

基于姿势字典学习的人体行为识别

蔡加欣^{1,2} 冯国灿^{1,2} 汤 鑫^{1,2} 罗志宏³

¹ 中山大学数学与计算科学学院, 广东 广州 510275
² 广东省计算科学重点实验室, 广东 广州 510275
³ 中山大学信息科学与技术学院, 广东 广州 510275

摘要 提出一种基于人体轮廓表达的姿势学习框架来进行人体行为识别。通过一种基于 Procrustes 形状分析和局部保持投影的姿势特征表示方法, 从人体运动视频中提取具有平移、旋转和放缩不变性的姿势特征, 在保留人体姿势的局部流形结构的同时尽量提取子空间上的判别信息。针对该特征还提出了一种基于姿势字典学习的人体行为识别框架, 对每类行为分别学习一个对应于该类的字典, 通过串联所有类的字典来得到整个姿势字典; 并通过最小重构误差准则来分类测试视频。在 Weizmann 和 MuHAVi-MAS14 数据集上的实验结果证实了该方法的识别率高于大部分经典方法。特别是在 MuHAVi-MAS14 数据集上的识别率对比已有的结果上有巨大的提升。

关键词 图像处理; 行为识别; Procrustes 形状分析; 局部保持投影; 稀疏表示; 字典学习

中图分类号 TP391 **文献标识码** A **doi**: 10.3788/AOS201434.1215002

Human Action Recognition by Learning Pose Dictionary

Cai Jiixin^{1,2} Feng Guocan^{1,2} Tang Xin^{1,2} Luo Zhihong³

¹ School of Mathematics and Computing Science, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China
² Guangdong Province Key Laboratory of Computational Science, Guangzhou, Guangdong 510275, China
³ School of Informational Science and Technology, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China

Abstract A framework for human action recognition by learning pose dictionary based on human contour representation is proposed. A new pose feature based on Procrustes analysis and local preserving projection is proposed, which can extract shape information from human motion video which is invariant to translation, scaling and rotation. Moreover, it can extract discriminative subspace information when preserving local manifold structure of human pose. After the pose feature is extracted, a human action recognition framework based on pose dictionary learning is proposed. Class-specific dictionaries are trained individually on all training frames of each class to build the whole pose dictionary by concatenating all class-specific dictionaries. The test video is classified with the minimum reconstruction error on the learned dictionary. Experimental results on Weizmann and MuHAVi-MAS14 dataset demonstrate proposed method outperforms most classical methods. Especially, classification rate of this method on MuHAVi-MAS14 dataset achieves a considerable boost compared with that of state-of-the-art approaches.

Key words image processing; behavior recognition; procrustes shape analysis; local preserving projection; sparse representation; dictionary learning

OCIS codes 150.0155; 100.4999; 100.5010; 140.1135

1 引言

基于视频内容的人体行为识别最近得到了研究

者的许多关注, 已经成为了机器视觉领域的一个研究热点。人体行为识别在公共环境下的智能监控、

收稿日期: 2014-04-03; **收到修改稿日期**: 2014-08-01

基金项目: 国家自然科学基金(61272338)

作者简介: 蔡加欣(1988—), 男, 博士研究生, 主要从事机器视觉和医学图像处理方面的研究。

E-mail: caijixin@mail2.sysu.edu.cn

导师简介: 冯国灿(1962—), 男, 博士, 教授, 博士生导师, 主要从事机器视觉和生物特征识别方面的研究。

E-mail: mcsfgc@mail.sysu.edu.cn

基于视觉的人机交互和体感游戏等方面都具有巨大的应用潜力。但是由于人体是个复杂的非刚性结构体,其姿势和外观具有多变性,提高人体行为识别的准确性和稳健性仍然是一个具有难度的挑战。

人体行为识别的框架通常可以分为两个部分:行为表示的特征提取部分和行为模式的分类部分。人体行为的特征提取方法通常可以分为基于全局表示的方法和基于局部表示的方法。基于全局表示的特征提取方法直接在整个视频上提取特征,不考虑行为的动态信息。其优点是无需提取和跟踪视频中的兴趣点,计算量低,对噪声和干扰的稳健性好,可以使用简单的距离匹配方法进行分类。文献[1]提出了运动能力图像和运动历史图像来指示和描述短时间内行为事件的发生。文献[2]提出了一种像素改变历史图像来描述像素级的动态变化历史和探测动态事件的发生。研究者们还提出了动作能量图^[3]和平均运动能量^[4]来表示一些带有周期性非匀速运动的复杂人体动作。文献[5]提出了一种运动历史体积来建立多摄像机拍摄环境下具有视角无关性的三维(3D)人体行为表示。在先前的工作中^[6],提出了一种方向性整体特征,来描述人体的往返运动,可以更好地区分外观相似的人体动作。基于局部表示的特征提取则从视频中提取局部特征的时间序列。常用的局部特征包括人体外观表示、兴趣点、空时兴趣点和空时体体积等。文献[7]提出了一种无监督学习方法从视频的空时兴趣点序列中提取视觉语义。文献[8]将空时体体积和旋转图在近邻图中进行量化和融合,以此作为人体行为的特征。文献[9]采用空时体体积上的3D梯度直方图来描述空时体体积。文献[10]提出一种3D尺度不变特征变换描述子来描述视频内空时兴趣点之间的关系。文献[11]提出了一种时空语义特征来描述人体交互行为。

最近,稀疏编码已经成为了信号处理和计算机视觉领域的一个研究热点,并已被广泛应用于图像超分辨^[12]、图像融合^[13]、高光谱数据分类^[14]和图像重建^[15-16]中。稀疏表示一般是把所有训练样本作为超完备字典,对测试样本求解基于该超完备字典的稀疏线性表达。而字典学习则是从训练样本中学习一个新的用于稀疏表示的超完备字典。目前基于稀疏表示的行为识别方法都是建立在视频的空时特征上,如空时兴趣点^[17]、空时体体积^[18]等,需要进行复杂的兴趣点提取和跟踪操作,计算量大且对噪声敏感。与这些已有的方法不同,本文提出了一种基于人体形状表示的姿势字典学习框架来进行人体行

为识别。首先,提出了一种基于 Procrustes 形状分析和局部保留投影的姿势特征提取方法:该方法从人体剪影图像中提取人体轮廓作为人体姿势的表示,在通过下采样到同一长度后,采用 Procrustes 形状分析法^[19]计算不同轮廓之间的距离,该距离与人体轮廓的平移、旋转和放缩无关;之后对人体轮廓建立 k 近邻图,计算轮廓间的局部相似度矩阵,并采用局部保留投影(LPP)^[20]将人体轮廓非线性投影到一个低维流形上,作为最终姿势特征。其次是学习姿势字典并进行人体行为识别:对每一类行为,在训练视频的所有姿势帧上学习一个对应于该类的子姿势字典;学习完每一类的子字典后,整个姿势字典就可以通过串联所有类的子字典的方法来得到;在构建完姿势字典后,求解测试视频中的每一帧在姿势字典下的稀疏表示,及其对应每个类的重构误差;最后分别通过 max pooling 和 sum pooling 策略计算测试视频对应每个类的重构误差,将测试视频划分到重构误差最小的类。在 Weizmann 和 MuHAVi-MAS14 测试数据集上的实验结果证实了该方法的有效性,特别是在 MuHAVi-MAS14 数据集上的识别率对比已有经典方法的结果有了巨大的提升。

2 相关工作

2.1 基于姿势的人体运动识别

基于姿势的人体运动分析近期吸引了研究者的关注。人体姿势通常可以通过两种方式来表达,一种是使用骨架构型来描述;另一种是使用人体外轮廓的形状来表示。骨架模型通过对人体的主要部位(如头部、肩部、躯干和手臂等),或各个主要关节的几何构型信息(坐标、角度和角速度等)进行建模和跟踪来表征人体姿势。文献[21]采用一种“头部-肩部-上身”模型来检测视频中的人体,将躯干和双手的二维(2D)和3D坐标转换到一个规范化特征空间中作为人体姿势的描述,并用粒子滤波器进行跟踪。文献[22]提出了一种用人体的21个特征点和35条特征线来表示人体姿势的方法。文献[23]结合人体目标的骨架、颜色和动态信息来对人体各部位进行跟踪并估计3D人体姿势。文献[24]通过范例学习对深度图像提取出的人体骨架姿势进行校正。文献[25]根据扩散张量场下的近邻相似度对骨架特征进行分裂,并采用多核支持向量机进行人体行为识别。骨架模型的优点在于可以同时有效表达人体的静态姿势信息和动态运动信息,其缺点在于:

各节点的自由度较高造成描述特征具有较大的维数,使得人体检测、模型拟合和目标跟踪的运算量巨大,且在跟踪和拟合过程中容易受遮挡和噪声干扰,难以得到精确的人体构型信息。

人体剪影图像的轮廓曲线是另一种简单但是非常有效的人体姿势表示方式。文献[26]使用核主成分分析从人体剪影图像中提取姿势特征,并用因子条件随机场对人体行为进行识别。文献[27]学习具有判别性的关键轮廓作为人体的姿势特征。文献[28]在多摄像头拍摄环境下提取关键姿势序列作为人体姿势特征,并用动态时间卷曲进行识别。文献[29]从人体剪影中使用 Kendall 统计形状提取形状变化信息作为人体姿态特征,并分别使用线性动态系统模型和动态时间卷曲来度量形状序列的相似性。文献[30]提出了一种将轮廓特征与从生物启发特征相结合来描述人体行为的方法。轮廓模型的优点在于,剪影的提取通常比较简单,在背景固定或摄像机没有运动的情况下,通过背景建模或差分图像法就可以很容易地得到运动人体剪影。其缺点则是难以表示出各个身体部位的运动信息,在动态场景下往往难以实现有效的人体前景分割。本文采用基于轮廓的姿势特征,使用 Procrustes 形状分析来获得不同轮廓间的相似度。与已有的轮廓特征相比,提出方法采用局部保留投影对姿势特征进行维数约简,可以在保留人体姿势的局部流形结构的同时尽量提取子空间上的判别信息,以增加不同动作的人体姿势之间的相异性。

2.2 稀疏表示和字典学习

稀疏编码具有强大的信号表示能力,已经成为了信号处理和计算机视觉的一个研究热点。给定一个过完备的训练集 $\mathbf{X} \in \mathbf{R}^{m \times n}$, 对一个新来的测试样

本 \mathbf{y} , 它可以通过 l_0 正则化用训练样本的一个稀疏线性组合来表示。稀疏表示的目标是在求解线性表示 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ 或逼近最佳线性重构 $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_F^2$ 的同时使得表示系数 $\boldsymbol{\beta}$ 尽量稀疏。它可以表示为以下的最优化问题:

$$\begin{cases} \min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \\ \text{s. t. } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} \end{cases} \quad (1)$$

此优化问题可以转化为以下形式:

$$\begin{cases} \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_F^2 \\ \text{s. t. } \|\boldsymbol{\beta}\|_0 < C \end{cases} \quad (2)$$

其求解可以采用正交投影寻踪(OMP)算法^[31]来实现。

字典学习的目标是直接从数据 \mathbf{Y} 学习一个用于稀疏表示的过完备字典 \mathbf{X} 。它在保持系数矩阵 \mathbf{B} 的每一列稀疏的同时,从数据 \mathbf{Y} 中产生过完备字典 \mathbf{X} 。字典学习问题可以由以下的最优化问题来表示:

$$\begin{cases} \min_{\mathbf{X}, \mathbf{B}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 \\ \text{s. t. } \|\mathbf{b}_i\|_0 < C, \quad \forall i \end{cases} \quad (3)$$

式中 \mathbf{b}_i 表示系数矩阵 \mathbf{B} 的第 i 列。上面的优化问题可以通过 K-SVD 算法^[32]来迭代求解。

3 姿势特征提取

采用从人体剪影图像中提取的轮廓点序列作为人体姿势的表达。给定一个行为视频上的某一帧图像,首先使用背景建模法将图像前景中的人体区域检测出来^[33],得到一个人体剪影图像。图 1 给出了一些从人体动作视频中提取出的单帧人体剪影图像的示例。

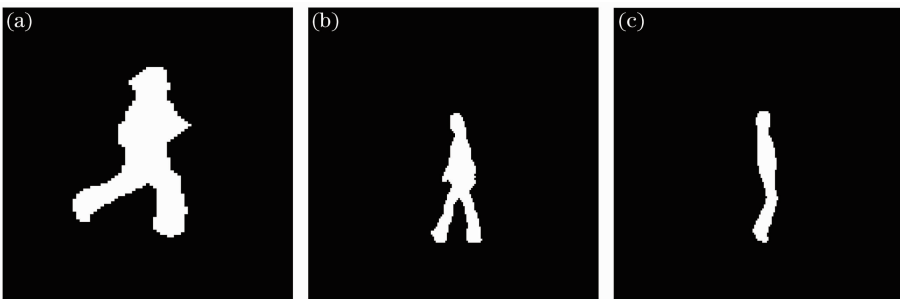


图 1 视频中提取出的不同动作人体剪影图像示例

Fig. 1 Example images of human silhouettes of different actions extracted from videos

对剪影图像提取闭合边界,可以得到一个由边界点序列表示的人体轮廓 $\{[x(i), y(i)]\}_{i=1}^n$ 。其中 $x(i)$ 和 $y(i)$ 分别是轮廓上第 i 个轮廓点的横轴和纵

轴坐标, n 是轮廓的长度。人体轮廓可以进一步用一个复数序列来表示:

$$\{z(i) = x(i) + jy(i)\}_{i=1}^n \quad (4)$$

为了方便对不同长度的轮廓进行匹配,对所有轮廓通过下采样归一化到相同长度。记轮廓的下采样长度为 L ,对轮廓的下采样方法如下:

$$\begin{cases} \hat{x}(i) = x \left(\left\lfloor i * \frac{n}{L} \right\rfloor \right) \\ \hat{y}(i) = y \left(\left\lfloor i * \frac{n}{L} \right\rfloor \right) \end{cases} \quad (5)$$

下采样后的所有轮廓具有相同长度。采用复数序列的形式来表示下采样后的轮廓特征:

$$\{\hat{z}(i) = \hat{x}(i) + j\hat{y}(i)\}_{i=1}^n. \quad (6)$$

本文提出了一种基于 Procrustes 形状分析和局部保留投影的姿势特征提取方法:提取人体行为视频中所有帧图像对应的人体轮廓,在通过下采样归一化到相同长度后,采用 Procrustes 形状分析法计算不同轮廓之间的距离。该距离与轮廓的平移、旋转和放缩无关。在得到不同轮廓间的距离后,建立一个近邻图,并计算轮廓间的相似度矩阵。然后采用 LPP 将轮廓非线性投影到一个低维流形上。低维流形上的数据表达将尽量保留原空间中的局部相似关系。将不同轮廓间的相似度矩阵经过局部保留投影后获得的低维数据作为最终的姿势特征。

3.1 Procrustes 形状分析

Procrustes 形状分析是一种统计形状分析方法,本文用它来估计不同轮廓之间的距离,建立具有

平移、放缩和旋转不变性的特征。将一个下采样后的人体轮廓记为 $z = [z(1), z(2), \dots, z(L)]^T$ 。给定两个人体轮廓 z_1 和 z_2 ,如果它们其中一个经过一系列若干步旋转、平移和放缩的组合操作后可以转化为另一方,则称两者在 Procrustes 拟合下是相等的,即:

$$\begin{cases} z_1 = \alpha \mathbf{I}_L + \beta z_2 \\ \beta = |\beta| \exp(j\angle_\beta) \end{cases}, \quad (7)$$

式中 α 和 β 是复数, $|\beta|$ 和 \angle_β 分别是 β 的幅度值和幅角主值, \mathbf{I}_L 是一个 L 维的单位向量。 $\alpha \mathbf{I}_L$ 对 z_2 进行平移操作, β 则将 z_2 放缩 $|\beta|$ 倍并旋转角度 \angle_β 。

对任意两个轮廓 z_1 和 z_2 ,它们之间的 Procrustes 距离定义为

$$d(z_1, z_2) = \min_{\alpha, \beta} \|z_1 - \alpha \mathbf{I}_L - \beta z_2\|. \quad (8)$$

通过求解最佳的平移因子 α 和旋转放缩因子 β 来确定两个轮廓间的最小匹配距离,得到的距离与轮廓本身的平移、旋转和放缩无关。

3.2 局部保持投影

在得到不同轮廓间的距离后,采用 k 近邻建图法建立一个无向连接图,图的顶点表示不同的轮廓,连接顶点的边的权重取为轮廓间的局部相似度。两个顶点 z_i 和 z_j 间的局部相似度定义如下:

$$w_{ij} = \begin{cases} \exp\left[-\frac{1}{2\epsilon^2} \|d(z_i, z_j)\|^2\right], & \text{if } z_i \text{ is among } k \text{ nearest neighbors of } z_j \text{ or } z_i \text{ is among } k \text{ nearest neighbors of } z_j \\ 0, & \text{otherwise} \end{cases}, \quad (9)$$

式中 w_{ij} 表示 z_i 和 z_j 的局部相似度, k 是近邻图的最大近邻数, $d(z_i, z_j)$ 表示顶点 z_i 和 z_j 之间的欧氏距离, ϵ 是高斯函数的参数。

接下来,使用 LPP 将姿势特征投影到一个低维流形上,并保留原数据的几何结构和局部相似关系。给定 L 维的原数据 \mathbf{X} 的局部相似度矩阵 \mathbf{W} ,基于图的非线性降维的目标是通过求解以下优化问题来获得最优的 l 维 ($l \leq L$) 表示 \mathbf{Y} :

$$\min_y \frac{1}{2} \sum_{i,j} w_{ij} \|y_i - y_j\|^2, \quad (10)$$

式中 y_i 是原数据 x_i 在低维空间中的表示。该目标的目的是使在高维空间中相似的两个点 x_i, x_j 所对应的低维空间表示 y_i, y_j 也是相似的。

LPP 采用线性投影。用 e 表示一个线性投影向量,则原数据 x_i 在低维空间上的投影为 $e^T x_i$ 。将

$y_i = e^T x_i$ 代入目标(10)式中,则有:

$$\begin{aligned} & \frac{1}{2} \sum_{i,j} w_{ij} \|y_i - y_j\|^2 = \\ & \frac{1}{2} \sum_{i,j} w_{ij} \|e^T y_i - e^T y_j\|^2 = \\ & \sum_i e^T x_i d_{ii} x_i^T e - \sum_{i,j} e^T x_i w_{ij} x_j^T e = \\ & e^T \mathbf{X} (\mathbf{D} - \mathbf{W}) \mathbf{X}^T e = e^T \mathbf{X} \mathbf{L}_p \mathbf{X}^T e, \end{aligned} \quad (11)$$

式中 \mathbf{D} 是一个对角阵,其对角线上的元素是 \mathbf{W} 对应列元素之和,即 $D_{ii} = \sum_j W_{ij}$ 。 $\mathbf{L}_p = \mathbf{D} - \mathbf{W}$ 是拉普拉斯矩阵, $\mathbf{X} = [x_1, x_2, \dots, x_n]$ 是高维空间的原数据矩阵。增加约束条件 $y^T \mathbf{D} y = 1$, 即 $e^T \mathbf{X} \mathbf{D} \mathbf{X}^T e = 1$, 则(11)式的优化问题退化为

$$\begin{cases} \min_e e^T \mathbf{X} \mathbf{L}_p \mathbf{X}^T e \\ \text{s. t. } e^T \mathbf{X} \mathbf{D} \mathbf{X}^T e = 1 \end{cases}, \quad (12)$$

该目标优化问题可以转化为一个求解广义特征值分解的问题：

$$\mathbf{X}\mathbf{L}_p\mathbf{X}^T\mathbf{e} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{e}. \quad (13)$$

记 $\mathbf{e}_0, \mathbf{e}_1, \dots, \mathbf{e}_{l-1}$ 是其前 l 个特征值 $\lambda_0 < \lambda_1 < \dots < \lambda_{l-1}$ 所对应的特征向量, 则局部保持投影所求得的最优投影即为 $\mathbf{E} = [\mathbf{e}_0, \mathbf{e}_1, \dots, \mathbf{e}_{l-1}]$ 。原数据 \mathbf{x}_i 的最优嵌入则为 $\mathbf{y}_i = \mathbf{E}^T \mathbf{x}_i$ 。

3.3 姿势特征提取

人体姿势特征的提取步骤如下：

1) 从人体行为视频中通过前景建模提取人体剪影图像, 并分割出人体轮廓。用轮廓点序列表示人体姿势, 按(5)式将其下采样到长度为 L , 得到 L 维数据 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$;

2) 根据(8)式计算所有轮廓间的两两距离。采用 k 近邻建图法建立无向连接图, 图的顶点表示不同的轮廓。根据(9)式得到顶点间的局部连接度矩阵 \mathbf{W} ;

3) 计算拉普拉斯矩阵 $\mathbf{L}_p = \mathbf{D} - \mathbf{W}$, 求解(13)式的广义特征值分解问题, 得到最优投影 $\mathbf{E} = [\mathbf{e}_0, \mathbf{e}_1, \dots, \mathbf{e}_{l-1}]$;

4) 得到最终的 l 维姿势特征 $\mathbf{Y} = \mathbf{e}^T \mathbf{X}$, 作为下一步姿势字典学习的输入特征;

按以上方法构建的姿势特征具有平移、旋转和尺度不变性, 在保留姿势的局部流形结构的同时尽量提取子空间上的判别信息。

4 姿势字典学习

最近, 稀疏表示和字典学习引起了研究者的广泛关注, 已经成为了信号处理和计算机视觉领域的一个热点。本文提出了一种新的人体行为识别框架, 通过从基于剪影表示的人体姿势中学习姿势字典, 来进行识别。在获得第 3.3 节中提取出的 l 维姿势特征 $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbf{R}^{n \times l}$ 后, 采用基于类的字典学习来对提取出的姿势特征进行稀疏表示, 并通过计算测试视频在已学习好的字典表示下的重构误差来进行行为视频的识别。

假设已有一个从若干个人体行为视频中的所有帧图像里提取出的 l 维局部姿势特征 $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbf{R}^{n \times l}$ 。给定数据 \mathbf{Y} , 可以通过求解以下的局部最优问题来学习过完备字典 \mathbf{D} 以使得 \mathbf{Y} 在该字典上的表达是稀疏的。

$$\begin{cases} (\mathbf{D}^*, \mathbf{A}^*) = \underset{\mathbf{D}, \mathbf{A}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{D}\mathbf{A}\|_F^2 \\ \text{s. t. } \|\mathbf{a}_i\|_0 \leq C, \quad \forall i \end{cases}, \quad (14)$$

式中目标项中的 $\|\cdot\|_F$ 表示矩阵的 Frobenius 范数, 该项用来保证样本在字典线性表示的重构误差足够小。 \mathbf{a}_i 是表示矩阵 \mathbf{A} 的第 i 列, 它代表样本 \mathbf{y}_i 在字典 \mathbf{D} 下的表示系数。在这里, 约束项中的 l_0 范数 $\|\cdot\|_0$ 的作用是强制使线性表达系数 \mathbf{a}_i 是稀疏的。参数 C 用来控制线性表示系数的稀疏程度。

假设人体行为的类别数为 K 。对于第 k 类的人体行为, 可以从训练帧中通过(14)式学习一个对应于该类的字典 \mathbf{D}_k 。采用 K-SVD 算法来求解(14)式中的优化问题, K-SVD 算法分为交替迭代的两步: 1) 在固定字典的情况下求稀疏系数; 2) 在固定稀疏系数的情况下求字典。其中在 K-SVD 算法的编码阶段采用 OMP 算法进行稀疏编码。学习完每一类的字典 $\mathbf{D}_k, k=1, 2, \dots, K$ 后, 整个姿势字典 \mathbf{D} 就可以通过串联所有类的字典的方法来获得, 即:

$$\mathbf{D} = [\mathbf{D}_1 | \mathbf{D}_2 | \dots | \mathbf{D}_K], \quad (15)$$

图 2 给出了在 Weizmann 数据库^[34]上学习到的各个类的姿势字典。其中每张子图各对应一个类的姿势字典, 共有 10 个类, 按次序分别为: bend, jack, jump, pjump, run, side, skip, walk, wavel 和 wave2。每类字典包含 20 个原子, L 设置为 100, C 为 10, (9)式中的参数 ϵ 取为 5, 建立 k 近邻图时的近邻个数为 5, 姿势特征的维数被约简到 20 维。

记测试视频为 $\check{\mathbf{Y}} = [\check{\mathbf{y}}_i]_{i=1}^T$ 。其中 T 是视频的帧数。在构建完姿势字典 \mathbf{D} 后, 对于测试视频中的某一帧 $\check{\mathbf{y}}_i$, 它在姿势字典下的稀疏表示可以通过求解以下的优化问题来获得。

$$\begin{cases} \mathbf{a}_i^* = \operatorname{argmin} \|\mathbf{a}_i\|_0 \\ \text{s. t. } \|\check{\mathbf{y}}_i - \mathbf{D}\mathbf{a}_i\|_F^2 \leq C \end{cases}, \quad (16)$$

采用 OMP 算法来求解这个优化问题。在得到测试帧 $\check{\mathbf{y}}_i$ 的稀疏表示后, 它对应第 k 类的重构误差 $r_k(\check{\mathbf{y}}_i)$ 可以通过下式来计算。

$$r_k(\check{\mathbf{y}}_i) = \|\check{\mathbf{y}}_i - \mathbf{D}\delta_k(\mathbf{a}_i^*)\|_2, \quad (17)$$

式中函数 $\delta_k(\mathbf{a}_i^*)$ 的作用是将系数向量 \mathbf{a}_i^* 中所有不与第 k 类对应的行置为 0。对系数向量 \mathbf{a}_i^* , 若其第 j 行 \mathbf{a}_{ij}^* 是对应第 k 类字典所属原子的表示系数, 则 $\delta_k(\mathbf{a}_i^*)$ 在 j 行上的值就等于原值 \mathbf{a}_{ij}^* , 否则就等于 0。

在得到测试视频所有帧对应每一类的重构误差后, 分别采用 max pooling 和 sum pooling 的策略来对测试视频进行基于最小重构误差的分类。

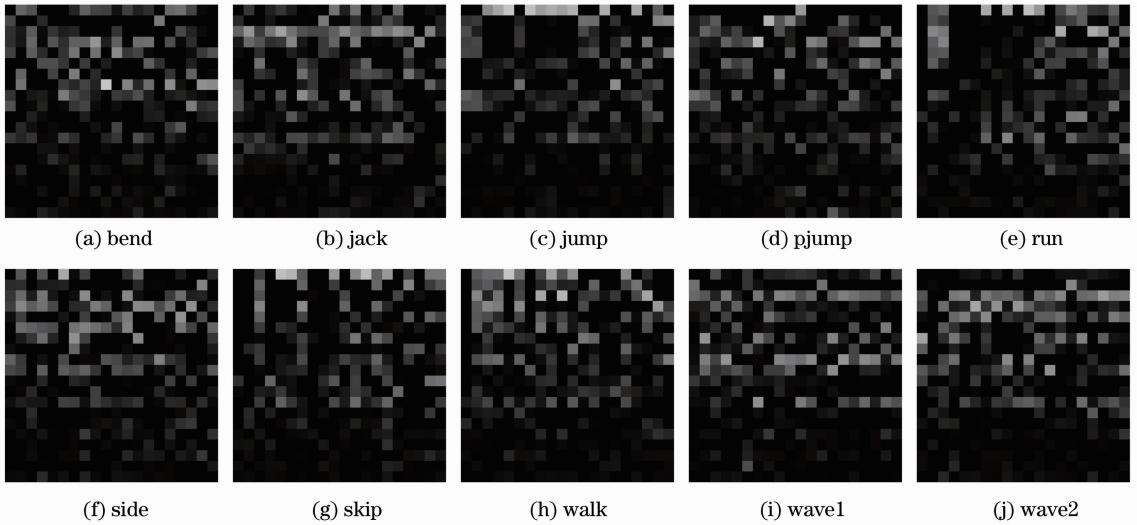


图 2 Weizmann 数据集学习到的各个类对应的姿势字典

Fig. 2 Learned pose dictionary of each class from Weizmann dataset

在 max pooling 策略中,首先对第 k 类,计算测试视频中所有帧对应该类的重构误差的最小值 $\min_{i=1}^T r_k(\tilde{y}_i)$,作为测试视频对应该类的重构误差。计算完测试视频对应所有类的重构误差后,测试视频的行为分类结果 \hat{k} 就取为重构误差最小的类。

$$\hat{k} = \operatorname{argmin}_k \min_{i=1}^T r_k(\tilde{y}_i). \quad (18)$$

另一方面,在 sum pooling 策略中,对第 k 类,计算测试视频中所有帧对应该类的重构误差的和 $\sum_{i=1}^T r_k(\tilde{y}_i)$,作为测试视频对应该类的重构误差。计算完测试视频对应所有类的重构误差后,测试视频的行为分类结果 \hat{k} 就取为重构误差最小的类。

$$\hat{k} = \operatorname{argmin}_k \sum_{i=1}^T r_k(\tilde{y}_i). \quad (19)$$

基于姿势字典学习框架的人体行为识别过程如图 3 所示,其具体步骤总结如下:

- 1) 按第 3.3 节介绍的步骤,提取姿势特征;
- 2) 在每一类行为的姿势特征上通过求解(14)式各建立一个子姿势字典;
- 3) 根据(14)式,将所有类的子姿势字典串联起来,组成整个姿势字典;
- 4) 根据(16)式,计算测试视频所有帧在姿势字典下的稀疏表示;
- 5) 根据(17)式,计算测试视频所有帧对应每个类的重构误差;
- 6) 分别根据(18)式和(19)式对测试视频进行分类。

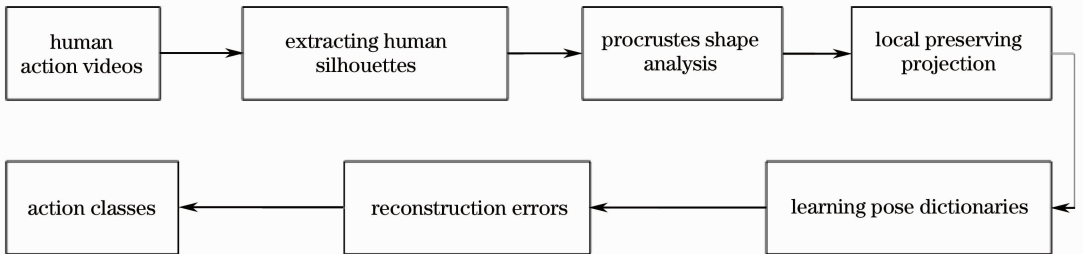


图 3 提出算法的框架

Fig. 3 Proposed algorithm framework

5 实验结果与分析

5.1 Weizmann 数据库

Weizmann 数据库^[34]是人体行为识别中常用的一个单视角户外背景的数据集,本文采用它作为实验数据集,来验证提出方法的有效性。Weizmann

数据库共有 10 种人体行为,包括 bend, jack, jump, pjump, run, side, skip, walk, wave1 和 wave2 等。每种动作分别由 9 个不同的人完成。数据库中一共包含 93 个不同的视频,每个视频的分辨率为 144 pixel×180 pixel,帧速率为 25 frame/s。

在实验中,默认设置 L 为 100, C 为 10,对每一类行为,对应该类的子字典的原子个数取为 20,(9)式中的参数 ϵ 取为 5,建立 k 近邻图时的近邻个数为 5,LPP 降维后的维数为 10。采用留一法交叉验证来对样本进行识别并统计识别率。在 Weizmann 数据库上对提出方法进行测试,得到 max pooling 的识别率为 97.85%,sum pooling 的识别率为 97.85%,分类结果的混淆矩阵分别如图 4、5 所示。还将提出方法的测试结果与其他经典方法进行比较,比较结果如表 1 所示。注意文献[6]原来采用了将数据集划分为固定训练集和测试集的方法进行试验,而文献[8]原本仅使用了 92 个视频进行测试,这里统一采用本文的实验设置对文献[6]和[8]的方法进行测试。实验结果表明提出方法的识别结果优于所列的其他经典方法,证实了提出方法的有效性。

表 1 各方法在 Weizmann 库上实验结果比较

Table 1 Comparison of proposed method with other methods on Weizmann dataset

Methods	Accuray /%
Ref. [6]	89.25
Ref. [7]	90.00
Ref. [8]	89.25
Ref. [9]	84.30
Ref. [10]	82.64
Ref. [35]	90.32
Proposed method (max pooling)	97.85
Proposed method (sum pooling)	97.85

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1	0	0	0	0	0	0	0	0	0
jack	0	1	0	0	0	0	0	0	0	0
jump	0	0	1	0	0	0	0	0	0	0
pjump	0	0	0	1	0	0	0	0	0	0
run	0	0	0	0	1	0	0	0	0	0
side	0	0	0	0	0	1	0	0	0	0
skip	0	0	0	0	0	0	1	0	0	0
walk	0	0	0	0	0	0	0	1	0	0
wave1	0	0.11	0	0	0	0	0	0	0.89	0
wave2	0	0.11	0	0	0	0	0	0	0	0.89

图 4 提出方法(max pooling)在 Weizmann 数据库上分类结果的混淆矩阵

Fig. 4 Confusion matrix of proposed method (max pooling) on Weizmann dataset

另外还进行了提出方法在 Weizmann 数据库上的参数实验。在固定其他参数为默认值时,比较提出方法的识别率与控制稀疏度的参数 C 的关系。

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1	0	0	0	0	0	0	0	0	0
jack	0	1	0	0	0	0	0	0	0	0
jump	0	0	1	0	0	0	0	0	0	0
pjump	0	0	0	1	0	0	0	0	0	0
run	0	0	0	0	1	0	0	0	0	0
side	0	0	0	0	0	1	0	0	0	0
skip	0	0	0	0	0	0	1	0	0	0
walk	0	0	0	0	0	0	0	1	0	0
wave1	0	0.11	0	0	0	0	0	0	0.89	0
wave2	0	0.11	0	0	0	0	0	0	0	0.89

图 5 提出方法(sum pooling)在 Weizmann 数据库上分类结果的混淆矩阵

Fig. 5 Confusion matrix of proposed method (sum pooling) on Weizmann dataset

实验结果如图 6 和图 7 所示,分别对应采用 max pooling 和 sum pooling 策略时本文方法识别率与 C 的关系。实验结果表明,在 C 取不同值的时候,提出方法都表现出不错的识别能力。且提出方法的识别率对参数 C 不敏感,在实践中容易寻找有效的选值。同样的,在固定其他参数不变时,比较提出方法的识别率与每类字典原子个数的关系。实验结果如图 8 和图 9 所示,分别对应采用 max pooling 和 sum pooling 策略时提出方法识别率与每类字典原子个数的关系,实验结果表明,当每类字典原子个数取为 20 时识别率最高。此外,还在固定其他参数的情况下比较了提出方法识别率与 LPP 降维后维数的关系。实验结果如图 10 和图 11 所示,分别对应采用 max pooling 和 sum pooling 策略时提出方

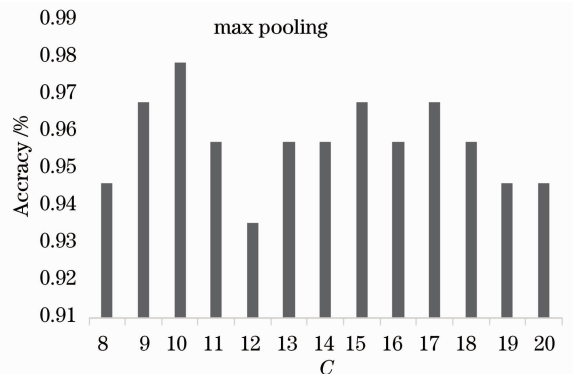


图 6 提出方法(max pooling)在 Weizmann 数据库上识别率与参数 C 的关系

Fig. 6 Relation between accuracy of proposed method (max pooling) and parameter C on Weizmann dataset

法识别率与 LPP 降维后维数的关系,实验结果表明,当 LPP 降维后维数取为 10~20 时可以得到最佳的识别率。以上参数实验说明提出方法在大部分参数下都具有很好的识别效果,对参数选择具有稳健性。

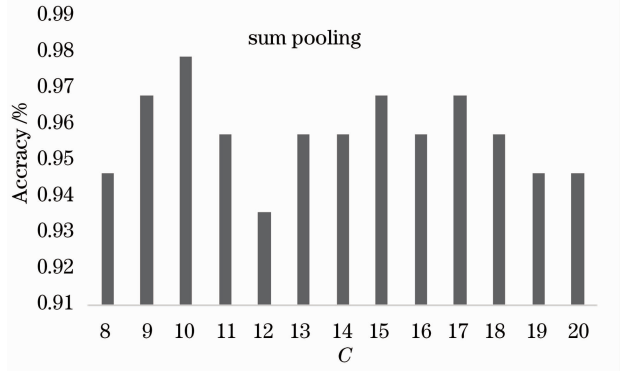


图 7 提出方法(sum pooling)在 Weizmann 数据库上识别率与参数 C 的关系

Fig. 7 Relation between accuracy of proposed method (sum pooling) and parameter C on Weizmann dataset

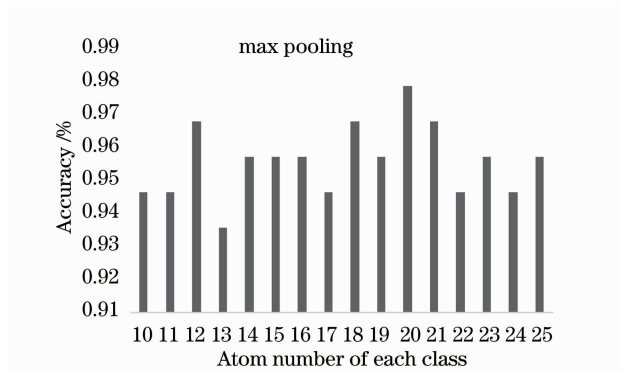


图 8 提出方法(max pooling)在 Weizmann 数据库上识别率与每类原子个数的关系

Fig. 8 Relation between accuracy of proposed method (max pooling) and atom number of each class on Weizmann dataset

5.2 MuHAVi-MAS14 数据库

MuHAVi-MAS14 数据库^[36]是一个更具挑战性的多视角数据集。将其作为本文方法的另一个验证数据集。MuHAVi-MAS14 数据集包含 14 个人体行为,分别是 Collapse left, Collapse right, Guard to kick, Guard to punch, Kick right, Punch right, Run left to right, Run right to left, Stand up left, Stand up right, Turn back left, Turn back right, Walk left to right 和 Walk right to left 等。所有行为分别由 2 个不同的人完成,并分别在 0°和 45°这两个不同视角下进行拍摄。所有行为在每个视

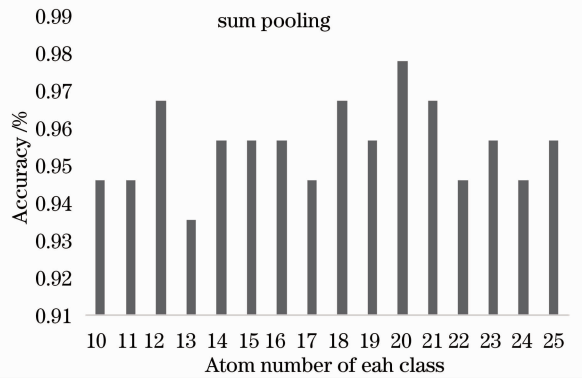


图 9 提出方法(sum pooling)在 Weizmann 数据库上识别率与每类原子个数的关系

Fig. 9 Relation between accuracy of proposed method (sum pooling) and atom number of each class on Weizmann dataset

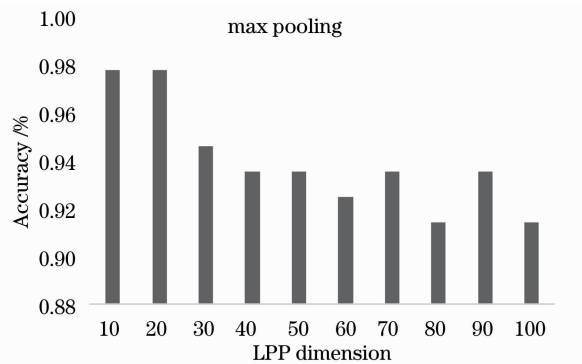


图 10 提出方法(max pooling)在 Weizmann 数据库上识别率与降维后维数的关系

Fig. 10 Relation between accuracy of proposed method (max pooling) and LPP dimension on Weizmann dataset

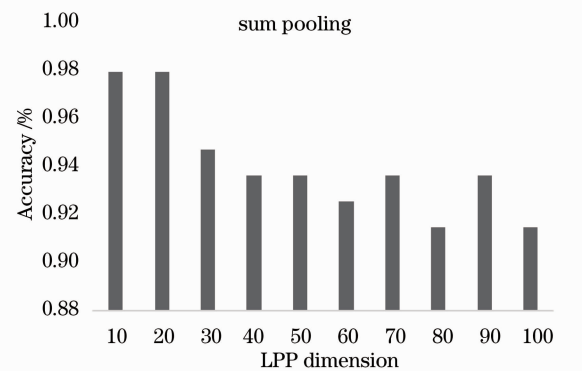


图 11 提出方法(sum pooling)在 Weizmann 数据库上识别率与降维后维数的关系

Fig. 11 Relation between accuracy of proposed method (sum pooling) and LPP dimension on Weizmann dataset

角下各拍摄了 68 个视频,共计 136 个视频。每个视

频的分辨率为 720 pixel×576 pixel, 视频内容中背景较为复杂, 并伴有光照变化。

在实验中, 默认设置 L 为 100, C 为 14, 对每一类行为, 对应该类的子字典的原子个数取为 11, (12) 式中的参数 ϵ 取为 5, 建立 k 近邻图时的近邻个数为 5, LPP 降维后的维数为 20。采用留一法来对样本进行识别并统计识别率。在 MuHAVi-MAS14 数据库上对提出方法进行测试, 得到 max pooling 的识别率为 97.79%, sum pooling 的识别率为 98.53%, 分类结果的混淆矩阵分别如图 12 和图 13 所示。还将提出方法的测试结果与其他经典方法进行比较, 比较结果如表 2 所示。实验结果表明, 提出方法在 MuHAVi-MAS14 数据集上的识别率结果对比已有经典方法的结果有巨大的提升, 证明了提出方法的优越性。

表 2 各方法在 MuHAVi-MAS14 库上实验结果比较
Table 2 Comparison of proposed method with other methods on MuHAVi-MAS14 dataset

Methods	Accuracy / %
Ref. [21]	86.03
Ref. [25]	82.35
Proposed method (max pooling)	97.79
Proposed method (sum pooling)	98.53

	CL	CR	GK	GP	KR	PR	RLR	RRL	SL	SR	TL	TR	WLR	WRL
CL	8/8													
CR		8/8												
GK			16/16											
GP				16/16										
KR					16/16									
PR						15/16				1/16				
RLR							8/8							
RRL								7/8						1/8
SL									4/4					
SR										8/8				
TL											4/4			
TR												8/8		
WLR													8/8	
WRL														8/8

图 12 提出方法(max pooling)在 MuHAVi-MAS14 数据库上分类结果的混淆矩阵

Fig. 12 Confusion matrix of proposed method (max pooling) on MuHAVi-MAS14 dataset

另外还进行了提出方法在 MuHAVi-MAS14 数据库上的参数实验。在固定其他参数为默认值时, 比较本文的识别率与控制稀疏度的参数 C 的关系。实验结果如图 14 和图 15 所示, 分别对应采用

	CL	CR	GK	GP	KR	PR	RLR	RRL	SL	SR	TL	TR	WLR	WRL
CL	8/8													
CR		8/8												
GK			16/16											
GP				16/16										
KR					16/16									
PR						15/16								1/16
RLR							8/8							
RRL								7/8						1/8
SL									4/4					
SR										8/8				
TL											4/4			
TR												8/8		
WLR													8/8	
WRL														8/8

图 13 提出方法(sum pooling)在 MuHAVi-MAS14 数据库上分类结果的混淆矩阵

Fig. 13 Confusion matrix of proposed method (sum pooling) on MuHAVi-MAS14 dataset

max pooling 和 sum pooling 策略时提出方法识别率与 C 的关系。实验结果表明, 在 C 取不同值的时候, 提出方法都表现出不错的识别能力, 并且提出方法的识别率对参数 C 变化不敏感, 结果稳定。同样的, 在固定其他参数不变时, 比较提出方法的识别率与每类字典原子个数的关系。实验结果如图 16 和图 17 所示, 分别对应采用 max pooling 和 sum pooling 策略时本文方法识别率与每类字典原子个数的关系, 实验结果表明, 当每类字典原子个数取为 11 时识别率最高。此外, 还在固定其他参数的情况下比较了提出方法识别率与 LPP 降维后维数的关

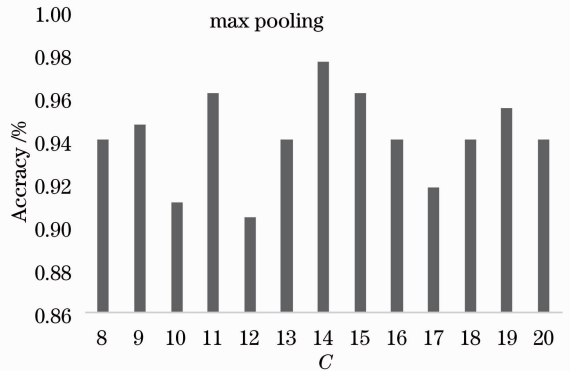


图 14 提出方法(max pooling)在 MuHAVi-MAS14 数据库上识别率与参数 C 的关系

Fig. 14 Relation between accuracy of proposed method (sum pooling) and parameter C on MuHAVi-MAS14 dataset

系。实验结果如图 18 和图 19 所示, 分别对应采用 max pooling 和 sum pooling 策略时提出方法识别

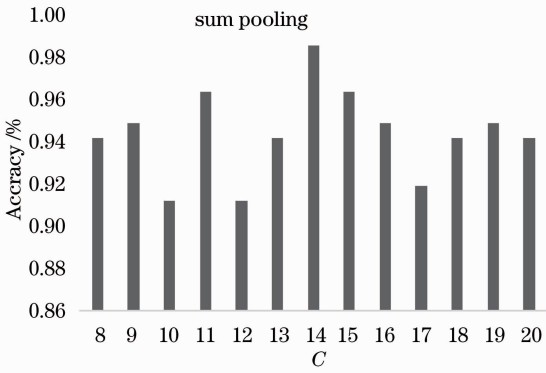


图 15 提出方法(sum pooling)在 MuHAVi-MAS14 数据库上识别率与参数 C 的关系

Fig. 15 Relation between accuracy of proposed method (sum pooling) and parameter C on MuHAVi-MAS14 dataset

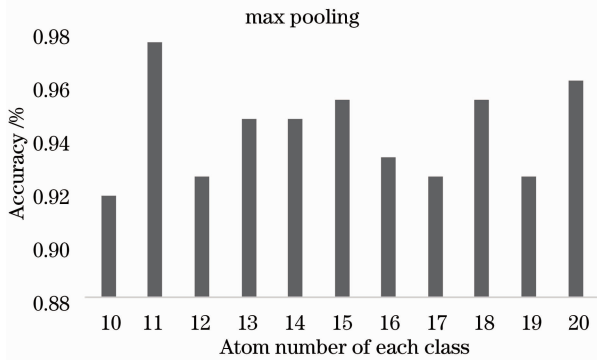


图 16 提出方法(max pooling)在 MuHAVi-MAS14 数据库上识别率与每类原子个数的关系

Fig. 16 Relation between accuracy of proposed method (max pooling) and atom number of each class on MuHAVi-MAS14 dataset

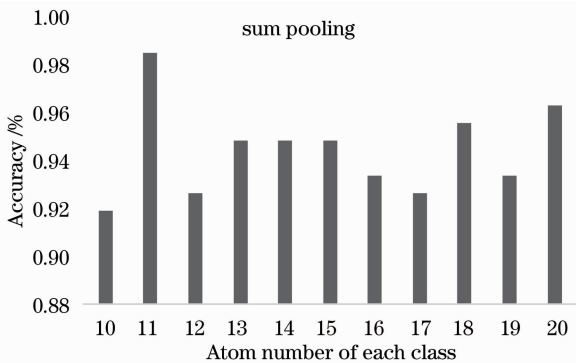


图 17 提出方法(sum pooling)在 MuHAVi-MAS14 数据库上识别率与每类原子个数的关系

Fig. 17 Relation between accuracy of proposed method (sum pooling) and atom number of each class on MuHAVi-MAS14 dataset

率与 LPP 降维后维数的关系,实验结果表明,当 LPP 降维后维数取为 20 时可以得到最佳的识别率。以上参数实验说明提出方法在大部分参数下都具有很好的识别效果,对参数选择具有稳健性。

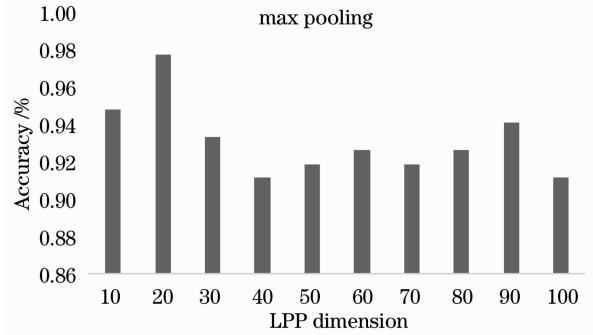


图 18 提出方法(max pooling)在 MuHAVi-MAS14 数据库上识别率与降维后维数的关系

Fig. 18 Relation between accuracy of proposed method (max pooling) and LPP dimension on MuHAVi-MAS14 dataset

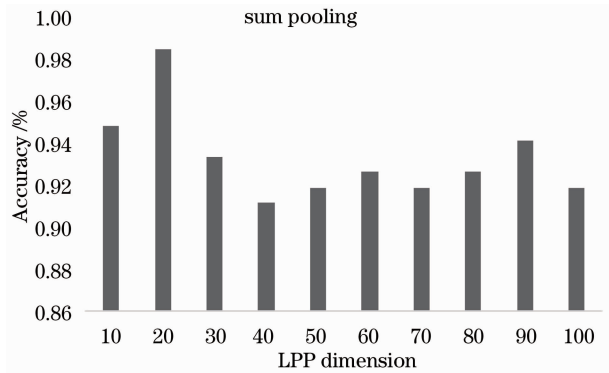


图 19 提出方法(sum pooling)在 MuHAVi-MAS14 数据库上识别率与降维后维数的关系

Fig. 19 Relation between accuracy of proposed method (sum pooling) and LPP dimension on MuHAVi-MAS14 dataset

6 结 论

提出了一种基于人体形状表示的姿势字典学习框架来进行人体行为识别。提出了一种基于 Procrustes 形状分析和局部保留投影的姿势特征提取方法,该特征具有平移、旋转和尺度不变性。之后对人体轮廓学习姿势字典并进行人体行为识别。对每一类,在训练视频的所有姿势帧上学习一个对应于该类的字典;学习完每一类的字典后,整个姿势字典就可以通过串联所有类的字典的方法来得到。在构建完姿势字典后,求解测试视频对应每个类的重构误差,将测试视频划分到重构误差最小的类。在

测试数据库上的实验结果证实了此方法的有效性。特别是在 MuHAVi-MAS14 数据集上的识别率比已有经典方法的结果有巨大的提升。

参 考 文 献

- 1 A F Bobick, J W Davis. The recognition of human movement using temporal templates [J]. *IEEE Trans Pattern Anal Mach Intell*, 2001, 23(3): 257–267.
- 2 T Xiang, S Gong. Beyond tracking: modelling activity and understanding behaviour [J]. *International Journal of Computer Vision*, 2006, 67(1): 21–51.
- 3 V Chandrashekhar, K Venkatesh. Action energy images for reliable human action recognition [C]. *New Delhi; Proc of Asian Symp on Information Display*, 2006. 484–487.
- 4 L Wang, D Suter. Informative shape representations for human action recognition [C]. *Hong Kong; 18th International Conference on Pattern Recognition*, 2006. 1266–1269.
- 5 D Weinland, R Ronfard, E Boyer. Free viewpoint action recognition using motion history volumes [J]. *Computer Vision and Image Understanding*, 2007, 104(2-3): 249–257.
- 6 J X Cai, G C Feng, X Tang. Human action recognition using oriented holistic feature [C]. *Melbourne; 20th IEEE International Conference on Image Processing*, 2013. 2420–2424.
- 7 C Niebles, H C Wang, F F Li. Unsupervised learning of human action categories using spatial-temporal words [J]. *International Journal of Computer Vision*, 2008, 79(3): 299–318.
- 8 J G Liu, S Ali, M Shah. Recognizing human actions using multiple features [C]. *Anchorage; Computer Vision and Pattern Recognition*, 2008. 1–8.
- 9 K Alexander, M Marcin, S Cordelia. A spatio-temporal descriptor based on 3D-gradients [C]. *Leeds; British Machine Vision Conference*, 2008. 995–1004.
- 10 P Scovanner, A Saad, S Mubarak. A 3-dimensional sift descriptor and its application to action recognition [C]. *Beijing; Proceedings of the 15th international conference on Multimedia*, 2007. 357–360.
- 11 Jin Biao, Hu Wenlong, Wang Hongqi. Moving-objects interaction recognition based on the spatial-temporal semantic information [J]. *Acta Optica Sinica*, 2012, 32(5): 0515002.
金 标, 胡文龙, 王宏琦. 基于时空语义信息的视频运动目标交互行为识别方法[J]. *光学学报*, 2012, 32(5): 0515002.
- 12 Zhang Hui, Xu Hui, Lin Liangkui. Super-resolution method of closely spaced objects based on sparse reconstruction using single frame infrared data [J]. *Acta Optica Sinica*, 2013, 33(4): 0411001.
张 慧, 徐 晖, 林两魁. 基于稀疏重构的空间邻近目标红外单帧图像超分辨率方法[J]. *光学学报*, 2013, 33(4): 0411001.
- 13 Yin Wen, Li Yuanxiang, Zhou Zeming, *et al.*. Remote sensing image fusion based on sparse representation [J]. *Acta Optica Sinica*, 2013, 33(4): 0428003.
尹 雯, 李元祥, 周则明, 刘世前. 基于稀疏表示的遥感图像融合方法[J]. *光学学报*, 2013, 33(4): 0428003.
- 14 Song Lin, Cheng Yongmei, Zhao Yongqiang. Hyper-spectrum classification based on sparse representation model and auto-regressive model [J]. *Acta Optica Sinica*, 2012, 32(3): 0330003.
宋 琳, 程咏梅, 赵永强. 基于稀疏表示模型和自回归模型的高光谱分类[J]. *光学学报*, 2012, 32(3): 0330003.
- 15 Li Mengjie, Li Jing, Sun Yi. Sparse angular differential phase-contrast computed tomography reconstruction using L^1 -norm and curvelet constraints [J]. *Acta Optica Sinica*, 2014, 34(1): 0111003.
李梦婕, 李 镜, 孙 怡. 基于 L^1 范数和曲波系数双约束的稀疏角度微分相位衬度计算机层析成像重建方法[J]. *光学学报*, 2014, 34(1): 0111003.
- 16 Li Jing, Sun Yi. L^1 -norm-based differential phase-contrast computerized tomography reconstruction algorithm with sparse angular resolution [J]. *Acta Optica Sinica*, 2012, 32(3): 0311002.
李 镜, 孙 怡. 基于 L^1 范数的微分相位衬度 CT 稀疏角度重建算法[J]. *光学学报*, 2012, 32(3): 0311002.
- 17 T Guha, R Ward. Learning sparse representations for human action recognition [J]. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(8): 1576–1588.
- 18 A Castrodad, G Sapiro. Sparse modeling of human actions from motion imagery [J]. *International Journal of Computer Vision*, 2012, 100(1): 1–15.
- 19 I L Dryden, K V Mardia. *Statistical Shape Analysis* [M]. *New York; John Wiley & Sons Press*, 1998. 87–99.
- 20 Guo Jinyu, Yuan Weiqi. Palmprint recognition based on locality preserving projection [J]. *Acta Optica Sinica*, 2008, 28(10): 1920–1924.
郭金玉, 苑玮琦. 基于局部保持投影的掌纹识别[J]. *光学学报*, 2008, 28(10): 1920–1924.
- 21 F Huo, E Hendriks, P Paclik, *et al.*. Markerless human motion capture and pose recognition [C]. *IEEE Workshop on Image Analysis for Multimedia Interactive Services*, 2009. 13–16.
- 22 I Leong, J Fang, M Tsai. Automatic body feature extraction from a marker-less scanned human body [J]. *Comput-Aided Des*, 2007, 39(7): 568–582.
- 23 S Ke, J Hwang, K Lan, *et al.*. View-invariant 3D human body pose reconstruction using a monocular video camera [C]. *IEEE International Conference on Distributed Smart Cameras*, 2011. 1–6.
- 24 W Shen, K Deng, X Bai, *et al.*. Exemplar-based human action pose correction and tagging [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1784–1791.
- 25 M Y Sang, K Arjan. Human action recognition based on skeleton splitting [J]. *Expert Systems with Applications*, 2013, 40(17): 6848–6855.
- 26 L Wang, D Suter. Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1–8.
- 27 S Cheema, A Eweawi, C Thureau, *et al.*. Action recognition by learning discriminative key poses [C]. *IEEE International Conference on Computer Vision Workshops*, 2011. 1302–1309.
- 28 A A Chaaoui, P P Climent, R F Florez. Silhouette-based human action recognition using sequences of key poses [J]. *Pattern Recognition Letters*, 2013, 34(15): 1799–1807.
- 29 A Veeraraghavan, A K R Chowdhury, R Chellappa. Matching shape sequences in video with applications in human movement analysis [J]. *IEEE Transactions Pattern Anal Mach Intell*, 2005, 27(12): 1896–1909.
- 30 K Schindler, L V Gool. Action Snippets: How many frames does human action recognition require? [C]. *Anchorage; Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1–8.
- 31 Y C Pati, R Rezaifar, P S Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition [C]. *IEEE Conference on Signals, Systems and Computers*, 1993. 40–44.
- 32 R Rubinstein, T Peleg, M Elad. Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model [J]. *IEEE Transactions on Signal Processing*, 2013, 61(3): 661–677.
- 33 Ming Ying, Jiang Jingjue. Background modeling and moving-objects detection based on cauchy distribution for video sequence

- [J]. *Acta Optica Sinica*, 2008, 28(3): 587–592.
明 英, 蒋晶珏. 基于柯西分布的视频图像序列背景建模和运动目标检测[J]. *光学学报*, 2008, 28(3): 587–592.
- 34 B Mosh, G Lena, S Eli, *et al.*. Actions as space-time shapes [C]. *IEEE International Conference on Computer Vision*, 2005. 1395–1402.
- 35 Q Zhao, H S Horace. Unsupervised approximate-semantic vocabulary learning for human action and video classification [J]. *Pattern Recognition Letters*, 2013, 34(15): 1870–1878.
- 36 S Singh, S Velatin, H Ragheb. MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods [C]. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010. 48–55.

栏目编辑: 张浩佳