计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(12):158−164 [doi: 10.15888/j.cnki.csa.007199]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 C3D 的足球视频场景分类算法①

程　萍, 冯　杰, 马汉杰, 许永恩, 王　健

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 马汉杰, E-mail: mahanjie@zstu.edu.cn

摘　要: 足球视频整场比赛持续时间较长, 许多视频内容并非广大观众的兴趣所在, 因此足球视频场景分类成为了近几十年来研究界的一项重要课题, 许多机器学习方法也被应用于这个课题上. 本文提出的基于 C3D (三维卷积神经网络) 的足球视频场景分类算法, 将三维卷积运用于足球视频领域, 并通过实验验证了本文算法的可行性. 本文实验的流程如下: 首先, 基于帧间差分法和徽标检测法检测法对足球视频场景切换进行检测, 实现镜头分割. 在此基础上, 提取分割镜头的语义特征并将其进行标记, 然后通过 C3D 对足球事件进行分类. 本文将足球视频分为 7 类, 分别为远镜头、中镜头、特写镜头、回放镜头、观众镜头、开场镜头及 VAR (视频助理裁判) 镜头. 实验结果表明, 该模型在足球视频数据集上的分类准确率为 96%.

关键词: 三维卷积; 足球; 镜头检测; 语义标注; 场景分类

引用格式: 程萍,冯杰,马汉杰,许永恩,王健.基于 C3D 的足球视频场景分类算法.计算机系统应用,2019,28(12):158–164. http://www.c-s-a.org.cn/1003-3254/7199.html

Soccer Video Scene Classification Algorithms Based on C3D

CHENG Ping, FENG Jie, MA Han-Jie, XU Yong-En, WANG Jian

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Football video lasts for a long time, and many video content is not the interest of audience. Therefore, football video scene classification has become an important research topic in recent decades, and many machine learning methods have also been applied to this topic. In this study, a soccer video scene classification algorithm based on 3D (three-dimensional) convolution neural network is proposed. The 3D convolution is applied to the field of soccer video, and the feasibility of this algorithm is verified by experiments. The flow of this experiment is as follows. Firstly, football video scene switching is detected based on frame difference method and logo detection method, and shot segmentation is realized. On this basis, the semantic features of shot segmentation are extracted and tagged, and then football events are classified by C3D. In this study, football videos are divided into seven categories: long shot, medium shot, close-up shot, playback shot, audience shot, opening shot, and VAR (Video Assistant Referee) shot. The experimental results show that the classification accuracy of the model is 96% on football video datasets.

Key words: three-dimensional convolution; football; shot detection; semantic annotation; scene classification

引言

随着网络技术的普及、多媒体信息的爆炸性增长、社会生活节奏不断加快, 人们往往没有足够的时间观看整段视频节目, 而是希望根据自己的需求观看特定的部分. 足球作为全球最受欢迎的一项体育运动, 有着广泛的收看群体. 足球比赛的持续时间较长, 人们

① 基金项目: 国家自然科学基金 (61501402)
Foundation item: National Natural Science Foundation of China (61501402)
收稿时间: 2019-04-26; 修改时间: 2019-05-21, 2019-06-03; 采用时间: 2019-06-14; csa 在线出版时间: 2019-12-10

感兴趣的内容却是不同的, 有人喜欢看精彩镜头 (射门、点球等), 有人喜欢看中场配合. 然而, 面对海量的视频数据, 依靠传统人工剪辑分类方式, 不仅极大地浪费人力资源而且也不能保证工作的及时性与可靠性.

目前, 足球视频的场景分类面临的主要问题包括: 场景切换检测的查全率和查准率不够高, 无法满足实际需求, 其次, 在人工制定足球语义方面, 需要耗费大量的人力资源. 为了解决上述问题, 国内外学者进行了系统、深入的研究, 并取得了一定的成果.

在解决场景切换检测准确率问题上, 陆思烨等[1]提出基于双阈值灰度直方图的场景检测算法, 通过比较相邻帧的灰度直方图差与高低阈值的大小, 针对场景可能的渐变情形, 比较非相邻帧的直方图差与高低阈值的大小来判断是否发生了场景切换. 方宏俊等[2]结合数字电视图像处理芯片中硬件算法设计的低复杂度要求, 介绍了一种基于动态阶数控制直方图分布的场景检测优化设计算法. 孙桃等[3]对动画帧图像分块并提取其 HSV 颜色特征, 然后将连续帧的相似度存入一个固定长度的缓存队列中, 最后基于动态 Bayesian 决策判定是否有场景切换. 段淑玉等[4]提出一种应用于帧率提升系统的, 根据内插帧各匹配块的均值 SAD 为检测依据的场景切换检测算法, 解决场景切换时 ME/MC 算法因匹配失误产生严重块效应的问题.

语义方面, 早期的场景分类研究中, 文献[5,6]是基于图像特征的, 即通过描述颜色、纹理和形状等底层特征来实现分类. 之后, 用融合多种特征的方法来描述不同内容的图像场景, Naveed 等[7]利用混合特征进行训练以预测人类活动. 他们使用 HOG、SIFT、LBP 等作为训练系统的特征集. Kang 等人[8]在对视频底层特征分析的基础上, 提取音视频关键字作为中级特征, 基于隐马尔科夫模型来检测精彩视频片段. Ekin 等[9]通过提取视频的中低级特征, 提出了一种有效的足球视频摘要生成框架, 能够生成慢速运动、进球和基于对象特征分类的慢速运动 3 类摘要. 文献[10,11]结合网络直播文本对体育视频事件进行检测, 实验表明检测到的事件类型更加丰富, 准确率也得到较大提升. 但网络直播文本的获取和文本事件与视频事件的对齐是此方法的关键和难点. 于俊清等[12]利用足球比赛中观众情绪波动情况, 建立情感激励曲线并对曲线尖峰进行检测, 但基于尖峰检测误差较大, 检测性能无法满足实际需要.

相较于传统的场景分类方法, 文献[13]利用卷积神经网络 (Convolutional Neural Network, CNN) 进行场景分类, 通过自学习的方式来“识别”图像, 利用反馈网络实现实现分类. Jiang 等[14]首先将视频分解为关键帧, 然后将这些关键帧的 CNN 特征传递给 RNN 进行分类. Tjondronegoro 等[15]对不同类型的事件进行统计分析, 选择 6 个具有区分性的特征, 根据统计结果建立一系列规则把足球视频事件分为进球、射门和犯规 3 类. 但过程比较繁杂, 人力耗费较大. Ji 等[16]提出基于三维卷积神经网络的体系结构, 该算法捕捉多个相邻帧中编码的运动信息, 从输入帧中生成多个信息通道. 最终的特征表示为结合了所有通道的信息. 但该算法应用场景是机场监控视频中人的行为识别. 目前, 对于足球视频中的活动识别, 模型仅限于处理二维数据的输入实现分类. 本文提出了一种基于三维 CNN (C3D) 的足球视频分类模型, 模型通过执行三维卷积, 从空间和时间维度中提取特征, 从而捕获编码在多个相邻帧中的运动信息. 具体算法流程如图 1 所示.

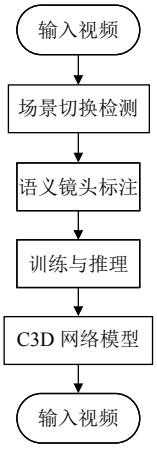

图 1 算法流程图

本文的算法以视频作为输入, 在场景切换检测的基础上将视频分割成小片段, 根据预定义的类别将片段进行标记, 然后, 将小片段送入 C3D (三维卷积) 网络中进行训练.

1 算法设计

1.1 场景切换检测

本文提出的算法建立在三维卷积的神经网络模型上, 模型以视频作为输入. 首先, 对足球视频的场景切换进行检测, 提取不同场景的视频片段以实现镜头分

割, 通过边界检测算法将视频中每个镜头的边界帧检测出来, 然后通过这些边界帧将完整的视频分割成一系列独立的镜头.

根据足球视频场景变换的特点, 整场比赛中突变镜头的情况较多, 在综合考虑了效率和准确率之后, 本文选取了基于像素比较的镜头分割方法: 帧间差分法, 如式 (1) 所示:

$$D(k,k+1)=\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}|I_k(x,y)-I_{k-1}(x,y)| \quad (1)$$

其中, $I_k(x,y)$ 和 $I_{k+1}(x,y)$ 分别表示第 k 帧和第 k+1 帧在 (x,y) 处的亮度值, M 和 N 分别表示该帧图像的高度和宽度. $D(k,k+1)$ 的值表示两帧之间的变化. 当 $D(k,k+1)$ 大于某一设定的阈值时则认为这两帧分别属于两个不同的镜头.

对于回放镜头, 由于镜头切换频率较低, 文献[12]通过实验验证了基于 Logo 的回放镜头检测方法的可行性, 因此, 本文采取上述基于徽标检测的方法对回放场景进行检测.

1.2 语义镜头标注

文献[17]通过提取特征值并结合决策树来对镜头进行分类和语义标注. 本文结合决策树算法及人工规则对分割好的镜头进行语义标注. 决策树流程如图 2 所示.

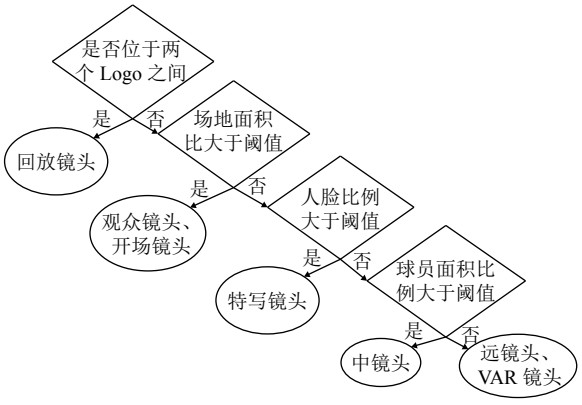

图 2 决策树流程图

在决策树的第一层首先区分的是回放镜头和非回放镜头, 决策树的第二层则是通过颜色直方图来判断场地面积, 以此将非回放镜头分为场内镜头和场外镜头 (观众镜头、开场镜头); 第三层通过计算人脸比例提取场内镜头中的特写镜头; 第四层通过计算场地和球员面积比例将剩下的场内镜头分为远镜头、VAR 镜头和中镜头.

本文将镜头场景分为 7 类, 分别为: 远镜头、中镜头、特写镜头、回放镜头、观众镜头、开场镜头及 VAR 镜头. 各镜头语义代表帧如图 3 所示.

(a) 远镜头

(b) 中镜头

(c) 特写镜头

(d) 回放镜头

(e) 观众镜头

(f) 开场镜头

(g) VAR 镜头

图 3 镜头代表帧

1.3 C3D 模型设计

本文优化了经典的 C3D 网络的结构. 经典的 C3D 网络结构是由 8 个 3D 卷积层 (Convolution)、5 个 3D 最大池化层 (MaxPooling)、2 个全连接层 (Fully-Connect) 构成. 优化后的网络结构减少卷积层的个数, 新的网络结构为 5 个卷积层, 5 个最大池化层, 3 个全连接层, 2 个 Dropout 层及 1 个 Softmax 层组成, 如图 4 所示.

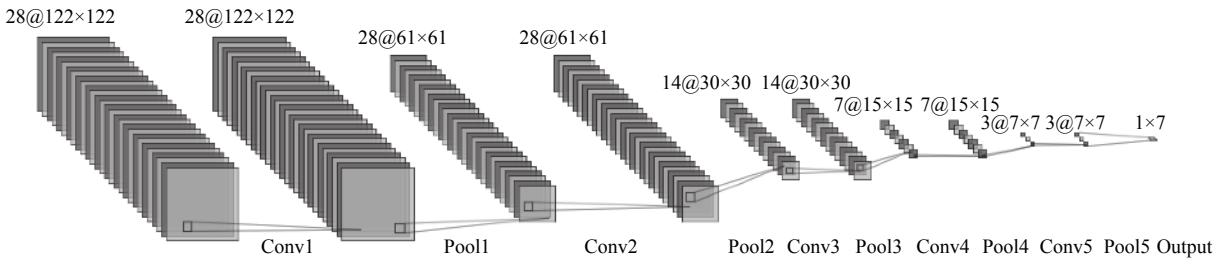

图 4 C3D 网络模型图

1) 输入层

输入层的数据为一个个视频片段, 但由于视频的长度长短不一, 我们选用最近邻插值的方式, 对视频片段少于采样长度的数据进行填补, 以达到每次采样的长度为 28 帧. 同时, 我们还将视频序列帧的大小统一尺寸为 122×122.

2) 卷积层与池化层

本文采用三维卷积, 在视频的空间和时间维上对相邻帧进行卷积操作以提取特征. 这些特征保存了空间信息和时间信息, 便于对视频中物体的运动进行检测. 针对单通道, 与 2D 卷积不同之处在于, 输入图像多了一个 depth 维度, 故输入大小为 (1, depth, height, width), 卷积核也多了一个 k_d(depth) 维度, 因此卷积核在输入 3D 图像的空间维度 (height 和 width 维) 和 depth 维度上均进行滑窗操作, 每次滑窗与 (k_d, k_h, k_w) 窗口内的 values 进行相关操作, 得到输出 3D 图像中的一个 value. 如图 5 所示.

图 5 2D 卷积

针对多通道, 输入大小为 (3, depth, height, width), 则与 2D 卷积的操作一样, 每次滑窗与 3 个 channels 上的 (k_d, k_h, k_w) 窗口内的所有 values 进行相关操作, 得到输出 3D 图像中的一个 value. 如图 6 所示.

我们的视频片段每帧的大小为 $c\times l\times w\times h$, 其中 c 为图像的通道数, l 为视频序列的长度, 即我们的采样帧数, w 和 h 为每帧的宽和高. 卷积层的核大小为 $d\times k\times k$, d 为卷积核的时间深度, k 为核的空间大小, 本文卷积层的核大小为 3×3×3, 所有池化层都是最大池化, 内核大小为 2×2×2(第一层除外), 步长为 1. 第一层池化层的内核大小为 1×2×2, 其目的是不过早地合并时间信号, 同时满足 28 帧的剪辑长度.

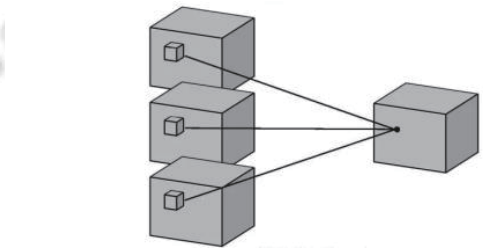

图 6 3D 卷积

3) Dropout 层

为了防止模型过拟合, 我们引入了 Dropout 层, 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. 网络模型将 Dropout 率为 p, 即一个神经元被保留的概率为 $1–p$. 当一个神经元被丢弃时, 无论输入的是什么、相关的参数是多少, 它的输出值都会被设置 0. Dropout_1 和 Dropout_2 层的 p 值初始设置为 0.5.

4) Flatten 层

Fatten 层的作用是将数据“拍平”, 即将多维的数据一维化, 作为从卷积层到全连接层的过渡. Flatten 层的处理不会影响批处理 batch_size 的大小, 因此, 本文的数据在经过 Flatten 层处理后, 数据大小为 256×9=2304.

5) 损失函数

损失函数是衡量我们的网络结构在我们的数据集上训练的好坏的一项指标. 当训练集的预测大部分为错误时, 则对应输出较大的损失函数值. 当模型的输出结果较好, 则损失函数的输出也将较低. 如果我们想要改变算法的某些部分来提高我们模型的性能, 损失函数的输出将作为我们的参考标准. 本文在架构中使用了交叉熵损失函数, 如式 (2) 所示:

$$H(p,q) = -\sum p(x)\log q(x) \quad (2)$$

通过概率分布 q 来表达概率分布 p 的困难程度, p 代表正确答案, q 代表的是预测值, 交叉熵越小, 两个概率的分布越接近.

在此基础上, 利用 Softmax 函数求出每个类的概率. Softmax 函数如式 (3) 所示:

$$S_x = \frac{e^{c_x}}{\sum_y e^{c_y}}, \forall x \text{ in } \{1,2,\cdots,M\} \quad (3)$$

其中, S 是每个可能结果 M 的分类概率得分. 假设我们有一个具有 M 种可能结果的分类问题, 当我们输入一幅图像进行分类时, 我们根据每个结果得到分类分数 $S_1, S_2, \cdots, S_M$. 在得到预测后, 将某一特定框架的分类分数除以所有指数分数之和后, 得到基于最小损失的实际类, 该类的概率最大.

2 实验分析

2.1 实验环境

实验环境为 Ubuntu16.0, 运行内存为 16 GB, GPU 型号为 NIVIDIA 1080 Ti, 内存为 12 GB. 本文用 MXNet 框架搭建模型.

2.2 数据集

针对足球比赛视频, 目前还未形成一个公开的数据集, 为了训练本文的分类算法, 必须自行收集数据, 并根据需要对镜头的场景进行标记. 因此, 本文也为足球数据集的生成做出了贡献. 在本文的分类中包含 7 种类型, 包括开场镜头、观众镜头、远镜头、中场镜头、回放镜头、特写镜头、以及 VAR 镜头. 每个类别的场景镜头约 600 个, 每个镜头的平均持续时间为 7 秒, 视频的帧率为 25 秒/帧. 本文的数据集包括 5 场世界杯比赛, 每场比赛为 90 分钟. 我们将数据的训练集与测试集按 4:1 的比例分配.

2.3 训练过程

由于本文的数据集有限, 在采用 3D CNN 提取特征时, 容易导致模型过度拟合, 因此, 我们引入了 Dropout 层, 并设置不同的 Dropout 参数进行训练. 训练准确率与测试准确率如表 1 所示.

本文发现, 模型在 Dropout 值为 0.7 时, 效果最好, 因此, 我们用这个 Dropout 值对网络进行训练.

在训练过程中, 网络在迭代周期为 1000 时, 损失函数的值最小, 此后, 损失函数开始收敛, 如图 7 所示.

表 1 不同的 Dropout 值对应的准确率

Dropout	Training accuracy	Testing accuracy
0.5	0.85	0.60
0.6	0.93	0.82
0.7	0.98	0.962
0.8	0.88	0.78

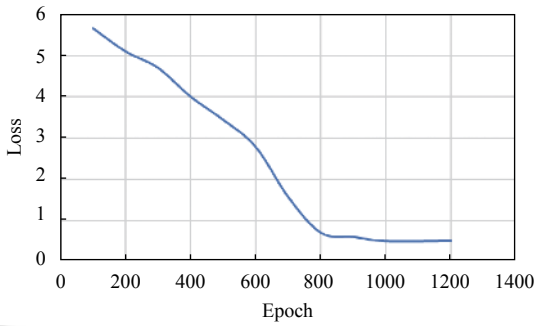

图 7 迭代周期与准确率

与此同时, 模型的优化函数采用自适应估计 (Adam) 梯度优化函数, 其他参数配置如表 2 所示.

表 2 超参数配置

相关参数	Value
Input	(122,122,3)
Convolution3D	(3,3,3)
MaxPooling3D	(2,2,2)
Learning_rate	0.03
Momentum	0.9

2.4 实验结果

本文算法将足球视频场景分为 7 类, 分类结果如表 3 所示, 表格最右列表示未分类 (漏检) 镜头数, 对角线的位置表示正确分类的镜头数. 由表 3 可以看出, 本文算法针对特写镜头的场景分类准确率最高 96%, 与文献[18]对比, 场景类别更加丰富. 同时, 在特写镜头分类时, 准确率高于文献[18]的基于 HMM 的事件检测分类算法. 与文献[19]提出的基于贝叶斯网络的足球事件检测算法相比, 虽然远镜头分类准确率低于其算法, 但特写镜头的准确率高于其算法. 与 Jiang[14]所提出的 2D CNN 网络比较, 特征提取更为有效, 分类的准确率较高. 与 Chen[15]相比, 数据集更丰富, 比较的类别也更多. 与基于 LSTM 的 2D CNN 相比, 3D CNN 同时捕获了二维 CNN 的时空特征, 取得了较好的效果. 本文数据对比如表 4 所示.

3 结束语

2D CNN 在空间特征的学习效果较好, 但无法对

视频的时间特征进行处理. 本文提出的基于 C3D 的足球视频场景分类算法, 分别对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, 本文在特写镜头分类时, 准确率提高了 2%. 本文算法在场景切换检测的基础上, 根据预定义的类别将各个场景片段进行标记, 利用 3D CNN 学习时空特征, 然后将其进行分类. 下一步工作计划是进行实时的足球视频场景分类, 与此同时, 扩展场景类别以识别足球视频中更复杂的场景.

表 3 查全率与准确率实验结果

实际	预测							
	远镜头	中镜头	特写镜头	回放镜头	观众镜头	开场镜头	VAR 镜头	漏检镜头
远镜头	35	2	0	0	0	1	0	2
中镜头	1	28	1	0	0	0	0	1
特写镜头	0	0	46	0	0	0	1	3
回放镜头	0	0	1	29	0	0	1	0
观众镜头	0	0	0	0	12	1	0	2
开场镜头	1	0	0	0	2	24	0	0
VAR 镜头	0	0	0	2	0	0	17	0
查全率	0.88	0.90	0.92	0.93	0.80	0.89	0.89	-
准确率	0.95	0.93	0.96	0.94	0.86	0.92	0.89	-

表 4 特写镜头分类结果对比

Algorithm	Accuracy(%)
谢文娟 [18]	91.04
Tavassolipour *et al.*[19]	86.27
Jiang *et al.* [14]	94.29
Khan *et al.* [20]	94.51
Chen *et al.* [15]	94.11
本文	96

参考文献

1 陆思烨, 李鸿燕, 孙健昊, 等. 基于双阈值灰度直方图的场景切换检测算法及实现. 上海工程技术大学学报, 2018, 32(1): 91–94. [doi: 10.3969/j.issn.1009-444X.2018.01.019]

2 方宏俊, 宋利, 杨小康. 适配分辨率动态变化的低复杂度视频场景切换检测方法. 计算机科学, 2017, 44(2): 290–295. [doi: 10.11896/j.issn.1002-137X.2017.02.049]

3 孙桃, 谢振平, 梅向东, 等. 基于在线 Bayesian 决策的动画场景切换检测方法. 计算机工程与应用, 2016, 52(22): 164–168. [doi: 10.3778/j.issn.1002-8331.1501-0376]

4 段淑玉, 陈艳. 基于块匹配运动估计的视频场景切换检测算法. 广西科技大学学报, 2018, 29(4): 92–98.

5 Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 2001, 42(3): 145–175. [doi: 10.1023/A:1011139631724]

6 Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA. 2008. 1–8.

7 Naveed H, Khan G, Khan AU, *et al*. Human activity recognition using mixture of heterogeneous features and sequential minimal optimization. International Journal of Machine Learning and Cybernetics, 2019, 10(9): 2329–2340. [doi: 10.1007/s13042-018-0870-1]

8 Kang YL, Lim JH, Kankanhalli MS, *et al*. Goal detection in soccer video using audio/visual keywords. Proceedings of 2004 International Conference on Image. Singapore. 2005. 1629–1632.

9 Ekin A, Tekalp AM, Mehrotra R. Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing, 2003, 12(7): 796–807. [doi: 10.1109/TIP.2003.812758]

10 Xu CS, Zhang YF, Zhu GY, *et al*. Using webcast text for semantic event detection in broadcast sports video. IEEE Transactions on Multimedia, 2008, 10(7): 1342–1355. [doi: 10.1109/TMM.2008.2004912]

11 Xu CS, Wang JJ, Wan K, *et al*. Live sports event detection based on broadcast video and web-casting text. Proceedings of the 14th ACM International Conference on Multimedia. Santa Barbara, CA, USA. 2006. 221–230.

12 于俊清, 张强, 王赠凯, 等. 利用回放场景和情感激励检测足球视频精彩镜头. 计算机学报, 2014, 37(6): 1268–1280.

13 Sharif Razavian A, Azizpour H, Sullivan J, *et al*. CNN features off-the-shelf: An astounding baseline for recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, OH, USA. 2014. 512–519.

14 Jiang HH, Lu Y, Xue J. Automatic soccer video event detection based on a deep neural network combined CNN and RNN. Proceedings of the 2016 IEEE 28th International Conference on Tools with Artificial Intelligence. San Jose, CA, USA. 2017. 490–494.

15 Tjondronegoro DW, Chen YPP. Knowledge-discounted event detection in sports video. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 2010, 40(5): 1009–1024. [doi: 10.1109/TSMCA.2010.2046729]

16 Ji SW, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1):

221–231. [doi: 10.1109/TPAMI.2012.59]
17 刘阳, 罗安平. 基于足球比赛事件检测的视频分析方法. 沈阳工业大学学报, 2018, 40(4): 415–419. [doi: 10.7688/j.issn.1000-1646.2018.04.10]
18 谢文娟. 足球视频精彩进球事件检测[硕士学位论文]. 西安: 西安电子科技大学, 2012.
19 Tavassolipour M, Karimian M, Kasaei S. Event detection and summarization in soccer videos using bayesian network and copula. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(2): 291–304. [doi: 10.1109/TCSVT.2013.2243640]
20 Khan MZ, Saleem S, Hassan MA, *et al*. Learning deep C3D features for soccer video event detection. Proceedings of the 2018 14th International Conference on Emerging Technologies. Islamabad, Pakistan. 2018. 1–6. [doi: 1109/ICET.2018.8603644]