Computer Science and Application 计算机科学与应用, 2020, 10(6), 1277-1285
Published Online June 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.106132

Hans汉斯

Multilayer Recurrent Neural Network for Action Recognition

Wei Du

North China University of Technology, Beijing
Email: duwei2012com@163.com

Received: Jun. 8th, 2020; accepted: Jun. 21st, 2020; published: Jun. 28th, 2020

Abstract

Human action recognition is a research hotspot of computer vision. In this paper, we introduce an object detection model to typical two-stream network and propose an action recognition model based on multilayer recurrent neural network. Our model uses three-dimensional pyramid dilated convolution network to process serial video images, and combines with Long Short-Term Memory Network to provide a pyramid convolutional Long Short-Term Memory Network that can analyze human actions in real-time. This paper uses five kinds of human actions from NTU RGB + D action recognition datasets, such as brush hair, sit down, stand up, hand waving, falling down. The experimental results show that our model has good accuracy and real-time in the aspect of monitoring video processing due to using dilated convolution and obviously reduces parameters.

Keywords

Action Recognition, Dilated Convolution, Long Short-Term Memory Network, Deep Learning

多层循环神经网络在动作识别中的应用

杜　激

北方工业大学，北京
Email: duwei2012com@163.com

收稿日期：2020年6月8日；录用日期：2020年6月21日；发布日期：2020年6月28日

摘　要

人体动作识别是目前计算机视觉的一个研究热点。本文在传统双流法的基础上，引入目标识别网络，提出了一种基于多层循环神经网络的人体动作识别算法。该算法利用三维扩张卷积金字塔处理连续视频图

文章引用：杜激. 多层循环神经网络在动作识别中的应用[J]. 计算机科学与应用, 2020, 10(6): 1277-1285.
DOI: 10.12677/csa.2020.106132

像，结合长短期记忆网络，给出了一种能够实时分析人体动作行为的金字塔卷积长短期记忆网络。本文利用NTU RGB + D人体动作识别数据库，对五种人体动作，如梳头、坐下、起立、挥手、跌倒等动作进行识别。试验结果表明算法由于采取了扩张卷积，参数量明显降低，在监控视频处理方面具有较好的准确性和实时性。

关键词

人体动作识别，扩张卷积，长短期记忆网络，深度学习

1. 简述

在人体动作识别任务中，包含两个基本步骤：一是确定人体完整动作的起始和终止时间，二是识别该动作。本文仅对动作识别进行讨论。

在动作识别的方法中，目前较为流行的方法有三维卷积、基于人体姿态识别和双流法等动作识别方法。

从直观出发，基于人体姿态识别的方法更容易被理解和接受，利用连续帧之间人体姿态变化的规律实现人体动作的识别。这类方法以人体姿态识别[1]为基础，因此其研究重点集中于人体姿态识别的准确性和可靠性。同时，人体姿态识别的结果直接影响动作识别的准确性。

区别于基于人体姿态识别的方法，双流法[2] [3] [4] [5]使用两个并列的分支网络，分别从彩色视频图像和光流图像中提取动作信息。从不同的角度提取人体动作特征，并融合多角度特征进行动作识别。但目前的双流法网络无法实现读取连续视频图像。

三维卷积解决了连续视频图像读取的问题。三维卷积[4] [6]在二维卷积的基础上增加了提取前后图像关联特征的能力，这让三维卷积可以在获取单幅图像特征的同时，也能获取前后帧图像之间的关联信息。唯一不足的是三维卷积无法灵活地读取长短不一的连续视频图像。

本文通过引入目标检测网络，构造了一种新的金字塔卷积长短期记忆网络(Pyramid Convolutional Long-Short Term Memory, P-ConvLSTM)。将视频图像中的运动前景分离出来，而后通过深层长短期记忆网络进一步识别视频中运动的人体。之后通过前后帧运动人体之间的联系，利用金字塔形的设计获取视频图像在不同维度下的特征信息，捕捉人体动作特征，进而对如跌倒、头痛等人体动作进行识别分类。由于采用了多层循环神经网络，因此在处理序列长短有变化的数据上更加灵活。

2. 相关工作

双流法于 2014 年由 Simonyan 和 Zisserman [2]提出。该方法采用两个并行的分支网络，分别提取空间信息和时间信息。空间信息网络利用三通道彩色视频图像提取图像中的人、事、物等空间信息，而时间信息网络利用视频的光流信息提取视频中随时间变化产生的时间信息。

两个分支网络相互独立，仅在信息融合阶段进行信息地整合。

当前的双流法研究中，多以使用较为成熟的图像识别网络为分支网络的主要结构，并在信息融合的位置上和方法上进行改良。这样可以大幅减少训练的时间和成本，但无法实现连续视频图像序列的读取。

我们再利用双流法提取不同角度特征这一特点，并结合三维卷积与多层循环神经网络，实现对连续视频序列的读取与处理，达到人体动作识别的目的。

未解决双流法中连续视频图像序列读取的问题，我们使用三维卷积[6]替换二维卷积。三维卷积以二维卷积为基础进行扩展，扩展后的三维卷积可以接收连续图像序列，并提取图像序列中每幅图像的特征信息及前后图像之间的时间信息。但三维卷积无法灵活读取长短不一的图像序列，因此我们结合多层循环神经网络，实现灵活读取图像序列的目的。

我们所使用的多层循环神经网络是卷积长短期记忆网络(Convolutional Long Short-Term Memory, ConvLSTM)，该方法由 Xingjian Shi [7]等人提出。ConvLSTM 是一种长短期记忆网络(Long Short-Term Memory，LSTM)的变形，它不仅利用自身的卷积操作获取视频图像中的信息，还利用 LSTM 的结构特点获取图像序列的长期依赖关系，更好地获取视频序列的特征。因此，ConvLSTM 被广泛应用于动作识别方法中[8] [9]。

3. 金字塔卷积长短期记忆网络 (Pyramid Convolutional Long-Short Term Memory, P-ConvLSTM)

与典型的双流法的工作不同，我们向双流法中引入了一个基于循环神经网络的目标识别网络作为网络的主干结构。目标检测网络可以将视频图像中运动前景与不动后景进行分离，让 P-ConvLSTM 的注意力集中在运动的人体上。

该目标识别网络由 Hongmei Song 等人[10]于 2018 年提出，网络中设计了金字塔结构的扩张卷积模块和深层双向 ConvLSTM 模块为目标识别网络提供了更强的空间识别能力和时间信息获取能力。我们在 Hongmei Song 等人的工作上进行了简单的改动，不仅让网络能更好的适应人体动作识别任务，还能读取连续视频图像数列。

P-ConvLSTM 中有两个主要的模块：扩张卷积金字塔模块(Dilated Convolution Pyramid Module, DCP)和金字塔卷积长短期记忆模块(Pyramid Convolutional Long-Short Term Memory Module, Pyramid ConvLSTM)。

3.1. 扩张卷积金字塔模块

在引入的目标识别网络中，Hongmei Song 等人使用多个并列的二维扩张卷积在保留图像原始特征的情况下实现多尺度特征提取，同时避免了下采样操作并降低了网络中的参数。

我们将二维扩张卷积改造为三维扩张卷积，实现视频图像序列的多尺度特征提取。在 DCP 模块中，我们使用了四个尺寸相同、扩张度分别为 2、4、8、16 的三维卷积，它们之间相互并列构成金字塔模型。

而扩张卷积实现多尺度特征提取则是充分利用了其特点。以尺寸为 3*3 的卷积核为例，当扩张度为 1 时，卷积核仅能作用于图像上 3*3 的范围，进行卷积操作的像素数量为 3*3；若扩张度为 2，当卷积核在作用于图像上时，卷积核相邻元素之间距离变为 2，卷积核作用范围扩大为 5*5，但其中进行卷积操作的像素数量仍为 3*3。图 1 中形象展示了扩张卷积。通过改变尺寸相同的卷积核扩张度可以有效实现多尺度特征提取，这样不仅保留图像的原始特征信息，还减少使用下采样操作。

3.2. 金字塔卷积长短期记忆模块

在 Hongmei Song 等人的工作中，使用双向循环神经网络进一步提取视频图像序列的前后帧之间的关联信息。这样的操作可以获取图像序列中某一图像来自前后两个方向的信息，更好的进行目标检测。图 2 展示了双向循环神经网络的基本结构。

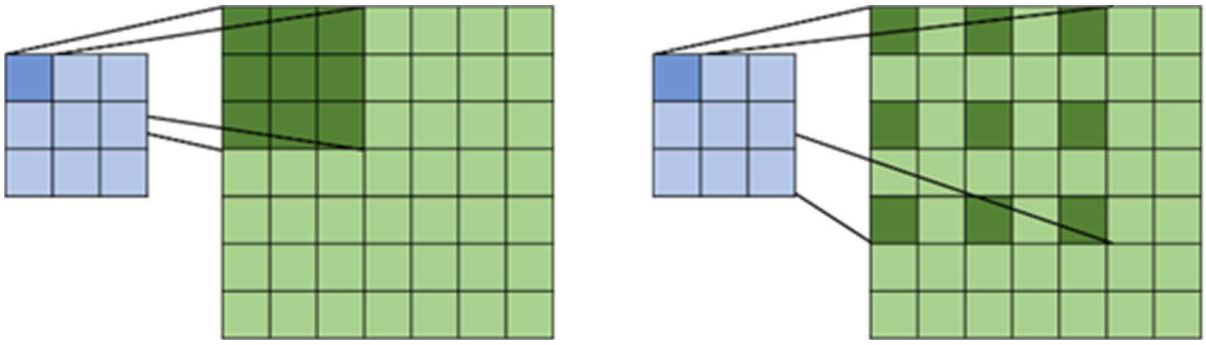


Figure 1. Dilated convolution. (a) When the dilated is 1, the 3*3 convolution kernel acts on the image, and the adjacent elements in the convolution kernel act on the image together; (b) When the dilated is 2, the distance between the adjacent elements of the 3*3 convolution kernel is pulled apart, and the area of convolution is expanded
图 1. 扩张卷积示意图。(a) 当扩张度为 1 时，3*3 卷积核作用于图像时卷积核中相邻元素靠在一起作用于图像上；(b) 当扩张度为 2 时，3*3 卷积核的相邻元素之间的距离被拉开，卷积作用的区域扩大

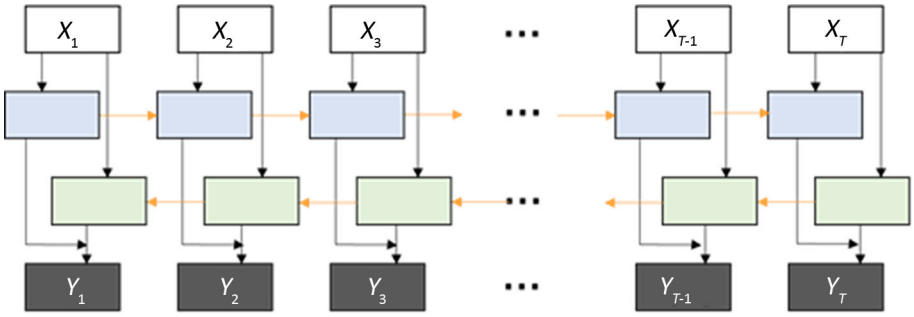


Figure 2. Bidirectional recurrent neural network
图 2. 双向循环神经网络示意图

而我们在 Pyramid ConvLSTM 模块中，使用两个扩张度不同的双层 ConvLSTM 网络分别对多尺度特征信息进行深层语义信息提取，获取人体动作的深层语义信息和视频图像的长期依赖关系。最后，将两个 ConvLSTM 的提取的特征信息进行融合用于最后的动作识别。而使用扩张度不同的 ConvLSTM 网络是为了获取不同尺寸下视频的语音信息和长期依赖关系，提升识别能力。

3.3. 改进模型

在数据预处理阶段，我们对视频数据进行分帧处理，并提取视频的稠密光流信息。之后，将彩色图像和光流信息图像进行尺寸调整。在输入到 P-ConvLSTM 中时，随机读取连续的是 10 帧视频图像。

P-ConvLSTM 有两个结构相同的分支，其中一个分支的输入为彩色视频图像，另一个分支的输入为光流信息图像。我们引用标准 ResNet-50 [11]模块并将其中的二维卷积修改为三维卷积，使其成为三维卷积 ResNet-50。然后我们使用 PDC 模块进行多尺度特征提取，同时将 PDC 模块原本使用的二维卷积修改为三维卷积，实现视频序列多尺度特征提取。之后，我们将多尺度特征与 ResNet-50 的特征进行融合输入到金字塔 ConvLSTM [12]模块中。在金字塔 ConvLSTM 模块中，融合后的特征信息将分别通过扩张度为 1 和扩张度为 2 两个平行的双层 ConvLSTM，进行多尺度特征提取。最后通过平均池化层、全连接层、softmax 得到分类向量。

通过将 ResNet-50 和 PDC 模块中的二维卷积操作替换为三维卷积，我们的网络 P-ConvLSTM 不仅可以读取连续的视频图像，还可以获取到所输入的连续图像的之间联系的特征信息。另外，将二维卷积修

改为三维卷积时，还需对步长、归一化等参数也许进行修改，确保不会出现融帧、丢帧的情况。图 3 展示了 P-ConvLSTM 的具体结构和模块之间的连接。

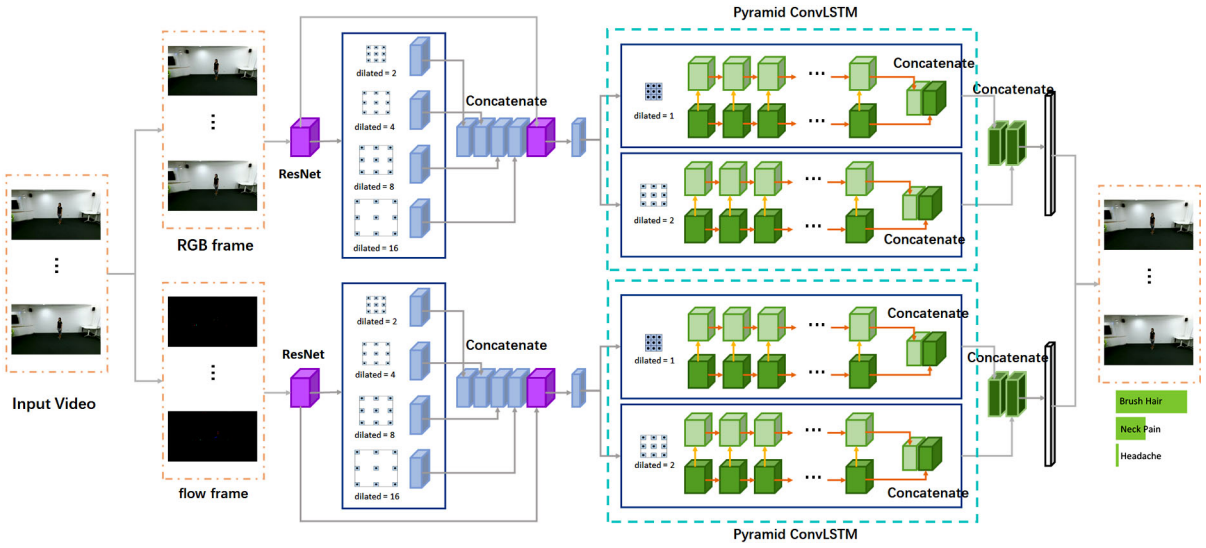


Figure 3. P-ConvLSTM
图 3. P-ConvLSTM 结构示意图

4. 实验

本文使用的数据来自南洋理工大学的 NTU RGB + D 人体动作识别数据库[13]。该数据库中包含多种类型人体动作数据，包括彩色视频数据、3D 人体骨骼数据、深度信息等。涵盖了 40 种日常动作、19 中医学动作及 11 种多人动作，每种动作包含 948 个数据。我们使用该数据库中的彩色视频数据，并从中挑选了 5 种常见动作进行试验。这 5 种动作包括梳头、坐下、起立、挥手及跌倒。

4.1. 实验环境及设置

我们使用 GTX 1060 对整个训练过程进行加速，运行平台是 win 10。整个 P-ConvLSTM 使用 Pytorch 框架进行搭建。在数据处理的阶段，我们从每一类中抽取 10%的数据作为测试集，63%作为训练集，23%作为验证集。训练时，我们每次读取两个视频，每个视频长度为 10 帧。初始设置的学习率为 0.001，每训练五十轮，学习率乘以 0.1。在训练过程中，每训练 100 次，记录一次模型的损失值；每进行完一轮训练后，即进行一次验证操作并记录模型的准确率。

在训练过程中，通过观察损失值和准确率的变化情况，我们发现：在训练进入到第五十轮以后，损失值和准确率的变化明显变慢；当训练进入六十轮以后，损失值趋于稳定，准确率基本保持不变。

之后我们调取第五十轮保存的模型数据，并调整训练集和验证集中数据，将训练集中部分数据与验证集中数据进行交换；将学习率调整为初始值，并再次进行训练，结果并无明显变化。这时，我们判断模型训练已经达到极限。

4.2. 实验结果

在测试阶段，我们首先对 P-ConvLSTM 及各个分支进行了准确度测试。在表 1 中，展示了 P-ConvLSTM 的准确度以及在各个类上的准确度。从结果中我们可以看出，时间信息分支的准确度要高于空间信息分支的准确度，P-ConvLSTM 的准确度高于任意一个分支。同时，P-ConvLSTM 在跌倒、挥手这两个动作

的准确度明显高于其他动作。

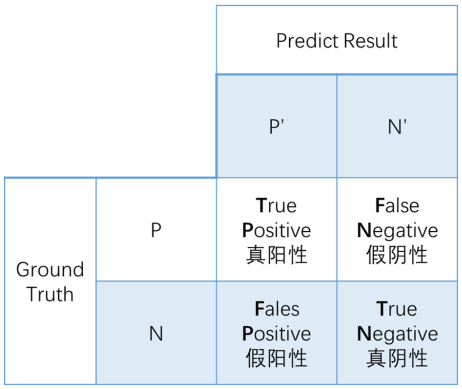
Table 1. P-ConvLSTM and its branch accuracy
表 1. P-ConvLSTM 及其分支准确度

准确度	P-ConvLSTM	空间信息分支	时间信息分支
总体	76.10	61.48	71.46
梳头	76.25	65.00	71.25
坐下	65.56	42.22	60.00
起立	77.17	46.74	72.83
挥手	83.13	78.31	77.11
跌倒	79.07	77.91	76.74

为进一步分析 P-ConvLSTM 的性能，我们计算了 P-ConvLSTM 的混淆矩阵，利用混淆矩阵对 P-ConvLSTM 进行进一步的分析。通过混淆矩阵，我们可以直观地了解到测试集中某一个动作，如跌倒动作正确分类和错误分类的数据，还可以知道其他动作的数据有多少被分类为这一跌倒动作。

我们以跌倒这一动作为例对混淆矩阵进行说明。假设 P-ConvLSTM 仅能识别跌倒和非跌倒。P-ConvLSTM 接收训练集中的数据后，对每一个数据进行处理并给出预测标签。这时会出现四种情况：真实标签为跌倒、预测标签也为跌倒的真阳性(True Positive, TP)；真实标签为跌倒、预测标签为非跌倒的假阴性(False Negative, FN)；真实标签为非跌倒、预测标签为跌倒的假阳性(False Positive, FP)；以及真实标签为非跌倒，预测标签也为非跌倒的真阴性(True Negative, TN)。根据这四种情况，对 P-ConvLSTM 的分类结果进行统计并将结果记录到与图 4 中类似的表格中，而表格中数据即为混淆矩阵中元素。将非跌倒进行更细致地划分可以得到多分类混淆矩阵。

在多分类混淆矩阵中，真阳性按序被记录在对角线上。某一个类的假阳性按照它们的真实标签依序被记录在该类的真阳性的同一列上；该类的假阴性依照预测标签的顺序依次记录在该类的真阳性的同一行；而混淆矩阵的其他部分是该类的真阴性。



$$FP\ rate = \frac{FP}{FP + TN}$$

$$TP\ rate = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

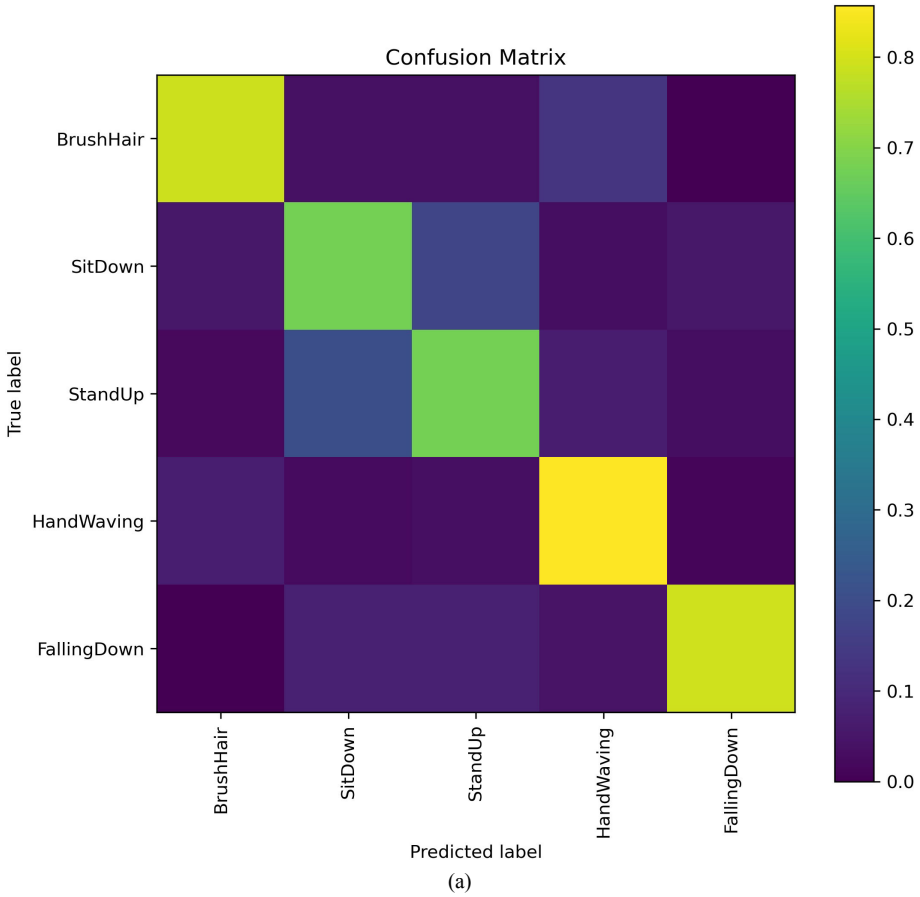
$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

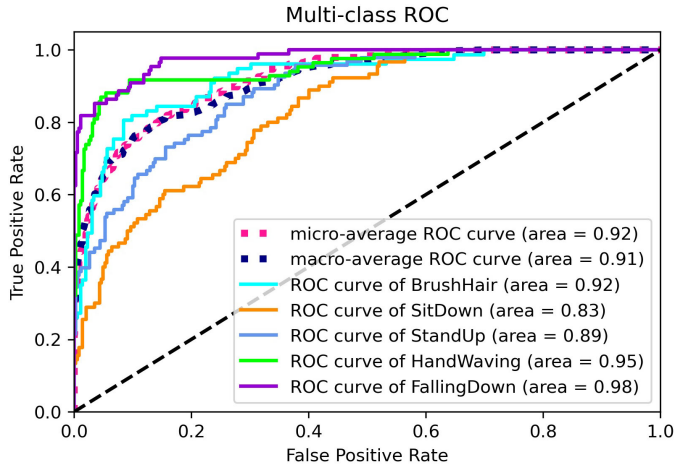
Figure 4. Confusion matrix and common performance metrics calculated from it [14]
图 4. 混淆矩阵及相关通用性能指标[14]

通过计算获得混淆矩阵后，经过归一化及上色将 P-ConvLSTM 的混淆矩阵绘制为图像，并在图 5(a)中展示。我们期望混淆矩阵对角线颜色尽可能的接近黄色，其他部分尽可能接近深蓝色，这样代表 P-ConvLSTM 可以较为准确地区分每一个类。其中，挥手这一动作的对角线颜色最接近黄色；梳头、跌

倒的对角线颜色都偏向黄色；坐下、起立的对角线颜色偏向绿色。同时，坐下、起立中容易出现混淆。P-ConvLSTM 可以有效地识别五类动作。



(a)



(b)

Figure 5. Confusion matrix and ROC curve
图 5. 混淆矩阵和 ROC

进一步对预测标签进行分析，利用真阳性率和假阳性率绘制 ROC 曲线。在某一类人体动作的数据中，真阳性率反映了 P-ConvLSTM 正确识别该动作的性能，假阳性率反映了 P-ConvLSTM 将其他人体动作的

数据中被识别为该类的情况。以真阳性率为纵坐标，假阳性率为横坐标绘制 ROC 曲线。我们希望 ROC 曲线尽可能的接近(0, 1)点，这代表混淆矩阵中的假阳性和假阴性都为零，P-ConvLSTM 能将所有数据完美分类。而 ROC 曲线越接近原点(0, 0)与(1, 1)点之间的直线，则表示分类结果接近随机分类。使用 ROC 曲线时，除通过直接观察 ROC 曲线是否接近(0, 1)点，还可通过曲线与 fp rate = 1、横坐标轴所围区域面积的大小进行分析，所围面积越接近 1 越好。在图 5(b)中展示了 P-ConvLSTM 的 ROC 曲线。其中，macro-averaging、micro-averaging 分别表示了算术平均和加权平均下 P-ConvLSTM 的 ROC 曲线。我们可以看到跌倒的 ROC 曲线最接近理想情况，但所有 ROC 曲线下的面积在 0.8 以上。

我们在训练过程中，跟踪记录了 top-1、top-3 的正确率。但在训练后，发现所记录的数据无法反馈 P-ConvLSTM 的性能。因此，我们在测试阶段绘制了 CMC 曲线(Cumulative Match Characteristic curve)。通过图 6 中 CMC 曲线看出 P-ConvLSTM 的 Top-1 准确率略高于 0.75，而 Top-2 准确率陡然上升到 0.9。

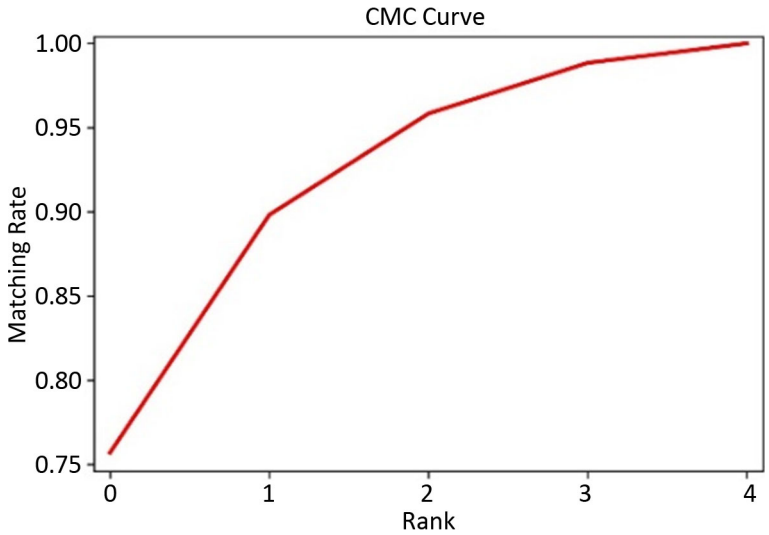


Figure 6. CMC curve
图 6. CMC 曲线

5. 结果分析

P-ConvLSTM 在挥手、跌倒、梳头等动作的表现都较为让人满意。但在坐下、起立两个动作上表现差强人意。从 CMC 曲线可以看出，P-ConvLSTM 给出的概率向量中，真实标签的概率位于高位，但存在其它选项影响判断。我们判断出现上述情况的主要原因有两点：一是在数据经过预处理后，图像分辨率降低，人物动作高度模糊，动作范围缩小；二是 P-ConvLSTM 只能接收连续的十帧视频图像，接收的视频过短无法感知完整的动作。

与其他的双流法模型相比，P-ConvLSTM 可以接收连续的视频图像。我们比对多个双流法模型的代码，发现这些模型在数据读取时，读取的并非连续的视频图像，而是视频中的某一帧图像。这些方法将视频中每一帧图像与真实标签建立联系，进而实现视频中动作的识别。P-ConvLSTM 使用三维卷积和 ConvLSTM 相结合的方法，可以一次读取多帧连续图像，获取每帧图像的信息以及图像之间相互联系，进一步获取动作整体信息，并将其与真实标签建立联系。

目前，P-ConvLSTM 能接收的图像尺寸较小，且无法接收更长的连续图像。在之后的工作中，我们将在图像预处理阶段进行改进，将原始图像中与人物动作无关的部分截去，令人物占据 P-ConvLSTM 的输入图像的主要范围。另外，进一步修改目标识别网络的识别能力让其能在更多情境下有更好的表现，

进而提高 P-ConvLSTM 的识别能力。

6. 总结

我们向双流法中引入目标识别模型，这种尝试证明是可行的，但所引入的目标识别模型的识别准确度会影响我们人体姿态识别的准确度。同时，每次只能读取 10 帧视频图像直接影响 P-ConvLSTM 获取人体动作的长期依赖关系，影响动作识别的准确度。

同时，因使用光流信息，且一般提取光流信息的方法速度较慢，而使用 GPU 给光流提取进行加速的方法需要进行复杂的配置过程，对一般使用者来说较难实现。若无法对光流提取进行加速，那么将降低双流法在实际应用中的实时性。

在下一步的工作中，我们将尝试不同的目标识别模型以探索目标识别网络性能对人体动作识别结果的影响，还将尝试读取更多视频图像查看对动作识别的影响。

参考文献

[1] Xiao, B., Wu, H. and Wei, Y. (2018) Simple Baselines for Human Pose Estimation and Tracking. *The European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 466-481. https://doi.org/10.1007/978-3-030-01231-1_29

[2] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, **27**, 568-576.

[3] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D. and Tang, X. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *The European Conference on Computer Vision*, Amsterdam, 8-16 October 2016, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[4] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6299-6308. https://doi.org/10.1109/CVPR.2017.502

[5] Feichtenhofer, C., Pinz, A. and Zisserman, A. (2016) Convolutional Two-Stream Network Fusion for Video Action Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 1933-1941. https://doi.org/10.1109/CVPR.2016.213

[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. *The IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[7] Shi, X., Chen, Z., Wang, H. and Yeung, D. (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, **28**, 802-810.

[8] Majd, M. and Safabakhsh, R. (2019) A Motion-Aware ConvLSTM Network for Action Recognition. *Applied Intelligence*, **49**, 2515-2521. https://doi.org/10.1007/s10489-018-1395-8

[9] Zhu, G., Zhang, L., Shen, P. and Shah, S.A.A. (2019) Continuous Gesture Segmentation and Recognition Using 3DCNN and Convolutional LSTM. *IEEE Transactions on Multimedia*, **21**, 1011-1021. https://doi.org/10.1109/TMM.2018.2869278

[10] Song, H., Wang, W., Shen, J., Zhao, S. and Lam, K.M. (2018) Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection. *The European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 715-731. https://doi.org/10.1007/978-3-030-01252-6_44

[11] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[12] Kim, S., Hong, S., Joh, M. and Song. S. (2017) DeepRain: ConvLSTM Network for Precipitation Prediction Using Multichannel Radar Data. Climate Informatics Workshop. arXiv:1711.02316 [cs.LG]

[13] Shahroudy, A., Liu, J., Ng, T. and Wang, G. (2016) NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *The IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 1010-1019. https://doi.org/10.1109/CVPR.2016.115

[14] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010