行为识别中一种基于融合特征的改进 VLAD 编码方法

罗会兰,王婵娟
(江西理工大学信息工程学院,江西赣州 341000)

摘 要: 本文提出了一种新的基于融合特征的改进 VLAD(Vector of Locally Aggregated Descriptors)编码方法,该方法命名为 IVLAD(Improved Vector of Locally Aggregated Descriptors),将其应用于行为识别算法中,得到了较好的性能提升. 针对单一特征描述符在描述视频空间信息的不足,提出将位置信息映射到特征空间中进行融合编码得到表示向量. 在编码阶段为了克服传统 VLAD 方法只考虑特征与聚类中心距离的不足,提出在其基础之上另外计算每个聚类中心与其最相似特征的差值. 为了进一步提高识别准确度,本文还提出对表征向量自身串联用以升维. 另外本文还研究了不同词典大小及归一化方法对于识别算法的影响. 在两个大型数据库 UCF101 及 HMDB51 上的实验比较表明,本文提出的方法比传统 VLAD 方法具有较大的性能提升.

关键词: 行为识别; 位置信息; 级联; 表示向量
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)01-0049-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.01.007

An Improved VLAD Coding Method Based on Fusion Feature in Action Recognition

LUO Hui-lan, WANG Chan-juan
(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi* 341000, *China*)

Abstract: A novel coding method IVLAD (Improved Vector of Locally Aggregated Descriptors) based on the fusion of features was proposed in this paper. It obtained good performance in behavior recognition. In order to solve the problem that single feature descriptor cannot express space information well, location information was mapped into feature space and then jointly coded to get the video expression vector. In order to avoid the deficiency of the traditional VLAD methods which only consider the distances of features and clustering centers, the distance between each cluster and its most similar feature was also used in the coding stage. Finally concatenating the video expression vector with itself was proposed to raise the dimension of vectors to further improve the recognition accuracy. Furthermore, the influences of the visual dictionary size, the location dictionary size and the normalization method on the recognition accuracy were studied. The experimental results on two large databases UCF101 and HMDB51 have shown that the proposed method had better performance than the traditional VLAD method.

Key words: action recognition; position information; concatenate; expression vector

1 引言

视频中的行为识别[1~5]作为计算机视觉的分支,应用领域非常广泛,比如智能监控[6]、人机交互、基于内容的视频搜索等. 现实环境录制的视频往往存在背景杂乱、相机抖动、尺度及视角变化等多种问题. 且相同动作类之间可能存在差异、不同动作类之间又具有相似性. 比如不同人走路的步子大小存在差异,蹲下和坐下这两个不同动作又有较大相似性. 这些问题使得行为识别一直是计算机视觉领域一个非常具有挑战性的难题.

当前行为识别研究领域大体可以分为两类:基于深

收稿日期:2017-10-18;修回日期:2018-02-12;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61862031, No. 61462035);江西省自然科学基金“视觉特征表达的自我深度学习模型研究”(No. 20171BAB202014)

度网络的方法[7~12]和传统的人工设计特征提取方法[13~17].深度学习中用于行为识别的主流方法是卷积神经网络 CNN(Convolutional Neural Networks)及其衍生方法[8,11,12].CNN 最先应用于图像识别领域并取得较大成功.但视频和图像不同,图像是静止的,视频是动态的.因此 Annane 等人[7]提出构建双流 CNN 网络用于行为识别,具体做法是将视频看做一段图像序列,空间流计算图像帧的 CNN 特征,时间流计算若干图像帧间的光流 CNN 特征,最后再将两者进行融合.这种方法虽然将立体的视频识别问题转化为了平面的图像识别问题,但却丢失了动作的时间关联信息.为了弥补双流架构在时间信息上的丢失,Wang L 等人[8]提出了三流 CNN 架构.该架构在双流架构的基础之上将时间流进一步细分,分为局部时间流和全局时间流.动作图像特征和光流特征分别作为空间流和局部时间流的输入,通过学习运动叠差图像 MSDI(Motion Stacked Difference Image)的 CNN 特征作为全局时间流的输入.在 UCF101 及 HMDB51 数据库上的实验表明,基于三流 CNN 架构的识别准确度比双流 CNN 方法[7]分别高了 1.7% 和 1.9%.

传统的方法就是对视频特征进行手动提取,然后再训练模型进行预测分类.相较于深度学习的端到端的方法,传统方法由于所提取特征可能并不能完全满足后续的分类任务要求,所以目前在性能上无法与端到端的方法相比.但手动提取的特征针对性更强,且在训练速度和对训练数据量的要求上都比基于深度模型的方法有优势.更何况经过多年的研究,许多优秀的特征提取和描述方法被开发出来并获得较好的识别效果.

文献[18]中提出利用时空兴趣点 STIP(Space Time Interest Points)来描述视频,STIP 特征[19]是利用角点探测器获得兴趣点进行跟踪并提取描述符信息(包括 HOF(Histogram of Flow)和 HOG(Histogram of Gray)).文献[20]提出用稠密轨迹 DT(Dense Trajectories)来表示视频,DT 特征是对视频进行稠密采样,捕捉运动轨迹,并沿着光流方向提取轨迹的描述符信息(包括 HOF、HOG、MBH(Motion Boundary Histograms)).后来 Wang[21]又提出改进的稠密轨迹 IDT(Improved Dense Trajectories),改进版中对人物进行了框定,消除了相机抖动及背景杂乱的影响.基于 IDT 特征的行为识别方法得到的识别准确度一度达到世界领先水平.但因为其沿着光流进行稠密采样,计算量较大,识别效率不高,Liu 等人[22]提出了改进的 HOG、HOF、MBH 特征提取方式,其方法较原始方法在速度上有较大提升.

基于手动提取特征的传统行为识别方法中,最常用的模型是视觉词袋模型.这类方法主要分为四步:特征选取、特征编码、向量归一化及分类预测.研究者除了在特征构造上下功夫之外,也在不断探索如何对提取到的特征进行编码从而获取更高效的向量表示.许多有效的编码方法被开发出来,比如说局部软分配(Local Soft Assignment)[23]、稀疏编码[24]、局部控制线性编码[25]等.后来研究者们又提出表达描述符与所属聚类中心差的编码方法,比如说:费舍尔编码[26]、超级向量编码[27]以及 VLAD(Vector of Locally Aggregated Descriptors)编码[28].文献[13]提出分别对位置信息和时空信息进行稀疏编码得到时空向量 SDV 和位置向量 SLV,然后再将 SDV 和 SLV 组合起来得到视频的稀疏向量 SSCV.文献[17]提出按照兴趣点位置信息的相似性对视觉特征进行聚类,然后进行编码得到视频的时空位置向量,最后将其与对视觉特征编码得到的向量串联用以表示视频.该方法将兴趣点的位置信息映射到了特征空间,丰富了特征表示,但是其编码方法弱化了聚类中心的作用.

受到以上文献[13,17,28]以及近年来多流 CNN 网络结构模型的启发,本文也利用位置信息对视频进行补充描述,构造了由时空向量、位置向量组成的多流手动特征结构作为视频的立体表示.旨在通过融合时空信息及位置信息进行编码,对传统方法进行改进,探索传统方法的新思路.同时编码方法选择高效的 VLAD 替代文献[13]中的稀疏编码,为了强化聚类中心的作用,本文还对 VLAD 做了改进,后期的实验证明,改进后的编码方法获取到的视频表示表征能力更强.

2 IVLAD 用于行为识别

本文提出了一种新的 IVLAD 行为识别算法,将位置信息映射到描述符中构建视频的立体表示向量进行行为识别,算法结构框图如图 1 所示.首先提取视频的兴趣点,获得兴趣点的视觉特征及位置信息;在此基础上,对视觉信息及位置信息分别构建视觉词典 VD_1(Visual Dictionary)及位置词典 LD_2(Location Dicyionary);然后将视觉特征在视觉词典上进行编码得到视觉向量.为了将位置信息映射到特征空间中,为视频采样得到的每个兴趣点分配一个残差特征和一个组特征.兴趣点的残差特征是视觉向量计算时该兴趣点的视觉特征与其所属聚类中心的残差;兴趣点的组特征是视觉词典生成时该兴趣点所属的组类信息.依据位置特征的聚类策略分别对残差特征和组特征进行聚类,通过这种映射方式,就将位置信息传递到了特征空间中.然后各自编码得到相对应的残差向量和组特征向量,将残差向量和组特征向量串联起来就是视频的位置向量.再将视觉向量和位置向量串联得到视频的全局表示向量 IVLAD,然后进行归一化处理;最后将归一化后的编码向量与其自身进行串联升维,输入到线性 SVM 中进行分类预测.

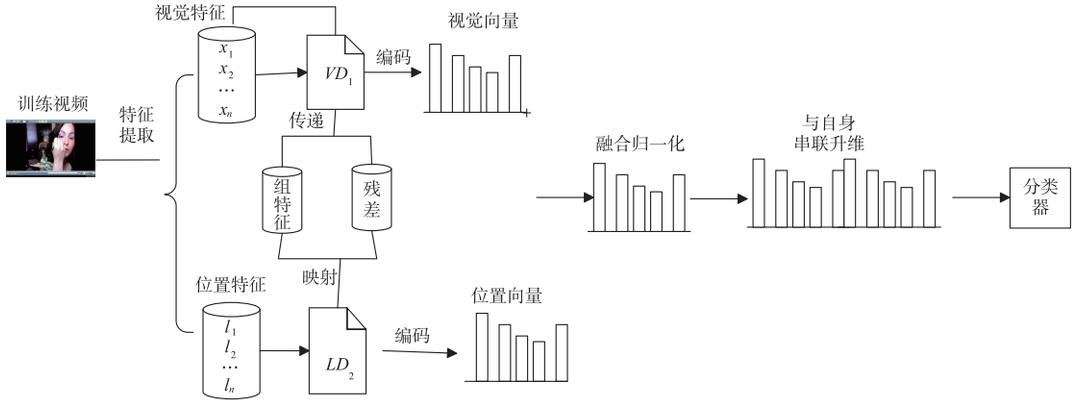

图1　IVLAD算法结构框图

2.1　特征提取

STIP 特征和 IDT 特征是人工设计特征提取方法中最常用的两种特征. STIP 特征是使用角点检测器(比如 Harris 和 Forstner[19])对视频的时空兴趣点进行检测,然后对检测到的兴趣点进行描述得到特征向量. 提取的特征信息有兴趣点的方向梯度直方图 HOG 和光流方向直方图 HOF 以及兴趣点的位置信息. 图 2 示例了打篮球的行为片段中时空兴趣点的变化,其中黄色圆圈表示角点检测器伴随着运动的进行检测到的角点. 可以很明显的看到,随着运动的发生,兴趣点的位置发生了变化.

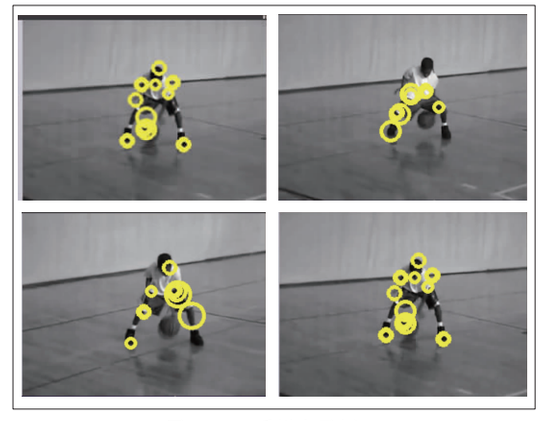

图2　STIP采样点示例

改进稠密轨迹 IDT 特征是对人体进行稠密采样得到运动轨迹点,然后对轨迹点提取特征. 提取的特征信息有轨迹点的方向梯度直方图 HOG、光流方向直方图 HOF、运动边界直方图 MBH 以及轨迹点的位置信息. 图 3 示例了一个女孩在画眼妆的行为片段中轨迹点的变化. 其中红色小点表示采样点,绿色小点表示有显著运动的采样点. 因为采样方式的不同,IDT 所检测到的角点比 STIP 要密集得多,所包含的信息量更大. 图 3 中绿色区域的变化则明显表明,即使是画眼妆这种动作幅度较小的行为,特征点的位置也会发生变化.

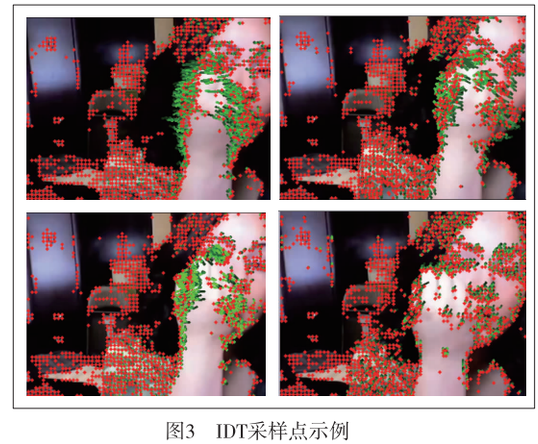

图3　IDT采样点示例

在以往的行为识别方法[16,20,21,29,30]中,选取特征对视频进行描述时只考虑了描述视频外观和运动信息的描述符(HOG、HOF 和 MBH),忽略了位置信息对视频的表示. 而由图 2 和图 3 可以看出,伴随着动作的发生,采样点的位置也会发生相应变化,因此将采样点的位置信息作为位置特征与视觉特征联合表示视频,将有助于完善视频表征. 图 4 示例了本文用于描述视频的特征构造过程. 本文提出不但使用兴趣点的视觉特征,也使用兴趣点的相对位置信息来构造视频的特征表达. 其中位置特征 P(Position)表示为 $P=[x/Nr, y/Nc, t/Nf]$. 式中的 x、y、t 分别表示兴趣点的横轴、纵轴和时间轴上的值,而 Nr、Nc 和 Nf 分别表示视频的高、宽和帧数. 这种规范化操作使得对于输入的任何视频,位置特征值都在相同的区间范围[0,1]内.

2.2　视频表示向量构造

由 2.1 节可知,假设在一段视频中提取到 n 个兴趣点,那么这段视频的视觉特征可以表示为 $\chi=\{x_1,\cdots,x_j,\cdots,x_n\}$,相应的位置特征可以表示为 $P=\{p_1,\cdots,p_j,\cdots,p_n\}$,其中 x_j 和 p_j 分别表示第 j 个兴趣点的视觉特征和位置特征. 为了获取更加高效的视觉表示,需要对底

层特征进行编码得到视频的向量级表示. 图 5 示例了视频向量级表示的生成过程. 2.2.1 和 2.2.2 将分别阐述视觉向量和位置向量的构造过程.

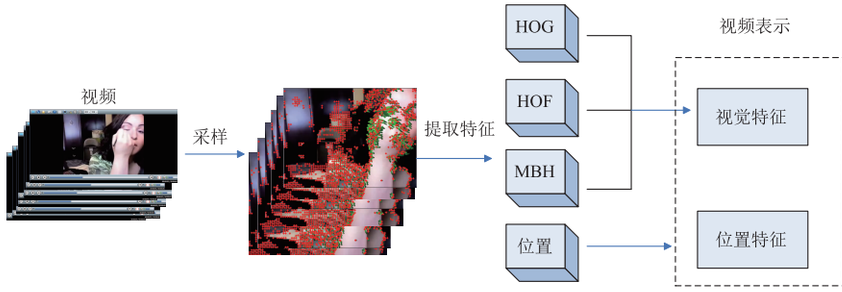

图4 用以描述视频的特征构造过程

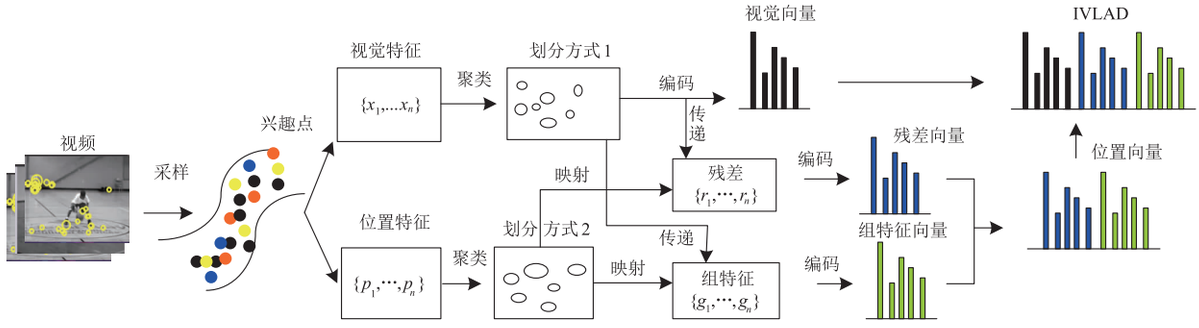

图5 用以描述视频的向量构造过程

2.2.1 视觉向量的构造

首先利用 k 均值聚类对视觉特征进行聚类,生成视觉词典. 假设聚类得到的视觉词典 VD_1 大小为 k_1,则 VD_1 表示为 $VD_1=\{C_1,\cdots,C_i,\cdots,C_{k_1}\}$,其中 C_i 表示视觉词典中第 i 个聚类中心. 原始的 VLAD 编码是计算每个聚类中心与其所属元素的差值之和,即第 i 个聚类中心的编码向量表示为

$$VLAD_i=\sum_{j=1}^{N_i}(x_j-C_i) \tag{1}$$

其中 x_j 表示聚类中心 C_i 所包含的第 j 个视觉特征,N_i 表示聚类中心 C_i 所包含的视觉特征个数. 为了解决突发性问题,文献[17]提出对 VLAD 方法中的聚类中心与其所属元素的残差之和做平均池化,本文也采用这种方法,同时为了增强视觉特征中聚类中心的作用,本文还提出增加对聚类中心与其最相似元素的残差的计算. 所以本文中第 i 个聚类中心的 VLAD 计算公式为

$$VLAD_{C_i}=\frac{1}{N_i}\sum_{j=1}^{N_i}(x_j-C_i)+(x_t-C_i) \tag{2}$$

其中 x_t 表示聚类中心 C_i 所包含的视觉特征中与其最相似的视觉特征. 综上所述,当视觉字典大小为 k_1 时,视频的视觉编码向量可以表示为

$$VLAD_v=[VLAD_{C_1},\cdots,VLAD_{C_{k_1}}] \tag{3}$$

2.2.2 位置向量的构造

由上文可知,兴趣点的位置特征和视觉特征一一对应,即每个兴趣点 j 都对应一个视觉特征 x_j 和一个位置特征 p_j. 为了将位置信息映射到视觉特征空间中,为每个兴趣点分配一个残差特征,兴趣点的残差特征值为视觉向量计算时该兴趣点的视觉特征与其所属聚类中心的残差. 也就是说我们把上一步中第 j 个兴趣点与其所属聚类中心的残差赋值给 r_j,作为第 j 个兴趣点的残差表示. 假设对位置特征进行聚类得到一个大小为 k_2 的位置词典 LD_2,p_j 是 LD_2 中第 i 个聚类中心 L_i 所包含的第 j 个位置特征,则将对应的 r_j 分配给第 i 个残差聚类中心,依照这种划分策略,就将所有兴趣点的残差特征聚类得到了 k_2 个聚类中心. 因为残差特征传递的是兴趣点的视觉特征与其所属聚类中心的差值,所以不再强化聚类中心的作用,则基于位置特征聚类策略的第 i 个聚类中心的残差向量可以编码表示为

$$VLAD_{R_i}=\frac{1}{N_i}\sum_{j=1}^{N_i}(r_j-R_i) \tag{4}$$

其中 R_i 表示依据位置特征聚类得到的对应的第 i 个残差聚类中心,N_i 表示第 i 个聚类中心 R_i 所包含的残差数,r_j 表示聚类中心 R_i 包含的第 j 个残差. 则视频基于位置特征划分的残差向量表示为

$$VLAD_r=[VLAD_{R_1},\cdots,VLAD_{R_{k_2}}] \tag{5}$$

为了保留上一步计算视觉向量时的聚类信息,为视频所提取的 n 个兴趣点分配一个组特征(group features). 也就是假设第 j 个兴趣点在计算视觉向量时属

于第 i 个聚类中心,那么给第 j 个兴趣点分配一个 g_j 表示第 j 个兴趣点的组类特征. g_j 定义为一个 k_1 维的向量 $g_j=[01000\cdots000]$,其中 1 位元素为 1,用以区分所属组别,其余元素均为 0. 因为组特征传递的是兴趣点基于视觉特征相似性的分类信息,存在多个兴趣点的组特征相同的情况,所以编码时也不额外增加聚类中心与其最相似特征的计算. 即第 i 个聚类中心 G_i 的编码向量表示为

$$VLAD_{G_i}=\frac{1}{N_i}\sum_{j=1}^{N_i}(g_j-G_i) \tag{6}$$

其中 G_i 表示依据位置特征聚类策略得到的对应的第 i 个组特征聚类中心,N_i 表示第 i 个聚类中心 G_i 所包含的组特征数,g_j 表示聚类中心 G_i 包含的第 j 个组特征. 将组特征向量和残差特征向量串联起来,即为位置向量. 则视频基于位置特征划分的组特征向量表示为

$$VLAD_g=[VLAD_{G_1},\cdots,VLAD_{G_{k_1}}] \tag{7}$$

位置向量表示为

$$VLAD_l=[VLAD_r,VLAD_g] \tag{8}$$

将编码得到的视觉向量和位置向量串联得到视频动作的表征向量 IVLAD(Improved Vector of Locally Aggregated Descriptors)

$$IVLAD=[VLAD_v,VLAD_l] \tag{9}$$

2.3　向量归一化

视频表征的好坏直接影响识别的结果,如何构造有效的视频表征一直是行为识别研究领域的热点课题. 对表征向量进行归一化处理能减小不同特征维度间幅值差别的影响,有效提升向量的表达能力,从而提高识别准确度. 行为识别中常用的归一化方式有 L1 正则、L2 正则和 Power Noemalization.

假定 $IVLAD=[x_1,\cdots,x_n]$,则经过 L1 归一化后的向量为

$$IVLAD_{L1}=\left[\frac{x_1}{|x_1|+\cdots+|x_n|},\cdots,\frac{x_n}{|x_1|+\cdots+|x_n|}\right] \tag{10}$$

经过 L2 归一化后的向量为

$$IVLAD_{L2}=\left[\frac{x_1}{\sqrt{x_1^2+\cdots+x_n^2}},\cdots,\frac{x_n}{\sqrt{x_1^2+\cdots+x_n^2}}\right] \tag{11}$$

经过 Power Normalization 后的向量表示为

$$IVLAD_{\text{Power}}=\text{sign}(IVLAD).*(abs(IVLAD).\hat{}\,\alpha) \tag{12}$$

其中 α 为规范参数,且满足条件 $0\leqslant\alpha\leqslant1$. Peng 等人[15]通过实验观察得到 L2 正则比 L1 正则更适合线性 SVM,且 Peng[15]指出 Power Normalization 能平滑直方图特征,使得具有代表性的少量特征得到有效表达.

所以本文提出在对 2.2 节得到的 IVLAD 向量进行归一化处理时,选择 L2 正则和 Power Normalization 组合处理方式. 即先对 IVLAD 进行 L2 归一化,再对归一化后的向量采用 Power Normalization 策略二次归一化,并将二次归一化的向量与其自身叠加串联作为视频的最终表示向量,以期通过提升表示向量维度达到增强向量表示能力的效果. 然后将最终表示向量输入到线性 SVM 中进行分类预测. 后续实验还进一步分析了不同归一化策略组合对识别性能的影响.

3　实验与分析

为了验证算法的有效性和鲁棒性,本文采用了目前行为识别算法研究中最常用的两个数据集:UCF101 数据库和 HMDB51 数据库. 在这两个数据集上比较分析了本文算法的分类性能,并进一步分析了不同词典大小对于算法性能的影响.

3.1　实验数据

UCF101[31]是目前用于行为识别的最大的数据库,包含 101 个动作类共计 13320 个视频片段,视频总时长超过 27 个小时,其视频数据大致可以分为五类:人物交互、身体行为、人人交互、演奏乐器和其他运动. 该数据集的每个动作类又分为 25 组,每组包含 4 ~ 7 个小片段,同一组的小片段视频有一些共同的特征,比如背景或者行为人是一样的. 每个视频片段的帧率和分辨率分别为 25FPS 和 320 × 240. 此数据库的一些动作示例图如图 6 所示.

图6　UCF101数据库视频截图

HMDB51[32]数据库有 51 个动作类共计 6766 个视频片段,每个动作类都具有超过 100 个视频片段. 该数据内的所有视频片段均从真实世界场景中搜集,比如电影或者 YouTube 视频库. 该数据库的类内差异比较大,比如视觉、尺度、背景、光照等. 因此在该数据库上的行为识别难度比较大. 此数据库的一些动作示例图如图 7 所示.

3.2　实验设置

UCF101 数据库上的训练和测试数据设置:参照数

图7 HMDB51数据库视频截图

据库 UCF101[31]的推荐设置,分三次对数据库进行识别,取每次的识别准确度的均值作为本算法的识别准确度.其中第一次是取 101 个类中每个动作类的第 8 ~ 25 组为训练组,第 1 ~ 7 组为测试组;第二次是取 101 个类中每个动作类的 1 ~ 7 组和 15 ~ 25 组为训练组,8 ~ 14 组为测试组;第三次是取 101 个类中每个动作类的 1 ~ 14 组和 22 ~ 25 组为训练组,15 ~ 21 组为测试组.

HMDB51 数据库上的训练和测试数据设置:参照数据库 HMDB51[32]的推荐设置,分三次对数据库进行识别,取三次的识别准确度的均值为本算法的识别准确度.每次识别实验时,都需测试每类动作的识别准确度(其中每个动作类中有 70 个视频片段用于训练,30 个视频片段用于测试).然后计算所有类的识别准确度的均值作为该次实验的识别准确度.

选取了两种常用特征 STIP 和 IDT 作为视频的特征表示.基于 IDT 特征的位置词典和视觉词典大小设置为 1000,位置词典大小设置为 50;基于 STIP 特征的视觉词典大小设置为 8000,位置词典大小为 400.编码方式都采用了本文所提出的改进 VLAD 编码方法,在编码后对表示向量进行了 L2 归一化和 Power Normalization,其中 Power Normalization 的规范参数 α 设置为 0.5;并将归一化后的向量与其自身串联作为视频的向量表示.

3.3 实验结果及分析

3.3.1 与基于传统手动特征方法的分类准确度比较

为了验证本文提出的算法具有良好的分类性能,本文在 UCF101 数据库和 HMDB51 数据库上分别和同类优秀算法[13~15,17]的实验结果进行了对比.

在 UCF101[31]数据库上,分析比较了基于两种不同特征(STIP 和 IDT)本文算法与当前识别准确率较高的一些方法[13~15,17]的识别准确度,结果如表 1 所示.

表 1 UCF101 数据库上识别准确度对比

特征	算法				
	SSCV[13]	ST-VLAD[17]	BOVW[15]	SFV[14]	本文算法
STIP	75.96	–	84.13	85.63	87.18
IDT	82.31	83.0	87.94	88.35	90.60

参与比较的方法 SSCV[13]提出对描述符和位置信息进行稀疏编码分别构造描述符稀疏向量 SDV(Super Discriptors Vector)和位置稀疏向量 SLV(Super Location Vector),然后将 SDV 和 SLV 组合得到 SSCV(Super Sparse Coding Vector)作为视频的表示向量进行分类预测.它基于两种特征的识别率分别为 75.96% 和 82.31%,分别比本文提出的算法低了 11.22% 和 8.29%;虽然其提出的算法也用到了位置信息作为视频表示的补充,但是其采用的是稀疏编码,本文提出的改进 VLAD 编码方法得到了更好的实验结果.

文献[17]虽然也提出了利用位置信息作为视频的补充表示,但因为其编码方法弱化了聚类中心的作用,且其在向量归一化阶段只采用了 L2 正则方法,其基于 IDT 特征的识别准确度在 UCF101 上比本文低了 7.6%.

文献[15]提出一种基于视觉词袋 BOVW(Bag Of Visual Words)的算法,它是通过对特征数据进行降维及白化操作来提高识别能力.该算法基于两种特征的识别率分别为 84.13% 和 87.94%,分别比本文提出的算法低了 3.05% 和 2.66%;虽然该文也采用了 VLAD 编码方法,且在编码之前对数据进行了降维白化处理,使得特征更加独立且维度更低.但是由于本文采用了位置信息作为视频的补充表示,且在编码过程中强调了聚类中心的作用,得到了更好的识别结果.

文献[14]提出对特征进行二次卷积费舍尔编码,并将卷积向量与原始费舍尔向量进行组合.该算法命名为 SFV(Stacked Fisher Vector).SFV 基于两种特征的识别率分别为 85.63% 和 88.35%,分别比本文提出的算法低了 1.55% 和 2.25%.相较 VLAD 编码,虽然费舍尔编码统计的信息更加高阶,但 SFV 算法的识别准确度仍然低于本文提出的改进 VLAD 编码方法的结果.说明相较于费舍尔编码,本文提出的 IVLAD 编码所统计的高阶信息,及补充的位置信息能更好表征视频中的动作.

本文也在 HMDB51 数据库上分析比较了基于两种不同特征(STIP 和 IDT)情况下本文算法与当前识别准确率较高的一些方法[13~15,17]的识别准确度,实验结果如表 2 所示.

表 2　HMDB51 数据库上识别准确度对比

特征	算法				
	SSCV[13]	ST-VLAD[17]	BOVW[15]	SFV[14]	本文算法
STIP	37.4	–	38.82	39.35	40.24
IDT	47.90	59.0	60.22	66.79	69.17

SSCV[13]基于特征 STIP 和特征 IDT 的识别准确度分别为 37.4% 和 47.90,分别比本文提出算法低了 2.84% 和 21.27%;ST-VLAD[17]基于特征 IDT 的识别准确度比本文低了 10.17%;BOVW[15]基于特征 STIP 和特征 IDT 的识别准确度分别为 38.82% 和 60.22%,分别比本文提出算法低了 1.42% 和 8.95%;SFV[14]基于特征 STIP 和特征 IDT 的识别准确度分别为 39.35% 和 66.79%,分别比本文所提算法低了 0.89% 和 2.38%. 表 2 结果表明在 HMDB51 数据库上本文所提出算法优于所较方法.

3.3.2　与 CNN 方法的分类准确度比较

为了验证本文算法有较好的识别性能,本文还与近期比较主流的一些 CNN 方法的识别结果进行了比较,结果如表 3 所示.

表 3　UCF101 和 HMDB51 数据集上本文方法与 CNN 方法比较

方法	UCF101		HMDB51	
	准确度	特征	准确度	特征
本文方法	90.6	IDT	69.17	IDT
双流 CNN[7]	88.0%	双流 CNN 特征	59.4%	双流 CNN 特征
三流 CNN[8]	92.1%	三流 CNN 特征	67.2%	三流 CNN 特征

由表 3 实验结果可以看出,在两个数据集上本文所提方法的识别准确度均比双流 CNN 模型的识别准确度高,在 UCF101 数据集上本文所提方法比三流 CNN 模型的识别准确度要低,在 HMDB51 数据集上的识别准确度要高于三流 CNN 模型. 说明对于更为复杂的视频,依据轨迹稠密采样的手动提取特征表示能力比基于三流 CNN 模型学习到的特征表示能力更强. 但随着深度学习的发展,基于深度网络自动学习特征的方法的表现已开始超越传统方法的表现. 目前已有研究者通过构建四流 CNN 模型并将网络层数加深至 101 层[10]得到了高于手动提取特征表示方法的识别准确度,在 UCF101 和 HMDB51 数据集上分别获得了 96.0% 和 74.9% 的识别准确度.

3.3.3　视觉词典及位置词典大小对算法性能影响分析

在研究不同词典大小对算法识别性能影响时,实验特征选用了计算量相对小一些的 STIP 特征,数据集则选用了 UCF101 数据集,并分别将视觉词典大小设置为 500、1000、1500,将位置词典大小设置为 25、38、44、50、75、100、200;并将视觉词典大小参数和位置词典大小参数依次组合,评估其对识别算法性能的影响. 实验的主要内容包括:

方式 1　固定视觉词典大小,比较不同位置词典大小对算法分类性能的影响.

方式 2　固定位置词典大小,比较不同视觉词典大小对算法分类性能的影响.

实验结果如图 8 所示.

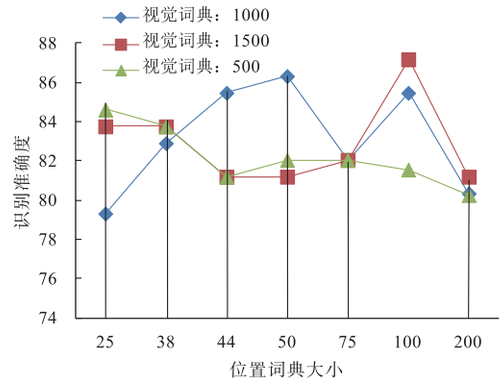

图8　不同词典大小识别准确度对比

当固定视觉词典大小为 1000 时,位置词典大小为 25 时的识别准确度为 79.32%、随着位置词典增大,识别准确度呈上升趋势. 识别效果在位置词典为 50 时达到最好,识别准确度为 86.32%. 当位置词典继续增大时,识别准确度反而有所降低. 说明当固定视觉词典大小时,识别准确度并不随位置词典的增大而提高. 相反当位置词典大小超过一定点之后,识别准确度可能会随着词典的增大而减小. 这可能是因为位置词典过大,会造成同一动作幅度大小不同时表征的差别大,影响了视觉特征的准确表达,从而降低了识别准确度.

当固定位置词典大小为 25 时,视觉词典大小为 500、1000 和 1500 时的识别准确度分别为 84.6154%、79.3158% 和 83.7607%. 视觉词典越大,识别准确度反而降低了,说明视觉词典较大时削弱了位置特征的表达能力,影响了识别的表现.

从本实验结果分析可知,视觉词典与位置词典存在一个最佳比例使得算法性能最佳. 为了兼顾效率,适当的选取视觉词典与位置词典的大小比例能得到很好的识别效果.

3.3.4　不同归一化方法对算法的影响

本实验分析了在 UCF101 数据库上对 STIP 特征采取不同归一化方法对算法性能的影响,实验的主要内容包括:

方式 1　对特征集、词典、单个样本特征及编码向量均做 L2 正则和 Power Normalization 处理.

方式 2　对词典、单个样本特征及编码向量做 L2

正则和 Power Normalization 处理.

方式3 只对编码向量做 L2 正则和 Power Normalization 处理.

方式4 只对编码向量做 L2 正则.

实验结果如图 9 所示.

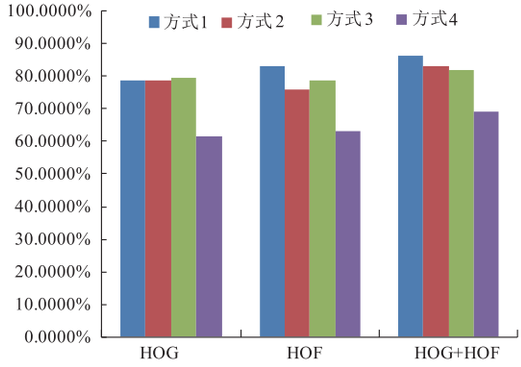

图9 不同归一化方法对算法性能的影响

从图 9 的实验结果可以看出,对基于 HOG、HOF 及 HOG + HOF 三个特征组合来说,方式 4(只对编码向量做 L2 正则)的识别准确度都是最差的. 识别准确度分别为 61.5385%、63.2479% 和 69.2308%,比各自最好的识别准确度分别差了 17.9487%、19.6581% 和 17.6940%. 因为方式 4 只采用了 L2 正则策略,而其他三种方式采用了 L2 正则和 Power Normalization 的组合策略. 对 HOF 和 HOG + HOF 特征组合来说,识别效果最好的是方式 1(对特征集、词典、单个样本特征及编码向量均做 L2 正则和 Power Normalization 处理),识别准确度分别达到了 82.9060% 和 86.3248%. 但对 HOG 特征来说,方式 3(只对编码向量做 L2 正则和 Power Normalization 处理)效果最好,识别准确度为 79.4872. 虽然三个特征对象达到识别准确度最好的归一化方式不同,但方式 1、方式 2 和方式 3 在三个特征对象上的识别准确度相差不大. 因为方式 1、方式 2 和方式 3 都采用了 L2 正则和 Power Normalization 的组合策略,仅在归一化对象上有所区别.

以上结果表明归一化处理的阶段对识别效果的影响不显著,归一化处理选择的策略对识别准确度的影响显著,L2 正则和 Power Normalization 的合并处理效果最佳.

3.3.5 向量串联升维前后的识别准确度对比

本实验分析了向量串联升维对识别准确度的影响,比较了参考文献及本文方法的视频表示向量串联升维前后的识别准确度. 实验结果如图 10 所示.

向量 SDV[13] 串联前后的识别准确度分别为 80.9211% 和 81.25%,串联后识别准确度提升了 0.3289%. 向量 SLV[13] 串联前后的识别准确度分别为

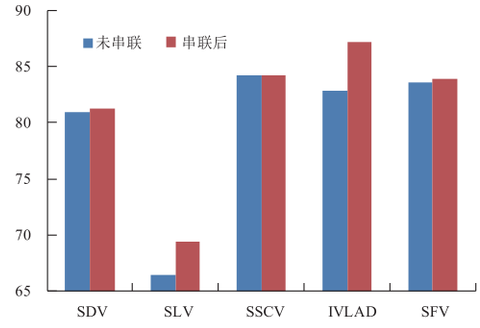

图10 不同向量串联前后识别准确度

66.4474% 和 69.4079%,串联后识别准确度提升了 2.9605%. 向量 SSCV[13] 串联前后的识别准确度分别为 84.2105% 和 84.2398%,串联后识别准确度提升了 0.0293%. 向量 IVLAD 串联前后的识别准确度分别为 82.906% 和 87.1795%,串联后识别准确度提升了 4.2735%. 向量 SFV[14] 串联前后的识别准确度分别为 83.6316% 和 83.9605%,串联后识别准确度提升了 0.3289%.

结果表明对于参考文献[13,14]中所提方法的表示向量及本文所提算法表示向量,将向量与其自身串联升维有助于提升识别效果. 在实验所用的数据集上准确度平均提高 1.99%. 但是也要注意,虽然对向量串联升维能提升识别准确度,但是模型训练时间以及预测分类时间也会因为计算量变大而有所增加. 例如在 MATLAB2016b 上用支持向量机对 UCF101 数据集的 8 个动作类 1079 个视频表示向量(其中 775 个样本用于训练,304 个样本用于测试,向量维度为 48996)进行识别时,向量未串联时,训练模型耗时 25.53s,分类耗时 4.35s;对向量串联升维后训练模型耗时 26.99s,分类耗时 4.57s. 向量串联前后训练及分类所需时间分别增长了 1.46s 和 0.22s. 随着数据样本的增多以及表示向量的维度增加,串联后训练分类耗时会增加的更多. 研究者可根据具体识别任务考虑是否对向量串联升维.

图 11 所示是在不同词典大小情况下,本文所提出的视频表征向量 IVLAD 串联前后对动作识别性能影响的实验比较结果.

由图 11 可以看出,所有实验词典大小下的 IVLAD 向量与其自身串联后的识别准确度都较未串联之前有所提高,且不同词典大小下的 IVLAD 与自身串联后所提升的识别准确度不一样. 当视觉词典为 1500 时,在一定范围内(位置词典小于或等于 75 时),IVLAD 与其自身串联后提升的识别准确度随着位置词典的增大而增大,但当位置词典超过一定范围时(位置词典大于 75 时),IVLAD 与其自身串联后提升的识别准确度随着位置词典的增大而减小了. 这也从侧面证明了位置特征

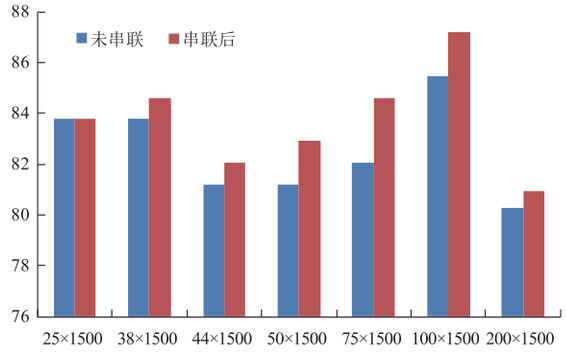

图11　不同词典大小的IVLAD向量串联前后识别准确度

对视觉特征的补充作用有助于提升视频的识别准确度,但也要注意位置词典不宜设置过大,从而误把同一动作判定为不同动作.

综上所述,对表示向量进行串联升维操作有助于获取更丰富的信息,提高视频表征能力进而提升识别准确度.

4　结论

本文提出了一种基于 IVLAD 编码向量的行为表示算法. 该算法将位置信息作为视觉特征的补充联合表征视频,并在编码方式上对传统 VLAD 编码提出了改进. 通过增加聚类中心与其最相似特征的差值计算增强聚类中心的影响,从而突出有效特征的表达. 最后还提出对表征向量进行自身串联升维,获取更高的识别准确度.

在实验中验证了本文算法的有效性,并分析了视觉词典和位置词典大小对算法性能的影响. 发现位置词典不宜过大,避免影响视觉特征的表达.

在实验中还分析了不同归一化方式对算法识别结果的影响,发现归一化处理的策略比归一化处理的阶段对识别结果的影响要大,且 L2 正则和 Power Normalization 的组合归一化的识别结果比对编码向量做单一 L2 正则处理的识别效果更好. 下一步研究将探索不同的向量融合方式及归一化策略的效果.

参考文献

[1] ZHU F, SHAO L, XIE J. From handcrafted to learned representations for human action recognition[J]. Image and Vision Computing, 2016, 55(2): 42 – 52.

[2] 杜友田,陈峰,徐文立,李永彬. 基于视觉的人的运动识别综述[J]. 电子学报,2007,35(1):84 – 90.
DU You-tian, CHEN Feng, XU Wen-li, LI Yong-bin. A survey on the vision-based human motion recognition[J]. Acta Electronica Sinica, 2007, 35(1): 84 – 90. (in Chinese)

[3] BOYER E. A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition[M]. Elsevier Science Inc, 2011.

[4] 苏松志,李绍滋,陈淑媛,蔡国榕,吴云东. 行人检测技术综述[J]. 电子学报,2012,40(4):814 – 820.
SU Song-zhi, LI Shao-zi, CHEN Shu-yuan, CAI Guo-rong, WU Yun-dong. A survey on pedestrian detection[J]. Acta Electronica Sinica, 2012, 40(4): 814 – 820. (in Chinese)

[5] DAWN D D, SHAIKH S H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector[J]. Visual Computer, 2016, 32(3): 289 – 306.

[6] 田国会,尹建芹,闫云章,李国栋. 基于混合高斯模型和主成分分析的轨迹分析行为识别方法[J]. 电子学报,2016,44(1):143 – 149.
TIAN Guo-hui, YIN Jian-qin, YAN Yun-zhang, LI Guo-dong. Gaussian mixture models and principal component analysis based human trajectory behavior recognition[J]. Acta Electronica Sinica, 2016, 44(1): 143 – 149. (in Chinese)

[7] SIMONYAN K, ZISSERMAN A. Two-Stream convolutional networks for action recognition in videos[A]. Proceedings of International Conference on Neural Information Processing Systems[C]. USA: MIT Press, 2014. 568 – 576.

[8] WANG L, GE L, LI R, et al. Three-stream CNNs for action recognition[J]. Pattern Recognition Letters, 2017, 92(C): 33 – 40.

[9] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[10] BILEN H, FERNANDO B, GAVVES E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, PP(99): 1 – 1.

[11] GKIOXARI G, GIRSHICK R, MALIK J. Contextual action recognition with R * CNN[A]. Proceedings of International Conference on Computer Vision[C]. USA: IEEE, 2015. 1080 – 1088.

[12] CHERON G, LAPTEV I, SCHMID C, et al. P-CNN: Pose-based CNN features for action recognition[J]. Proceedings of International Conference on Computer Vision[C]. USA: IEEE, 2015. 3218 – 3226.

[13] YANG X, TIAN Y L. Action recognition using super sparse coding vector with spatio-temporal awareness [A]. Recognizing Complex Events in Videos by Learning Key Static-Dynamic Evidences[C]. Berlin: Springer, 2014. 727 – 741.

[14] PENG X, ZOU C, QIAO Y, et al. Action recognition with stacked fisher vectors[A]. European Conference on Com-

puter Vision[C]. Berlin:Springer,2014. 581 – 595.
[15] PENG X,WANG L,WANG X,et al. Bag of visual words and fusion methods for action recognition:Comprehensive study and good practice[J]. Computer Vision & Image Understanding,2016,150(C):109 – 125.
[16] ARANDJELOVIC R,ZISSERMAN A. All about VLAD[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2013. 1578 – 1585.
[17] DUTA I C,IONESCU B,AIZAWA K,et al. Spatio-Temporal VLAD Encoding for Human Action Recognition in Videos[M]. USA:MultiMedia Modeling,2017.
[18] WANG H,ULLAH M M,KLÄSER A,et al. Evaluation of local spatio-temporal features for action recognition[A]. Proceedings of British Machine Vision Conference(BMVC2009)[C]. London,UK:DBLP,2009. DOI:10.5244/C. 23. 124.
[19] LAPTEV I. On space-time interest points[A]. Proceedings of IEEE International Conference on Computer Vision[C]. USA:IEEE,2005. 107 – 123.
[20] WANG H,KLASER A,SCHMID C,et al. Action recognition by dense trajectories[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2011. 3169 – 3176.
[21] WANG H,SCHMID C. Action recognition with improved trajectories[A]. Proceedings of IEEE International Conference on Computer Vision[C]. USA:IEEE,2014. 3551 – 3558.
[22] UIJLINGS J,DUTA I C,SANGINETO E,et al. Video classification with Densely extracted HOG/HOF/MBH features:an evaluation of the accuracy/computational efficiency trade-off[J]. International Journal of Multimedia Information Retrieval,2015,4(1):33 – 44.
[23] LIU L,WANG L,LIU X. In defense of soft-assignment coding[A]. Proceedings of IEEE International Conference on Computer Vision[C]. USA:IEEE,2011. 2486 – 2493.
[24] YANG J,YU K,GONG Y,et al. Linear spatial pyramid matching using sparse coding for image classification[A]. computer Vision and Pattern Recognition[C]. USA:IEEE,2009. 1794 – 1801.
[25] WANG J,YANG J,YU K,et al. Locality-constrained linear coding for image classification[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2010. 3360 – 3367.
[26] PERRONNIN F,SÁNCHEZ J,MENSINK T. Improving the fisher kernel for large-scale image classification[A]. European Conference on Computer Vision[C]. Berlin:Springer-Verlag,2010. 143 – 156.
[27] ZHOU X,YU K,ZHANG T,et al. Image classification using super-vector coding of local image descriptors[A]. European Conference on Computer Vision[C]. Berlin:Springer,2010,6315:141 – 154.
[28] JEGOU H,DOUZE M,SCHMID C,et al. Aggregating local descriptors into a compact image representation[A]. Computer Vision and Pattern Recognition[C]. USA:IEEE,2010. 3304 – 3311.
[29] SOMASUNDARAM G,CHERIAN A,MORELLAS V,et al. Action recognition using global spatio-temporal features derived from sparse representations[J]. Computer Vision & Image Understanding,2014,123(7):1 – 13.
[30] ZHANG B,WANG H. Encoding scale into fisher vector for human action recognition[A]. Visual Communications and Image Processing[C]. USA:IEEE,2016. 1 – 4.
[31] SOOMRO K,ZAMIR A R,SHAH M. UCF101:A Dataset of 101 Human Actions Classes From Videos in the Wild[DB/OL]. http://crcv. ucf. edu/data/UCF101. php.
[32] KUEHNE H,JHUANG H,GARROTE E,et al. HMDB:A large video database for human motion recognition[A]. Proceedings of International Conference on Computer Vision[C]. USA:IEEE,2011. 2556 – 2563.

作者简介

罗会兰 女. 1974 年 9 月生,江西上高人. 2008 年浙江大学获工学博士学位. 现为江西理工大学图像处理实验室教授、硕士生导师. 主要从事机器学习、模式识别等方面的研究.
E-mail:luohuilan@ sina. com

王婵娟 女. 1992 年 5 月生,江西鄱阳人. 2015 年进入江西理工大学,在读硕士研究生. 主要从事计算机视觉、机器学习技术方面的有关研究.
E-mail:909748120@ qq. com