Computer Science and Application 计算机科学与应用, 2023, 13(3), 369-377
Published Online March 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.133035

Hans汉斯

基于CNN-BiGRU的足球视频片段分类方法

苏　航，张胜男

沈阳工业大学软件学院，辽宁　沈阳

收稿日期：2023年2月16日；录用日期：2023年3月15日；发布日期：2023年3月23日

摘　要

基于深度学习的视频分类是体育视频研究的一个重要方向。针对目前视频事件类型识别率低的问题，本文提出了一种基于CNN-BiGRU网络的足球视频事件分类方法。该方法首先利用PySceneDetect工具的场景切换检测功能对完整足球视频进行镜头分割，在此基础上构建包含五类足球事件的数据集；随后通过实验对比，选择将目前主流的卷积神经网络VGG16与BiGRU结合构建分类模型。实验结果表明，CNN与RNN的结合，解决了视频中时间维度利用不足的问题，更有效的整合足球视频中时间维度和空间维度的动态信息，实现比传统技术更高的精度和更快的速度。目前该模型对足球视频数据集上的某单一事件识别率最高达到97.4%。

关键词

足球视频，视频切分，视频分类，深度学习

A Classification Method of Soccer Video Based on CNN-BiGRU

Hang Su, Shengnan Zhang

School of Software, Shenyang University of Technology, Shenyang Liaoning

Received: Feb. 16th, 2023; accepted: Mar. 15th, 2023; published: Mar. 23rd, 2023

Abstract

Video classification based on deep learning is an important direction of sports video research. Aiming at the problem of low recognition rate of video event types, this paper proposes a football video event classification method based on CNN-BiGRU network. It first uses the scene switching detection function of PySceneDetect tool to segment the complete football video, and builds a data set containing five types of football events on this basis, then, through experimental comparisons,

文章引用: 苏航, 张胜男. 基于 CNN-BiGRU 的足球视频片段分类方法[J]. 计算机科学与应用, 2023, 13(3): 369-377.
DOI: 10.12677/csa.2023.133035

combine the current mainstream convolutional neural network VGG16 with BiGRU to construct a classification model. The experimental results show that the combination of CNN and RNN solves the problem of insufficient utilization of the time dimension of videos, more effectively integrates the dynamic information of two dimensions of time and space in football videos, and achieves higher accuracy and faster speed than traditional technologies. At present, the model has a maxi-mum recognition rate of 97.4% for a single event on the football video dataset.

Keywords

Football Video, Video Segmentation, Video Classification, Deep Learning

1. 引言

足球是世界第一大运动，有着广泛的收视群体，但一场足球比赛的时间较长，要从海量的视频数据中快速找到用户关注的内容，仅仅依靠传统的人工剪辑分类是十分困难的。早期的足球视频分类领域的研究主要是通过人工制定规则结合机器学习的方式来进行事件检测，此方法受制于人为设定的经验参数且不具备可扩展性。目前随着计算机视觉的发展，利用深度学习的方法处理足球视频问题已取得了重大进展。

在足球语义规则的处理上，传统的机器学习方法广泛应用于视频分类检测中，常见方法有支持向量机、贝叶斯网络、隐马尔可夫模型等，这些方法基于多种人工设定的特征进行场景分类，如图像特征、用颜色、纹理和形状等底层特征。此外，研究人员还常借助视频相关的文本、音频、回放镜头等多种信息形成多模态特征以实现事件的检测[1] [2] [3]。Naveed 等将混合特征用于模型训练，使用 HOG、SIFT、LBP 等作为训练系统的特征集[4]。Pandya 等提出一种基于精确边界预测的时序动作检测方法，以光流的变化来获取足球视频中的事件[5]。

相较于传统的场景分类方法，深度学习能够通过一些简单模型将接收到的原始数据转化为更易于人类理解的语义特征，进而能够实现更有效的视频分类。针对深度学习在动作识别领域的研究，Ji 等提出了一种在视频的时间和空间上卷积的三维卷积方式，将多个卷积层和下采样层串联构成动作识别网络[6]。Song 等结合视频帧序列和光流序列，利用 I3D 网络对视频单元进行分类，输出每个视频单元的预测概率值，在此基础上再利用分组方法将相邻的片段组合在一起以定位事件的边界[7]。Cheng 等采用 3DCNN 和 CNN 分别提取足球视频特征和音频特征，并进行多模态融合[8]。文献[9] [10] [11] [12]针对动作检测问题分别提出新的 CNN 结构，使用 CNN 对动作类型进行识别，利用滑动窗口实现事件边界确定。

可见在足球视频处理中，CNN 的卷积层能够很好地感知图像的局部特征，感知数据点与周围数据点之间的关系[13]。但数据在 CNN 中只能单向流动且仅考虑每个时间步的当前输入，可能导致之前退化信息的丢失，而将 CNN 与 RNN 结合可以有效提取被 CNN 忽略的时序特征，提高特征提取的准确度。本文即是基于 CNN-BiGRU 网络训练了一个事件分类模型，实现足球赛事的视频片段分类。

2. 算法设计

视频事件类型识别以切分好的视频片段为研究对象，通过搭建不同神经网络(GoogleNet、ResNet50、

VGG16)对准确率和召回率对比，最终选择 VGG16 对数据集中的视频片段的帧进行处理，提取出单帧的空间特征。使用双向循环神经网络 BiGRU 对特征序列提取动态信息，预测视频片段的事件类型，分类流程如图 1 所示。

Figure 1. Flow chart of classification
图 1. 分类流程图

图 1 中利用 PySceneDetect 检测切换镜头，然后将完整足球视频切分成片段，再根据预定义的事件类型将视频片段分类标记构建数据集，最后将数据集传入 CNN-BiGRU 网络训练事件分类模型。

2.1. 视频镜头切换检测

检测模型输入为足球视频片段，因此首先需要根据预先设定的事件类型将视频分割构建数据集，通过边界检测算法将视频中每个镜头的边界帧检测出来，然后再通过这些边界帧将完整的视频分割成一系列独立的镜头。

本文的边界检测利用改进的帧间插值法完成。算法将每个解码帧的颜色空间从 RGB 转换为 HSV，根据前后两帧的视频数据，计算出它们不同的区域大小。如果区域尺寸大于某个预先设定的值，就认为场景已经切换。这里 PySceneDetect 工具利用了 OpenCV 中提供的阈值函数。帧间差分法流程如图 2 所示。

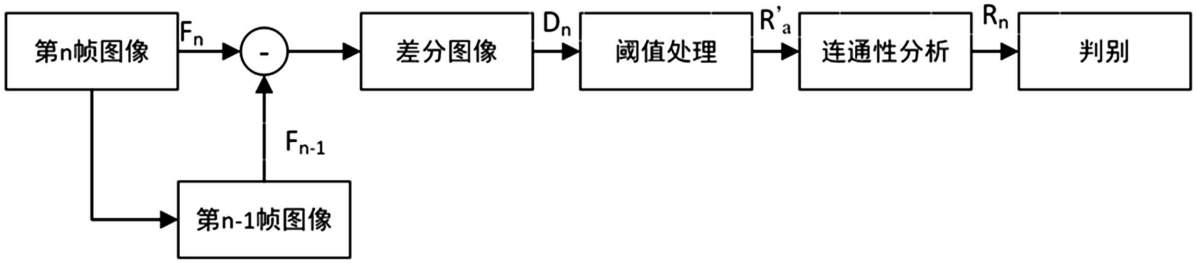

Figure 2. Iinter-frame difference method
图 2. 帧间插值法

视频序列中第 n 帧和第 $n-1$ 帧图像记为 f_n 和 f_{n-1}，两帧对应像素点的灰度值记为 $f_n(x,y)$ 和 $f_{n-1}(x,y)$，按照式(1)将两帧图像对应像素点的灰度值进行相减，并取其绝对值，得到差分图像 D_n：

$$D_n(x,y)=\left|f_n(x,y)-f_{n-1}(x,y)\right| \tag{1}$$

设定阈值 T，按照式(2)逐个对像素点进行二值化处理，得到二值化图 R'_n，其中灰度值为 255 的点为镜头转换点，灰度值为 0 的点为非镜头转换点，对图像 R'_n 进行连通性分析，最终可得到含有完整运动目标的图像 R_n。

$$R'_n(x,y)=\begin{cases}255, D_n(x,y)>T\\0, \text{others}\end{cases} \tag{2}$$

根据定义的事件类型通过人工的方法构建了一个包含射门、进球、点球、角球、黄牌、红牌的数据集，其中射门视频 156 个，进球视频 64 个，点球视频 114 个，角球视频 129 个，红/黄牌视频各 107 个。充足的数据集为事件分类模型做好准备。

2.2. CNN-BiGRU 模型设计

事件类型识别以切分好的视频片段为研究对象，通过对视频单帧图像上空间特征的提取，以及对帧序列时间维度上动态信息的整合，完成视频片段的整体事件类型的识别。本文模型中单帧上空间特征的提取使用了 CNN，特征序列动态信息的提取采用了基于 LSTM 的改进型循环神经网络 BiGRU。

相比传统 LSTM，GRU 模型的门函数由 3 种减少为 2 种：更新门和重置门，虽然参数比 LSTM 更少，但却能够达到与 LSTM 相当的功能，且不存在 LSTM 中的细胞状态，可以显著提高训练效率。此外，由双向 GRU (BiGRU)代替普通 GRU，可以利用 BiGRU 的特有结构加强对时序序列的敏感程度，提高足球事件的预测准确率。CNN-BiGRU 网络模型图如图 3 所示。

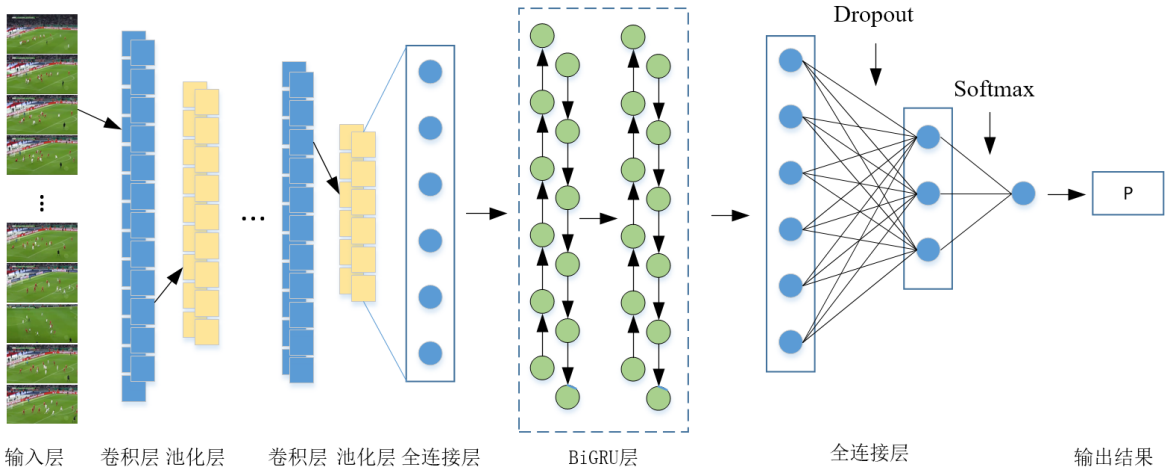

Figure 3. Network model graph of CNN-BiGRU
图 3. CNN-BiGRU 网络模型图

1) 输入层

在使用 CNN-BiGRU 网络训练时间分类模型的问题中，首先需要使用 CNN 提取单张图片特征信息，由于数据集中存储的均为已标记好的足球事件视频，因此需要对数据集进行预处理，对每个视频提取出视频帧序列，并存放在指定命名规则的文件夹中供后续训练模型使用。

2) CNN 层

在特征提取网络的选择上，本文主要考虑了提取速度以及准确性两大因素，综合考虑了 VGG、GoogleNet 以及 ResNet 的优缺点。其中，VGG 网络将多个较小的卷积核串联(卷积核中不插入池化层)，以此关联与大卷集合相同面积的连接区域，如图 4 所示。由于大卷积层后只能使用一个非线性层，VGG 将大卷积核拆分成多个小核卷积层，各个卷积层后能插入非线性层。通过增加整个网络的非线性处理单元，使得网络对特征的学习能力更强。此外，多个 3 × 3 的串联卷积层比一个大尺寸的卷积层拥有更少的网络参数。VGG 共有 A-E 五种网络结构，每个网络都是由 5 个卷积组、2 个 4096 维的全连接层和一个分类层组成，每段卷积组有 2~3 个卷积层。五个网络随着层数的增加识别准确率也逐渐增加，当深度达到 16 时效果有较大提升，因此本文中选 VGG-16 作为特征提取网络。

3) GRU 层

GRU 是循环神经网络 RNN 的一种。和 LSTM 一样，也是为了解决长期记忆和反向传播中的梯度等问题而提出来的，其网络结构如图 5 所示。图中，GRU 层的输出不仅和前一层的输入数据有关，而且还和上一时间步的输出有关，通过在同一层 GRU 单元之间建立有向连接赋予 GRU 对过去数据记忆能力。

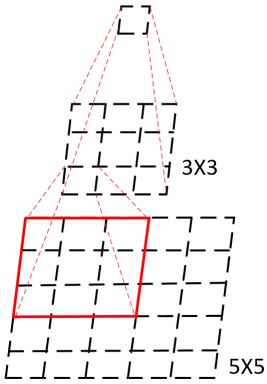

Figure 4. Performance improvement principle of VGG
图 4. VGG 网络的性能提升原理

与 GRU 模型不同的是，BiGRU 模型由前向 GRU 层和后向 GRU 层组成，因此可以在前向和后向两个方向上处理序列，两个方向均具有独立的隐藏层。BiGRU 的网络结构如图 6 所示。

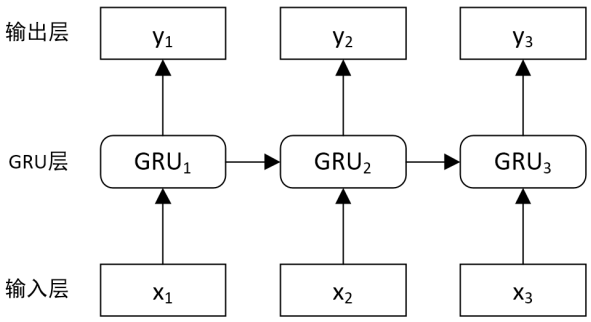

Figure 5. Network structure of GRU
图 5. GRU 网络结构

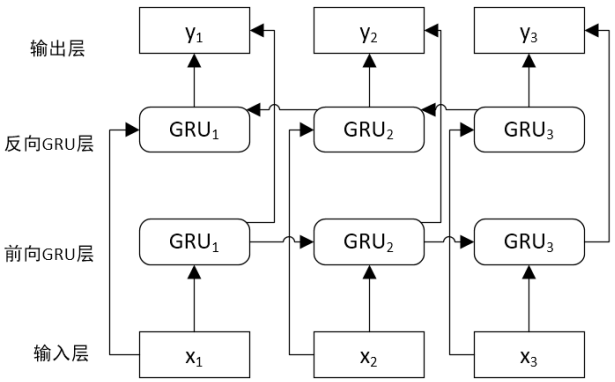

Figure 6. Network structure of BiGRU
图 6. BiGRU 的网络结构

这种结构可以使得每个 GRU 隐藏层在特定的时间步长内都可以同时捕获前向和后向的信息，因此可以提取出更加准确的足球特征信息，提升模型训练效果。

4) Dropout 层

为了防止模型过拟合，本文引入了 Dropout 层。Dropout 是神经网络最有效也最常用的正则化方法之

一，它将深度神经网络模型作为一个集成的模型进行训练，然后将所有值取平均，而不只是训练单个模型。当一个神经元被丢弃时，无论输入以及相关参数是多少，它的输出值都会被设置 0。

5) Flatten 层

将最后的输出张量输入到一个密集连接分类器网络中，即 Dense 层的堆叠。由于分类器只可以处理 1D 向量，而当前的输出是 3D 张量，因此需要将 3D 输出展平为 1D，然后在上面添加几个 Dense 层。Flatten 层是用来对数组进行展平操作的，其作用是将多维的输入一维化，常用在卷积层到全连接层的过度。同时 Flatten 层不会影响 batch_size 的大小。

6) 损失函数

损失函数是反应模型在数据集上训练好坏的重要指标，其值与输出结果负相关。本文在架构中使用了交叉熵损失函数。

在训练 BiGRU 的过程中，使用 softmax loss 来对网络进行优化。对于一个输入事件样本片段 $x, y \in R^c$ 是该样本对应的事件类型标签，若该样本属于事件 c，则 $y_c = 1$，否则 $y_c = 0$。网络在时刻 t 的损失计算如公式(3)：

$$L(t,p,x,y) = -\sum_{c} y_c \log\left(p_{t,c}\right) \tag{3}$$

其中 $p_{t,c}$ 即为 BiGRU 在当前时刻 t 预测输入片段属于事件类型 c 的 softmax 概率。模型在每一个时刻 t 都进行反向传播，所以对于所有的训练样本 λ，总的损失为：

$$L_{train} = \sum_{(x,y)\in\lambda} \sum_{t=1}^{T} L(t,p,x,y) \tag{4}$$

3. 实验分析

3.1. 实验环境

实验环境为 Ubuntu17.0，处理器为 Intel Core i9 12900k，运行内存 16GB，GPU 型号为 Tesla V100，显存为 16 GB。本文使用 Keras 框架来搭建模型，BiGRU 结构的细节主要使用 Keras 工具箱实现，编程语言为 Python。

3.2. 数据集

近些年来随着 GPU 算力的提高，有关动作以及体育视频类型的数据集越来越多，如规模较大的 UCF101 [14]包含 13,320 个视频，视频内容主要为单人事件。Sports-1M 包含大约 120 万个视频，487 种运动类型。此外文献[15]中 NCAA 提出了篮球数据集，包含了 257 个篮球视频，共分为 11 个篮球事件。目前在足球领域还没公开数据集，为此，本文从 2018~2020 赛季西甲、德甲、英超联赛收集数据并构建数据集，其中事件类型定义为 5 类，分别为射门、点球、角球、黄牌、红牌，其中射门视频 156 个，进球视频 64 个，点球视频 114 个，角球视频 129 个，红/黄牌视频各 107 个。训练集与验证集比例为 7:3，数据来源如表 1 所示。

3.3. 训练过程

本文主要考虑了提取速度以及准确性两大因素，分别将 VGG、BN-Inception、ResNet 与 BiGRU 组建神经网络训练模型，并对比结果。训练时 Batch 样本数量设置为 64，初始学习率为 0.001，使用 SGD 更新网络参数，Epoch 迭代周期为 50，学习率每迭代完一个周期更新为当前的 0.1。

Table 1. Datasources
表 1. 数据来源

联赛名称	镜头总数	射门	进球	点球	角球	红/黄牌
西班牙足球甲级联赛	150	49	21	-	43	37
德国足球联邦联赛	145	53	23	-	39	30
英格兰足球超级联赛	171	54	30	-	47	40
UCF101	114	-	-	114	-	-
总计	570	156 (120:36)	64 (48:16)	114 (87:27)	129 (99:30)	107 (81:26)

由于本文构建的数据集有限，在采用 BiGRU 整合视频帧对应的特征序列并提取特征序列中包含的事件时，导致无法训练出能够泛化到新数据的模型，导致模型过拟合。为解决这一问题，首先在网络中进行数据增强，即是从现有的训练样本中生成更多的训练数据，利用多种能够生成可信图像的随机变换来增加样本，尽可能减少过拟合。为进一步降低过拟合，还需要向模型中添加一个 Dropout 层。

由于网络层数的增加会随着 BiGRU 层数的增加而增加，进而导致网络过拟合。因此 dropout 参数的设定以及 BiGRU 的层数对训练结果密切相关。BiGRU 分别取 1~3 层，Dropout 参数设置为 0.1、0.2、0.3、0.4、0.5，共 15 种情况，实验结果如图 7 所示。

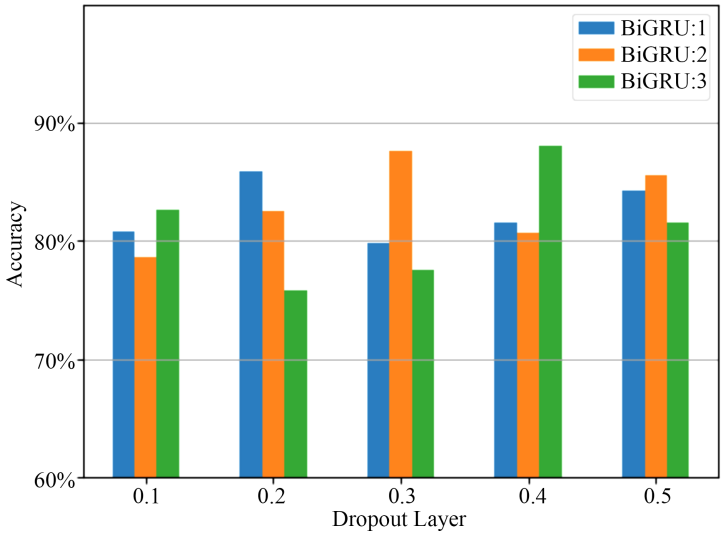

Figure 7. BiGRU layers and Dropout parameters
图 7. BiGRU 层数和 Dropout 参数

由图 7 可以看出，当 Dropout 参数和 BiGRU 层数分别设置为 0.3，2 和 0.4，3 的情况下模型的 Accuracy 最高，后者在数值上略高于前者；但随着 BiGRU 层数增多会增加训练成本，延长训练时间，因此 BiGRU 的层数设置为 2，Dropout 参数设置为 0.3。模型的其他超参数设置如表 2 所示。

Table 2. Super parameters setting of CNN-BiGRU
表 2. CNN-BiGRU 模型超参数设置

超参数名称	优化器	卷积层层数	BiGRU 层数	Dropout	Batchsize	Learning rate	Epoch	训练集与验证集比例
参数值	RMSProp	6	2	0.3	128	0.005	50	7:3

3.4. 实验结论

在类型识别任务上，目前普遍使用的评测指标有查准率(Precision)、查全率率(Recall)，如式(5)和式(6)。准确率和召回率的计算，一般需要先设定一个阈值，通过阈值对预测结果切分后计算准确率和召回率，不同的阈值对应的准确率和召回率不同。

$$\text{percision} = \frac{\text{correct}}{\text{correct} + \text{false}} \tag{5}$$

$$\text{recall} = \frac{\text{correct}}{\text{correct} + \text{miss}} \tag{6}$$

本文主要考虑了提取速度以及准确性两大因素，分别将 VGG、GooleNet、ResNet 与 BiGRU 组建神经网络训练模型，并对比结果。在不同特征提取网络下的基于识别结果如表 3 所示。

可见在事件识别模型中，VGG-16 的效果明显好于另外两个模型，因此选用 VGG-16 作为特征提取网络。

本文算法将足球视频事件定义为 5 类，从表 3 可以看出模型在对点球、角球、红/黄牌事件的准确率较高，这是因为角球及点球事件画面中有较为明显的球场边界特征区域，红/黄牌事件也有明显颜色卡片出现。表 3 也显示本文模型在比较复杂的进球视频的识别率上偏低，其主要原因是足球在视频帧图像中很小，且容易被球员遮挡，或者受制于拍摄角度当球门背景是场外的观众席时，白色的足球很容易跟背景混淆，难以判断球是否进门导致识别率偏低；其次，运动战进球和射门在感知上的差异仅仅在于球是否进入球门，不易做出判断；此外，VGG16 在提取图片特征时产生的误差，后期可以通过完善训练样本和模型参数设计进一步提高识别率。

Table 3. Recognition results of event type
表 3. 事件类型识别结果

事件	VGG-16		ResNet-50		GooleNet	
	查准率	查全率	查准率	查全率	查准率	查全率
进球	0.601	0.457	0.145	0.126	0.182	0.391
射门	0.745	0.793	0.677	0.701	0.565	0.675
点球	0.873	0.915	0.783	0.751	0.893	0.716
角球	0.835	0.803	0.780	0.800	0.887	0.813
红/黄牌	0.936	0.901	0.878	0.825	0.853	0.864

4. 结语

对视频进行快速分类以满足不同观众的兴趣和爱好，是目前多媒体领域的研究重点。本文提出的足球视频场景分类算法，分别对视频的时间特征和空间特征进行了有效的提取，实现了比传统技术更好的精度和更快的速度，其中对点球以及红黄牌的识别率都达到了 95%以上。由于 GRU 对不同长度的事件片段具有识别能力，下一步的工作计划是在事件类型识别模型的基础上，向 BiGRU 中加入边界确定模块，目的是当一段长视频传入模型时，模型具有独立判断事件边界并分类的功能。

参考文献

[1] Doman, K., Tomita, T., Ide, I., Deguchi, D. and Murase, H. (2014) Event Detection Based on Twitter Enthusiasm Degree for Generating a Sports Highlight Video. *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, 3-7 November 2014, 949-952. https://doi.org/10.1145/2647868.2654973

[2] Kolekar, M.H. and Sengupta, S. (2015) Bayesian Network-Based Customized Highlight Generation for Broadcast Soccer Videos. *IEEE Transactions on Broadcasting*, **61**, 195-209. https://doi.org/10.1109/TBC.2015.2424011

[3] Arbat, S., Sinha, S.K. and Shikha, B.K. (2014) Event Detection in Broadcast Soccer Video by Detecting Replays. *International Journal of Scientific &Technology Research*, **3**, 282-285.

[4] Naveed, H., Khan, G., Khan, A.U., Siddiqi, S. and Khan, M.U.G. (2019) Human Activity Recognition Using Mixture of Heterogeneous Features and Sequential Minimal Optimization. *International Journal of Machine Learning and Cybernetics*, **10**, 2329-2340. https://doi.org/10.1007/s13042-018-0870-1

[5] Pandya, D.S. and Zaveri, M.A. (2017) Frame Based Approach for Automatic Event Boundary Detection of Soccer Video Using Optical Flow. 2017 *IEEE International Conference on Signal and Image Processing Applications* (*ICSIPA*), Kuching, 12-14 September 2017, 402-406. https://doi.org/10.1109/ICSIPA.2017.8120644

[6] Ji, S., Xu, W., Yang, M. and Yu, K. (2012) 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 221-231. https://doi.org/10.1109/TPAMI.2012.59

[7] Song, C. and Rasmussen, C. (2019) Multi-Camera Temporal Grouping for Play/Break Event Detection in Soccer Games. In: Bebis, G., *et al*., Eds., *Advances in Visual Computing*. *ISVC* 2019. *Lecture Notes in Computer Science*, Vol. 11844, Springer, Cham, 231-243. https://doi.org/10.1007/978-3-030-33720-9_18

[8] 程萍. 基于多模态融合的足球视频精彩事件检测[D]: [硕士学位论文]. 杭州: 浙江理工大学, 2020. https://doi.org/10.27786/d.cnki.gzjlg.2020.000217

[9] Lea, C., Flynn, M.D., Vidal, R., Reiter, A. and Hager, G.D. (2017) Temporal Convolutional Networks for Action Segmentation and Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 156-165. https://doi.org/10.1109/CVPR.2017.113

[10] Shou, Z., Chan, J., Zareian, A., Miyazawa, K. and Chang, S -F. (2017) CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 5734-5743. https://doi.org/10.1109/CVPR.2017.155

[11] Khan, M.Z., Saleem, S., Hassan, M.A. and Usman Ghanni Khan, M. (2018) Learning Deep C3D Features for Soccer Video Event Detection. 2018 14*th International Conference on Emerging Technologies* (*ICET*), Islamabad, 21-22 November 2018, 1-6. https://doi.org/10.1109/ICET.2018.8603644

[12] Rongved, O.A.N., Hicks, S.A., Thambawita, V., *et al*. (2020) Real-Time Detection of Events in Soccer Videos using 3D Convolutional Neural Networks. 2020 *IEEE International Symposium on Multimedia* (*ISM)*, Naples, 2-4 December 2020, 135-144. https://doi.org/10.1109/ISM.2020.00030

[13] Vanderplaetse, B. and Dupont, S. (2020) Improved Soccer Action Spotting using both Audio and Video Streams. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 14-19 June 2020, 896-897. https://doi.org/10.1109/CVPRW50498.2020.00456

[14] Ren, L., Sun, Y., Wang, H. and Zhang, L. (2018) Prediction of Bearing Remaining Useful Life with Deep Convolution Neural Network. *IEEE Access*, **6**, 13041-13049. https://doi.org/10.1109/ACCESS.2018.2804930

[15] Ramanathan, V., Huang, J., Abu-El-Haija, S., *et al*. (2016) Detecting Events and Key Actors in Multi-Person Videos. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 3043-3053. https://doi.org/10.1109/CVPR.2016.332