DOI:10.13203/j.whugis20140581 文章编号:1671-8860(2016)04-0468-07

多层独立子空间分析时空特征的人体行为识别方法

瞿 涛[1] 邓德祥[1] 刘 慧[2] 邹 炼[1] 刘弋锋[1]

1 武汉大学电子信息学院,湖北 武汉,430072
2 厦门大学通信工程系,福建 厦门,361005

摘 要:人体行为识别在视频监控、医疗诊断等领域都有重要的意义。目前人体识别的主要方法是将人为设计的二维特征扩展到三维空间,或利用运动轨迹,提取出时空特征。基于深度学习的思想,直接在三维空间中构建多层神经网络,从大量的视频数据中学习不同行为的时空特征。首先,采用独立子空间分析(independent subspace analysis, ISA)方法,构造两层卷积叠加神经网络,从训练视频中学习网络权重。然后,对特征使用K-means聚类,转化为视觉单词,根据视觉单词频率直方图计算支持向量机模型(support vector machine, SVM)判决超平面,最后对待分析视频进行动作分类。使用该方法对Hollywood2数据库的12种行为进行实验,结果表明,ISA学习到的特征权重与Gabor滤波器类似,对图像频率和方向具有明显的选择性,对相位变化具有鲁棒性,能够显著提高认为识别的正确率,符合人眼的视觉特征。

关键词:卷积叠加;独立子空间分析;多层网络;无监督学习;深度学习;人体行为识别
中图法分类号:P208;TP391 **文献标志码**:A

人体行为识别一直是计算机视觉的研究热点。在视频监控、运动分析、视频检索、体育赛事分析、医疗诊断等领域,行为识别都具有广泛的应用前景和巨大的经济价值[1-2]。

目前行为识别采用的方法主要有两类,基于时空特征提取和基于运动轨迹分析。基于时空特征提取的方法将二维图像特征扩展到三维空间,原有的二维特征是针对某些具体应用设计的,例如尺度不变特征转换(scale invariant feature transform, SIFT)和加速鲁棒特征(speeded-up robust features, SURF)[3]常用作基于角点特征的二维图像拼接,哈伯(Gabor)梯度方向直方图(histograms of oriented gradients, HOG)、Gabor和局部二值模式(local binary pattern, LBP)主要用于提取人的面部特征。在三维空间,这些方法扩展为Extended SURF、Hessian和HOG3D等,但是上述方法不是为三维人体行为识别设计的,扩展性较差。对运动轨迹进行特征描述的方法中,Jain等人提取DCS特征预测运动残差进行分类;Theusner等描述姿态空间的运动状态实现运动检测[4],该类方法的提取效果同样受人为设计特征的影响,通用性差。实验表明,没有一个通用的手工提取特征的方法能够运用于所有数据库,如果从视频本身直接学习特征更加有效。

近年来,计算机运算能力不断提高,为更复杂的处理算法提供了硬件基础。同时,深度学习在目标识别方面表现优异,已成为当前研究热点,例如去噪自编码器(denoising auto-encoder)[5]、卷积神经网络[6],以及神经自回归分布估计器(neural autoregeressive distribution estimator, NADE)等[7]。这些方法直接从数据本身学习特征,因此通用性更好,在图像识别领域获得了广泛的应用。但这些方法目标方程比较复杂,往往需要2~3 d才能完成训练。

本文使用了一种完全无监督的方法,从大量的视频数据中直接学习不同行为的时空特征。其核心是使用ISA算法,构造两层卷积叠加神经网络,学习样本的深层不变性特征。实验证明,ISA提取的非线性特征,对特定频率和方向的样本具有明显的选择性,对于相位变化,也具备传统稀疏编码所不具备的鲁棒性。同时,ISA的特征具有自我学习性,扩展能力也比传统手工提取特征的方法好。本文所使用的卷积叠加方法只需要进行矩阵的矢量积和卷积操作,对高维样本的训练速

收稿日期:2015-02-02
项目资助:国家自然科学基金(61072135)。
第一作者:瞿涛,博士,主要从事图像处理、深度学习、目标跟踪与识别研究。aboutyoucsm@126.com
通讯作者:邹炼,博士,副教授。zoulian@whu.edu.cn

度比其他深度神经网络方法快很多。最后,本网络从海量数据中提取特征,也使基于大数据的自动特征提取方法成为可能[2]。

1　多层独立子空间特征提取

本文对局部视频特征提取并分类,主要步骤为局部时空样本提取、特征学习和动作分类。通过密集采样得到视频时空样本后,使用两层 ISA 学习特征权重,第一层对小图像块进行训练,第二层对多个小图像块特征卷积叠加后再次训练,得到更大图像样本的特征。然后使用 K-means 方法进行特征聚类,建立视觉词汇表及视觉单词(聚类中心)。统计每个训练视频各种单词出现的频率直方图,将其作为 SVM 分类器的输入,并得到每类动作的判决平面。测试样本以同样的方法得到特征后,根据已建立的词汇表,计算频率分布直方图,最终送入 SVM 分类,判定所属的动作类别。整个流程如图 1 所示。

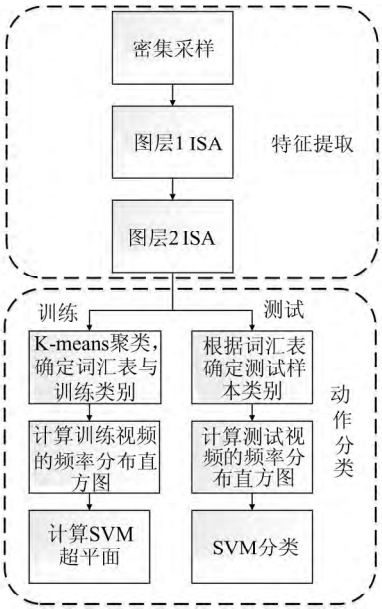

图 1　算法流程图

Fig. 1　Flowchart of Algorithm

1.1　局部时空样本提取

三维时空样本提取的方法有 Harris3D、Cuboid、Hessian 和密集采样(dense sampling)等。其中,密集采样能得到大量的视频时空样本,处理效果优于其他方法。但密集采样得到的样本数量是其他方法的 15～20 倍,对后续处理提出了更高的要求。

密集采样先在时空域进行视频缩放,然后从随机位置提取视频块,得到最终样本。每个样本包含 5 个维度的属性,即(x, y, t, σ, τ)。其中(x, y)是空域位置,t 是时域位置,σ 是时域缩放尺度,τ 是空域缩放尺度。对于每一个样本,σ、τ 为确定值,计算以(x, y, t)为中心的 3D 视频块,得到视频基本样本,然后使用主成分分析(principal component analysis,PCA)将特征向量投影到低维空间,减小计算量[8]。

1.2　基于时空样本的 ISA

ISA 是一种无监督学习方法,具有两层网络的生成模型,可以有效模拟人类视觉系统 $V1$ 区简单细胞(simple cells)与复杂细胞(complex cells)感受野的层次化响应模式。网络结构如图 2 所示,模型第一层学习线性变换的权重 $\boldsymbol{W}$,第二层将同一子空间元素合并,执行固定的非线性变换 V,得到对相位变化响应不变的特征。

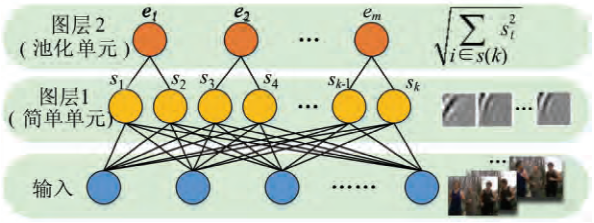

图 2　ISA 网络结构

Fig. 2　Architecture of ISA Network

对每个输入样本 Z,ISA 首先进行线性变换:

$$s_i = \sum_{i=1}^{n} W_{ki} z_i^t \tag{1}$$

式中,z_i^t 是 PCA 降维后均值为 0 的白化数据,t 取值从 1 到 T,协方差矩阵为单位矩阵;$\boldsymbol{W} \in \boldsymbol{R}^{k \times n}$,是输入数据与第一隐层简单单元(simple units)之间的权重;s_i 是对应的线性特征响应,类似于视神经中的简单细胞输出。然而,传统的独立成分分析方法(independent component analysis,ICA)、稀疏编码等线性系统,无法模拟复杂细胞的相位不变性特点。因此,ISA 第二层对上一层特征进行非线性变换:

$$e_k = \sqrt{\sum_{i \in S(k)} s_i^2} \tag{2}$$

式中,e_k 是每个子空间 $S(k)$ 的特征输出,其计算包含累加求和操作,因此该过程也称作合并(pooling);k 为子空间包含的特征个数,$k>1$;ISA 中每个子空间 $S(k)$ 由 k 个特征组合而成,其与线性代数子空间的概念非常接近。特征之间的组合的意义在于,能够增强同一图像块在位置(相位)变化时响应的一致性,得到相位不变性特征。

ISA 的关键为计算第一层的权值 $\boldsymbol{W}$,从而最终得到输出 e_k。其满足如下约束条件:

$$p_i(z^t;\boldsymbol{W},\boldsymbol{V})=\sqrt{\sum_{k=1}^{m}\boldsymbol{V}_{ik}\left(\sum_{j=1}^{n}\boldsymbol{W}_{kj}z_j^t\right)^2} \tag{3}$$

$$\min_{\boldsymbol{W}}\sum_{t=1}^{T}\sum_{i=1}^{m}p_i(z^t;\boldsymbol{W},\boldsymbol{V}) \tag{4}$$

使得 $\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}=\boldsymbol{I}$

式中，$\boldsymbol{V}\in\boldsymbol{R}^{m\times k}$，是简单单元与合并单元(pooling units)的固定权重；n 是输入维数；k 是简单单元的个数；m 是合并单元的个数。正交约束(orthonormal constraint)$\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}=\boldsymbol{I}$ 保证了特征互异性，即特征之间不相关。

权重 w_i 类似于一组参数相似的 Gabor 滤波器，同一子空间内的特征方向、频率相似，相位不同；表现为特征之间的形状基本相似，但是相互有微小的位移。每个子空间的特征对方向和频率变化更敏感，比相位变化更具鲁棒性。相位不变性是简单细胞与复杂细胞的一个重要区别，也使 ISA 比一般的 ICA 或者稀疏编码具有更好的效果[9]。

1.3 两层卷积叠加 ISA 网络

在特征训练时，输入网络的样本尺寸越大，特征的概括能力也就越强。然而，标准 ISA 训练过程在输入数据维数 n 较大时效率不高，这是因为在投影梯度下降时，每一步都会执行正交化：

$$\boldsymbol{W}\leftarrow(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}})-\frac{1}{2}\boldsymbol{W} \tag{5}$$

正交化的计算复杂度为 $O(n^3)$。为了处理高维大规模数据，我们依次使用包含 ISA 和 PCA 的两层卷积神经网络学习方法[10]，将大的图像块拆分为不同的小块进行单独训练，以减少每次 ISA 的计算维数，加快训练速度。整个过程如图 3 所示。

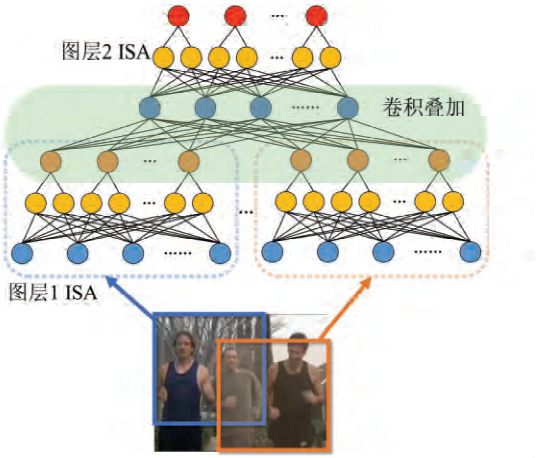

图 3 卷积叠加 ISA

Fig. 3 Stacked Convolutional ISA Network

第一层网络训练时，首先对小的输入图像块进行 PCA 降维，使用 ISA 算法学习权重矩阵，再将其与更大的图像块进行卷积，即把大图像拆分成不同的子块，单独计算子块特征，再将所有特征合并输出。第二层计算时，先用 PCA 对第一层的合并特征进行降维预处理，将结果作为 ISA 输入，计算最终的样本特征。整个过程中 PCA 白化和卷积叠加均减少了计算量，使第二层网络只需要处理低维特征，因此，整个网络对大尺寸的样本同样有较快的训练速度。

训练时，叠加模型与其他深度学习结构所使用的逐层贪婪(greedily layerwise)方法相同。第一层训练直至收敛之后，再训练第二层，同样大大缩小了同时训练的计算量[11]。

2 特征分类与行为判别

当得到所有局部视频样本特征后，首先将其量化为视觉单词(visual words)。量化的第一步为建立词汇表，本文使用 K-means 聚类方法对训练特征进行聚类操作，每个聚类中心对应一个单词。以词汇表为基准，计算每个特征与聚类中心的欧氏距离，将其与距离最近的单词中心归为一类，确定每个样本所属关系。

一个视频包含众多样本，量化后每个样本对应一个视觉单词。那么，每个视频可进一步描述为多种视觉单词的组合。属于不同动作的视频包含的视觉单词的频率分布直方图也必不相同，这为我们后续的动作分类提供了依据。

训练 SVM 时，所有视频样本作为 SVM 分类器的输入[8]，同时输入的还包括视频对应的样本动作标签。本文使用非线性的 SVM(non-linear support vector machine)与 RBF-χ^2 核方法作为动作分类器。对于多分类问题，我们将每种类别划分为属于或不属于该类的二分类问题(one-against-rest)。

得到所有动作种类各自的分类平面后，即可对测试样本进行分类，其输入同样为测试视频的视觉单词频率分布直方图，输出即为视频是否包含分类动作。

3 实验与分析

3.1 实验环境与 Hollywood2 数据库

本文算法验证所使用的软件为 MATLAB 2014a-64bit，电脑配置为 Intel(R) Core(TM) i7-

4770 CPU @ 3.40 GHz,32 GB RAM。

本文使用的行为分类数据库为 Hollywood2,该数据库提供了 12 类人体行为和 10 类场景,共计 3 669 个视频样本,总时长接近 20.1 h。Hollywood2 数据库的视频片段来自 69 部电影,提供了非常有挑战性的一个真实环境下综合测试标准,环境和动作的复杂性决定了其较大的分类难度。数据库包含有干扰噪声的训练样本,完全正确的训练样本,以及人工检验过的测试样本。

我们使用 Hollywood2 数据库中的人体行为库作为测试对象,训练样本为 823 个正确标注的视频。测试样本为 884 个已作好标签的视频片段。所有样本共有 12 个动作,分别为接电话、开车、吃、打架、下车、握手、拥抱、接吻、跑步、坐下、坐起和站起。

3.2　实验参数

本文实验所选的第一层 ISA 输入为 16×16 大小连续 10 帧的图像块,经过 PCA 降维后白化数据为 300 维,ISA 每个子空间大小为 2,即包含两个特征。第二层重新提取大小为 20×20,连续 14 帧,PCA 降至 200 维,第二层 ISA 每个子空间大小为 4。ISA 独自完成权重 W 的训练后,即得到完整的两层神经网络。计算特征时 K-means 聚类中心数量为 5 000,SVM 最终的 12 个动作各自判决超平面。

测试过程与训练过程类似,样本大小为 20×20,连续 14 帧,得到每个视频的单词及其统计分布,比较 SVM 判决结果与正确标签的匹配程度,计算准确率,对于 12 个动作类别,执行完上述操作后,平均准确率即为最终测试结果。

3.3　调谐曲线特征分析与可视化

为了能更好地分析测试效果,将训练得到的第一层 ISA 的权重 W 的子空间进行可视化。

图 4 为两组子空间的特征,每个子空间包含两个特征检测子,对应图中的第 1、2 行,3、4 行,每个特征检测子 W_i 的大小为 16×16,10 帧。可以发现,同一特征检测子内每幅图都类似于一个 Gabor 滤波器,但连续帧之间存在着微小的变化,类似于视频中动作连续变化的过程。对于同一子空间内的两个特征,其图像也非常相似,本质为同一子空间内特征权重 W 具有较高的依赖性(非独立性),而不同子空间的特征独立性较强,相似度也较差。

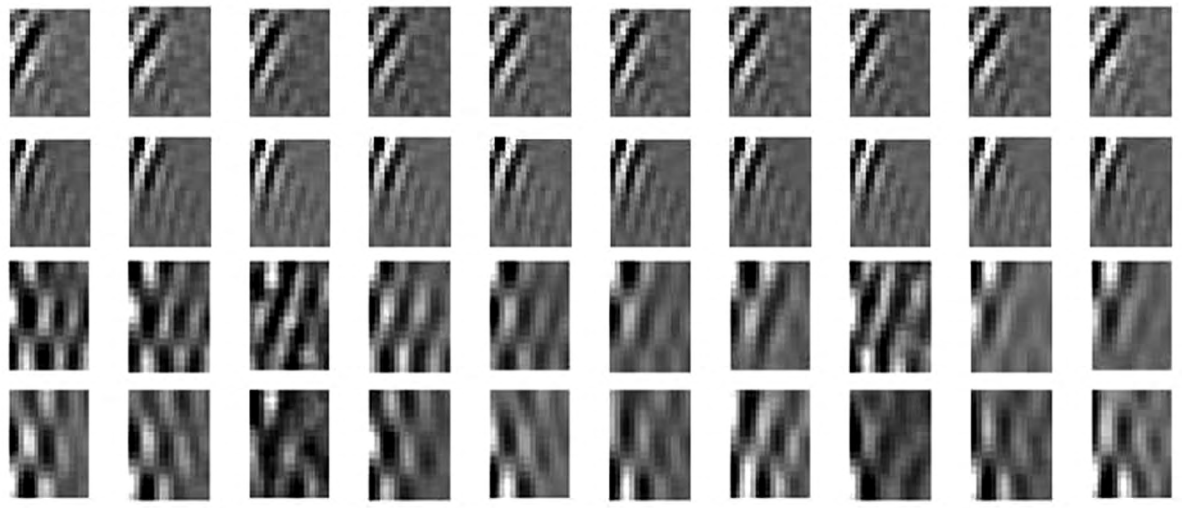

图 4　ISA 特征检测子

Fig. 4　ISA Feature Descriptor

为了进一步分析 ISA 神经网络的性质,使用视神经学(visual neuroscience)所用的调谐曲线(tuning curves)方法,计算输入为正弦光栅时的输出响应,正弦函数为:

$$f(x,y)=\sin(2\pi\alpha(\sin\theta\cdot x+\cos\theta\cdot y)+\beta) \tag{6}$$

式中,θ 是振幅的方向;α 是正弦信号的频率;β 为光栅图相位;$(x,\ y)$ 代表空间坐标。通过改变式(6)中的 α、θ 和 β,计算对应光栅图的特征响应 e_k,其最大值所对应的 α、β 和 θ 即为最佳频率、相位和方向。图 5 为图 4 第一组子空间特征的调谐曲线响应曲线。

调谐曲线中的相位分量,可以认为是对应物体在空间中位置有微小变化时对应的特征响应;其频率分量可以认为图像进行大小缩放时对应的特征响应;其方向分量可以认为是图像进行方向

旋转时对应的特征响应。无论是子空间特征 e_k，还是单独的特征 W_i 响应，其频率和方向曲线都只在某个范围内具有较大响应，这表明特征对频率和方向具有选择性，即对于某些范围内的频率和方向具有较大响应，对于其他值则处于抑制状态。同理，在视频动作边缘的某个缩放倍数和方向下具有较高的响应，体现出特征检测子对视频动作的过滤和选择作用。

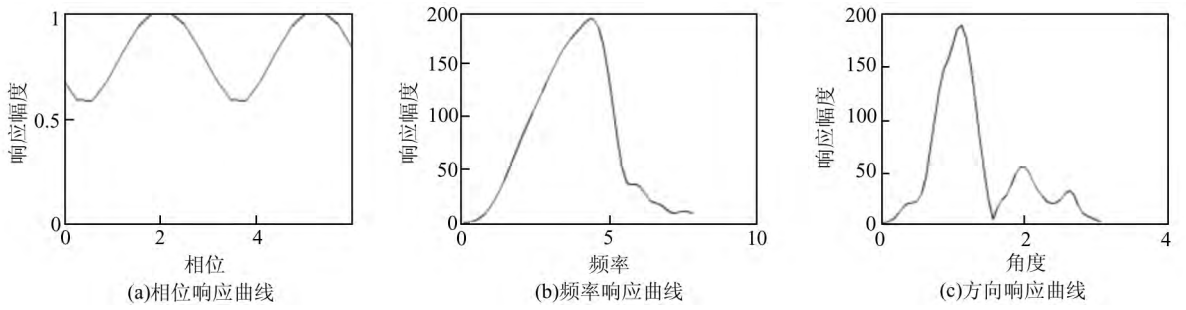

图5 ISA调谐曲线响应曲线

Fig. 5 Tuning Curves of the ISA Features

但是对于相位，子空间检测子 e_k 的选择性则不如频率和方向那样明显，也不如单独线性特征 W_i 的相位响应曲线那样有较大变化。这说明本文的模型初步具备了视神经中复杂细胞的一个关键属性，即相位不变性。对应视频中的行为，表现为同样的物体，其空间内的小位移并不会改变特征输出结果，空间位置变化的鲁棒性较好。

3.4 综合测试结果

高层ISA特征可视化难度较大，其输入为1 200维低层特征，但是我们可以通过比较单层ISA与多层ISA的分类结果部分动作，分析多层神经网络对最终结果的影响。表1中分别对比了单层ICA和双层ICA、单层ISA和双层ISA的分类结果。对于每个动作，计算其正确分类的比例，为了能对分类结果进行综合评判，本文对12个动作分类精度的求取平均值，将其作为最终衡量标准。另外，对于每类动作的最高分类精度，用黑体标出。

表1 测试结果/%

Tab. 1 Experiment Results/%

动作	单层ICA	双层ICA	单层ISA	双层ISA
接电话	21.1	12.7	20.0	**29.2**
开车	83.3	86.0	82.2	**88.1**
吃	52.7	56.9	47.4	**60.9**
打架	56.5	72.8	65.5	**77.2**
下车	24.7	33.0	29.1	**38.8**
握手	17.3	12.1	19.3	**45.2**
拥抱	25.9	26.1	28.2	**43.7**
接吻	49.6	56.9	51.8	**59.6**
跑步	61.7	65.9	63.7	**73.1**
坐下	36.2	46.3	35.3	**57.7**
坐起	12.5	11.3	18.3	**36.1**
站起	41.5	46.1	45.0	**63.3**
平均值	40.3	43.8	42.1	**56.1**

比较发现，对于单层网络，ICA和ISA的平均精度都比较低，分别为40.3%和42.1%。可以理解为单层网络只能表达简单细胞所具有的功能，对于复杂的形状和非线性特征不具有好的表达能力。当测试网络为双层网络时，ICA和ISA平均精度都有明显的提升，但是双层ICA提升不大，因为仅仅将两层ICA简单叠加，其本质仍为线性变换，精度提升主要表现在卷积叠加部分，并不能体现出高层特征。但是对于双层ISA，精度56.1%的提升更为明显，这也表现出高层特征的非线性变换不仅对于特征具有很好的选择性，对于空间位置的变化也具有更好的概括性。同时，在绝大部分的动作中，双层ISA的精度均为最高，也高于Han等[12]和Gilbert等[13]的方法，如图6所示。

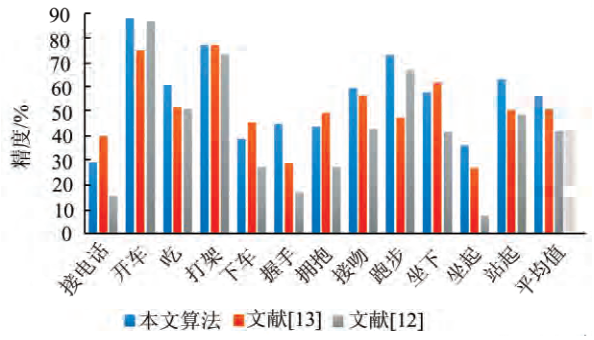

图6 不同算法动作分类准确度结果对比

Fig. 6 Comparison of Different Algorithm

另一方面，不同动作之间的准确度也不相同，原因在于，动作所处的环境背景复杂度，以及该行为本身是否具有较明显的动作变化，均会对分类结果造成影响。例如，接电话的动作分类准确度只有29.2%，这是因为接电话的动作往往是站立或坐

下完成的,这与“开车”“吃”“坐下”“坐起”动作有很大的重合,并且接电话本身姿势比较固定,没有明显的动作变化,对“接电话”这一行为的分类均造成了干扰。同样,对于“开车”“打架”这些动作变化幅度较大的行为,分类精度为 88.1%和 77.2%,均明显高于其他动作。这也说明,神经网络提取的特征,能否对一类动作进行准确抽象,将极大地影响分类结果。

4　结　语

本文使用无监督学习方法,结合多层独立子空间分析与卷积叠加技术,实现多种人体行为的分类。基于 Hollywood2 数据库的实验表明,本方法精度要优于 ICA 和单层 ISA,达到56.1%,为更精确的动作识别奠定了基础。同时,本文直接对原始时空样本进行处理,可进一步利用各类动作自身的运动轨迹,获取更多的特征描述;另外,如何进一步提取更具概括能力的表达,增强特征鲁棒性,也是下一步的研究重点。

参　考　文　献

[1] Liu Yu, Kang Chaogui, Wang Fahui. Towards Big Data-driven Human Mobility Patterns and Models [J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 274-277(**刘瑜, 康朝贵, 王法辉. 大数据驱动的人类移动模式和模型研究**[J]. **武汉大学学报・信息科学版**, 2014, 39(6): 274-277)

[2] Li Deren, Yao Yuan, Shao Zhenfeng. Big Data in Smart City [J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(6): 613-640(**李德仁, 姚远, 邵振峰. 智慧城市中的大数据**[J]. **武汉大学学报・信息科学版**, 2014, 39(6):613-640)

[3] Liu Hui, Li Qingquan, Gao Chunxian, et al. Moving Target Detection Using C_SURF Registration[J]. *Geomatics and Information Science of Wuhan University*, 2014, 39(8): 951-955(**刘慧, 李清泉, 高春仙, 等. 利用** C_SURF **配准的空基视频运动目标检测**[J]. **武汉大学学报・信息科学版**, 2014, 39(8): 951-955)

[4] Jain M, Jégou H, Bouthemy P. Better Exploiting Motion for Better Action Recognition[C]. Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 2013

[5] Yin S, Liu C, Zhang Z, et al. Noisy Training for Deep Neural Networks in Speech Recognition[J]. *Eurasip Journal on Audio Speech&Music Processing*, 2015,1:1-14

[6] Deng L, Abdel-Hamid O, Yu D. A Deep Convolutional Neural Network Using Heterogeneous Pooling for Trading Acoustic Invariance with Phonetic Confusion[C]. Acoustics, Speech and Signal Processing (ICASSP), Vancouver, USA, 2013

[7] Wang Kai, Shu Ning, Li Liang, et al. Weighted Hyperspectral Image Target Detection Algorithm Based on ICA Orthogonal Subspace Projection[J]. *Geomatics and Information Science of Wuhan University*, 2013, 38(4): 440-444(**王凯, 舒宁, 李亮, 等. 利用** ICA **正交子空间投影加权的高光谱影像目标探测算法**[J]. **武汉大学学报・信息科学版**, 2013,38(4): 440-444)

[8] Wang H, Ullah M M, Klaser A, et al. Evaluation of Local Spatio-temporal Features for Action Recognition[C]. The BMVC 2009-British Machine Vision Conference, London, 2009

[9] Hyv R A, Hurri J, Hoyer P O. Natural Image Statistics: A Probabilistic Approach to Early Computational Vision [M]. Berlin:Springer, 2009

[10] Le Q V. Building High-level Features Using Large Scale Unsupervised Learning [C]. The Acoustics, Speech and Signal Processing (ICASSP), Vancouver, USA, 2013

[11] Le Q V, Zou W Y, Yeung S Y, et al. Learning Hierarchical Invariant Spatio-temporal Features for Action Recognition with Independent Subspace Analysis [C]. The Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, Colorado Springs, US, 2011

[12] Han D, Bo L, Sminchisescu C. Selection and Context for Action Recognition[C]. Computer Vision, 2009 IEEE 12th International Conference on, Kyoto, Japan, 2009

[13] Gilbert A, Illingworth J, Bowden R. Action Recognition Using Mined Hierarchical Compound Features[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2011, 33(5): 883-897

Extracting Spatio-temporal Features via Multi-layer Independent Subspace Analysis for Action Recognition

QU Tao[1] *DENG Dexiang*[1] *LIU Hui*[2] *ZOU Lian*[1] *LIU Yifeng*[1]

1 School of Electronic Information, Wuhan University, Wuhan 430072, China

2 Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education, Xiamen University, Xiamen 361005, China

Abstract: Human action recognition plays an important role in the field such as video supervision and medical diagnosis. Current methods are based on the expansion from two-dimension artificial design features to three-dimensions, ones or extracting spatio-temporal features via trajectories. Based on deep learning methods, this paper proposes a multilayer neural network in three-dimensional space, learning rich spatio-temporal features from large amount of videos. First, we use independent subspace analysis to build a two layer stacked convolutional neural network, obtaining weights from training database. Spatio-temporal features are then quantized into visual words with K-means clustering. Non-linear support vector machine(SVM) were used to classify frequency histograms of visual words into different action groups. We apply our algorithm to Hollywood2 database, extracting spatio-temporal features from 12 human action groups. Result shows that the feature weights trained by ISA network are similar with those by Gabor filter, which have obvious selectivity of frequency and direction, robustness to phase variation, conforming to the human visual system.

Key words: stack and convolution; independent subspace analysis; multi-layer neural network; unsupervised learning; deep learning; human action recognition

First author: QU Tao, PhD, specializes in image processing, deep learning, target tracking and identification. E-mail: aboutyoucsm@126.com

Corresponding author: ZOU Lian, PhD, associate professor. E-mail: zoulian@whu.edu.cn.

Foundation support: The National Natural Science Foundation of China, No. 61072135.