BNU_HUAWEI_远程教育项目文档

刘洋

一、表情识别相关工作

1 人脸识别

作为表情识别任务的前置工作，人脸识别主要用于识别并截取人脸区域，输入到表情识别模型中识别表情分类。

在该项任务中，项目主要使用 Wider Face 数据集对 yolov5 模型进行训练，具体来看，该数据集在以往实验中主要用于人脸识别工作，模型共使用该数据集中的 61 个场景下的 16106 张图像，其中，训练集包括 12880 张图像，测试集包括 3226 张图像，每张图像在识别难度、尺度、姿态、遮挡、表情和光照条件等方面均存在差别，如图 1，从而有利于对不同情景下的人脸进行识别。

图 1 Wider Face 数据集示例

yolov5 模型使用 yolov5s 网络结构（网络图可通过 github 下载“yolov5s.pt”文件，通过 netron，由于该网络图过长，文档中不进行展示），通过训练 108 个 epoch，最终 map0.5 达到 0.75，如图 2，能够达到人脸准确识别的效果，同时，目前正在测试 yolov5x 网络结构，测试是否能够达到更精准的效果。

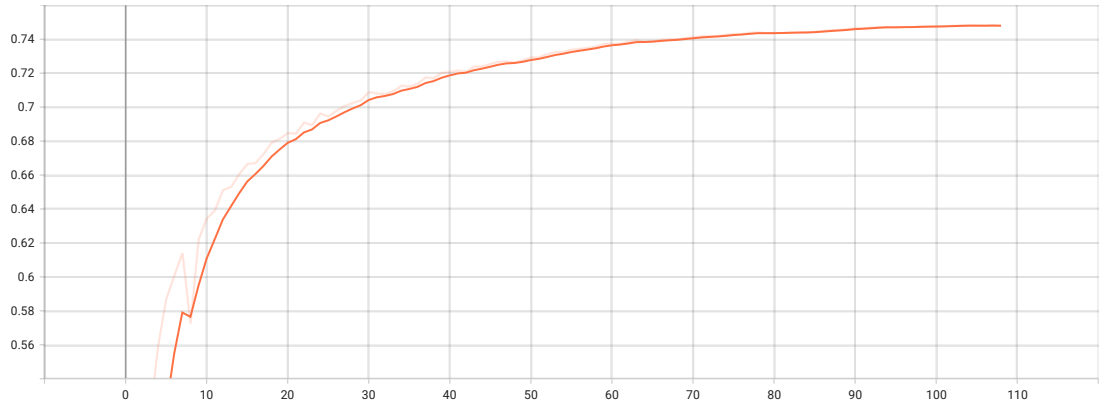

图 2 108 个 epoch 下人脸识别模型的 map0.5 的数值变化图

2 表情识别

在上述成果的基础上，表情识别任务通过对人脸识别的人脸截图输出进行识别，完成表

情的输出，其中人脸识别输出结果示例如图 3。

图 3 人脸识别输出结果示例图

而表情识别模型是利用 VggNet16 对 FER2013 数据集进行训练，具体来看，模型共使用 FER2013 数据集的 35887 张图像，其中训练集包括 28709 张图像，测试集包括 7178 张图像，该数据集图像是大小固定为 48×48 的灰度图像，如图 4，共包含 anger、disgust、fear、happy、sad、surprised 和 normal 共七类表情。

（注：最近发现 FER2013 存在部分标签标注错误的情况，后续会使用优化后的 FER2013 plus 数据集重新进行训练）

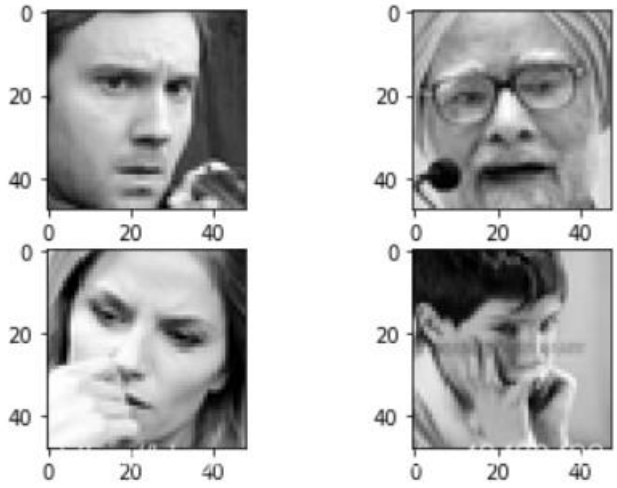

图 4 FER2013 数据集图片示例

本研究通过利用 VGGNet16 网络，网络结构见图 5，其分为 13 个卷积层，3 个全连接层和 5 个池化层，该网络通过加深卷积块提高训练效果。

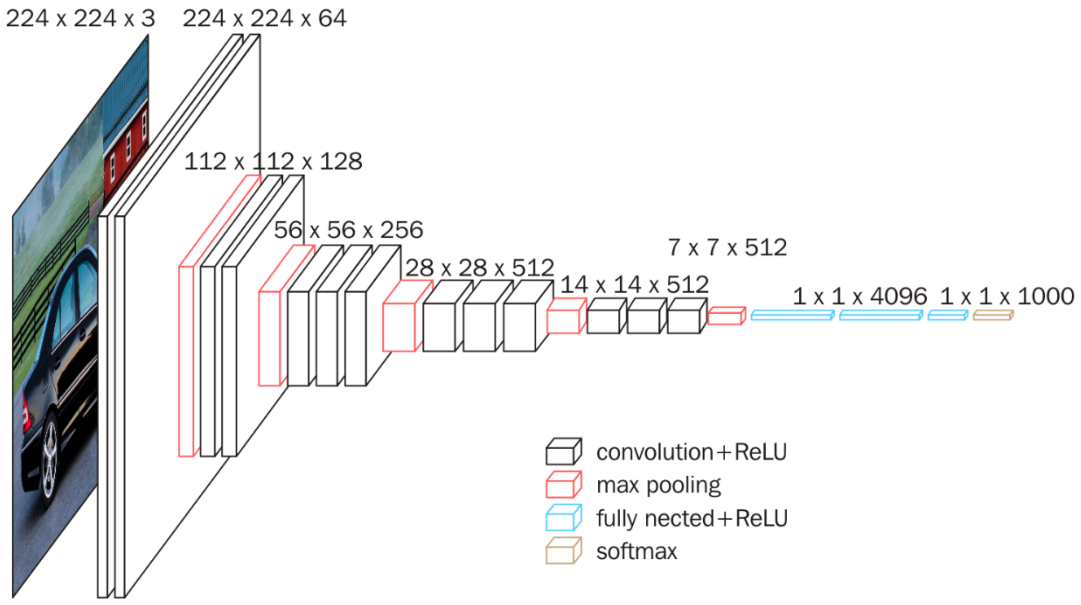

图 5 VGG16 网络结构图

之后，研究针对上述数据集训练 200 个 epoch,最终验证集准确率达到 67.6%，准确率较高，后续研究通过更改数据集为 FER2013 plus 或者将算法更换为 resnet 等网络结构重新训练，争取准确率上升到 75%+，目前应用到视频效果如图 6。

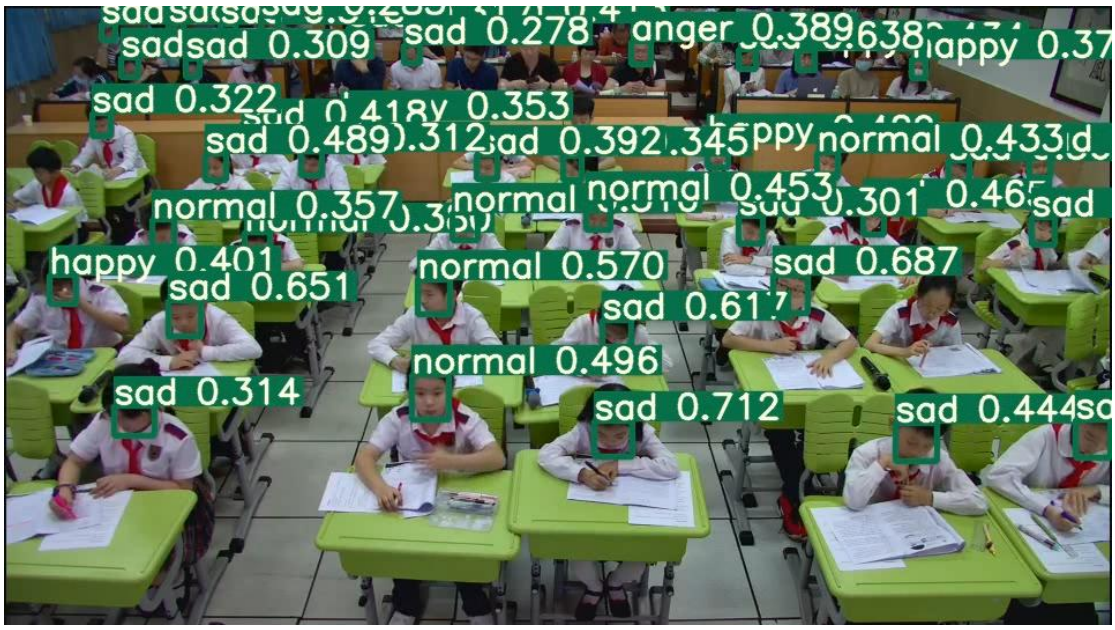

图 6 表情识别在具体教学场景下的应用

二、动作识别相关工作

动作识别任务主要分为三部分工作，即实体识别、目标追踪和动作（姿态）识别。简单来说，实体识别用于识别人体区域，多目标追踪是为每一个人体赋予特殊 ID,满足当人体发生移动时能够重新定位特定目标的需求（此部分为附加实现，若与动作识别结果结合，设想可以实现特定 ID 人的连续动作输出，但对视频拍摄有要求，后续内容会提到），动作识别是在实体识别的基础上完成特定动作的识别。

1 实体识别

研究针对实体识别任务使用 MS COCO 数据集，它提供了 80 个类别的图像数据用于训练模型，其中训练集包括 118287 张图像，验证集包括 5000 张图像，测试集包括 40670 张图像，官网示例图片见图 7。

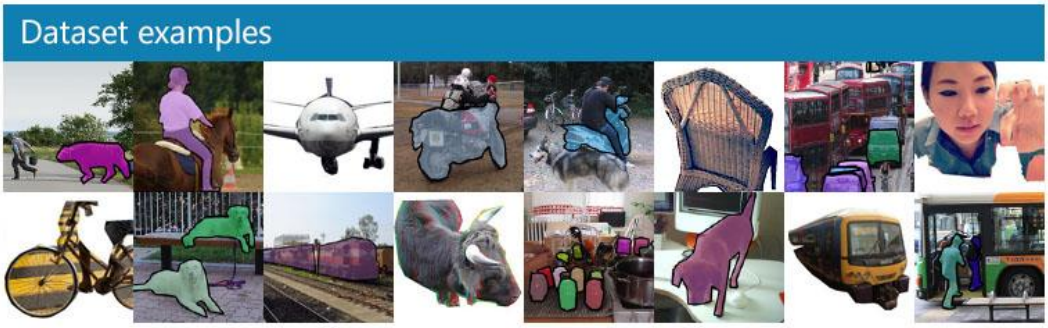

图 7 COCO 数据集官网示例图

实体识别同人脸识别同样使用 yolov5 模型中的 yolov5s 网络结构，筛选出其中的 person 目标，最终 map0.5 达 0.5378（为演示效果，训练了 100 个 epoch 后停止，具有提升空间），效果图如图 8，现在同样正在更改为 yolov5x 网络结构进行训练，预计达到达到更高的 map。

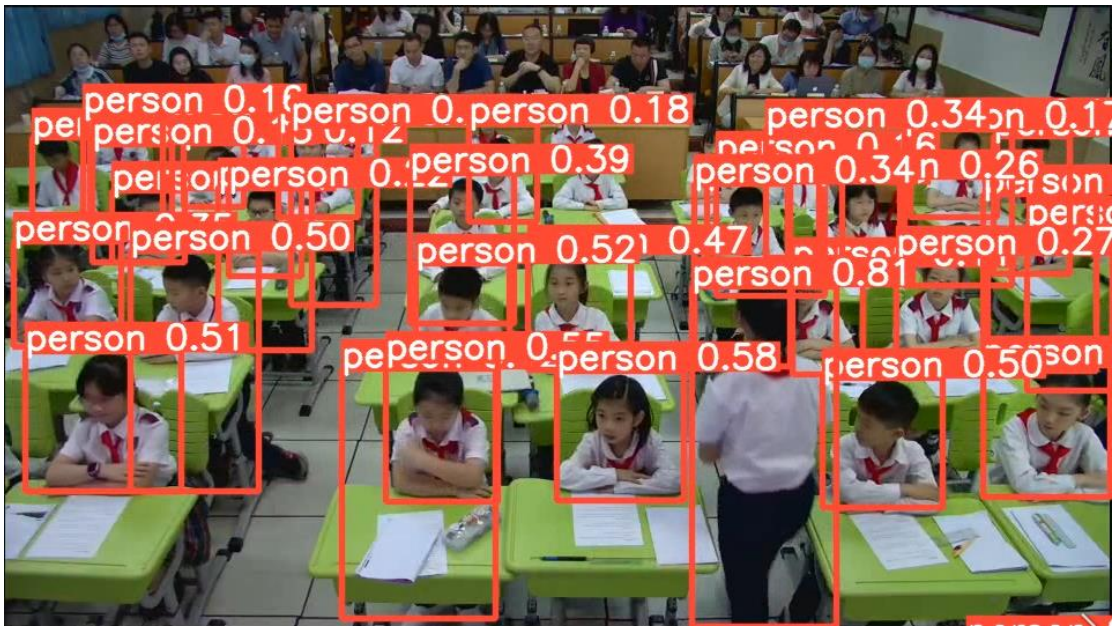

图 8 实体识别效果图

2 多目标追踪

目标追踪任务解决从第一帧到最后一帧，里面有多个教师或学生，不断地随着摄像头转换有出有进摄像区域，目的是对每个目标，能跟其他目标区分开，维持他们的 ID。研究中采用 Deepsort 算法针对 mot16 数据集进行训练。

首先，关于 mot16 数据集，该数据集包含总共有 14 个视频，训练集和测试集各 7 个，这些视频每个都不一样，按照官网的说法，它们有些是固定摄像机进行拍摄的，有些是移动摄像机进行拍摄的，而且拍摄的角度各不一样（低、中、高度进行拍摄），拍摄的条件不一样，包括不同的天气，白天或者夜晚等，还具有非常高的人群密度，数据集中具体的参数介绍可见链接，图 9 显示了其中的第 13 个展示视频示意图。

图 9 mot16 数据集视频截图

Deepsort 算法是在以卡尔曼滤波算法和匈牙利算法为主体的 sort 算法的改进版，在 sort 算法的基础上，该算法加入了级联匹配和新轨迹确认（简单来说，轨迹可分为确认态和不确认态，轨迹与检测结果连续匹配三次或者联系失配三次才会进行状态的转化），算法流程见图 10，具体操作见链接中讲解内容。

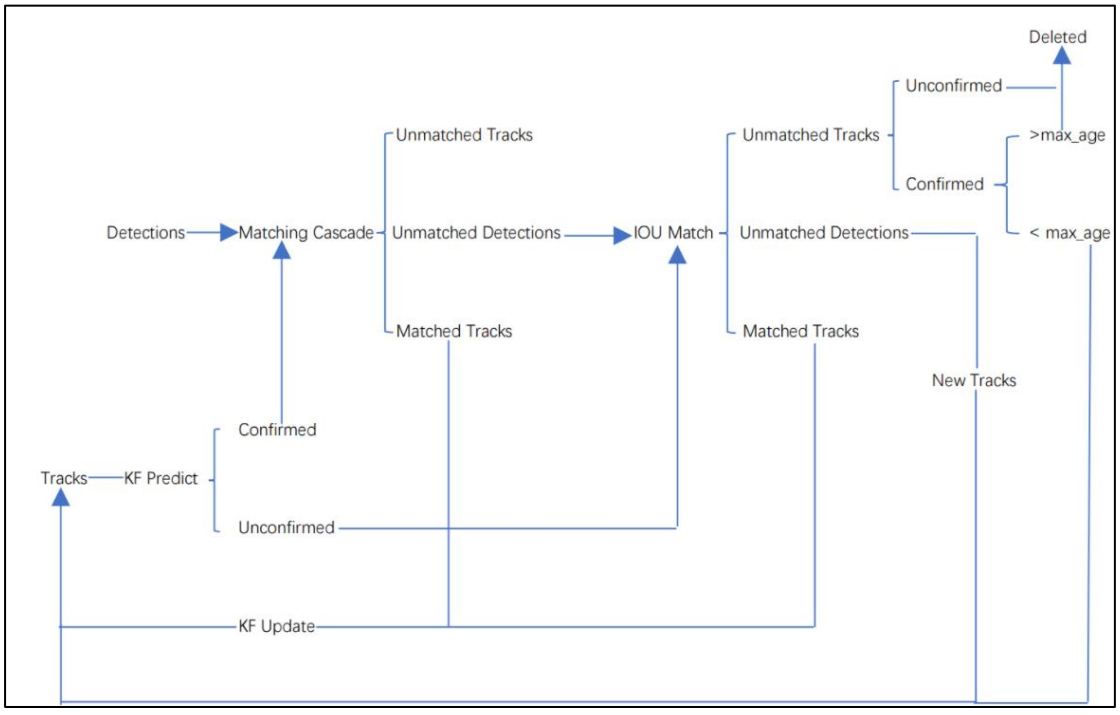

图 10 算法流程图

目前的效果视频是使用网上的预训练权重产生的，因此之后使用实验产生的训练权重目标追踪效果会更加优异。该模型对视频有一定要求：此模型最好在一种连续的单屏空间内进

行，且一切目标的消失出现均是由于摄像头的角度转化而造成的，也就是说需要长镜头（一镜到底）场景，而不是毫无逻辑的突然剪辑转变，所以若视频经常出现单屏、双屏和四屏的转化，该模型存在的意义不大，可以考虑取消。具体效果如图 11（存在实体识别不全的问题，之后进行优化解决）。

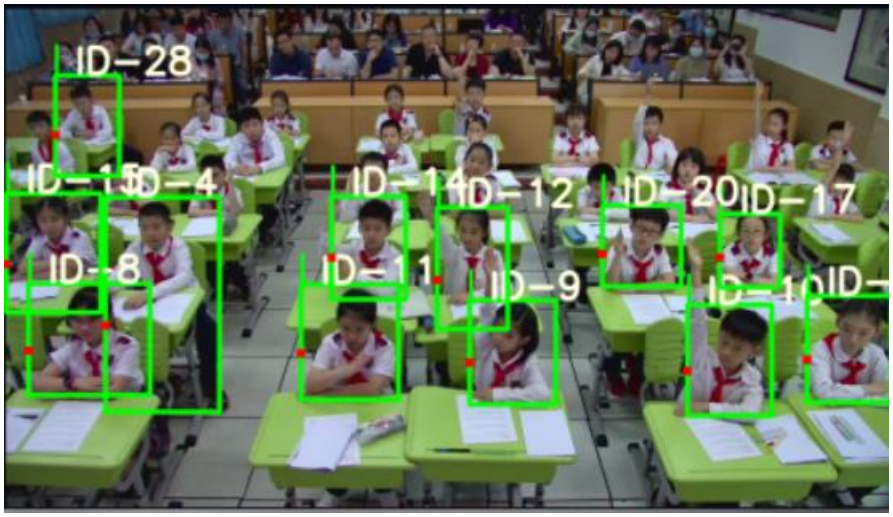

图 11 多目标追踪效果图

3 动作（姿态）识别

该模型的前置工作主要是依托实体识别的结果实现的（与目标追踪无关联）.该任务依托 ava 数据集，使用 slowfast 模型实现。

首先，ava 数据集准确来讲共有 80 类原子动作，80 类标签分为三类：person movement, object manipulation, person interaction，该数据集中的视频由电影片段构成，每个视频只针对第 15-30 分钟的视频关键帧进行标注，而关键帧简单理解就是每隔一秒取一帧，该数据集庞大，视频集占 150 多 G，视频帧占 390 多 G，因此训练时间较长。

对于 slowfast 模型，该算法原理可见图 12，简单来讲，这是一种用于分类视频动作的快慢结合的网络，Fast 网络的区别主要体现在模型对于视频的运动信息采用的策略是视频输入帧的选取间隔 stride 较小，因此该层网络的输入是高帧率信息，相反，Slow 网络则以较大的 stride 采取空间语义场景视频帧，使用低帧率信息。但是，Fast 网络的结构较 Slow 网络的结构复杂度低，这样的目的是为了实现两层网络的运算时间相同，不至于出现一方输出等待另一方输出时间过久的情况，最终两方结果进行整合实现输出，上述讲解内容可参考图 13 的论文网络结构截图。

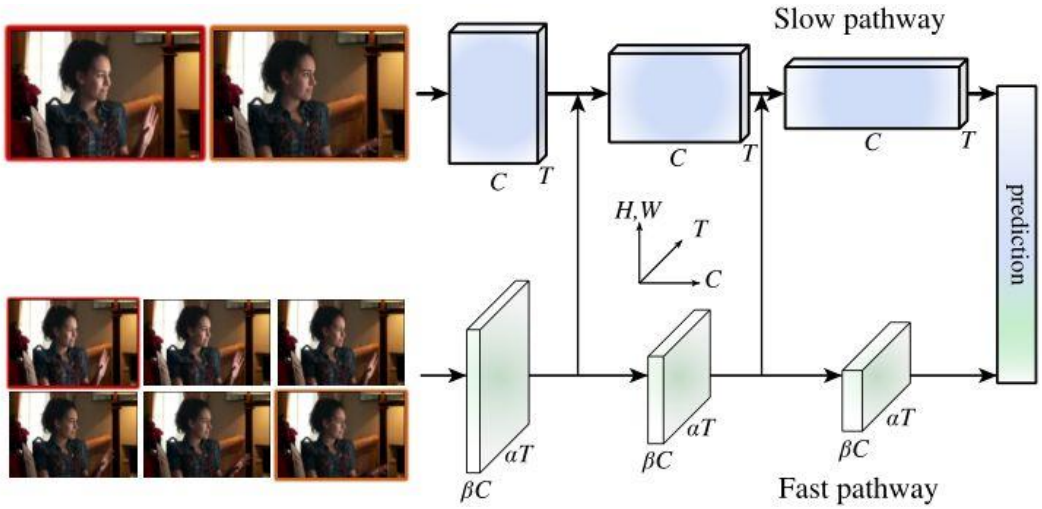

图 12 slowfast 网络图

stage	*Slow* pathway	*Fast* pathway	output sizes $T\times S^2$
raw clip	-	-	64×224^2
data layer	stride 16, 1^2	stride **2**, 1^2	*Slow* : 4×224^2 *Fast* : 32×224^2
conv_1	1×7^2, 64 stride 1, 2^2	5×7^2, 8 stride 1, 2^2	*Slow* : 4×112^2 *Fast* : 32×112^2
pool_1	1×3^2 max stride 1, 2^2	1×3^2 max stride 1, 2^2	*Slow* : 4×56^2 *Fast* : 32×56^2
res_2	$\begin{bmatrix}1\times1^2,64\\1\times3^2,64\\1\times1^2,256\end{bmatrix}\times3$	$\begin{bmatrix}3\times1^2,8\\1\times3^2,8\\1\times1^2,32\end{bmatrix}\times3$	*Slow* : 4×56^2 *Fast* : 32×56^2
res_3	$\begin{bmatrix}1\times1^2,128\\1\times3^2,128\\1\times1^2,512\end{bmatrix}\times4$	$\begin{bmatrix}3\times1^2,16\\1\times3^2,16\\1\times1^2,64\end{bmatrix}\times4$	*Slow* : 4×28^2 *Fast* : 32×28^2
res_4	$\begin{bmatrix}3\times1^2,256\\1\times3^2,256\\1\times1^2,1024\end{bmatrix}\times6$	$\begin{bmatrix}3\times1^2,32\\1\times3^2,32\\1\times1^2,128\end{bmatrix}\times6$	*Slow* : 4×14^2 *Fast* : 32×14^2
res_5	$\begin{bmatrix}3\times1^2,512\\1\times3^2,512\\1\times1^2,2048\end{bmatrix}\times3$	$\begin{bmatrix}3\times1^2,64\\1\times3^2,64\\1\times1^2,256\end{bmatrix}\times3$	*Slow* : 4×7^2 *Fast* : 32×7^2
global average pool, concate, fc			# classes

图 13 slowfast 网络结构图

该模型目前使用预训练权重实现演示，预计采用 8 卡训练 40 个 epoch 需要一天左右时间,效果会更加优异，目前测试图如图 14（实体识别缺失问题还在解决中）。

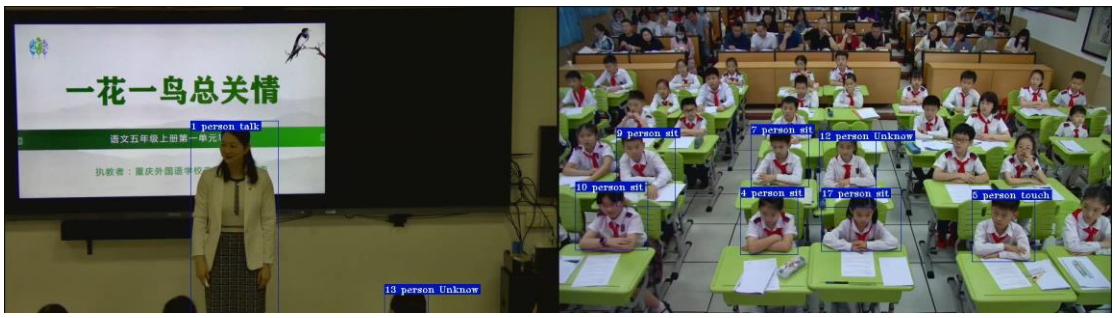

图 14 动作识别效果展示图

三、小节

以上为表情识别和动作识别两块主要任务的具体内容及进度介绍。