

RFC-Net: 基于残差结构的动作质量评估网络

周娴玮, 赖 坚, 陈伟涛, 阮 乐, 李振丰, 余松森

(华南师范大学 软件学院, 广东 佛山 528010)

摘 要: 动作质量评估是视频分析中一个重要且具有挑战性的问题, 动作质量评估是指对特定动作(如跳水、体操等)的完成质量进行评分, 分数评估模型是通过将视频特征回归到该领域专家提供的真实分数来进行学习。现有的大多数方法是直接使用动作识别任务的模型如(C3D 和 I3D)来解决问题。为了增强网络模型的特征提取效果, 从而提高分数回归的准确性, 该文提出了一种基于残差结构的动作质量评估网络模型 RFC-Net。该网络由特征提取器和特征聚合器组成, 在特征提取器中使用 I3D 网络对视频特征进行提取, 在特征聚合器中对特征提取器最后一层卷积得到的视频特征分别进行平均的全局池化和残差卷积操作, 对得到的结果进行特征融合, 最后输出视频的分数的表示。在动作质量评估领域公开的 MTL-AQA 数据集上, 该方法取得的斯皮尔曼相关性系数为 0.946 3。为进一步验证模型在不同背景下、动作差异较大时的泛化能力, 制作了羽毛球运动视频数据集, 并在此基础上进行了不同模型之间的对比实验。

关键词: 动作质量评估; 视频特征提取; 视频特征聚合; 神经网络; 斯皮尔曼相关性系数

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2022)11-0146-08

doi:10.3969/j.issn.1673-629X.2022.11.022

RFC-Net: Action Quality Assessment Network Based on Residual Structure

ZHOU Xian-wei, LAI Jian, CHEN Wei-tao, RUAN Le, LI Zhen-feng, YU Song-sen

(School of Software, South China Normal University, Foshan 528010, China)

Abstract: Action quality assessment is an important and challenging problem in video analysis. Action quality assessment refers to scoring the completion quality of specific actions (e. g. diving, gymnastics, etc.) and score assessment models are learned by regressing video features on the true scores provided by experts in the field. Most existing approaches use models for action recognition tasks such as (C3D and I3D) directly to solve the problem. To enhance the feature extraction effect of the network model and thus improve the accuracy of the score regression, we propose a residual structure-based action quality assessment network model RFC-Net. The network consists of a feature extractor and a feature aggregator, in which the video features are extracted using the I3D network in the feature extractor, and the video features obtained from the last layer of convolution in the feature extractor are subjected to average global pooling and residual convolution operations, respectively. The proposed method has achieved a Spearman correlation coefficient of 0.946 3 on the MTL-AQA dataset, which is publicly available in the field of action quality assessment. In order to further validate the generalisation ability of the model under different contexts and big action differences, a badminton video dataset was produced and a comparison experiment between different models was conducted on this basis.

Key words: action quality assessment; video feature extraction; video feature aggregation; neural networks; Spearman's correlation coefficient

0 引 言

人类运动动作质量评估(Action Quality Assessment, AQA)指的是评估一个特定动作的执行情况, 为该动作进行打分。动作质量评估在现实中具有巨大的应用价值, 如运用在体育视频分析^[1-6]、外科手术训练^[7-8], 以及其他的一些技能评估中^[9-10]。

动作质量评估相较于人类动作识别(Human Action Recognition, HAR)更具有挑战性, 因为 HAR 是识别不同类别的动作, 其动作之间差别较大, 然而在 AQA 领域中, 处理对象基本上都是同一类别下的动作, 其动作间的差别较为细微, 难以区别。目前绝大部分的运动, 其中运动员比赛的评分(如跳水^[11]、滑

收稿日期: 2022-06-29

修回日期: 2022-08-30

作者简介: 周娴玮(1982-), 男, 博士, 讲师, 硕导, 研究方向为多智能体系统、机器视觉、多传感信息融合等; 通信作者: 余松森(1972-), 男, 博士后, 教授, 硕导, 研究方向为计算机视觉、多传感信息融合等。

冰^[12]等),都是由相关领域的专家评委根据运动员的表现给出相应的分数。在现实生活中,一名合格的专家评委是非常稀少的,因为他们必须经过长期的训练才能熟悉所有特定的动作,因此用自动评分系统取代教练评分是未来的一种趋势,另一方面,自动评分系统某种程度来说比较公平公正能够避免评分丑闻。

近几年里在 AQA 领域提出了许多方法^[13-15]试图解决分数预测问题,文献[16]通过对视频中的运动员进行追踪,只提取视频中运动员有关的特征。这种方法虽然一定程度上降低了背景因素的干扰,但是对于跳水运动而言,水花溅起的大小和高度,也是决定最后分数的关键因素,因此不根据运动特性而去除背景因素是不太合理的。文献[17]通过对比学习进行分数预测,在数据集中挑选部分视频作为范例视频,然后通过学习范例视频与输入视频之间的相似性来预测输入视频最终的评分。这类方法学习的相似度,通常会存在较大的误差,并且需要人工选择范例视频,这在某种程度上增加了预测的复杂度。文献[11,18-20]中的方法都是通过端到端之间的模型进行分数预测。在此类方法中,文献[11]通过增加卷积层的数量加强特征提取的效果,但是卷积层数的增加,会导致网络出现了退化,有效特征丢失。文献[21]使用 LSTM 作为特征聚合器,LSTM 在卷积的顶层只能获取高层次的动作而不能获取关键的低层次动作,作为特征聚合器不能起到一个很好的效果。

现有的方法不能有效地执行特征聚合,在 AQA 任务中需要一种简单、有效的特征聚合机制。为解决上述问题,该文提出了一种端到端的 RFC-Net (Residual Full Connection Network) 模型用于预测视频的分数。AQA 模型由特征提取器和特征聚合器两个部分组成,特征提取器是用于视频特征提取的网络,特征聚合器是用于特征聚合以进行分数回归的网络。3D Convolutional Networks (C3D)^[22]因为能提取出视频中物体信息、场景信息和动作信息的特征,不需要根据特定任务进行微调都可以取得不错的效果,被广泛用于动作质量评估领域。C3D 看似更适合做视频处理,但存在维度问题,经过 8 层卷积层到最后全连接层有 4 096 个输出单元,这样就会更难训练,并且不能有效地将 2D 网络的预训练权重迁移到 3D 网络。所以该文的特征提取器采用 Two-Stream Inflated 3D ConvNets (I3D)^[23],I3D 主要依据最优的图像网络架构实现,对它们的卷积和池化核从 2D 扩展到 3D,并选择使用它们的参数,最终得到了非常深的时空分类网络,并且分别采用了不同帧组成的 Clip。该文的特征聚合器由平均池化层和 RFC Block 组成,其中 RFC Block 参照残差网络 (Residual Networks) 的设计,卷积

操作的接受域范围有限,导致了长期依赖关系的损失,所以使用全连接层作为权重层,每层的全连接层之间加入激活函数,最后再恒等映射 (identity mapping) 聚合所有特征进行输出。

为进一步验证模型在不同背景下、动作差异较大时的泛化能力,该文制作了一类羽毛球运动的数据集简称 (BS dataset),其中大部分是从视频媒体共享平台收集而来的羽毛球运动的训练视频。把所收集到的视频交由专业的羽毛球教练根据不同难度的动作标准进行评分,最后参照 MTL-AQA 数据集的格式对其进行标签化操作。与现有的数据集相比,该数据集有着其他的一些特性。首先视频是教学的视频而不是比赛的视频,在教学中动作会较为缓慢,能更清楚识别出动作;其次羽毛球运动是人使用长柄网状球拍击打羽毛球的体育项目,不仅仅是要考虑人的动作是否标准,球拍的位置和握拍的方式也会影响动作的分数;最后该数据集包括了不同背景下 (不仅仅有羽毛球球场背景,还有居家背景) 的教练或学员打羽毛球的视频,增加了背景因素的影响,加大了预测分数的难度。

RFC-Net 方法在 MTL-AQA 数据集以及 BS-AQA 数据集上进行了测试。实验结果验证了该模型能够提高视频中的动作分数预测效果。该文的主要贡献可以总结如下:

(1) 针对动作质量评估中预测的分数误差较大问题进行改进。

(2) 提出了 RFC Block 用于视频的特征聚合,消融实验表明该模块能够提高特征聚合的效果。

(3) 提出了一种基于残差结构的 RFC-Net 模型,该模型由特征提取器和特征聚合器组成,模型在 MTL-AQA 数据集以及 BS-AQA 数据集中取得了较好的结果。

1 相关工作

动作质量评估领域与动作识别领域方法相类似,这个领域几乎所有的工作都将视频中人类动作表现分数的问题作为一个回归问题。文献[24-25]对方法做了一个概括,该文按照不同的视频处理对该领域的方法进行划分,分为基于视频的方法以及基于人体骨架的方法。

1.1 基于人体骨架的动作质量评估

基于人体骨架的方法表现为,使用姿态估计算法对视频中的人体骨架进行识别,通过对姿态的识别得到人体骨骼各个关键点的信息。因为是从视频中识别而来的数据,所以这些关键点的信息是一个个二维坐标,所得到的 2D 人体骨架信息大多数使用图卷积神经网络提取特征来训练回归器。

文献[26]提出了一种新的时空金字塔图卷积网络(ST-PGN),用于人体动作质量评估和姿态估计,此方法能够使用来自骨架特征层次的所有特征。作者的另外一篇文献在此基础上又提出了一种新的多任务框架,该框架利用图卷积网络主干把人类骨架关节之间的互联特性嵌入到所提取的特征中,然后根据不同的任务需要进行不同的处理^[27]。文献[28]建立了可训练的骨架关节关系图,并分析了其中的关节运动,提出了两个新的模块,联合共性模块和联合差异模块,用于关节运动学习,此外还有文献[29-31]也属于这类方法。

虽然骨骼数据相比视频数据,可以去除视频中背景因素的影响,专注于人体的动作姿态,能够针对于不同环境下的动作进行质量评估。但是一个动作往往和物品或环境有着交互的关系,对只关注于动作本身的骨架信息而言难以分辨出这其中的联系。

1.2 基于视频的动作质量评估

基于视频的方法表现为,直接把 RGB 视频流作为输入,使用特征提取网络来提取视频中的特征,这种特征包含人体的动作信息以及视频中的环境信息。这也可以概括为使用模型来学习视频和动作分数之间的直接映射,然后采用这种映射关系用于预测新的视频中的动作分数。这些工作大多数使用三维卷积神经网络来提取视频特征,然后使用回归方法来获得预测分数,不同的论文在分数回归阶段的处理方法也不尽相同。

随着视频分析领域的进展,动作质量评估领域采取更深层次的特征提取网络对视频中的特征进行提取,同时还用不同的方法进行分数的回归。文献[18]

中把动作识别中的框架引入到动作质量评估中,使用 C3D 以及 C3D-LSTM 的方法对特征进行提取,最后使用 SVM 回归预测分数,效果相较于之前论文中的单独 C3D 卷积有了一定的提高。此外,在文献[13]中还发布了一个全新的多任务动作质量评估(MTL-AQA)数据集,使用了两种不同的模型 C3D-AVG, MSCADC 验证多任务对动作质量评估的影响。文献[21]使用 C3D 提取特征序列,然后用多尺度卷积跳跃 LSTM(M-LSTM)和自注意 LSTM(S-LSTM)这两个独立模块进行处理,对特征进行聚合,用于预测分数。文献[32]引入了 LDL(标签分布学习)方法。在分数回归阶段设计了一种不确定性感知分数分布学习(USDL)方法来探索一个分数的组成,通过将给定的单个分数标签转移到高斯分布分数中进行学习,从而直接估计动作视频的得分分布。文献[33]中对其进行改进,解决了 USDL 不适用于具有连续标签的数据集以及在训练中需要固定的方差的问题,进一步开发了分布式自动编码器(DAE),DAE 同时具有回归算法和标签分布学习(LDL)的优点。

除了直接进行分数回归外,一些学者也把目光转向了另一种解决方案:文献[16]使用孪生网络在给定的动作视频与参考视频之间进行比较,计算视频之间的相似度,从而根据参考视频的分数评估出给定视频的分数。尽管对比回归框架可以预测相对分数,但是相对分数通常取值范围很广,因此文献[34]提出一个群意识(group-aware)回归树将对学习得到的相对分数做了更为细致的回归,该方法是目前 AQA 领域的 SOTA。

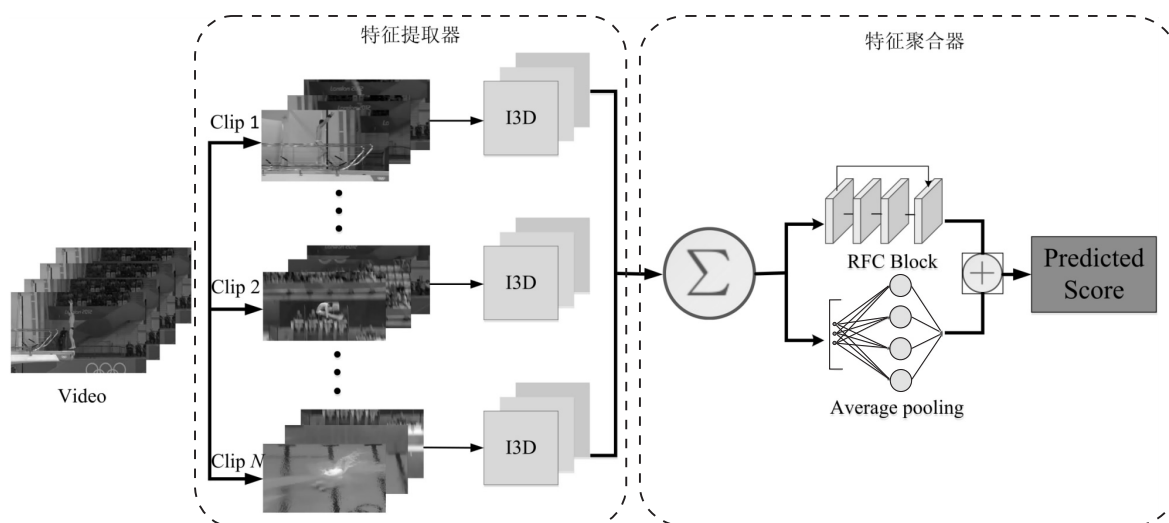


图 1 RFC 网络结构

2 基于残差结构的动作质量评估方法

该文使用特征提取器进行视频的特征提取,一个完整的视频帧数量太多,不能够一次输入到特征提取

器中,首先对视频进行分割处理,然后把分割完的视频片段分别输入特征提取器,特征提取器提取的特征向量作为特征聚合器的输入。为了使预测的结果更加接近真实评分,该文提出了一种由 RFC Block 和平均池

化层构成的特征聚合器,聚合的结果将作为预测的分数结果。

在本章节将会对 RCF-Net 模型进行详细的描述,其中内容包括 RFC-Net 网络结构,如何对视频特征进行提取以及 RFC 结构如何聚合视频特征。

2.1 特征提取器

在进行特征聚合之前必须得对视频的特征进行提取,对于特征提取网络的选择,之前大部分论文都是采取 C3D 作为特征提取器,但是一般 3D 网络的深度较浅和参数过多,这样影响了模型的表达能力和加大了训练的难度。而 I3D 作为较为优秀的一种特征提取架构被广泛用于动作识别以及动作质量评估中。它以最新的图片分类模型为基础结构,将 kernels 膨胀 (inflate) 结合到 3D Conv。可以从视频中学习到时空特征,同时成功把 ImageNet 中的预训练权重扩展到视频行为识别中,因此 RCF-Net 模型选取 I3D 网络作为视频的特征提取网络。

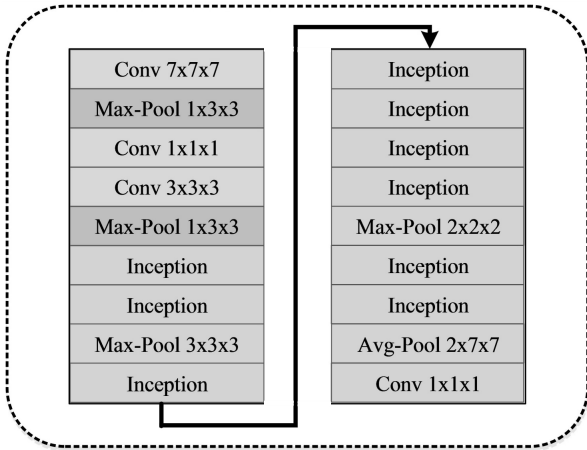


图 2 I3D 网络结构

该文所采用的 I3D 架构 (见图 2) 采用了 4 个卷积层、5 个池化层以及 Inception 模块,除最后一个卷积层之外,在每一个卷积后面都加入了 Batch Normalization 层和激活层 (ReLU)。

对于一个给定的具有 T 帧的 $V = \{F_t\}_{t=1}^T$ 视频,其中 F_t 代表第 t 帧,利用一个滑动窗口将其分割成 N 个不重叠的片段,其中每个片段包含 T/N 个连续的帧。收集到的 N 个片段即 $\{\text{Clip}_1, \text{Clip}_2, \dots, \text{Clip}_N\}$ 被进一步发送到 I3D,不同的片段之间得到 N 个特征,分别为 $\{f_1, f_2, \dots, f_N\}$,在 C3D-AVG 中表明了平均池化层具备良好的加快计算速度和防止过拟合的作用,把这 N 个特征通过平均池化对其进行聚合。

$$F(x) = \sum_{x=1}^N f_x \quad (1)$$

$F(x)$ 为视频通过 I3D 网络所提取的特征,输出是一个 1 024 维的特征向量,随后把 $F(x)$ 作为特征聚合器的输入。

2.2 特征聚合器

在动作质量评估中,一个视频中包含的运动的视频帧数较少,一些模型试图使用扩展更多的卷积进行更深层次特征提取。在一定程度上,网络越深越大表达能力就越强,提取的不同层次的信息便越多。但是随着特征提取网络层数的增加,会带来许多问题,网络出现了退化,有效特征丢失,这便导致网络的效果逐渐降低。在经过前面若干次卷积、激励、池化后,模型会得到一个高质量的全连接层,因此该文不再增加卷积用于特征的提取,而是把提取到的特征进行有效的聚合以提高分数的预测效果。RFC Block 网络结构如图 3 所示。

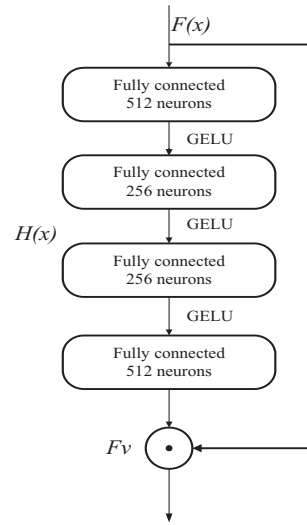


图 3 RFC Block 网络结构

在接受特征提取器输入的特征 $F(x)$ 后, $F(x)$ 分别输入到平均池化层以及 RFC Block 中。该文所提出的 RFC Block,参照残差网络结构设计而成,由于卷积操作的接受域范围有限,导致了长期依赖关系的损失,因此隐藏层由四层的全连接层组成,每层全连接层的特征值个数分别为 $\{512, 256, 256, 512\}$,并且在每层之间加入激活函数 GELU。由于隐藏层全部使用全连接层,会导致特征值数目过大,因此需要随机删除全连接中的部分特征值以减少参数量。Dropout 通过将一些激活数乘于 0 来规范化模型,ReLU 作为激活函数引入非线性,强化网络的学习能力,而 GELU 可以看作 Dropout 和 ReLU 的结合,在后续的实验部分 RFC Block 分别使用了这两类激活函数,验证这两类激活函数在本模型中的效果。GELU 激活函数公式如下:

$$\text{GELU}(x) = xP(X \leq x) = x\varphi(x), x \sim N(0, 1) \quad (2)$$

其中, x 是输入值, X 是具有零均值和单位方差的高斯随机变量。 $P(X \leq x)$ 是 X 小于或等于给定值 x 的概率, $\varphi(x)$ 是指高斯正态分布的累积分布。全连接层是一维列向量,经过了隐藏层和激活后得到的特征可

以与一开始输入的特征进行聚合, RCF Block 可以表示为:

$$F_{\text{RFC}}(x) = F(x) \odot F(x_l | w_l) \quad (3)$$

其中, $F(x_l | w_l)$ 表示残差块中隐藏层中的输出特征, x_l 为输入隐藏层之前的特征, w_l 为隐藏层学习到的权重, 其中 l 为隐藏层的层数 $l \in [1, 4]$, RFC Block 的输出 $F_{\text{RFC}}(x)$ 为两个通道数的合并, 使得描述图像的特征维度增加, 而每一维度特征下的信息量不变, $F(x)$ 经过平均池化层得到的 F_{avg} 特征值数为 512。故整个特征聚合模块的输出为:

$$F_V = F_{\text{RFC}} \oplus F_{\text{avg}} \quad (4)$$

F_V 为 RFC Block 与平均池化层聚合的特征, 采取对应元素位置相加的聚合方式, 在维度不变的情况下使描述图像的特征每一维下的信息量增多, 显然对最终的图像的分类是有益的。最后 F_V 进行回归得到该视频动作的预测分数。

2.3 损失函数

损失函数用来评价模型的预测值和真实值不一样的程度, 不同的模型用的损失函数一般也不一样。该文需要预测视频中的动作质量分数, 这可以看作一个分数回归的任务, 给定带有动作质量标签的输入视频, 基于输入视频预测动作质量:

$$\bar{S} = F_V(F(x)) \quad (5)$$

其中, \bar{S} 为预测分数, $F(x)$ 为特征提取器, F_V 为特征聚合器。动作质量评估的回归问题通过预测分数与真实分数之间的误差来解决, 损失函数可以表示为:

$$\ell_{\text{AQA}} = \frac{1}{n} \sum_{i=1}^n (S_i - \bar{S}_i)^2 \quad (6)$$

该文采用 MSE 作为损失函数, 用于评估模型的效果, 训练过程中均方误差越小则预测分数越接近真实得分。

3 实验

3.1 数据集

3.1.1 MTL-AQA dataset

这是一个于 2018 年发布的 AQA 领域数据集。它包含了 1 412 个视频样本, 是迄今为止该领域最大的 AQA 数据集。这个数据集关注跳水运动, 所有的样本都是来自于不同国际比赛中的跳水运动。这些视频包含了 103 帧。它们有不同的视角和相机角度。该数据集包含男女运动员的样本, 个人和同步跳水、3 米跳台和 10 米跳台跳水、奥运裁判的最终动作质量成绩、任务难度水平、赛事的评论, 以及细粒度的动作标签。

3.1.2 BS-AQA dataset

为进一步验证模型在不同背景下、动作差异较大

时的泛化能力, 该文制作了羽毛球视频数据集, 目前已经提出了一些 AQA 数据集, 如 AQA-7^[35]、MTL-AQA^[11] 以及 FD-10^[12] 数据集主要包含体操、跳水、滑冰的动作。由于球类运动的特殊性, 与体操、跳水、滑冰等运动不一样。在羽毛球等球类竞技比赛中, 对动作是否标准并无要求, 只要击败对手即可, 因此该文采用羽毛球运动训练阶段的视频来进行动作质量评估。

视频主要来源于各个视频网站中的羽毛球训练视频, 为每个视频进行编号, 并请羽毛球教练为每个视频进行评分。BS-AQA 数据集分数分布如图 4 所示。

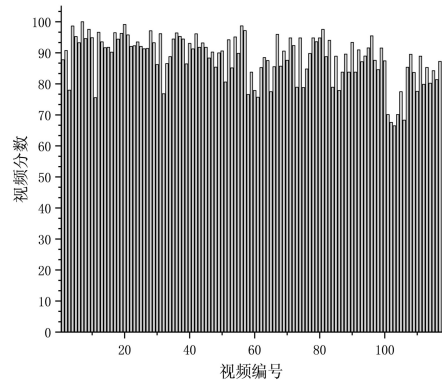


图 4 BS-AQA 视频得分分布

关于羽毛球运动数据集评分:

在羽毛球教练的建议下, 把羽毛球运动分为四个阶段, 不同的阶段在整个运动中所占的权重不一致。不同阶段动作的评分进行加权求和从而得到总体的评分, 这无疑比直接通过整个视频的直接评分更加合理。并且为了降低教练评分的主观性, 邀请多位教练分阶段对视频进行评分, 最后所得的分数为多位教练的评分取均值。用 $\text{Stage}_i, i \in [1, 4]$ 表示教练对 i 阶段的评分, 各个阶段的评分系数由教练根据经验得出。

$$\text{Score}_{\text{overall}} = \text{Stage}_1 * 0.4 + \text{Stage}_2 * 0.2 + \text{Stage}_3 * 0.3 + \text{Stage}_4 * 0.1 \quad (7)$$

3.2 评价指标

在动作质量评估领域, 主要采用斯皮尔曼等级相关系数 (Spearman's rank correlation, Sp. Corr) 作为评价指标, Spearman's rank correlation 是反映两组变量之间联系的密切程度, Spearman's rank correlation 用 ρ 来表示。对于两组大小为 n 的数据 X 和 Y , 将其转换为等级的数据 $x_i, y_i, i = 1, 2, \dots, n$, \bar{x} 和 \bar{y} 分别表示 x_i, y_i 的平均值, 相关系数 ρ 可以表示为:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (8)$$

3.3 实验细节

在整个实验中该文使用的是 Pytorch 框架, 并采用

在 Kinetics Dataset 上进行预训练的 I3D 模型作为特征提取器, I3D 中采用 ReLU 作为激活函数, 使用 MSE 损失函数以及 Adam 优化器, 学习率设置为 $1e-4$ 。把每个视频提取包含完整动作的 96 帧用作训练模型, 96 帧被分为 6 个 Clip 16 个帧剪辑或 3 个 Clip 32 个帧剪辑。视频帧较大直接输入网络会导致训练速度过慢等问题, 原始的视频帧的大小被调整到 171×128 , 随后裁剪后的视频帧大小为 112×112 。并且通过随机水平翻转来进行数据增强, 最后将 RGB 图像三通道的数值进行均值化、归一化处理, 3 个通道中的数据整理到 $[-1, 1]$ 区间, 得到 $\text{frames} \times 112 \times 112 \times 3$ 的输入。通过上述步骤的处理, 进一步降低了网络训练难度。

在 RFC-Net 中, 该文设置了消融实验用于验证 RFC Block 的有效性, 并且比较了不同的帧数的 Clip 对输出结果的影响。

3.4 实验结果

3.4.1 MTL-AQA 实验结果

(1) 与其他公开模型的结果对比。

由于 MTL-AQA 数据集包含动作的难度, 裁判将它们的分数与难度相乘得到最终分数, 该文选择将最后输出的分数与动作难度相乘。在 MTL-AQA 数据集中, 把近两年 (2020–2021) 文献 [19]、文献 [32] 和文献 [34] 所提出的四种模型 (ResNet34_ (2+1), MUSDL, USDL, CoRe) 与该文提出的模型进行对比, 结果如图 5 所示。

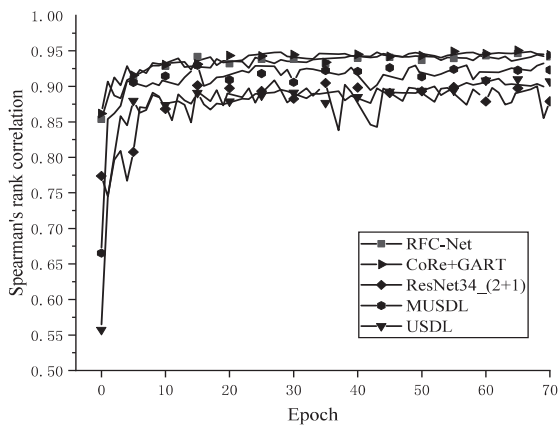


图 5 MTL-AQA 数据集实验结果对比

(2) 与 SOTA (CoRe+GART 模型) 对比。

从表 1 中结果可知, 除了文献 [34] 中模型外, 文中模型优于之前的所有模型, 在文献 [34] 中把对比学习用于动作质量评估, 通过比较两个不同分数的视频, 学习视频之间的差异, 最后使用群感知回归树来回归预测最终得分。该模型至今为止是 MTL-AQA 数据集中的 SOTA。文中模型对比于 SOTA 模型, Spearman's rank correlation 比文献 [34] 较低, 但是文献 [34] 采用的是对比学习方法, 预测视频所需要对比的

范例视频需要手动进行选择, 这使得模型变得更加复杂并且降低了模型预测的效率。在结果相差不是很大的情况下, 相比于对文献 [34] 的方法, 通过改进特征提取或特征聚合的方法进行端到端的学习更加简便以及更加贴合应用场景。

在 RFC-Net 中, 设置了消融实验用于验证 RFC Block 的有效性, 并且比较了不同帧数的 Clip 对输出结果的影响。

表 1 与 MTL-AQA 数据集上现有方法的性能比较

Method	Sp. Corr	Year
MSCADC-STL	0.847 2	2019
C3D-AVG-STL	0.896 0	2019
C3D-AVG-MTL	0.904 4	2019
USDL	0.923 1	2020
MUSDL	0.927 3	2020
ResNet34_ (2+1)	0.931 5	2021
CoRe+GART	0.951 2	2021
TSA-Net	0.942 2	2022
Ours RFC-Net	0.946 3	2022

(3) 消融实验结果对比。

为了进一步探究 RFC 模块是否能提升特征的聚合结果, 设置了消融实验, 即不加入 RFC Block 聚合特征而是直接对特征提取器提取的特征进行分数回归。此外参照文献 [19] 中的方法, 在 MTL-AQA 数据集上测试了不同帧数的 Clip 对结果的影响, 把 Clip 中的帧数分别设置为 16 与 32, 对比了 Clip 中不同帧数对实验结果的影响。Spearman's rank correlation 对比如图 6 所示, 可以看到在参数不变的情况下 32 帧比 16 帧的 Clip 效果相对较好一些。

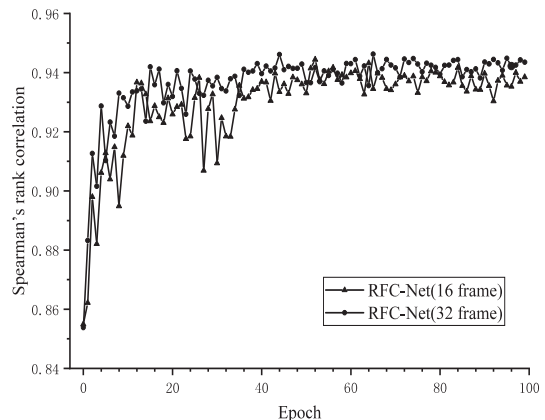


图 6 Clip 中不同帧数结果对比

RFC-Net 消融实验结果如表 2 所示。当 Clip 为 16 帧的时候, 加入 RFC Block 的结果优于没有 RFC Block 的结果, 这验证了 RFC Block 能够提升特征聚合的结果。文中模型在 Clip 为 32 帧的时候得到最优的结果。

表 2 RFC-Net 消融实验

16frame	32frame	RFC-Block	Sp. Corr
✓			0.937 5
✓		✓	0.944 4
	✓	✓	0.946 3

3.4.2 BS-AQA 数据集结果

为验证模型在不同背景下、动作差异较大时模型

的泛化能力,制作了羽毛球视频数据集,挑选了文献[11]和文献[19]中的三种采用端到端的分数回归方法模型(MSCADC, C3D-AVG, RestNet-34(2+1))与该文提出的 RFC-Net 方法进行对比。因为 CoRe+GART 模型需要人工挑选范例视频,而范例视频的挑选对实验结果影响较大,故在 BS-AQA 数据集中没有选择当前的 SOTA 模型进行对比。

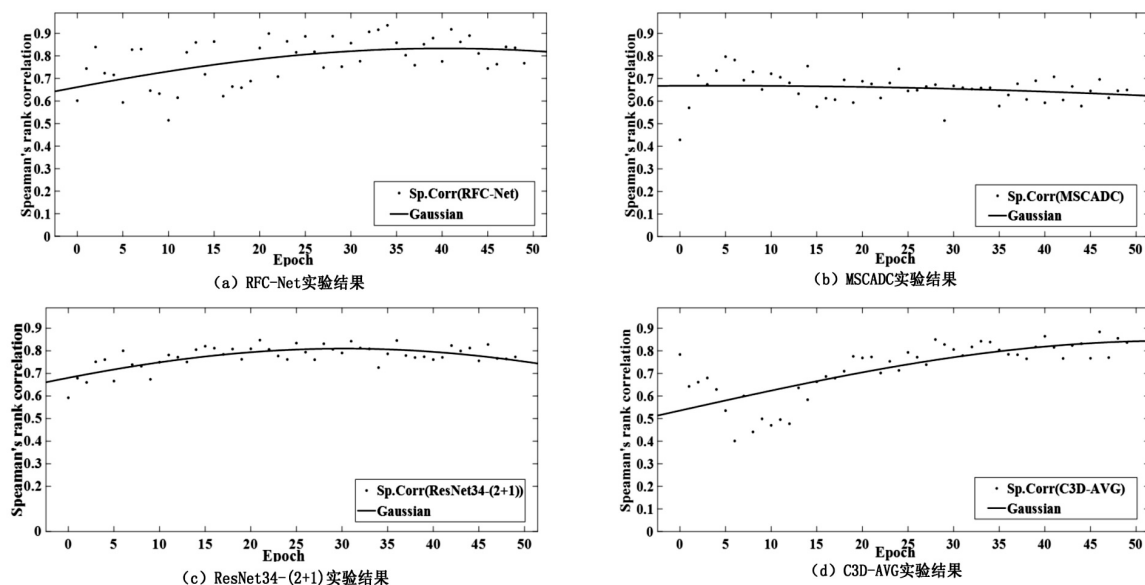


图 7 BS-AQA 数据集实验结果

四种模型在测试集上的结果如图 7 所示。散点为当前 Epoch 下的 Spearman's rank correlation,数据集中的视频背景复杂,动作差异较大,因此点的分布较为分散,为了找到一条线段来尽可能贴近地描述这些散点。该文使用非线性 Gaussian 函数对散点进行拟合,其中 Gaussian 函数有 8 种类型,选取其中基础类型,其函数表达式为:

$$a * \exp(c - ((x - b)/c)^2) \quad (9)$$

最后得到一条拟合线用于表示 Spearman's rank correlation 的回归结果。可以看到在 BS-AQA 数据集中 RFC-Net 的 Spearman's rank correlation 普遍高于其他三种模型。

表 3 BS-AQA 数据集性能比较

Method	Sp. Corr	MSE
MSCADC	0.830 7	27.82
C3D-AVG	0.883 9	18.58
ResNet34_(2+1)	0.863 7	4.52
Ours RFC-Net	0.935 5	6.89

表 3 给出了四种模型在 BS-AQA 数据集上的具体性能比较,比较结果显示在主要的评价指标 Spearman's rank correlation 中。RFC-Net 在四类模型中效果最优,而 MSE 略低于 RestNet_34(2+1),这表

明 RFC-Net 模型在不同的动作类别中仍然具备较好的结果,模型的泛化能力较强。

4 结束语

为了提高动作质量评分的准确性,提出了一种基于残差结构的动作质量评估网络模型。RFC-Net 由特征提取器和特征聚合器组成,在特征聚合器中采用了 RFC Block 和平均池化层进行特征聚合,通过消融实验表明, RFC Block 能够对提取的视频特征进行有效的特征聚合,更加准确地预测动作的得分, RFC-Net 模型在 MTL-AQA 数据集上取得了仅次于 SOTA 的结果。此外为了探究该模型的泛化能力而制作了 BS-AQA 数据集,实验结果表明在羽毛球运动动作质量评估中,与其他端到端的模型相比,该模型仍然表现出了具备竞争力的结果。

该方法仍具备改进的空间,未来将在以下两个方向进行研究:

(1) 在特征提取器中进行改进,特征提取是提取视频中所有的特征,但是和动作质量评估相关的特征在整个视频特征中占据较小的部分,如何精确提取运动员的运动特征是未来的一个研究方向;

(2) 如何在减少参数量的情况下提升其聚合效果也是未来的一个研究方向。

参考文献:

- [1] PIRSAVASH H, VONDRICK C, TORRALBA A. Assessing the quality of actions[C]//European conference on computer vision. Zurich: Springer, 2014: 556–571.
- [2] OGASAWARA T, FUKAMACHI H, AOYAGI K, et al. Archery skill assessment using an acceleration sensor[J]. IEEE Transactions on Human–Machine Systems, 2021, 51(3): 221–228.
- [3] LEI Q, ZHANG H, DU J. Temporal attention learning for action quality assessment in sports video[J]. Signal, Image and Video Processing, 2021, 15(7): 1575–1583.
- [4] SUKHWANI M, JAWAHAR C V. Tennisvid2text: fine-grained descriptions for domain specific videos[J]. arXiv: 1511.08522, 2015.
- [5] WANG T, JIN M, LI M. Towards accurate and interpretable surgical skill assessment: a video-based method for skill score prediction and guiding feedback generation[J]. International Journal of Computer Assisted Radiology and Surgery, 2021, 16(9): 1595–1605.
- [6] 艾新龔. 基于深度学习的跆拳道智慧教学算法与应用[D]. 新乡: 河南师范大学, 2020.
- [7] 马倩倩, 贺莉. 基于动作识别算法的健美操难度自动评分系统设计[J]. 西昌学院学报: 自然科学版, 2021, 35(2): 106–110.
- [8] ZHANG Q, LI B. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model[C]//Proceedings of the 2011 international ACM workshop on medical multimedia analysis and retrieval. Scottsdale: ACM, 2011: 19–24.
- [9] DOUGHTY H, MAYOL-CUEVAS W, DAMEN D. The pros and cons: rank-aware temporal attention for skill determination in long videos[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New York: IEEE, 2019: 7862–7871.
- [10] DOUGHTY H, DAMEN D, MAYOL-CUEVAS W. Who's better, who's best: skill determination in video using deep ranking[J]. arXiv: 1703.09913, 2017.
- [11] PARMAR P, MORRIS B T. What and how well you performed? a multitask learning approach to action quality assessment[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seoul: IEEE, 2019: 304–313.
- [12] LIU S, LIU X, HUANG G, et al. FSD-10: a dataset for competitive sports content analysis[J]. arXiv: 2002.03312, 2020.
- [13] DONG L J, ZHANG H B, SHI Q, et al. Learning and fusing multiple hidden substages for action quality assessment[J]. Knowledge-Based Systems, 2021, 229: 107388.
- [14] ZHANG S J, PAN J H, GAO J, et al. Semi-supervised action quality assessment with self-supervised segment feature recovery[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6017–6028.
- [15] GORDON A S. Automated video assessment of human performance[C]//Proceedings of AI-ED world conference on artificial intelligence in education. Kobe: [s. n.], 1995.
- [16] WANG S, YANG D, ZHAI P, et al. Tsa-net: tube self-attention network for action quality assessment[C]//Proceedings of the 29th ACM international conference on multimedia. Chengdu: ACM, 2021: 4902–4910.
- [17] JAIN H, HARIT G, SHARMA A. Action quality assessment using siamese network-based deep metric learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(6): 2260–2273.
- [18] PARMAR P, MORRIS B T. Learning to score olympic events[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Honolulu: IEEE, 2017: 20–28.
- [19] FARABI S, HIMEL H, GAZZALI F, et al. Improving action quality assessment using weighted aggregation[C]//Iberian conference on pattern recognition and image analysis. Aveiro: Springer, 2022: 576–587.
- [20] ZENG L A, HONG F T, ZHENG W S, et al. Hybrid dynamic-static context-aware attention network for action assessment in long videos[C]//Proceedings of the 28th ACM international conference on multimedia. New York: ACM, 2020: 2526–2534.
- [21] XU C, FU Y, ZHANG B, et al. Learning to score figure skating sport videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(12): 4578–4590.
- [22] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatio-temporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. Santiago: IEEE, 2015: 4489–4497.
- [23] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 6299–6308.
- [24] 张洪博, 董力嘉, 潘玉彪, 等. 视频理解中的动作质量评估方法综述[J]. 计算机科学, 2022, 49(7): 79–88.
- [25] 令安, 郑伟诗. 视频动作质量评估[J]. 中国传媒大学学报: 自然科学版, 2021, 28(5): 2935.
- [26] PARSA B, DARIUSH B. Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. Snowmass: IEEE, 2020: 1080–1090.
- [27] PARSA B, BANERJEE A G. A multi-task learning approach for human activity segmentation and ergonomics risk assessment[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. Waikoloa: IEEE, 2021: 2352–2362.