DOI:10.15923/j.cnki.cn22-1382/t.2017.3.12

基于深度运动图的人体行为识别

史东承， 李延林*

(长春工业大学 计算机科学与工程学院，吉林 长春 130012)

摘 要：将人体行为深度映射图(depth map)连续投影到 3 个互相垂直的笛卡尔平面，然后对投影做绝对差分，累积各自投影面的差分图像，得到完整的人体行为三维信息——深度运动图(Depth Motion Maps，DMMs)。利用 MSRAction dataset 和 3D Action Pairs dataset 进行训练以获取人体行为字典。在识别未知动作时，利用 Tikhonov 矩阵计算得出权重系数向量。最后，利用 L2 范式正则化协同表示对待识别动作进行分类。通过上述两个数据库的验证，分别达到了 95.3%和 83.8%的平均识别率，已经达到对 DMMs 的较高识别率。

关键词：人体行为；识别；深度运动图；L2 范式

中图分类号：TP 391 **文献标志码**：A **文章编号**：1674-1374(2017)03-0276-06

Human action recognition based on depth motion maps

SHI Dongcheng， LI Yanlin*

(School of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China)

Abstract: Human action depth maps are projected continuously to three perpendicular Descartes plane. The projections are absolute differentiated and cumulated to obtain complete 3D information of human action which is called the Depth Motion Maps (DMMs). With MSRAction dataset and 3D Action Pairs software, DMMs is trained to get the human action dictionary. When the unidentified human action is input, weight coefficients are calculated by using the Tikhonov matrix, and then L2-regularized collaborative representation classifier is used to classify the actions. Two data-set experiments indicate that the average recognition rates is 95.3% and 83.8% respectively.

Key words: human action; recognition; depth motion maps; L2-regularized.

0 引 言

人体行为识别[1]是计算机视觉领域的热门研究方向之一。传统方法在获取人体行为时，多使用二维视频，致使人体行为在三维空间中的动作信息在初始场景下就已经丢失了部分信息。并且由于传统摄像机拍摄的视频多受光照的影响，传统方法不得不对光照进行二次处理，造成人体行

收稿日期：2017-03-21

基金项目：吉林省教育厅“十三五”规划项目(吉教科合字[2016]第 349 号)

作者简介：史东承(1959－)，男，汉族，吉林长春人，长春工业大学教授，硕士，主要从事图像处理与机器视觉方向研究，E-mail:shidongchen@ccut.edu.cn. *通讯作者：李延林(1990－)，男，朝鲜族，吉林通化人，长春工业大学硕士研究生，主要从事图像处理与机器视觉方向研究，E-mail:balinshitou@163.com.

为信息的再次损失。随着 RGBD 摄像机以及深度传感器[2]的使用(如 Kinect 等)，在源头获取人体行为的三维信息变得十分便捷，所得深度图像序列也对光照不敏感，在信息的处理上不必处理光照产生的影响。正是由于人体行为信息由二维转向了三维，让人体行为的处理方式也更加多样和灵活化。

文中正是利用此种优势提出了 DMMs 进行人体行为识别。将人体行为特征的三维信息作为特征向量，并利用 L2 范式正则化协同表示分类器进行分类。算法流程如图 1 所示。

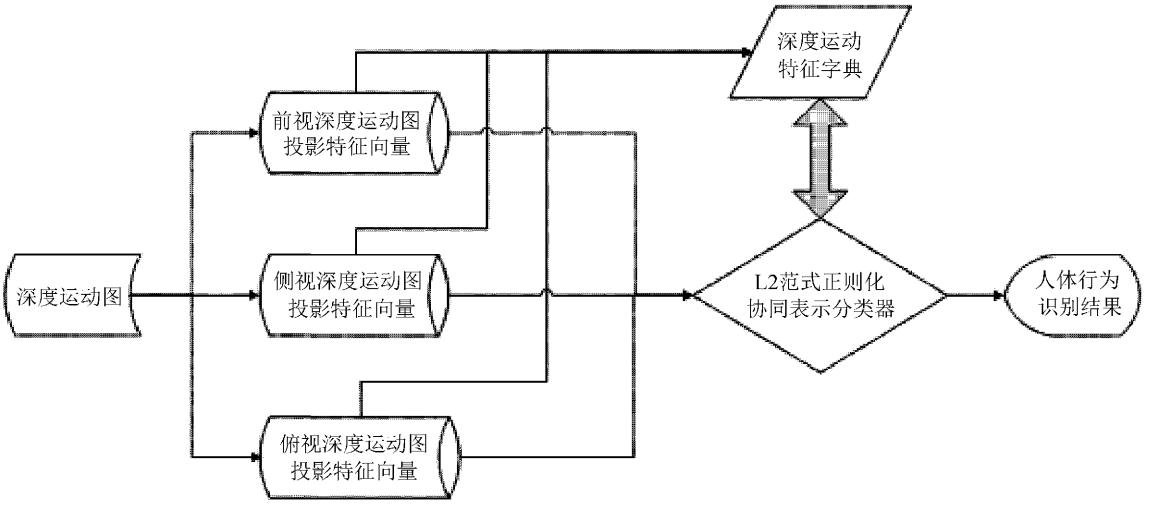

图 1　算法流程图

1　深度运动图

1.1　深度运动图的生成

深度映射图是由 RGBD 摄像机拍摄所得，以往常常用来构建物体的三维信息和三维结构。文中通过利用 RGBD 摄像机获取人体行为视频，使视频拥有人体行为的三维信息。通过 MSRAction3D[3]和 3D Action Pairs 两个数据库的深度运动图(DMM)，将 DMM 在 3 个互相垂直的笛卡尔平面上做二维投影，投影视角为前视(front view)、侧视(side view)和俯视(top view)，所得 DMM 记为 DMM_f、DMM_s 和 DMM_t，分别如图 2 和图 3 所示。

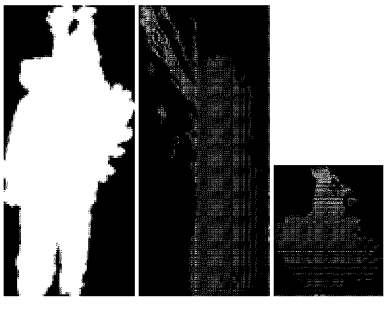

图 2　Tennis Serve

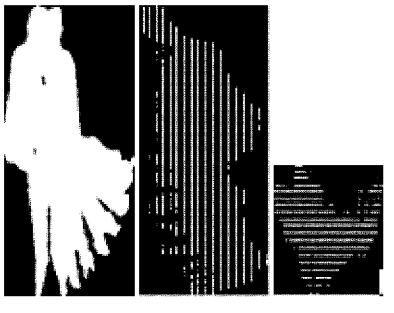

图 3　Side Kick

假设一个深度视频有 N 个帧组成，每一帧在其投影视角下的二维投影记为 Map_v，则

$$DMM_v = \sum_{i=a}^{b} \left| Map_v^i - Map_v^{i-1} \right| \qquad (1)$$

$$1 \leqslant a < b \leqslant N$$

式中：i ——视频序列在视角 v 下的的第 i 帧(注意：为了去除 DMM 的冗余信息，即视频帧的起始帧和结尾帧都含有一些动作幅度不是很大的帧，文中建议去除这些帧，以提升输入信息的有效性)。

1.2 深度运动图特征向量的生成

从图2和图3可以看出，DMM在3个维度上的投影综合在一起具有很高的辨识度。文中正是利用这一特性，将生成的DMM作为人体行为的特征向量。因拍摄人体时摄像机距离人体的远近及拍摄人物的高度、胖瘦各有不同，会造成拍摄的不同视频序列经投影变换后，所得DMM的尺寸各不相同。这里，将利用双三次插值法对DMM的尺寸进行重新调整。

DMM是利用深度图中的像素进行计算，为了避免因大像素值过大影响特征向量表征的均衡性。文中会将所有DMM特征向量进行归一化处理。一般将DMM_v归一化后的向量记为$\overline{DMM_v}$。设DMM_f的尺寸记为$m_f \times n_f$，DMM_s的尺寸记为$m_s \times n_s$，DMM_t的尺寸记为$m_t \times n_t$。则视频序列的DMM特征向量$\boldsymbol{h}$可记为

$$[\mathrm{vec}(DMM_f),\mathrm{vec}(DMM_s),\mathrm{vec}(DMM_t)]^T$$

vec(•)表示向量化操作符。一个特征向量$\boldsymbol{h}$的维度为$(m_f \times n_f + m_s \times n_s + m_t \times n_t) \times 1$。

2 L2范式正则化协同表示分类

2.1 稀疏表示方法

稀疏编码[4]是从人眼视觉系统研究开发得来，它是一种高效合理的编码方式。稀疏编码在人脸识别[5]及图像分类[6]中都获得了不错的成绩。稀疏编码分类的核心思想是使用训练样本生成过完备字典，并利用过完备字典对测试样本进行稀疏表示。最后，计算测试样本与稀疏编码的差值，最小差值即可表示其所表示的类别。

假设有n个类的训练样本$A=[A_1,A_2,\cdots,A_n]\in R^{d\times n}$，$A_j$表示一个训练样本，$d$表示每个训练样本的维度。$g\in R^{d\times 1}$表示一个测试样本，$g$可以利用训练样本的稀疏线性组合来进行表示，即

$$g = A\alpha \tag{2}$$

$\boldsymbol{\alpha}$是一个$n\times 1$的训练样本的系数向量，$\alpha_j(j=1,2,\cdots,n)$表示第j类训练样本A_j的系数。利用$L1$范式最小化[7]来计算测试样本g的系数向量$\hat{\boldsymbol{\alpha}}$

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha}\{\| g - A\alpha \|_2^2 + \theta \| \alpha \|_1\} \tag{3}$$

式中：θ——正则化尺度参数，是用来平衡稀疏项的影响。

测试样本g类别标签，则利用公式

$$e_j = \| g - A_j \hat{\alpha}_j \|_2 \tag{4}$$

进行计算，然后利用公式

$$\mathrm{class}(g) = \arg\min_j\{e_j\} \tag{5}$$

得出测试样本g所属类别。

2.2 L2范式正则化协同表示方法

正如式(3)所表达的L1范式[8]稀疏项，虽然能提高一些准确度，但计算量较大，实时性表现不佳。因此，文中使用L2范式来代替L1范式稀疏项。设Y_p为一个测试样本，设已训练完成的字典$A=[h_1,h_2,\cdots,h_n]$，n表示所用的训练样本数量。则

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\alpha}\{\| Y_p - A\alpha \|_2^2 + \lambda \| \boldsymbol{L}\alpha \|_2^2\} \tag{6}$$

λ表示正则化参数，$\boldsymbol{L}$[9]表示Tikhnov[10]正则化矩阵，$\boldsymbol{L}$的表达式如下

$$\boldsymbol{L} = \begin{pmatrix} \| Y_p - h_1 \|_2 & & 0 \\ & \ddots & \\ 0 & & \| Y_p - h_n \|_2 \end{pmatrix} \tag{7}$$

因此，向量系数$\hat{\boldsymbol{\alpha}}$可以利用公式

$$\hat{\boldsymbol{\alpha}} = (A^T A + \lambda \boldsymbol{L}^T \boldsymbol{L})^{-1} A^T Y_p \tag{8}$$

当利用测试样本Y_p计算出$\hat{\boldsymbol{\alpha}}$后，即可利用公式(4)，$e_j = \| \| Y_p - A_j \hat{\alpha}_j \|_2$，$j\in(1,2,\cdots,n)$，求出字典$A$中各元素与样本之间的差值。最后，利用式(5)即可求出g的类别。

3 实验结果

3.1 实验数据

文中使用Matlab进行算法仿真，使用MSRAction3D dataset进行算法准确度评估，并与当前主流算法进行比对。MSRAction3D dataset包含10个人，每个人做20种不同的动作，且每个人做的每一种动作都会重复2～3次，这么做的目的是为了提升训练后的类内多样性，以提升识别率。首先将MSRAction3D dataset分成3个组，每个组中所包含的动作见表1。

表 1　MSRAction3D dataset 3 组动作分组

Action set1(AS1)	Action set2(AS2)	Action set3(AS3)
Horizontal wave(2)	High wave(1)	High throw(6)
Hammer(3)	Hand catch(4)	Forward kick(14)
Forward punch(5)	Draw x(7)	Side kick(15)
High throw(6)	Draw tick(8)	Jogging(16)
Hand clap(10)	Draw circle(9)	Tennis swing(17)
Bend(13)	Two hand wave(11)	Tennis serve(18)
Tennis serve(18)	Forward kick(14)	Golf swing(19)
Pickup throw(20)	Side boxing(12)	Pickup throw(20)

表 1 中，我们将 3 组动作中的每组动作都进行测试 1 组、测试 2 组和交叉测试组试验。在测试 1 组中，每个人执行的第 1 个动作作为训练数据，后两个动作作为测试数据。测试 2 组，每个人执行的前两个动作作为训练数据，最后一个作为测试数据。在交叉测试组，1、3、5、7、9 这 5 个人的动作作为训练数据，2、4、6、8、10 这 5 个人的动作作为测试数据。由于不同人做不同动作时，频率、力度和幅度各有不同，易造成识别误差，因此，交叉测试组的结果比其他两个组更能表明算法的鲁棒性。

为了进一步说明本方法对相似动作的识别较有优势，文中还引入了另一个数据库 3D Action Pairs dataset。此数据库内动作由 10 个人做出，每人每个动作做 3 次，动作内容见表 2。

表 2　3D Action Pairs dataset 所包含的动作

3D Action Pairs dataset 动作类别	
Pickup a box	Put down a box
Lift a box	Place a box
Push a chair	Pull a chair
Wear a hat	Take off a hat
Put on backpack	Take off a backpack
Stick a poster	Remove a poster

实验中，将数据库里每个动作类别的前 5 个人所做的动作作为训练集，剩余 5 个人的动作作为测试集。部分 3D Action Pairs dataset 内的动作视图如图 4 所示。

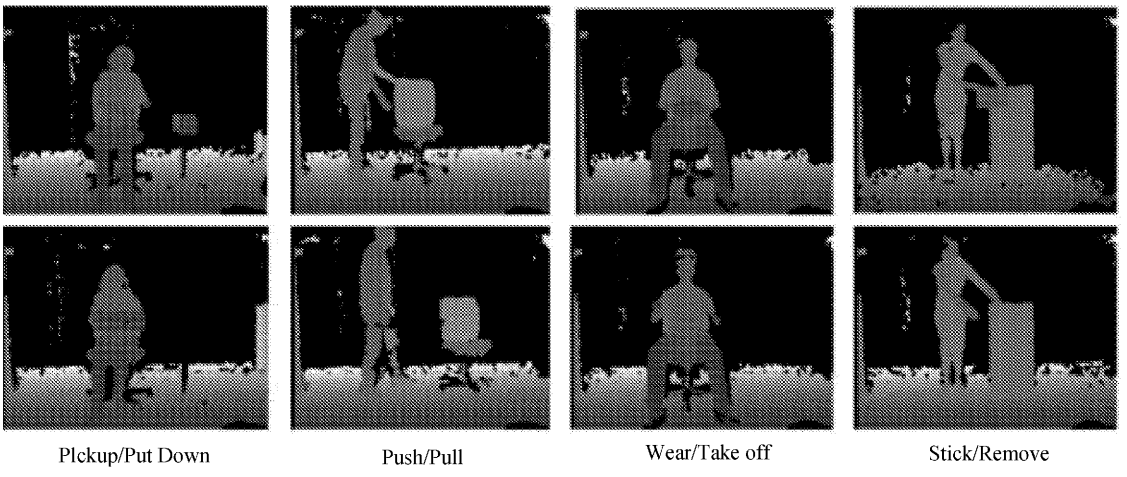

图 4　3D Action Pairs dataset 图像示例

3.2　实验结果

使用 MSRAction3D dataset 与主流的 3 种方法进行识别准确率评估。文中方法与主流方法识别率比对见表 3。

表3 文中方法与主流方法识别率比对

方法	测试1组				测试2组				交叉测试组			
	AS1	AS2	AS3	AVG	AS1	AS2	AS3	AVG	AS1	AS2	AS3	AVG
Lu et al.[11]	98.4	96.7	93.2	96.1	98.6	97.1	94.7	96.8	88.0	85.4	63.7	79.0
Yang et al.[12]	94.7	95.5	97.4	95.9	97.4	98.5	97.3	97.7	74.7	76.1	96.4	82.4
Li et al.[13]	89.5	89.2	96.4	91.7	93.2	93.0	96.5	94.2	73.0	71.9	79.1	74.7
文中方法(DMM)	97.3	97.5	98.1	97.6	98.6	98.7	100	99.1	94.3	82.3	91.0	89.2

从表3可以看出,文中提出的方法在3组测试中都有比较明显的优势,尤其是最具挑战性的交叉测试组,文中的平均识别率都明显优于其他3种方法,可见基于DMM。的人体行为识别在相似动作识别上依然有很强的鲁棒性。

为了进一步说明文中方法在相似动作中的识别优势,我们使用3D Action Pairs dataset作为数据库,使用算法分别为Skeleton+LOP和文中算法作比较。以下两张混淆矩阵即为这两种算法的比较,分别如图5和图6所示。

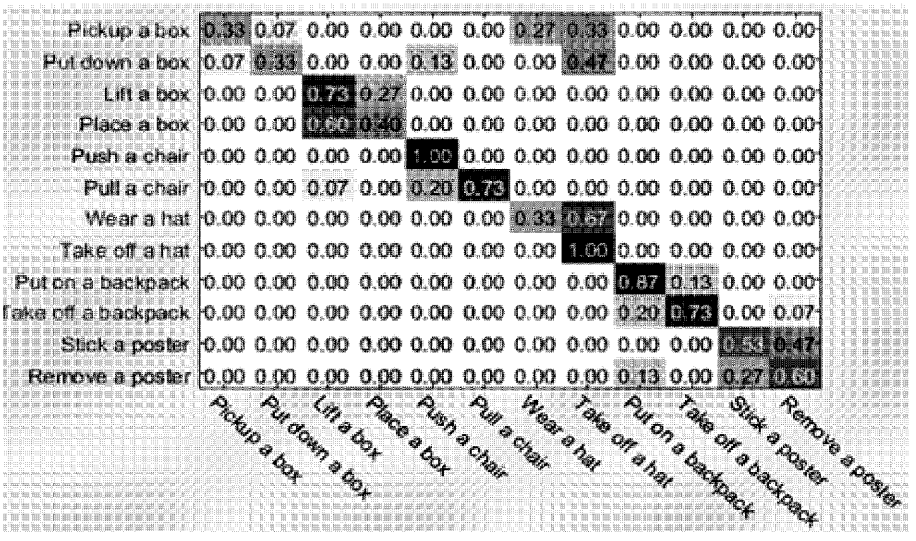

图5 Skeleton+LOP

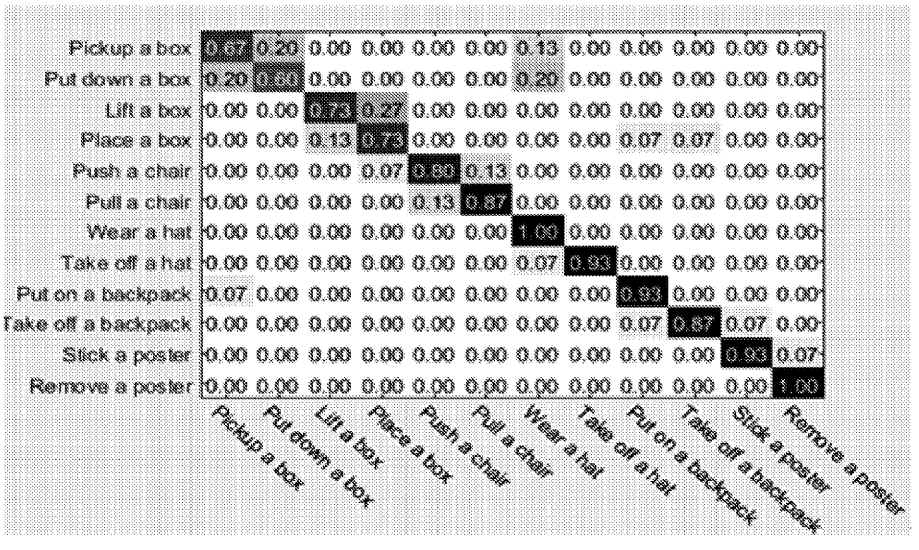

图6 文中方法

从图 5 和图 6 可以看出，以 Skeleton＋LOP 所形成的算法在 3D Action Pairs dataset 处于较为明显的劣势。文中所提的方法在 Lift a box，Push a chair，Take off a hat，Put on a backpack 以及 Take off a backpack 这 5 个动作中略微差一些，但在其他 7 个动作中都占有优势，尤其是在 Pickup a box，Put down a box 以及 Remove a poster 这 3 个动作中都占据绝对的识别优势。

可见本方法不仅在 MSRAction dataset 有着不错的识别率，在 3D Action Pairs dataset 这样动作十分相似的数据库实验中依然可以达到不错的识别率。足见文中提出的算法在类内多样性和类间区分上有着不错的鲁棒性。

参考文献：

[1] Cheng G，Wan Y，Saudagar A N，et al. Advances in human action recognition：A survey[J]. Computer Science，2015(1)：1-30.

[2] 陈万军，张二虎.基于深度信息的人体动作识别研究综述[J].西安理工大学学报，2015(3)：253-264.

[3] Wang J，Liu Z，Wu Y，et al. Mining actionlet ensemble for action recognition with depth cameras [C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2012：1290-1297.

[4] Wright J，Yang A Y，Ganesh A，et al. Robust face recognition via sparse representation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2009，31(2)：210-227.

[5] Gao S，Tsang W H，Chia L T. Kernel sparse representation for image classification and face recognition[C]//Computer Vision -ECCV，2010：1-14.

[6] Yang J，Yu K，Gong Y，et al. Linear spatial pyramid matching using sparse coding for image classification[C]//IEEE，2009：1794-1801.

[7] Wright J，Ma Y. Dense error correction via l1-minimization[C]// IEEE International Conference on Acoustics. IEEE Computer Society，2009：3033-3036.

[8] Lei Zhang，Meng Yang，Xiangchu Feng. Sparse representation or collaborative representation：Which helps face recognition [C]// International Conference on Computer Vision. IEEE Computer Society，2011：471-478.

[9] Chen C，Tramel E W，Fowler J E. Compressed-sensing recovery of images and video using multihypothesis predictions [C]// Conference on Circuits. IEEE，2011：1193-1198.

[10] Golub G H，Hansen P C，O′Leary D P. Tikhonov regularization and total least squares [J]. Siam Journal on Matrix Analysis & Applications，2010，21(1)：185-194.

[11] Lu Xia，Chia Chih Chen，Aggarwal J K. View invariant human action recognition using histograms of 3D joints [C]//Computer Vision and Pattern Recognition Workshops，2012.

[12] Yang X，Tian Y L. EigenJoints-based action recognition using Naive-Bayes-Nearest-Neighbor[J]. Perceptual & Motor Skills，2012，38(3c)：14-19.

[13] Li W，Zhang Z，Liu Z. Action recognition based on a bag of 3D points[J]. Advances in Artificial Intelligence，2016：3-14.