

# 一种部位语义感知的视频行为识别方法

刘晓东, 李松阳, 王 淼

(河南工程学院 计算机学院, 河南 郑州 451191)

**摘 要:** 针对传统视频行为识别中在建模高层次语义运动模式和语义运动演化方面的不足, 本文提出一种部位语义感知的行为识别方法. 该方法首先基于人体部位位置信息, 提取人体部位高层次语义特征; 然后基于人体部位语义特征和人体部位语义特征时间演变, 对人体行为进行识别. 我们基于 caffe 架构实现了该方法, 实验结果表明该方法具有很好的性能.

**关键词:** 行为识别; 深度学习; 卷积神经网络

**中图分类号:** TP393

**文献标识码:** A

**文章编号:** 1000-7180(2018)12-0012-04

## A Part Semantic Based Action Recognition Method

LIU Xiao-dong, LI Song-yang, WANG Miao

(School of Computer, Henan Institute of Engineering, Zhengzhou 451191, China)

**Abstract:** Aiming at the deficiency that model high level semantic motion pattern and semantic movement evolution in the research of traditional video action recognition, this paper proposes a part semantic based video action recognition method. First, it extracts high level semantic features based on the position information of the body parts. Second, we propose a feature descriptor for action recognition based on human part semantic and its movement evolution. The method is implemented on Caffe. The experiment results display good performance.

**Key words:** action recognition; deep learning; convolutional neural network.

## 1 引言

随着互联网和移动端智能设备的快速发展, 多媒体视频与图像应用越来越广泛, 包括国家/城市安全、医疗、教育、通讯、工业生产以及文化影视等, 这些视频应用每时每刻产生海量的视频数据<sup>[1]</sup>. 在国外, 仅 YouTube 一个视频网站就拥有 10 亿用户, 每分钟有 300 小时视频被上传. 中国大约有 3 000 万台监控摄像机, 每月将生成 60 EB 的视频数据. 对这些视频数据进行有效分析、理解, 可以使数据创造出更大的经济和社会效益.

视频行为识别是视频理解重要研究内容之一, 目前在视频行为识别方面主要采用基于深度学习的行为识别. 双流卷积神经网络<sup>[2]</sup> (Convolutional

Neural Networks, CNN) 架构通过学习来自静态帧的外观信息和视频帧之间的运动信息来对视频行为进行分类, 但是该架构在信息使用上存在一定的局限性, 第一, 该模型仅对输出的分类预测分数进行融合, 无法学习时空特征之间的像素级对应关系; 第二, 在空间 CNN 上仅采用单一的视频帧, 在时间 CNN 上仅使用 L 个时间相邻帧的光流, 时空信息不能被充分利用. 双流 CNN 在视频行为识别方面取得了很好的性能, 但是由于光流计算量大而无法进行实时处理, 香港中文大学研究人员利用从光流 CNN 学习知识迁移到运动矢量 CNN 来提高性能<sup>[3]</sup>. C3D<sup>[4]</sup> 采用三维深度卷积网络学习时空特征, 可以同时建模运动和外观信息, 在性能上要优于双流卷积神经网络, 该方法采用整个视频帧作为输入,

**收稿日期:** 2018-03-15; **修回日期:** 2018-04-01

**基金项目:** 国家自然科学基金(61501174); 河南省科技厅科技攻关项目(182102310919); 河南工程学院博士基金项目(D2015022); 河南省教育厅科学技术研究重点项目(16A520041)

无需进行任何预处理,可以很容易缩放到的数据集.最近方法利用 CNN 检测部分感兴趣区域,然后提取候选区域时空特征. P-CNN<sup>[5]</sup> 基于姿态估计将每帧图像的身体部位进行裁剪,每个部位补丁均调整为匹配 CNN 输入的固定大小,利用 CNN 提取每个部位外观和光流特征,聚合人体部位轨迹的运动和外观信息进行人体行为细粒度识别.该方法为每个部位定义和训练独立的 CNN 网络,从每个部分提取的特征联结为一个向量,虽然该方法成功捕获到局部信息,但是它并没有建立高层次语义运动模式、语义运动演化和特定行为之间的内在联系,缺乏用于建模目标部分语义的中间层,这不能充分发挥基于 CNN 的方法在识别方面利用部分语义的潜能.

本文针对传统方法在建模高层次语义运动模式、语义运动演化和特定行为关系方面的不足,提出一种语义感知的视频行为识别方法.该方法基于人体部位位置信息,提取人体部位高层次语义特征;利用人体部位语义空间特征和人体部位语义特征时间演变,对人体细粒度行为进行识别;基于 Caffe 架构实现了该方法,实验结果表明该方法性能优于已有研究成果.

## 2 部位语义感知的视频行为识别方法

部位语义感知的视频行为识别包括人体姿态估计和细粒度行为识别两个子网络来,如图 1 所示.

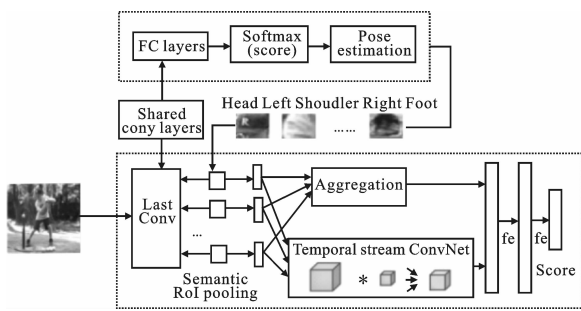


图 1 部位语义感知的视频行为识别

每个视频帧作为一个全卷积神经网络的输入,最后一层卷积层的输出同时作为姿态估计子网络和行为识别子网络的输入,实现计算共享.姿态估计子网络采用基于全卷积网络架构的人体姿态估计方法<sup>[6]</sup>,视频行为识别子网络主要包括人体部位语义特征提取和语义特征时空融合两个阶段.

### 2.1 人体部位语义特征提取

如图 1 所示,人体姿态估计子网络的输出作为

行为人体部位语义特征提取的一个输入,利用 ROI(region of interest) 语义池<sup>[7]</sup> 对最后一层卷积层进行最大池操作,将每个身体部位语义被划分为  $H \times W$  子窗口(在实验阶段选取  $H = W = 7$ ),获取每个部位的语义信息.在本文中 ROI 池是一个以人体部位为中心的一个矩形,每个 ROI 池可以用一个四元组  $\langle r, c, h, w \rangle$  来表示,其中  $\langle r, c \rangle$  为 ROI 池左上角坐标,  $\langle h, w \rangle$  为高度和宽度. ROI 池通过将  $h \times w$  语义窗口划分为  $H \times W$  子窗口来获得人体部位语义.假设选取的身体部位个数为  $P$ ,用  $f_{t_i}^k (k = 1, 2, \dots, P)$  来表示第  $k$  个部位在第  $t_i$  帧的语义特征.

### 2.2 语义特征时空融合

在视频语义特征融合阶段,对提取到的人体部位语义特征进行时空融合.在空间特征融合方面,针对每个部位  $k$ ,对视频帧采用聚合的方法得到固定长度的人体部位语义特征,用  $y_k$  来表示第  $k$  个部位聚合后的语义信息,则有:

$$y_k = g(f_{t_1}^k, f_{t_2}^k, \dots, f_{t_T}^k) \quad (1)$$

在时间特征融合方面,与表示视频运动信息的光流不同,本文采用高层次语义运动特征作为时间流网络的输入,我们给出以下两种语义运动模型:

(1) 获取身体各个部位语义的运动特征,然后将各部位语义运动特征进行连接,得到时间流融合特征.令

$$\Delta f_i^k = f_{t_{i+1}}^k - f_{t_i}^k \quad (2)$$

将矩阵  $[\Delta f_1^k, \Delta f_2^k, \dots, \Delta f_{T-1}^k] (k = 1, 2, \dots, P)$  作为时间流 CNN 的输入得到第  $k$  个部位语义信息的语义描述,记为  $s_k (k = 1, 2, \dots, p)$ ,然后将  $s_k (k = 1, 2, \dots, p)$  进行连接得到时间流融合特征.

(2) 对人体区域部位语义进行融合,比如将左肩、左肘、左手语义信息进行融合.从现实意义上考虑,当我们去识别人体行为时,往往依靠获取一些关键部位运动信息,比如胳膊、腿等.区域部位语义特征融合采用若干部分全链接层,第  $t$  帧第  $g$  个部分全链接方法可以表示为:

$$y_t^g = f\left(\sum_{i=k_i}^{k_j} W_i x_i\right) \quad g = 1, 2, \dots, G \quad (3)$$

式中,  $x_i$  为第  $i (i = 1, 2, \dots, p)$  个部位语义特征;  $W_i$  为与之对应的权重参数;  $G$  为组合个数,比如我们将人体部位组合为左胳膊、左腿、右胳膊、右腿,头部五个区域;  $k_i, k_j$  分别表示区域  $g$  的第  $i$  和第  $j$  个身体部位,令

$$\Delta y_i^q = y_{i+1}^q - y_i^q \tag{4}$$

将矩阵 $[\Delta y_1^q, \Delta y_2^q, \cdots, \Delta y_{T-1}^q](q = 1, 2, \cdots, G)$ 作为时间流 CNN 的输入得到第  $q$  个区域部位语义运动特征的语义描述, 然后再利用一个全链接层去构成一个完整的网络架构, 这使得向后传播可以到达每个部位, 在训练的时候调整各个部位权重, 计算方法为:

$$y = f(\sum_{q=1}^G W_q x_q) \tag{5}$$

式中,  $x_q$  为第  $q$  个区域语义特征;  $W_q$  为与之对应的权重参数.

将空间流 CNN 和时间流 CNN 得到的特征进行融合, 将融合后的特征经过三个全链接层得到一个  $n$  维向量, 其中  $n$  为视频行为个数.

### 3 性能评估

本节给出实验设置、实验结果并对实验结果进行分析.

#### 3.1 实验设置

我们在两个公开的视频行为识别数据集上对我们提出的方法进行性能评估: UCF101 数据集<sup>[9]</sup>和 HMDB51 数据集<sup>[7]</sup>. (1) UCF101 数据集包含 13 k 视频片段, 其中 9.5 k 训练数据集. 每个视频片段平均包含 180 视频帧, 标注为 101 个视频行为类别, 该数据集包含真实的用户上传视频, 包括摄像机运动和杂乱的背景. (2) HMDB51 数据集包含 6.8 k 视频片段, 其中 3.7 k 训练数据集, 标注为 51 个视频行为类别<sup>[10]</sup>.

#### 3.2 实验结果

在该实验中, 我们首先分别验证空间流融合和时间流融合的性能, 为了区别我们将 2.2 小节第(1)种身体各个部位语义运动特征称为部位运动特征, 将 2.2 小节第(2)种方法称为区域运动特征. 表 1 为本文与双流行为识别方法在空间流 CNN 和时间流 CNN 上的性能对比. 从表 1 可以看出, 在 UCF101 数据集和 HMDB51 数据集上, 本文时空流 CNN 性能均优于双流, 这是由于双流 CNN 的输入为原始视频帧或者视频帧运动特征, 而本文 CNN 的输入不仅为高层次语义特征, 且更加关注与视频行为紧密联系的人体部位.

接下来, 将本文空间 CNN 分别与部位运动特征和区域运动特征融合, 表 2 给出了融合后的性能与双流 CNN 及最近几种方法<sup>[11-13]</sup>的精确性验证结果.

表 1 时空融合性能

| 实验方法     | UCF101 | HMDB51 |
|----------|--------|--------|
| 双流空间 CNN | 72.7%  | 40.8%  |
| 本文空间 CNN | 83.1%  | 45.1%  |
| 双流时间 CNN | 81%    | 46.6%  |
| 部位运动特征   | 88.2%  | 62.3%  |
| 区域运动特征   | 88.7%  | 62.8%  |

表 2 精确性验证

| 实验方法                            | UCF101 | HMDB51 |
|---------------------------------|--------|--------|
| 双流                              | 86.9%  | 58%    |
| Bilen et al. <sup>[11]</sup>    | 89.1%  | 65.2%  |
| Fernando et al. <sup>[12]</sup> | 91.4%  | 66.9%  |
| Yang et al. <sup>[13]</sup>     | 91.6%  | 61.8%  |
| 部位运动特征融合                        | 92.7%  | 67.8%  |
| 区域运动特征融合                        | 93.1%  | 68.1%  |

从表 2 可以看出, 本文提出的部位语义感知的视频行为识别方法在 UCF101 数据集和 HMDB51 数据集上的性能均优于已有方法. 这是由于(1)本文对高层次语义运动模式进行了建模; (2)本文充分利用了部位语义信息. 同时, 我们也注意到区域运动特征融合方法在性能上要优于部位运动特征融合, 这说明在视频行为识别中部位不能划分过细而忽视部位之间的区域特征.

### 4 结束语

在本文中, 我们提出了一种部位语义感知的视频行为识别方法. 该方法利用人体姿态估计得到的人体部位位置和 ROI 池提取人体部位语义; 利用人体部位语义空间特征和人体部位语义时间演变, 对人体细粒度行为进行识别. 实验结果表面本文所提出的方法在性能上要优于已有方法.

#### 参考文献:

[1] Su W T, Lu Y H, Kaseb A S. Harvest the information from multimedia big data in global camera networks[C]//Multimedia Big Data (BigMM), 2015 IEEE International Conference on. Beijing, China, IEEE, 2015; 184-191.

[2] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. Montreal, Quebec, Canada, 2014; 568-576.

[3] Zhang Bowen, Wang Limin, Wang Zhe, et al. Real-time action recognition with enhanced motion vector

- CNNs[C]//2016 IEEE International Conference on Computer Vision and Pattern Recognition. Boston, Massac, 2016: 2718-2726.
- [4] 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. Houston, TX, USA, 2015: 4489-4497.
- [5] Chéron G, Laptev I, Schmid C. P-cnn: Pose-based cnn features for action recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. Houston, TX, USA 2015: 3218-3226.
- [6] Chu X, Ouyang W, Li H, et al. Structured feature learning for pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition 2016. Boston, Massac, 2016: 4715-4723.
- [7] Girshick R. Fast R-CNN[C]//IEEE International Conference on Computer Vision. Houston, TX, USA, 2015: 1440-1448.
- [8] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild [EB/OL]. [2018-03-10]. <http://arxiv.org/abs/1212.0402>.
- [9] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition [C]//IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563.
- [10] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition[J]. Acm Transactions on Information Systems, 2016, 22(1): 20-36.
- [11] Bilen H, Fernando B, Gavves E, et al. Dynamic image networks for action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 3034-3042.
- [12] Fernando B, Anderson P, Hutter M, et al. Discriminative hierarchical rank pooling for activity recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 1924-1932.
- [13] Yang X, Molchanov P, Kautz J. Multilayer and multi-modal fusion of deep neural networks for video classification [C]//ACM on Multimedia Conference 2016. Amsterdam, The Netherlands, 2016: 978-987.

### 作者简介:

**刘晓东** 男, (1981-), 博士, 博士后, 副教授. 研究方向为网络信息与多媒体. E-mail: liuxiaodongxht@qq.com.

**李松阳** 男, (1985-), 博士, 副教授. 研究方向为计算机图形学.

**王 森** 男, (1981-), 博士, 副教授. 研究方向为空间查询、空间推理.