 和 b,当图像分类网络输出的最大概率 p_i 大于 a 时,判定学生行为是最大概率对应的行为;当图像分类网络输出的最大概率在 b 与 a 之间时,将对应的关键点 p_m 坐标送入关键点辅助分类网络,如果关键点辅助分类网络输出的最大概率对应的行为与图像分类网络输出的最大概率对应的行为是一致的,则判定此为学生的行为。考虑到现实场景下,输出一个错误结果比检测不到学生行为所带来的影响更加恶劣,所以若图像分类结果与关键点分类结果不一致则舍弃该图像,不输出检测结果。具体流程如图 6 所示,其中 a 为 0.9,b 为 0.75。

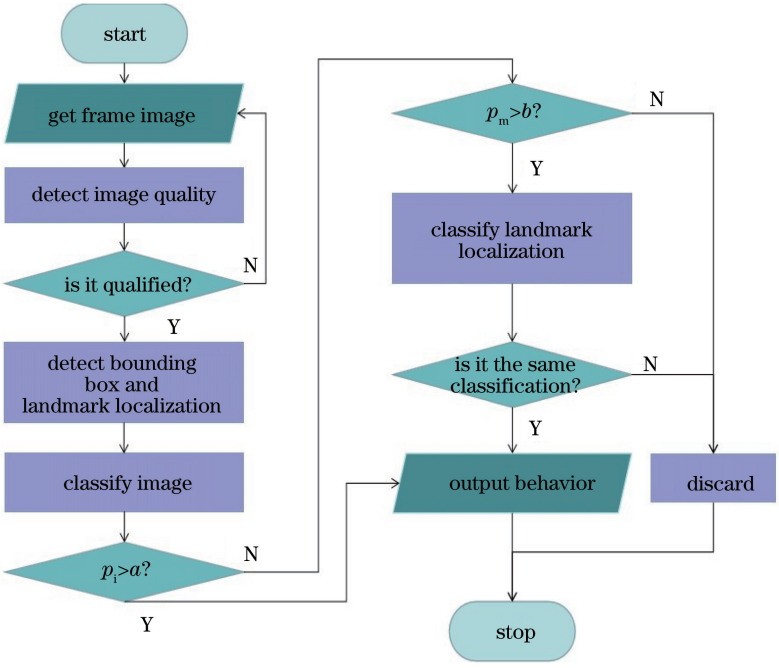

图 6 在 Jetson TX2 上的行为检测流程图

Fig. 6 Flow chart of behavior detection on Jetson TX2

3 实验结果与分析

表 2 给出了实验中不同的网络结构、激活函数以及损失函数在数据集上的准确率。由表 2 可以看出,相较于原始网络,改进的 MTCNN 网络在使用相同的损失函数和激活函数时准确率均有提升,最高提升 0.5%。准确率提升的原因在于改进的 MTCNN 增加了 MTCNN 网络的层数,同时利用尺寸为 1×1 的卷积核以及 inception 模块拼接并提取了同一图层的不同特征。在仅改变损失函数的情况下,本文提出的相对陡峭损失函数使原始模型的准确率提高了 0.6%。激活函数 bh-Swish 相较于 Prelu[22] 和 h-Swish,最高可使模型的准确率提升 2.9%。可以看出,激活函数的改变在本文中对准确率的提升贡献最大。通过不同改进措施的混合使用,单一改进措施对模型准确率的提升程度有所下降。

表 2 不同网络结构在不同激活函数以及损失函数下的准确率

Table 2 Accuracies of different network structures under different activation functions and loss functions unit: %

Network structure	Prelu		h-Swish		bh-Swish	
	SEUCLID	Relative steep loss	SEUCLID	Relative steep loss	SEUCLID	Relative steep loss
Original MTCNN	75.3	75.9	77.6	77.8	78.2	78.3
Improved MTCNN	75.8	76.2	77.8	78.1	78.4	78.6

图 7 给出了本文的算法检测五种不同行为的准确率。从图 7 中可以看出，本文的算法对站立行为的检测准确率较高，达到了 81%，而对睡觉行为的检测准确率较低，仅为 76%。通过分析可知，站立的姿势不易被遮挡，漏检率较少，因此检测准确率较高；睡觉姿势不容易被检测，准确率较低。整体上来说，行为检测系统的准确率为 78.6%，效果较为理想。

在检测同一教室场景视频条件下，各算法在数据集上的准确率及在 Jetson TX2 上的准确率如表 3 所示。从表 3 可以看出，在数据集上，SSD 表现最好，达到了 79.2%的准确率；本文算法的准确率为 78.6%，略低于 SSD 但高于其他算法，相较于两个轻量级网络具有较大优势。在 Jetson TX2 上，相较于在数据集上，算法准确率均减小，本文算法的准确率减小量是最小的，且准确率均高于其他算法。

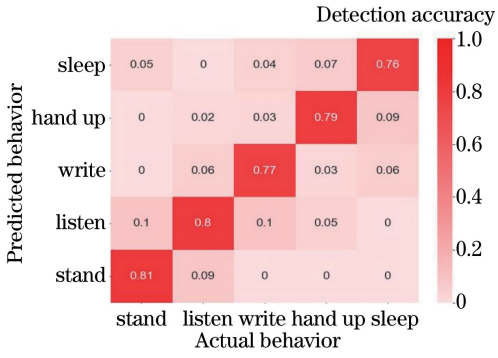

图 7 五类行为的检测准确率

Fig. 7 Detection accuracy of five kinds of behaviors

这是由于本文算法利用图像分类网络筛选了一部分行为分类不明显的预测框，再利用关键点分类网络进行了二次筛选，因此留下的预测框分类具有更高的准确率。

表 3 各算法的准确率

Table 3 Accuracy of each algorithm unit: %

Algorithm	YOLOv3	tiny-YOLOv3	SSD	MobileNetV2-SSD	MobileNetV3-SSD	Our algorithm
Dataset	78.2	66.7	79.2	75.8	77.2	78.6
Jetson TX2	74.5	60.4	77.1	73.3	74.1	77.9

表 4 为各算法在 Jetson TX2 上处理一帧图像的平均时间。从表 4 可以看出，本文算法在 Jetson TX2 设备上的处理速度快于轻量级网络 tiny-YOLOv3 和 MobileNetV3-SSD 等算法。这表明本文算法在 Jetson TX2 上具有良好的实时性。

表 4 各算法在 Jetson TX2 上处理 1 frame 图像的平均时间

Table 4 Average time of each algorithm to process 1 frame image on Jetson TX2 unit: ms

Algorithm	YOLOv3	tiny-YOLOv3	SSD	MobileNetV2-SSD	MobileNetV3-SSD	Our algorithm
Time	780	88	653	80	76	69

4 结 论

提出了一种基于改进的 MTCNN 网络并结合图像和关键点联合分类的教室行为检测算法。在改进 MTCNN 网络的同时提出了新的激活函数和损失函数，且改进的措施均增加了模型的准确率。另外，在 Jetson TX2 上实现了完整的教室行为检测系统。综合实时检测的准确率和检测速度可以看出，所提出的算法相比 YOLOv3 和 SSD 等算法具有一定的优势。之后的研究重心将通过添加或选取更加有效的关键点来减少由遮挡带来的漏检等问题，同时将制作含有更多图片及学生行为类型的数据集，为后期训练更准确的学生行为检测模型提供支持。

参 考 文 献

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [EB/OL]. (2014-11-12) [2019-12-21]. https://arxiv.org/abs/1406.2199.

[2] Ng JoeY H, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: deep networks for video classification [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 4694-4702.

[3] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition [M] // Leibe B, Matas J, Sebe N, et al. Computer Vision—— ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9912: 20-36.

[5] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional

networks[J]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 4489-4497.

[6] Zhou P P, Ding Q H, Luo H B, et al. Anomaly detection and location in crowded surveillance videos [J]. Acta Optica Sinica, 2018, 38(8): 0815007.
周培培, 丁庆海, 罗海波, 等. 视频监控中的人群异常行为检测与定位[J]. 光学学报, 2018, 38(8): 0815007.

[7] Hu T, Zhu X, Guo W, et al. Human action recognition based on scene semantics[J]. Multimedia Tools and Applications, 2019, 78(20): 28515-28536.

[8] Hu L Q, Cai Z Q, Xing L N, et al. Human action recognition via learning joint points information toward big AI system [J]. Journal of Visual Communication and Image Representation, 2019: 102688.

[9] Zhang K P, Zhang Z P, Li Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[10] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2818-2826.

[11] Ramachandran P, Zoph B, Le Q V. Swish: a self-gated activation function[EB/OL]. (2017-10-16)[2019-12-21]. https://arxiv.org/abs/1710.05941v1.

[12] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE, 2019: 1314-1324.

[13] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[EB/OL]. (2010-01-17)[2019-12-21]. https://wenku.baidu.com/view/7822feb5770bf78a65295450.html.

[14] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]// Leibe B, Matas J, Sebe N, et al. Computer Vision—— ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9905: 21-37.

[15] Redmon J, Farhadi A. Yolov3: An incremental improvement[EB/OL]. (2018-04-08)[2019-12-21]. https://arxiv.org/abs/1804.02767.

[16] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 4510-4520.

[17] Chen L L, Zhang Z D, Peng L. Real-time detection based on improved single shot MultiBox detector[J]. Laser & Optoelectronics Progress, 2019, 56(1): 011002.
陈立里, 张正道, 彭力. 基于改进SSD的实时检测方法[J]. 激光与光电子学进展, 2019, 56(1): 011002.

[18] Cui H, Dahnoun N. Real-time stereo vision implementation on Nvidia Jetson TX2[C]//2019 8th Mediterranean Conference on Embedded Computing (MECO), June 10-14, 2019, Budva, Montenegro. New York: IEEE, 2019: 1-5.

[19] Jose E, M G, Haridas M T P, et al. Face recognition based surveillance system using FaceNet and MTCNN on jetson TX2[C]//2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), March 15-16, 2019, Coimbatore, India. New York: IEEE, 2019: 608-613.

[20] Giubilato R, Chiodini S, Pertile M, et al. An evaluation of ROS-compatible stereo visual SLAM methods on a nVidia Jetson TX2[J]. Measurement, 2019, 140: 161-170.

[21] Wang Z W, Han J, Sun X B, et al. Method for orientation determination of transmission line tower based on visual navigation [J]. Laser & Optoelectronics Progress, 2019, 56(8): 081006.
王祖武, 韩军, 孙晓斌, 等. 基于视觉导航的输电线杆塔方位确定方法[J]. 激光与光电子学进展, 2019, 56(8): 081006.

[22] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]// 2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 15802053.