基于 $L_{2,1}$ 范数稀疏特征选择和超法向量的深度图像序列行为识别

宋相法 张延锋 郑逢斌
(河南大学计算机与信息工程学院 开封 475004)

摘 要 结合 $L_{2,1}$ 范数稀疏特征选择和超法向量提出了一种新的深度图像序列行为识别方法。首先从深度图像序列中提取超法向量特征;然后利用 $L_{2,1}$ 范数稀疏特征选择方法从超法向量特征中选择出最具判别性的稀疏特征子集作为特征表示;最后利用线性分类器 Liblinear 进行分类。在 MSR Action3D 数据库上的实验结果表明,所提方法使用2%的超法向量特征获得的识别率为 94.55%,并且具有比其他方法更高的识别精度。

关键词 行为识别,深度图像序列,超法向量,稀疏特征选择,$L_{2,1}$ 范数

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.052

Activity Recognition from Depth Image Sequences Based on $L_{2,1}$-norm Sparse Feature Selection and Super Normal Vector

SONG Xiang-fa ZHANG Yan-feng ZHENG Feng-bin
(School of Computer and Information Engineering, Henan University, Kaifeng 475004, China)

Abstract This paper presented a novel method of activity recognition from depth image sequences based on $L_{2,1}$-norm sparse feature selection and super normal vector. First, the super normal vector feature is extracted from depth image sequences. Then the most discriminative feature subset is selected from the whole super normal vector feature set based on the method of $L_{2,1}$-norm sparse feature selection. Finally, the classification is based on Liblinear classifier. Experimental results on MSR Action3D dataset show that the proposed method achieves 94.55% of recognition accuracy using only 2% of the whole super normal vector feature, and is superior to the state-of-art methods.

Keywords Activity recognition, Depth image sequences, Super normal vector, Sparse feature selection, $L_{2,1}$-norm

1 引言

近年来,人体行为识别是计算机视觉领域中最活跃的研究主题之一。它将使得计算机具备感知外部世界的能力,进而使计算机更加智能,人机关系更加和谐。人体行为识别在诸如人机交互、智能视觉监控、虚拟现实、智能家居、老年人看护等领域起着至关重要的作用,受到了研究人员的广泛关注[1-5]。

在过去的几十年里,研究人员主要利用可见光摄像机获取的RGB图像序列进行人体行为识别的研究[1-2],但是它们对颜色、光照变化和复杂的背景等因素比较敏感,识别结果容易受光照变化、阴影、物体遮挡及环境变化等因素的影响。当环境、光照发生变化时,识别精度会大幅度降低。虽然研究人员解决了一些难点问题,但是人体行为识别仍是一个具有挑战性的任务[3]。

随着廉价的微软 Kinect 传感器的出现,它携带的深度摄相机引起了研究人员的极大关注,在计算机视觉和机器人等领域取得了广泛的应用[3]。与传统的利用可见光摄像机获取的RGB图像序列相比,深度摄相机获取的深度图像序列具有诸多优势。例如,深度图像提供了场景的三维结构信息,这些信息为行为识别提供了比颜色和纹理更有力的判别信息;同时深度图像不受光照变化的影响等[3]。因此基于深度图像序列进行人体行为识别成为了研究热点。例如,文献[6]使用一个动作图谱对行为进行建模,并且使用三维词袋描述一系列显著姿势,最后使用隐马尔可夫模型对行为进行识别。文献[7]利用深度图像信息和骨架点信息进行行为识别。文献[8]把整个深度图像序列进行叠加,从而得到一幅深度映射,然后利用从该深度映射中提取的梯度描述子作为行为特征表示。文献[9]提出了一种基于时空深度长方体相似性特征的人体行为识别方法。文献[10]基于深度图像序列中的形状和运动信息,用曲面法向量的直方图分布来表示一个动作。文献[11]提出了基于超法向量编码的深度图像序列人体行为识别方法。文献[12]提出了基于深度信息和RGB信息相融合的人体行为识别方法。文献[13]提出了一个基于流形学习的深度图像序列人体行为识别方法。

到稿日期:2016-02-26 返修日期:2016-05-29 本文受国家自然科学基金(U1504611,61272282),河南省教育厅科学技术研究重点项目(15A520010)资助。

宋相法(1975-),博士,副教授,硕士生导师,CCF会员,主要研究方向为计算机视觉与机器学习,E-mail:xiangfasong@163.com;**张延锋**(1977-),硕士,讲师,主要研究方向为图像处理;**郑逢斌**(1963-),博士,教授,博士生导师,主要研究方向为智能信息处理。

虽然基于深度图像序列的人体行为识别在近些年已经取得一定进展，但是目前还没有一个公认的、鲁棒性好的方法来识别行为。因此，本文在上述工作的基础上，基于 $L_{2,1}$ 范数稀疏特征选择[14]和超法向量[11]提出了一种新的深度图像序列行为识别方法，实验表明文中所提方法使用 2% 的超法向量特征获得的识别率为 94.55%，与超法向量方法[11]相比不但减少了所使用的特征数量，而且提高了识别精度。

2　本文算法

2.1　超法向量特征

超法向量特征是由文献[11]提出的一种深度图像序列特征提取方法。该方法首先计算出深度图像序列中每个点的法向量描述子，然后采用 Fisher 向量提取方法[15]的一个简化的非概率方法来提取超法向量特征。该特征能够同时捕获局部运动和几何信息，因此具有较强的描述能力。

假定深度图像序列可以用式(1)的函数形式来表达[10]：

$$R^3\rightarrow R^1: z=f(x,y,t) \tag{1}$$

该函数构成了四维空间中的一个曲面 S，S 上的点 (x,y,t,z) 满足式(2)：

$$S(x,y,t,z)=f(x,y,t)-z=0 \tag{2}$$

其中，x,y,z 为空间坐标，t 为时间。

曲面 S 上的点 $S(x,y,t,z)$ 处的法向量 n 表示为[10]：

$$n=\nabla S=\left(\frac{\partial z}{\partial x},\frac{\partial z}{\partial y},\frac{\partial z}{\partial t},-1\right)^{\mathrm{T}} \tag{3}$$

点 $S(x,y,t,z)$ 的法向量描述子 p 由其局部时空邻域中的 L 个点的法向量级联而成，表示为[11]：

$$p=(n_1^{\mathrm{T}},\cdots,n_L^{\mathrm{T}})^{\mathrm{T}} \tag{4}$$

令 $P=\{p_1,\cdots,p_N\}\in R^M$ 表示一个训练集合，稀疏编码[15]的数学表达式为：

$$\min_{\boldsymbol{D},\boldsymbol{\alpha}}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|p_i-\boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2+\lambda\|\boldsymbol{\alpha}_i\|_1\right)$$
$$\text{s.t. } d_k^{\mathrm{T}}d_k\leqslant 1,\ \forall k=1,\cdots,K \tag{5}$$

其中，$\boldsymbol{D}\in\boldsymbol{R}^{M\times K}$ 是字典，它的每一列 $(d_k)_{k=1}^{K}$ 表示一个视觉单词，$\boldsymbol{\alpha}\in R^{K\times N}$ 是稀疏分解系数，λ 为正则参数。式(5)中的 $\boldsymbol{D}$ 和 $\boldsymbol{\alpha}$ 可以利用 SPAMS 工具箱[16]求解得到。

对于每个视觉单词 d_k，首先利用空间平均池化方法来汇聚第 t 帧中的量化误差，如式(6)所示：

$$u_k(t)=\frac{1}{|N_t|}\sum_{i\in N_t}\alpha_{k,i}(p_i-d_k) \tag{6}$$

其中，$u_k(t)$ 表示第 k 个视觉单词在第 t 帧中的池化误差向量。然后利用时间最大池化来汇聚 T 帧的量化误差，如式(7)所示：

$$u_{k,i}=\max_{t=1,\cdots,T}u_{k,i}(t),i=1,\cdots,M \tag{7}$$

其中，u_k 是第 k 个视觉单词在整卷中的向量表示，$u_{k,i}$ 表示 u_k 的第 i 个分量。把 K 个向量 u_k 级联起来得到最终向量 U，如式(8)所示：

$$U=(u_1^{\mathrm{T}},\cdots,u_K^{\mathrm{T}})^{\mathrm{T}} \tag{8}$$

为了使所提取的特征能更全面地反映时空特征属性，采用自适应时空金字塔方法把图像序列划分成若干块[11]，然后从每一块中提取特征向量 U_i，最后把 U_i 级联起来作为深度图像序列的超法向量特征 x，如式(9)所示：

$$x=(U_1^{\mathrm{T}},\cdots,U_{|V|}^{\mathrm{T}})^{\mathrm{T}} \tag{9}$$

其中，$|V|$ 是图像序列被划分的块数。超法向量特征 x 是一个长向量，多达 10^5 甚至 10^6 级别，存在大量冗余信息，同时也导致识别精度下降。因此本文利用 $L_{2,1}$ 范数特征选择方法[14]选择出稀疏且具有判别能力的特征子集，与超法向量方法[11]相比，不但减少了所使用的特征数量，而且提高了识别精度。

2.2　基于 $L_{2,1}$ 范数的稀疏特征选择

稀疏特征选择是使用各种不同的稀疏模型，使得被选择出来的特征尽可能稀疏且具有判别性，从而提高识别精度。其引起了计算机视觉与机器学习领域研究人员的广泛关注[14,17-20]。

在本文中，假设有 c 个类别，从深度图像序列训练集中提取出的超法向量特征矩阵 $X=[x_1,x_2,\cdots,x_n]\in R^{t\times n}$，其对应的类别标签矩阵 $Y=[y_1,y_2,\cdots,y_n]^{\mathrm{T}}\in R^{n\times c}$，$y_i=[y_{i1},\cdots,y_{ic}]^{\mathrm{T}}\in R^c$ 表示第 i 个样本所对应的类别信息，如果第 i 个样本属于第 j 类，那么 $y_{ij}=1$，否则 $y_{ij}=0$。

对一般的最小平方回归加上 $L_{2,1}$ 范数正则化约束，得到如下表达式[14]：

$$\min_W J(W)=\sum_{i=1}^{n}\|W^{\mathrm{T}}x_i-y_i\|_2+\rho\sum_{l=1}^{t}\|w_l\|_2$$
$$=\|X^{\mathrm{T}}W-Y\|_{2,1}+\rho\|W\|_{2,1} \tag{10}$$

其中，$W\in R^{t\times c}$ 是稀疏特征选择映射矩阵，ρ 为正则参数。式(10)也可以表示为：

$$\min_W\frac{1}{\rho}\|X^{\mathrm{T}}W-Y\|_{2,1}+\|W\|_{2,1} \tag{11}$$

令 $E=\frac{1}{\rho}(Y-X^{\mathrm{T}}W)$，$q=t+n$，$Z=\begin{bmatrix}W\\E\end{bmatrix}\in R^{q\times c}$，则有

$$\frac{1}{\rho}\|X^{\mathrm{T}}W-Y\|_{2,1}+\|W\|_{2,1}=\|E\|_{2,1}+\|W\|_{2,1}=$$
$$\|Z\|_{2,1} \tag{12}$$

且满足 $X^{\mathrm{T}}W+\rho E=Y$，令 $B=[X^{\mathrm{T}}\ \rho I]\in R^{n\times q}$，则式(10)可转化为：

$$\min_U\|Z\|_{2,1}\quad\text{s.t. } BZ=Y \tag{13}$$

式(13)的拉格朗日函数表达式为：

$$f(Z)=\|Z\|_{2,1}-\mathrm{Tr}(\Lambda^{\mathrm{T}}(BZ-Y)) \tag{14}$$

对式(14)执行梯度下降法，可以获得迭代解[14]：

$$Z=D^{-1}B^{\mathrm{T}}(BD^{-1}B^{\mathrm{T}})^{-1}Y \tag{15}$$

其中，D 是对角矩阵，其对角线上的元素为 $d_{ij}=\frac{1}{2\|z_i\|_2}$，$z_i$ 是 Z 的第 i 行向量。

Z 的前 t 行即为稀疏特征选择映射矩阵 $W=[z_1;z_2;\cdots;z_t]$，根据 W 从超法向量特征 x 中选择出稀疏且具判别能力的特征，然后使用 Liblinear 分类器[21]进行分类。

2.3　算法步骤

综上所述，本文算法的步骤描述如下：

step 1　根据式(4)计算深度图像序列中每个点的法向量描述子 p；

step 2　根据式(5)计算法向量描述子 $\{p_i\}$ 的稀疏编码系数 $\{\alpha_i\}$；

step 3 根据式(6)—式(8)计算向量特征$\{U_i\}$;

step 4 根据式(9)得到深度图像序列的超法向量特征x;

step 5 根据式(15)计算Z,从而得到稀疏特征选择映射矩阵W;

step 6 根据W从超法向量特征x中选择出稀疏且具判别能力的特征,然后使用 Liblinear 分类器进行分类。

3 实验结果与分析

为了对本文算法的识别效果进行评估,在深度图像序列权威数据库 MSR Action3D 上进行了测试,如图 1 所示。该数据库包含 20 种人体行为:high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw;每种行为由 10 个表演者重复表演 3 次。深度图像序列数据采集的速率为 15 帧/秒,图像分辨率为 640×480。为了保证比较的公平性,与文献[6-11]进行了相同的设置,一半对象的动作样例作为训练集,另一半对象的动作样例作为测试集。实验中,式(5)中的正则参数$\lambda=0.15$,字典$\boldsymbol{D}$的大小$k=100$,式(10)中的正则参数$\rho=10$,利用W从超法向量特征x中选择出 2%的特征子集。实验平台如下:CUP 为 3.6 GHz 的 Intel Core 四核 i7-4790,内存为 32G,系统为 64 位 Windows 7,软件平台为 64 位 Matlab2013b。

图 1 MSR Action3D 数据库上的部分样例图

表 1 列出了本文算法在 MSR Action3D 数据库上的识别结果及其他文献[6-11]在该数据集上的识别结果(最好结果用粗体标识)。由表 1 可知,本文算法在 MSR Action3D 数据库上的识别精度高于其他 6 种算法,识别精度提高了 1.46%~19.85%。

表 1 各种方法在 MSR Action3D 数据集上的实验结果

方法	识别精度/%
Bag of 3D Points[6]	74.70
Actionlet Ensemble[7]	88.20
Depth Motion Maps[8]	88.73
DSTIP[9]	89.3
HON4D[10]	88.89
Super normal [11]	93.09
本文方法	**94.55**

混淆矩阵(confusion matrix,CM)是衡量模式识别与机器学习中识别性能的一个经典方式。图 2 示出了本文算法在 MSR Action3D 数据库上的混淆矩阵。由图 2 可知:1)在 20 类动作中,有 15 类动作的识别精度达到了 100%;2)识别的错误率主要发生在 hand catch 和 high throw, draw x 和 hammer 以及 forword punch 和 hammer 之间,其原因是这些动作比较相似,该发现与文献[11]的发现是一致的。

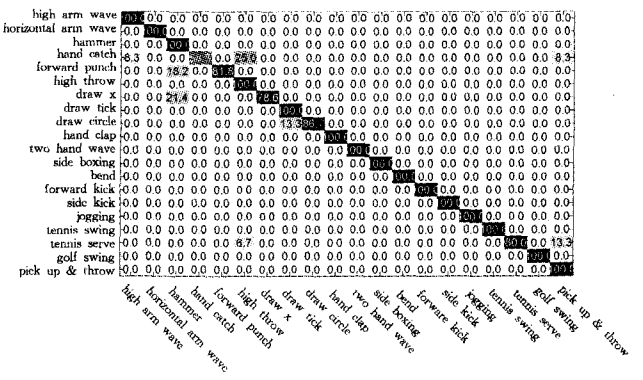

图 2 本文方法在 MSR Action3D 数据库上的混淆矩阵

为了考察不同个数的特征子集对识别结果所产生的影响,在 MSR Action3D 数据库上进行了实验,并给出了根据识别精度计算出的实验结果变化曲线图,如图 3 所示,其中 x 轴为所选特征子集占整体特征的百分比,取值范围为 1%~10%,步长为 1,y 轴为识别精度。从图 3 可以看出,当特征子集为 1%时,识别精度为 93.82%;当特征子集为 2%时,识别精度获得最大值 94.55%;当特征子集超过 3%时,识别精度开始呈下降趋势。

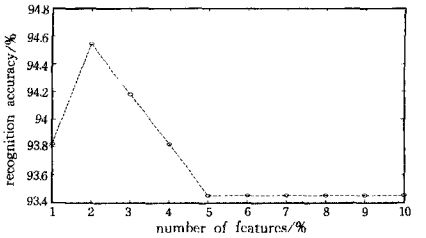

图 3 本文算法利用不同的特征子集在 MSR Action3D 数据库上的实验结果

另外为了考察稀疏编码式(5)中的正则参数 λ 的不同取值对识别结果所产生的影响,在 MSR Action3D 数据库上进行了实验,并给出了根据识别精度计算出的实验结果变化曲线图,如图 4 所示,其中 x 轴为正则参数 λ 的不同取值,取值范围为 0.1~0.5,步长为 0.05,y 轴为识别精度。从图 4 可以看出,当 λ 为 0.1 时,识别精度为 93.81%;当 λ 为 0.15 时,识别精度获得最大值 94.55%;当 λ 超过 0.2 时,识别精度开始呈下降趋势。

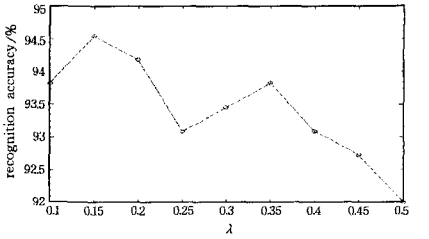

图 4 本文算法利用稀疏编码中的正则参数 λ 的不同取值在 MSR Action3D 数据库上的实验结果

结束语 针对仅利用超法向量特征导致深度图像序列行为识别精度较低的问题,本文提出了基于 $L_{1,2}$ 范数稀疏特征选择和超法向量的深度图像序列行为识别算法。在 MSR

(下转第 323 页)

[28] ZHENG Y R，YUAN J Z，LIU H Z. An algorithm of lane detection based on IPM-DVS[J]. Journal of Beijing Union University，2015，29(2)，41-46.

[29] ZE Z，CHEN E T. Automatic lane detection from vehicle motion trajectories[C]// 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2013：466-471.

[30] HEECHUL，JUNG，JUNGGON，et al. An efficient lane detection algorithm for lane departure detection[C]// 2013 IEEE on Intelligent Vehicles Symposium (IV). 2013：976-981.

(上接第 308 页)

Action3D 数据库上的实验结果表明，本文所提方法从超法向量特征中选择出 2%的特征子集获得的识别率为 94.55%，与超法向量方法[11]相比，不但减少了所使用的特征数量，而且提高了识别精度。

参 考 文 献

[1] AGGARWAL J K，RYOO M S. Human activity analysis：a review[J]. ACM Computing Surveys，2011，43(3)：1-43.

[2] HU Q，QIN L，HUANG Q M. A survey on visual human action recognition[J]. Chinese Journal of Computers，2013，36(12)：2512-2524. (in Chinese)
胡琼，秦磊，黄庆明. 基于视觉的人体动作识别综述[J]. 计算机学报，2013，36(12)：2512-2524.

[3] HAN J H，SHAO L，XU D，et al. Enhanced computer vision with microsoft kinect sensor：a review[J]. IEEE Transactions on Cybernetics，2013，43(5)：1318-1334.

[4] AGGARWAL J K，LU X. Human activity recognition from 3D data：a review[J]. Pattern Recognition Letters，2014，48(2)：70-80.

[5] LEI Q，CHEN D S，LI S Z. Advances on human action recognition in realistic scenes[J]. Computer Science，2014，41(12)：1-7. (in Chinese)
雷庆，陈锻生，李绍滋. 复杂场景下的人体行为识别研究新进展[J]. 计算机科学，2014，41(12)：1-7.

[6] LI W Q，ZHANG Z Y，LIU Z C. Action recognition based on a bag of 3D points[C]// Proceedings of the IEEE International Conference on Human Communicative Behavior Analysis. 2010：9-14.

[7] WANG J，LIU Z C，WU Y，et al. Mining actionlet ensemble for action recognition with depth cameras[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2012：1291-1297.

[8] YANG X D，ZHANG C Y，TIAN Y L. Recognition actions using depth motion maps-based histograms of oriented gradients[C]// Proceedings of ACM Conference on Multimedia. 2012：1057-1060.

[9] LU X，AGGARWAL J K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2013：2834-2841.

[10] OREIFEJ O，LIU Zi C. HON4D：Histogram of oriented 4D normals for activity recognition from depth sequences[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2013：716-723.

[11] YANG X D，TIAN Y L. Super normal vector for activity recognition using depth sequences[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2014：804-811.

[12] SHEN X X，ZHANG H，GAO Z，et al. Behavior recognition algorithm based on depth information and RGB Image[J]. Pattern Recognition and Artificial Intelligence，2013，26(8)：722-728. (in Chinese)
申晓霞，张桦，高赞，等. 基于深度信息和 RGB 图像的行为识别算法[J]. 模式识别与人工智能，2013，26(8)：722-728.

[13] WANG X，WO B H，GUAN Q，et al. Human action recognition based on manifold learning[J]. Journal of Image and Graphics，2014，19(6)：914-923. (in Chinese)
王鑫，沃波海，管秋，等. 基于流形学习的人体动作识别[J]. 中国图象图形学报，2014，19(6)：914-923.

[14] NIE F P，HUANG H，CAI X，et al. Efficient and robust feature selection via joint $L_{2,1}$-norms minimization[C]// Proceedings of International Conference on Neural Information Processing Systems. 2010：1813-1821.

[15] JORGE S，FLORENT P，THOMAS M，et al. Image classification with the fisher vector：theory and practice[J]. International Journal of Computer Vision，2013，105(3)：222-245.

[16] MAIRAL J，BACH F，PONCE J，et al. Online learning for matrix factorization and sparse coding[J]. Journal of Machine Learning Research，2010，11(1)：19-60.

[17] HE R，TAN T N，WANG L，et al. $L_{2,1}$ Regularized correntropy for robust feature selection[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2012：2504-2511.

[18] SHI X S，YANG Y J，GUO Z H，et al. Face recognition by sparse discriminant analysis via joint $L_{2,1}$-norm minimization[J]. Pattern Recognition，2014，47(7)：2447-2453.

[19] SHI C J，RUAN Q Q. Feature selection with enhanced sparsity for web image annotation[J]. Journal of Software，2015，26(7)：1800-1811. (in Chinese)
史彩娟，阮秋琦. 基于增强稀疏性特征选择的网络图像标注[J]. 软件学报，2015，26(7)：1800-1811.

[20] ZHOU P Y，LI J，SHEN N M，et al. BSFCoS：Fast co-saliency detection based on block and sparse principal feature extraction[J]. Computer Science，2015，42(8)：305-309. (in Chinese)
周培云，李静，沈宁敏，等. BSFCoS：基于分块与稀疏主特征提取的快速协同显著性检测[J]. 计算机科学，2015，42(8)：305-309.

[21] FAN R G，CHANG K W，HSIEH C J，et al. LIBLINEAR：A library for large linear classification[J]. Journal of Machine Learning Research，2008，9(8)：1871-1874.