

# 利用网络图像增强行为识别

闻 号

(安徽大学 电子信息工程学院,安徽 合肥 230601)

**摘 要:**鉴于商业视觉搜索引擎的日益成熟,网络数据可能是下一个扩大视觉识别的重要数据源。通过观察发现,动作名称查询到的网络图像具有歧视性的动作场景。网络图像的歧视性信息和视频的时间信息之间有相互补充的优势。在此基础上提出一种利用大量的网络图像来增强行为识别的方法。具体框架是:提取行为视频的密集轨迹特征,并与网络图像特征相结合后放入支持向量机中训练分类。该方法是一个跨域学习问题,为了有效地利用网络图像特征,引入了跨域字典学习算法来处理网络图像,以解决网络图像域和视频域之间存在的域差异问题。由于网络图像可以轻松地在网络上获取,所以该方法几乎零成本地增强行为识别。在 KTH 和 YouTube 数据集上的实验结果表明,该方法有效提高了人体行为识别的准确率。

**关键词:**网络学习;迁移学习;行为识别;密集轨迹;字典学习

**中图分类号:**TP39

**文献标识码:**A

**文章编号:**1673-629X(2019)01-0031-04

**doi:**10.3969/j.issn.1673-629X.2019.01.007

## Improvement of Action Recognition Using Web Images

WEN Hao

(School of Electronics and Information Engineering, Anhui University, Hefei 230601, China)

**Abstract:** In view of the growing maturity of commercial visual search engines, Web data may be the next important data source to expand visual recognition. It is observed that the Web images queried by the action name is discriminatory to the action scene. Clearly, there are complementary benefits between the temporal information available in videos and the discriminatory scenes portrayed in images. On the basis, we propose an algorithm which can enhance action recognition by using a large number of Web images. We extract the dense trajectory feature of behavior video and put it into support vector machine for training classification in combination with Web image feature. This algorithm is a cross-domain learning problem. In order to effectively use Web image features, we introduce a cross-domain dictionary learning algorithm to deal with Web images for solving the domain differences between Web image domain and video domain. Because the Web images can be easily obtained on the network, it can enhance action recognition with at almost zero cost. Experiment shows that the proposed algorithm can improve the accuracy of human action recognition effectively on KTH and YouTube datasets.

**Key words:** Web learning; transfer learning; action recognition; dense trajectory; dictionary learning

## 0 引 言

随着智能手机、动作相机、监控摄像机等的普及,网络上视频的数量已经超出了人们观看所有视频的能力。由于行为识别问题在视频监控、人机交互和视频内容分析等方面具有很大的潜力,视频中人体行为的识别受到了广泛关注。例如,Wang 等<sup>[1]</sup>提出了一种改进的密集轨迹算法。文献[2]使用了在做小码书情况下的多时空特征。文献[3]使用多种特征来描述行为的整体分布和局部变化。文献[4]使用能量函数对运动区域进行高斯取样,使样本点分布于运动剧烈的

区域。虽然这些方法已经在目标检测和跟踪方面取得了惊人的进展,但是从视频中检测出更多的抽象动作和事件仍然具有挑战性。

在训练人体行为模型时需要大量的训练数据来避免过度拟合,然而数据获取需要耗费大量人力物力。相比之下,从网络上收集和处理数据要便宜得多。而且观察到,通过动作名称查询的 Web 图像通常描述一个歧视性的动作场景,以此可以捕捉并突出显示视频中感兴趣的动作和事件。所以这是一个证明网络图像可以增强行为识别的有力证据。显然在视频中提供的

收稿日期:2018-01-20

修回日期:2018-05-24

网络出版时间:2018-09-21

基金项目:安徽省自然科学基金(1508085MF120)

作者简介:闻 号(1991-),男,硕士,研究方向为网络数据在行为识别上的应用、机器学习、迁移学习。

网络出版地址:<http://kns.cnki.net/kcms/detail/61.1450.TP.20180920.1536.028.html>

时间信息和图像中描绘的歧视性场景间存在互补优势。

提出的方法与 Web learning (网络学习) 息息相关。典型的工作有文献[5-6], 从这些研究内容可以看出, 网络数据域与目标域之间的域差异是个热点问题。域差异问题是一个跨域学习问题, 也是一个迁移

学习问题。因此, 试图通过跨域字典学习的方法, 同时对网络图像域和目标域进行字典学习来解决这个问题。

## 1 方法实现

设计的人体行为识别算法流程如图 1 所示。

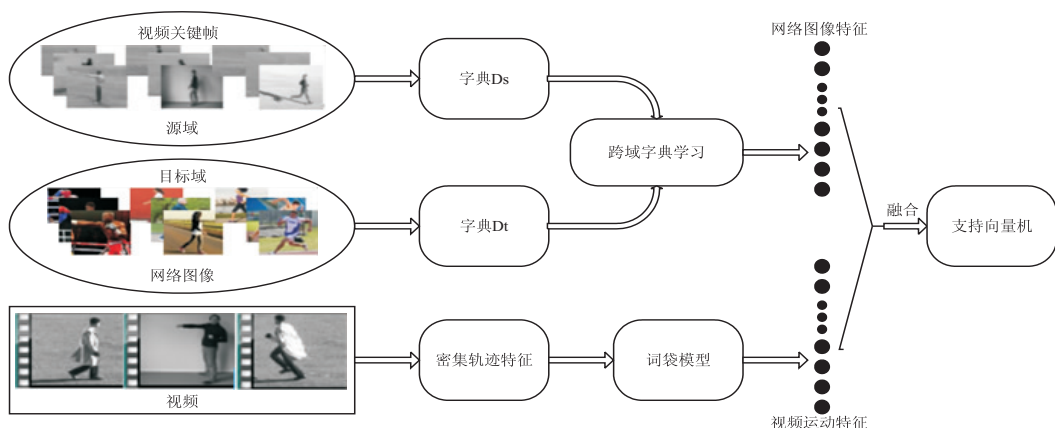


图 1 算法流程

获取网络图像作为目标域, 获取数据集中每个类视频的关键帧图像作为源域。使用 SIFT 算法提取的目标域和源域图像的底层特征描述子, 分别作为跨域字典学习算法的输入, 同时进行字典学习, 进而获得网络图像的特征表示; 使用文献[1]提出的密集轨迹算法提取数据集中视频的密集轨迹特征, 再通过字典学习、词袋模型编码得到视频中人体行为的特征表示。将两组特征进行长拼接, 把拼接后的特征向量放入支持向量机中进行训练分类。实验选择的数据集分别是 KTH<sup>[7]</sup> 和 YouTube<sup>[8]</sup>。

### 1.1 获取训练集图像

本节列出了收集和组织网络图像的步骤。借助 Google Image API, 可以轻松地以几乎零成本获取所需的动作图像。使用每个类别名称作为关键字在 Google 图片搜索服务中下载检索到的图像。使用照片过滤器删除不太可能出现在视频中的人造图像。收集了大约 15 000 张网络图像(如图 2 所示)分别用在 KTH 数据库中的六种人体行为和 YouTube 数据库的十一种人体行为的识别实验中。



图 2 网络图片(从左向右分别表示骑车、骑马、高尔夫、荡秋千、颠球)

### 1.2 跨域字典学习

首先引入一个基本问题, 设  $Y_i$  表示  $L$  个  $n$  维目标域输入信号,  $Y_s$  表示  $M$  个辅助域输入信号, 即  $Y_i = (y_i^1, y_i^2, \dots, y_i^L) \in R^{n \times L}$ ,  $Y_s = (y_s^1, y_s^2, \dots, y_s^M) \in R^{n \times M}$ , 可以通过式(1)来学习字典  $D_i, D_s$ 。

$$\begin{aligned} \langle D_i, D_s, X_i, X_s \rangle = \arg \min_{D_i, D_s, X_i, X_s} & \|Y_i - D_i X_i\|_2^2 + \\ & \|Y_s - D_s X_s\|_2^2 + \Phi([X_i, X_s]) \\ \text{s. t.} & \forall i, [\|X_i^i\|_0, \|X_s^i\|_0] \leq T \end{aligned} \quad (1)$$

其中,  $D_i$  表示目标域字典;  $X_i = (X_i^1, X_i^2, \dots, X_i^L)$  表示目标域的稀疏系数;  $D_s$  表示辅助域字典;  $X_s =$

$(X_s^1, X_s^2, \dots, X_s^M)$  表示辅助域的稀疏系数;  $\Phi(\cdot)$  表示相同类别的特征向量描述不同数据集之间的欧氏距离, 即两个相同动作在两个不同数据集中存在的差异性。

根据文献[9]中提出的一个策略:将不同视角拍摄的同一个行为投影到跨视角字典对时,鼓励其享有相同的特征表示。受该策略的启发,将  $\Phi([X_t, X_s])$  重写为  $\|X_t^T - AX_s^T\|_2^2, \|X_s^T - AX_t^T\|_2^2$  的值越小,在相似点之间共享相同标签的可能性越大。

根据 Zhu Fan 等<sup>[10]</sup>提出的方法,对式1转换:

$$\langle D_t, D_s, X_t, A, W \rangle = \arg \min_{D_t, D_s, X_t, A, W} \left\| \begin{pmatrix} Y_t \\ Y_s A^T \\ \sqrt{\alpha} Q \\ \sqrt{\beta} H \end{pmatrix} - \begin{pmatrix} D_t \\ D_s \\ \sqrt{\alpha} v \\ \sqrt{\beta} W \end{pmatrix} X_t \right\|_2^2 \quad (2)$$

$$\text{s. t.} \quad \forall i, \|X_t^i\|_0 \leq T$$

其中,  $W$  表示分类器  $f(x)$  的系数;  $H$  表示目标域类标签;  $v$  表示线性变换矩阵;  $\alpha$  和  $\beta$  作为权值系数分别表示  $\|Q - vX_t\|_2^2$  和  $\|H - WX_t\|_2^2$  的相对贡献。

也可以把式2转换为最简单的形式,上式因子可简写为:

$$\begin{cases} Y = (Y_t^T, (Y_s A^T)^T, \sqrt{\alpha} Q^T, \sqrt{\beta} H^T)^T \\ D = (D_t^T, D_s^T, \sqrt{\alpha} v^T, \sqrt{\beta} W^T)^T \end{cases} \quad (3)$$

优化问题目标函数简化为:

$$\langle D_t, X_t \rangle = \arg \min_{D_t, X_t} \|Y - DX_t\|_2^2 \quad (4)$$

$$\text{s. t.} \quad \forall i, \|X_t^i\|_0 \leq T$$

从而优化问题即可使用 K-SVD<sup>[11]</sup>算法通过迭代更新的方式求解。

### 1.3 词袋模型

根据文献[1]提出的密集轨迹算法获取行为视频的底层特征描述子。为了评估文中方法的性能,使用标准的词袋模型方法,为底层特征描述子构造了一个字典。根据经验将字典的可视化词语个数固定为4 000,使用 k-means 方法随机选择 100 000 训练特征进行聚类。初始化 k-means 8 次,以此提高精度,保证最低的误差结果。特征描述子会根据欧氏距离被分配到它们最接近的词汇,由此产生的视觉词汇直方图被用作视频中人体行为的特征表示。

## 2 实验

### 2.1 数据集

KTH 数据集包含六种人类运动行为:散步、慢走、跑、拳击、挥手和鼓掌(如图3所示)。每一种行为由25个人展示数次,分别拍摄在四个不同场景下。数据库总共有598个视频样本。根据文献[7]中的实验设置把样本中(2,3,5,6,7,8,9,10,22)9个人分为测试集,剩下的16人为训练集。

YouTube 数据集包含11种人类行为:骑车、跳水、高尔夫、颠球、蹦床、骑马、投篮、排球、秋千、网球和遛狗(如图4所示)。

### 2.2 实验结果分析

表1和表2分别列出了在 KTH 数据集和 YouTube 数据集中的实验结果。可以看出,文中方法比密集轨迹算法表现得更出色,在 KTH 数据集中准确率提高了1%,在 YouTube 数据集中提高了2.2%。在具有背景复杂、拍摄时摄像机移动等复杂视频的 YouTube 数据集中,文中方法明显优于其他方法。实验结果表明,该方法可以有效地增强视频中的动作识别能力。



图3 KTH 数据库视频实例

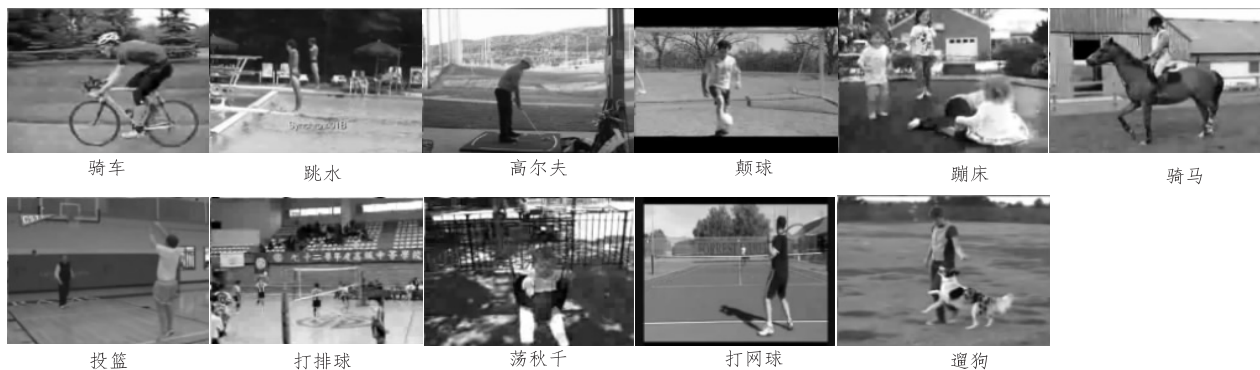


图4 YouTube 数据库视频实例

表 1 在 KTH 数据集 中的实验结果

方法	准确率/%
文献[7]	91.8
文献[12]	93.3
密集轨迹	93.1
文中方法	93.9

表 2 在 YouTube 数据集 中的实验结果

方法	准确率/%
文献[13]	71.2
文献[14]	75.2
密集轨迹	83.3
文中方法	85.5

在对网络图像进行跨域字典学习时引入了视频关键帧,所以不确定视频关键帧有没有对结果产生影响。对此进行了一组对比实验,如表 3、表 4 所示。第一个是只使用视频作为输入;第二个是视频与视频关键帧作为输入;第三个是视频加上视频关键帧和网络图片作为输入。实验结果表明,文中方法有效增强了密集轨迹算法对人体行为的识别能力。

表 3 使用不同的训练数据在 KTH 数据集 中的实验结果

方法	准确率/%
视频	92.7
视频+关键帧	92.7
视频+关键帧+Web images	93.9

表 4 使用不同的训练数据在 YouTube 数据集 中的实验结果

方法	准确率/%
视频	83.3
视频+关键帧	83.6
视频+关键帧+Web images	85.5

3 结束语

通过对网络数据学习理论的研究,提出了一种利用大量的网络数据作为辅助数据来增强密集轨迹算法对人体行为的识别能力的方法。实验结果表明,该方法有效提高了密集轨迹算法对人体行为的识别能力。特别对含有质量低、场景较复杂等复杂视频的 YouTube 数据库,其表现更突出。下一步的工作是解决图片的收集问题,不再是通过人为筛选图片,而是通过训练的人体行为模型自动筛选图片,这样会大大提高图片获取的速度和数量。

参考文献:

[1] WANG Heng, KLASER A, SCHMID C, et al. Action recog-

niton by dense trajectories[C]//IEEE conference on computer vision and pattern recognition. Providence, RI, USA: IEEE, 2011: 3169-3176.

[2] 宋健明, 张 桦, 高 赞, 等. 基于多时空特征的人体动作识别算法[J]. 光电子·激光, 2014, 25(10): 2009-2017.

[3] 秦华标, 张亚宁, 蔡静静. 基于复合时空特征的人体行为识别方法[J]. 计算机辅助设计与图形学学报, 2014, 26(8): 1320-1325.

[4] 刘雨娇, 范 勇, 高 琳, 等. 基于时空深度特征的人体行为识别算法[J]. 计算机工程, 2015, 41(5): 259-263.

[5] GAN Chuang, YAO Ting, YANG Kuiyuan, et al. You lead, we exceed: labor-free video concept learning by jointly exploiting web videos and images[C]//IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE, 2016: 923-932.

[6] GAN Chuang, SUN Chen, DUAN Lixin, et al. Webly-supervised video recognition by mutually voting for relevant web images and web video frames[C]//European conference on computer vision. [s. l.]: Springer International Publishing, 2016: 849-866.

[7] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies[C]//IEEE conference on computer vision and pattern recognition. Anchorage, AK, USA: IEEE, 2008: 1-8.

[8] GORELICK L, BLANK M, SHECHTMAN E, et al. Action as space-time shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 29(12): 2247-2253.

[9] ZHENG Jingjing, JIANG Zhuolin, PHILLIPS P J, et al. Cross-view action recognition via a transferable dictionary pair[C]//BMVC. [s. l.]: [s. n.], 2012: 1-11.

[10] ZHU Fan, SHAO Ling. Weakly-supervised cross-domain dictionary learning for visual recognition[J]. International Journal of Computer Vision, 2014, 109(1-2): 42-59.

[11] AHARON M, ELAD M, BRUCKSTEIN A. rmK-SVD: an algorithm for designing overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Processing, 2006, 54(11): 4311-4322.

[12] YUAN Junsong, LIU Zicheng, WU Ying. Discriminative sub-volume search for efficient action detection[C]//IEEE conference on computer vision and pattern recognition. Miami, FL, USA: IEEE, 2009: 2442-2449.

[13] LIU Jingen, LUO Jiebo, SHAH M. Recognizing realistic actions from videos in the wild[C]//IEEE conference on computer vision and pattern recognition. Miami, FL, USA: IEEE, 2009: 1996-2003.

[14] IKIZLERCINBIS N, SCLAROFF S. Object, scene and actions: combining multiple features for human action recognition[C]//European conference on computer vision. [s. l.]: [s. n.], 2010: 494-507.