ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 1,Jan. 2019,pp. 68－79
DOI:10. 16337/j. 1004-9037. 2019. 01. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

动态多视角复杂3D人体行为数据库及行为识别

王永雄　李　璇　李梁华

(上海理工大学光电信息与计算机工程学院,上海,200093)

摘　要: 提供了一个较大规模的基于RGB-D摄像机的人体复杂行为数据库DMV (Dynamic and multi-view) action3D,从2个固定视角和一台移动机器人动态视角录制人体行为。数据库现有31个不同的行为类,包括日常行为、交互行为和异常行为类等三大类动作,收集了超过620个行为视频约60万帧彩色图像和深度图像,为机器人寻找最佳视角提供了可供验证的数据库。为验证数据集的可靠性和实用性,本文采取4种方法进行人体行为识别,分别是基于关节点信息特征、基于卷积神经网络(Convolutional neural networks, CNN)和条件随机场(Conditional random field, CRF)结合的CRFasRNN方法提取的彩色图像HOG3D特征,然后采用支持向量机(Support vector machine,SVM)方法进行了人体行为识别;基于3维卷积网络(C3D)和3D密集连接残差网络提取时空特征,通过softmax层以预测动作标签。实验结果表明:DMV action3D人体行为数据库由于场景多变、动作复杂等特点,识别的难度也大幅增大。DMV action3D数据集对于研究真实环境下的人体行为具有较大的优势,为服务机器人识别真实环境下的人体行为提供了一个较佳的资源。

关键词: 人体行为识别;3D数据库;多视角

中图分类号: TP181　　**文献标志码:** A

Dynamic and Multi-view Complicated 3D Database of Human Activity and Activity Recognition

Wang Yongxiong, Li Xuan, Li Lianghua

(School of Optional-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China)

Abstract: In view of the fact that the existing 3D databases have fewer behavioral categories, few interactions with scenes, and single and fixed perspectives, this paper provides a large-scale human body complex behavior database DMV action3D based on RGB-D cameras, from two fixed perspectives and a mobile robot records human behavior from a dynamic perspective. There are 31 different behavioral classes in the database, including daily behaviors, interaction behaviors, and abnormal behaviors angles. Validated database collected more than 620 behavioral videos, about 600 000 frames of color images and depth images, to provide robots with optimal viewing. In order to verify the reliability and practicability of the data set, this paper adopts four methods for human behavior recognition, which are HOG3D features extracted by CRFasRNN method based on the information features of customs nodes, CNN and

基金项目: 国家自然科学基金(61673276,61603255)资助项目;上海青年科技英才扬帆计划(17YF1427000)资助项目。
收稿日期: 2018-06-23;**修订日期:** 2018-12-05

conditional random field (CRF), and then adopts SVM method for human behavior recognition. Spatial and temporal characteristics are extracted based on the three-dimensional convolutional network (C3D) and the 3D dense connection residual network, and the motion tags are predicted by softmax layer. The results show that DMV action3D human behavior database is characterized by a variety of scenes and complicated movements, and the difficulty of recognition is greatly increased. The DMV action 3D database has great advantages for studying human behavior in real environments, and provides a better resource for serving robots to recognize human behavior in real environments.

Key words: human activity recognition; 3D dataset; multi-view

引　言

人体行为识别一直是计算机视觉领域的研究热点之一,基于视觉的人体行为识别研究在人机交互、视频检索以及智能服务等方面都有着非常重要的应用价值,也是人机共融技术的主要难点和瓶颈。立体视觉和深度传感器的发展,为获得3D场景和人体动作提供了多样化的研究手段和方法。近年来,基于RGB+D的人体动作识别研究越来越多[1-3],但是依然存在诸多问题,距离实际应用还是有一定距离。作为研究人体动作识别所需要的人体动作识别数据库,存在着数量较少、与真实场景差距较大等问题,特别是基于RGB+D的数据库更少。

常见的人体行为数据库包括:KTH,MSR Action3D,MSR DailyActivity、康奈尔大学的CAD-60和CAD-120等。尽管这些数据库为人体行为识别研究提供了很多的便利,然而还有很多地方值得改进,例如人体行为识别的训练样本不足、动作类别少和场景单调等诸多问题。面临的问题主要包括两个方面:(1)实际中的视觉观测数据具有不确定性。各种遮挡现象、人体自身的遮挡、视角和光线以及人体的柔性等原因会引起错误估计和识别,目前还不能很好地解决人体的遮挡问题。现有的大部分数据库是基于固定的视角,训练和识别是基于某一个特定视角完成的,一旦视角变化,识别率将会大幅度下降,因此难以应用到实际当中;(2)人的行为具有较高的多歧义性。不同人的相同姿态或行为可能会有不同的含义,同一人的相同姿态或行为在不同情况、不同时间下也会有很多的含义。在前面提到的大部分数据库忽略了与人交互的物体和场景信息,或者数据库中的人体行为与环境、物体的交互不多,而这对人体行为的识别有着不可忽视的影响,降低了推理的准确性。例如,手持书本和手持刀是有明显的行为意图差异。人体行为与测试对象的年龄、性别和行为习惯有关,因此数据库中测试样本的多样性决定了数据库的适用性。同时,数据库人体行为的复杂性和多变性会使识别算法的复杂度大幅增加。

1　相关工作

人体行为识别和数据库是两个紧密联系的部分,录制的数据库为行为识别方法的实现提供了途径,不同行为识别方法为检验数据库的合理性提供了依据和指标,本节主要介绍现有的常用人体行为识别数据库和部分人体行为识别的方法。

1.1　现存数据库简介

随着传感器及硬件的发展,出现了许多应用于视觉研究的设备可以方便快速的获取彩色图像、深度信息和骨架信息。数据库从最初的黑白视频数据发展到拥有深度图像彩色图像和三维空间数据的大型人体行为数据库,这些人体行为识别数据库为人体行为识别的研究提供了可靠的数据来源,提高了研究者的效率。为分析不同数据库的特点和优劣,各个数据库对比情况如表1所示。KTH数据库由皇家理工学院的Schldt等[4]提出。KTH数据库包含了4个场景下的6种不同的行为,每一种动作行为由25人录制完成。该数据库一共包含了2 391个样本,每一帧图像的分辨率为160像素×120像素。KTH

的缺点在于仅仅提供了具有灰度信息的视频序列,场景中背景相对静止,摄像机的位置也是相对固定的。Weizman数据库是以色列Weizman科学研究所Gorelick等[5]所提供的人体行为识别数据库。Weizman数据库包含了10中不同的行为,这些行为由9个人录制,每一帧的分辨率为144像素×180像素。Weizman数据库与KTH相似,是经典的人体行为识别数据库,都只提供了灰度视频序列,分辨率较低,人体动作选取简单,相机、视角和背景都是相对静止的。

UCF运动数据[6-8]含有150段关于体育的视频序列,包含了13种不同体育动作项目。该数据库在人物、视角、光照和背景等方面有较大差异,提供了彩色视频序列,相机运动和背景变换使得数据库具有一定挑战性。

INRIA XMAS数据库是由Weinland等[9]提出,数据库从5个视角获得11个人的14种行为动作,室内4个方向和头顶一个方向共安装了5个摄像头,数据库包含14种行为。该数据库提供了人体轮廓和体积信息,5个视角的光照和背景基本不变,录制视频的人可随意选择位置和方向,存在较大的不统一性,这些都使得数据库的难度上升,适合从单视角和多视角两个方面研究。

表1 数据库对比

Tab. 1 Databases comparison

数据库	RGB	Depth	Skelection	静态	动态	动作分类	样本
KTH	√			1		6	2 391
Weizman	√			1		10	90
UCF	√			1		10	182
INRIA XMAS	√			5		14	2 310
MSR Action3D	√	√	√	1		20	567
MSR DailyActivity	√	√	√	1		16	320
CAD-60/CAD-120	√	√	√	1		12/10	60/120
DMV action3D	√	√	√	2	1	31	620

MSR Action3D数据库[10]是利用深度信息进行人体行为分析研究的最早数据集之一,该数据库的动作仅限于游戏动作包含了20组不同的行为,每种行为由10个人完成。每帧图像分辨率为320像素×240像素,该数据库包含彩色图深度信息和骨架信息等。

MSR DailyActivity数据库[11-12]是最具有挑战性的数据库之一,数据库中所有数据均是由Kinect采集。数据库一共采集了16个人类日常行为动作,共有10位实验者参与数据采集,每人依次完成这16个动作。320个日常运动样本的动作和背景具有较大的差异性,然而此数据库的局限性是样本数量较少,相机视角固定。

CAD-60和CAD-120数据库由康奈尔大学的Sung等[13-15]建立。CAD-60数据库所有动作视频均由Kinect传感器采集,该数据库中有60个RGB-D视频以及其对应的彩色图与人体三位骨架,包含了日常生活中的短视频动作。该数据库包含了4人12个日常行为的数据,数据库采用文本格式保存了关节点空间坐标数据,另外还有一个随机动作视频可用于评估分类效果。CAD-120数据库有4人的120个RGB-D视频以及对应的RGB图、深度图与人体三维骨架信息,大部分行为是在厨房和卧室完成的。

1.2 现有数据库存在的问题

虽然目前的数据库可以方便于人体行为研究,但不可忽视的是依然存在很多的问题。首先,现存

的数据库动作类别较少且部分行为数据不完整。KTH和Weizman数据库仅包含了灰度信息并且分辨率较低,MSR Action3D数据库仅限于游戏动作,没有与物体交互的行为。CAD数据库具有多背景的特点,但视频样本的数量有限,用来训练和分类的数量也相应地受到限制。其次,目前已经存在的数据库多数都为固定视角,却很少有数据库能做到多视角采集数据,多视角的数据能给研究者提供多角度信息的研究,以此来判断视角的优劣性。最后,目前采集的数据库背景都是静态的,无法验证跟随机器人动态识别人体行为和移动机器人寻找最佳视角的可靠性。

2 动态多视角数据库

针对上述问题,本文采集了一个大规模的人体行为数据库,视频总帧数超过60万帧,每个视频同时包含了彩色图像、深度图像和基于摄像机坐标系的三维人体骨架位置信息。

2.1 设备介绍

数据库录制主要用到的设备如下。

(1) Microsoft Kinect[16]

Kinect for Xbox 360,简称Kinect,由微软开发研制的成像设备,原理如图1所示。Kinect有3个镜头,中间的镜头是RGB彩色摄影机,用来采集彩色图像。左右两边镜头则分别为红外线发射器和红外线CMOS摄影机,构成了3D结构光深度感应器,可以用来采集深度数据(场景中物体到摄像头的距离)。彩色摄像头最大支持1 280像素×960像素分辨率成像,红外摄像头最大支持640像素×480像素成像。

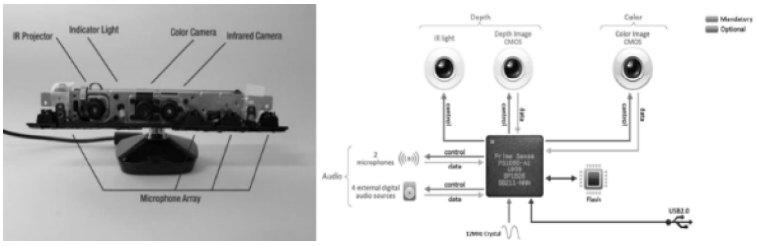

图1 Kinect硬件原理图

Fig.1 Kinect hardware schematic diagram

(2) Turtlebot机器人(Turtlebot Robot)[17]

Turtlebot是由Willow garage设计的移动机器人,其硬件主要有Yujin Kobuki移动底座、Kinect视觉传感器、2200mAh(或4400mAh)电池和可装卸的结构模块,使用著名的ROS(Robot operating system)作为操作系统,能实现3D地图导航、跟随等功能。

图2 Turtlebot移动机器人

Fig.2 Turtlebot mobile robot

2.2 场景构建

在本文的数据库中,所有的动作都是在复杂背景中使用Kinect从两个静态的视角和一个动态视角进行摄制完成的。动态Kinect摄像头由Turtlebot机器人携带,从正面到侧面以人为圆心,半径为2.5 m的轨道上连续录制,摄像机录制场景图如图3所示。Turtlebot机器人角速度为0.1 rad/s。图3中的两个静态摄像头的角度为90°,摄像头1从正面录制,摄像头2从侧面录制。

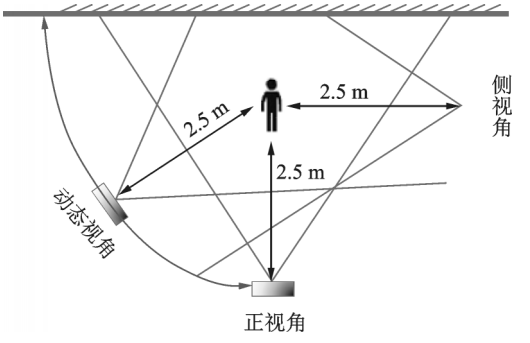

图3 模拟场景图

Fig.3 Simulation scenario

2.3 多视角交互人体行为识别数据库

数据库包含了3个不同场景下的31个不同的人体行为。每类行为包含20个人的动作,是一个多角度、交互式动态视角人体行为数据库。此20人年龄在21～25之间,其中男女身高范围在1.55～1.78 m之间。人体行为包含三大类,分别为基本动作类(10个)、与物体交互类(15个)和行为异常类(6个)。基本动作类有单手高举挥舞、扔东西、鼓掌、双手挥舞、向侧面出拳、慢跑、坐下起立、原地向上跳、自拍以及看手表。与物体交互类有读书、写字、擦汗、脱外套、穿鞋子/脱鞋子、戴眼镜/摘眼镜、踢箱子、从口袋里拿东西、打电话、喝水、吃零食、在黑板上写字、使用电脑、搬箱子以及搬椅子。行为异常类包含了摔倒、躺在地上、坐在地上、摔杯子、从椅子上跌落以及肚子疼。DMV action3D数据库既保证了多视角、多背景、多样本和多交互行为,还增加了一个动态视角,为实验者分析视角和寻找最佳角度提供了可供验证的数据库。作者采集的数据库图片如图4所示,其中包含了彩色图、深度图、关节点位置和时间信息等。

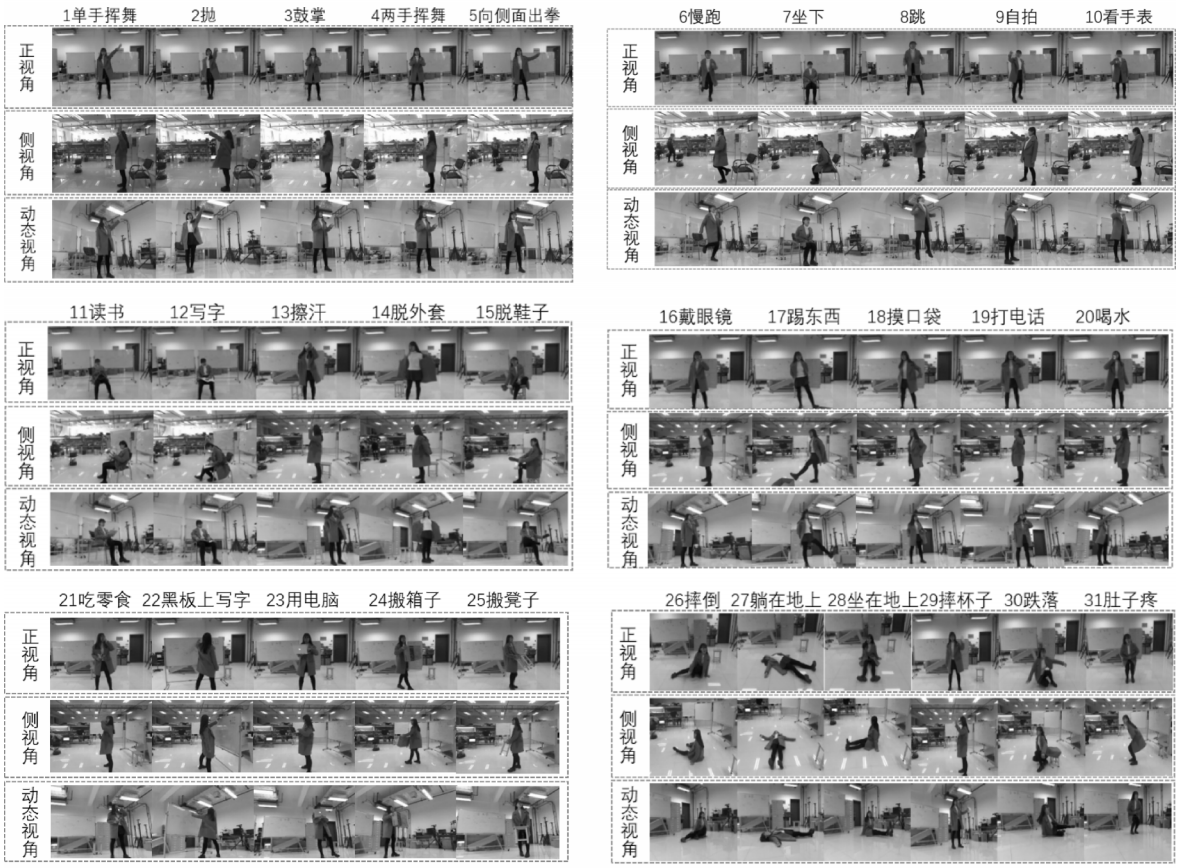

图4 DMV action3D数据库

Fig.4 DMV action3D database

3 人体行为特征提取和识别

为了对数据库进行测评,实验利用交叉验证方法的Libsvm程序和RBF-SVM[18]、C3D网络以及3D密集连接残差网络,对数据库录制的视频动作进行识别。采用了4种不同的特征提取方法,验证数据库的有效性:(1)利用数据库获取的关节点计算动能、势能和关节角等来构建BOW模型特征向量,并利用

SVM进行训练和识别。(2)利用深度学习技术的图像识别方法CRFasRNN[19]对数据库RGB图像进行人像分割,对分割后的人像提取HOG3D[20]特征,最后使用SVM进行训练和识别。(3)利用CNN对数据库视频进行每一帧的识别,考虑到连续帧间的运动信息,在卷积层进行3D卷积以捕捉时间和空间维度都具有区分性的特征,通过线性分类器对特征向量进行分类实现行为识别。(4)利用3D密集连接残差网络进行特征提取,将视频均匀分割成每16帧一段的短视频,经过3D密集连接残差网络提取特征后进行softmax层分类。

3.1 基于关节点特征提取

运动能量特征能够定量地表示运动能量信息,因此分别提取人体骨架的20关节点空间三维坐标、方向变化、关节点动能、姿态势能和关节角等特征构建人体空间特征矩阵。通过提取空间三维坐标$P_{n,t}$(n, t分别表示第n个关节点和t时刻)、方向变化特征$\varphi_{n,t}$、关节点动能特征$Ek_{n,t}$和人体姿态势能特征$E_{n,t}$等4类特征组成局部特征矩阵$\boldsymbol{Y}_t$,再对$\boldsymbol{Y}_t$特征向量进行归一化得到$\boldsymbol{Y}_t^*$,其定义如下

$$\boldsymbol{\varphi}_{n,t}=(x_{n,t}-x_{n,t-1},y_{n,t}-y_{n,t-1},z_{n,t}-z_{n,t-1}) \tag{1}$$

$$\boldsymbol{Ek}_{n,t}=k_n v_{n,t}^2 \tag{2}$$

$$\boldsymbol{E}_{n,t}=L_n\left|p_{n,t}-p_{1,t}\right| \tag{3}$$

$$\boldsymbol{Y}_t=\begin{bmatrix} p_{1,t} & \boldsymbol{\varphi}_{1,t} & \boldsymbol{Ek}_{1,t} & \boldsymbol{E}_{1,t} \\ p_{2,t} & \boldsymbol{\varphi}_{2,t} & \boldsymbol{Ek}_{2,t} & \boldsymbol{E}_{2,t} \\ \vdots & \vdots & \vdots & \vdots \\ p_{n,t} & \boldsymbol{\varphi}_{n,t} & \boldsymbol{Ek}_{n,t} & \boldsymbol{E}_{n,t} \end{bmatrix} \tag{4}$$

$$\boldsymbol{Y}_t^*=\begin{bmatrix} p_{1,t}^* & \boldsymbol{\varphi}_{1,t}^* & \boldsymbol{Ek}_{1,t}^* & \boldsymbol{E}_{1,t}^* \\ p_{2,t}^* & \boldsymbol{\varphi}_{2,t}^* & \boldsymbol{Ek}_{2,t}^* & \boldsymbol{E}_{2,t}^* \\ \vdots & \vdots & \vdots & \vdots \\ p_{n,t}^* & \boldsymbol{\varphi}_{n,t}^* & \boldsymbol{Ek}_{n,t}^* & \boldsymbol{E}_{n,t}^* \end{bmatrix} \tag{5}$$

式中:$v_{n,t}$为t时刻第n个关节点的速度;$x_{n,t},y_{n,t},z_{n,t}$分别为t时刻第n个关节点空间三维坐标;k_n和L_n为参数。然后采用BOW[21]模型构建特征向量,BOW模型构建方法如图5所示。

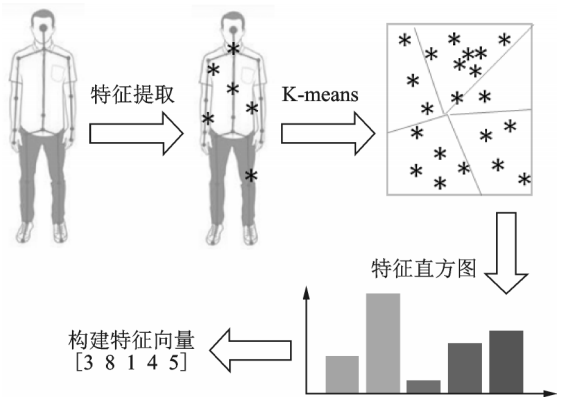

图5 Bag of word构建过程

Fig. 5 Bag of word building process

由于数据特征的数量较大,为了降低数组维度,对Y_t^*利用K-means算法[22]进行聚类,将所有特征向量映射到K个聚类中心,得到第F_t帧的$\mathbf{BOW}_t$特征如下

$$\mathbf{BOW}_t=\left[\mathrm{bin}_1,\cdots,\mathrm{bin}_k\right] \tag{6}$$

式中:bin_1为映射到聚类中心C1的个数,以此类推。其次,关节角可以直观有效的描述人体行为识别特征,选取6个人体关节角特征如图6所示,计算关节角的公式为

$$\theta_{n,t}=\arccos\left\{\frac{\boldsymbol{\alpha}\cdot\boldsymbol{\beta}}{|\boldsymbol{\alpha}||\boldsymbol{\beta}|}\right\} \tag{7}$$

式中$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$表示以关节角为原点对应的两个向量。

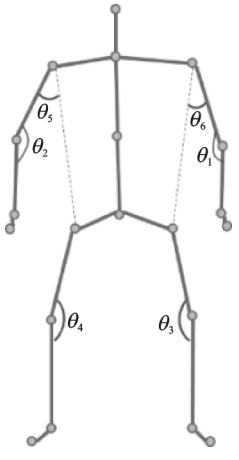

图6 人体骨架关节角

Fig. 6 Human skeleton joint angles

最后,利用人体关节点数据提取的关节点动能、势能、方向变化和空间信息特征,对BOW构建得到特征向量$\mathbf{BOW}_t$,而6个人体关

节角作为最具代表人体动作的特征个数较少，不需要进行降维处理。将6个关节角与$\mathbf{BOW}_t$一起构建成特征向量$\boldsymbol{AF}_t$为

$$\boldsymbol{AF}_t=\left[\mathbf{BOW}_t\ \theta^*_{1,t}\ \theta^*_{2,t}\ \theta^*_{3,t}\ \theta^*_{4,t}\ \theta^*_{5,t}\ \theta^*_{6,t}\right] \tag{8}$$

3.2 基于RGB的HOG3D特征提取

本数据库采用第三视角采取动态录制的方法，人和背景都是变化的，正常消除背景环境干扰的方法（例如帧差法、光流算法等）都不再适用。CRFasRNN的方法即使在图片模糊的情况下分割效果依然突出，边缘平滑，人像清晰，故采用此方法对人体进行分割，再提取人体运动HOG3D特征，方法如图7所示。

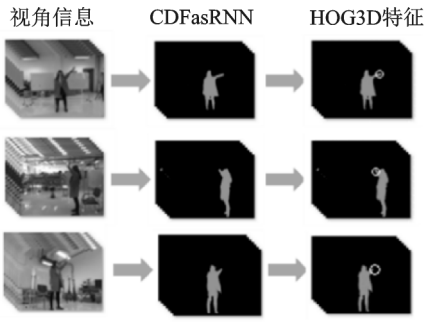

图7 基于RGB的HOG3D特征提取

Fig. 7 HOG3D feature extraction based on RGB

CRFasRNN是基于像素级标签，利用深度学习的图像识别方法，结合卷积神经网络CNN[23]和条件随机场CRF的优势，通过反向传播对整个网络进行端到端的训练。通过像素i的观测值y_i预测像素i的标签x_i，条件随机场满足吉布斯分布，公式为

$$P(X=x|I)=\frac{I}{Z(I)}\exp(-E(x|I)) \tag{9}$$

式中：I表示图片，随机变量x_i的取值与像素i相关，可以从预定义的标签$L=\{l_1,l_2,\cdots,l_L\}$中获取任何值。在训练期间，一张完整的图片作为输入，使用损失函数计算网络的每个像素输出误差。

3.3 基于3D卷积神经网络的时空特征提取

深度学习在图像识别、人体行为识别等方面表现很好。与2D ConvNet相比，3D ConvNet能够通过3D卷积和3D池化操作对视频进行建模时空信息。3D ConvNets卷积和池化操作在时空上执行，而2D ConvNets仅在空间上完成。为了更好地测试数据库在不同场景下的录制效果，选用C3D网络架构来获取时空特征。提取数据库时空特征中，视频被分割成16帧长的片段，两个连续片段间有8帧重叠。片段被传递到C3D网络，C3D网络有8个卷积层、5个最大池化层和2个全连接层，最后是softmax输出层，如图8所示。所有的3D卷积核均为3×3×3，在空间和时间上步长为1。为了保持早期的时间信息，设置pool1核大小为1×2×2、步长为1×2×2，其余所有3D池化层均为2×2×2，步长为2×2×2，每个全连接层有4 096个输出单元，网络经过两次全连接层和softmax层后得到分类结果。

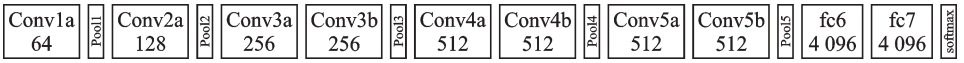

图8 C3D网络架构

Fig.8 C3D network architecture

3.4 基于3D密集连接残差网络的特征提取

3D密集连接残差网络由一个小型的密集连接网络和一个残差结构组合而成，用于提取视频的时空域特征。每个3D密集连接残差结构的输入是一个4层卷积（每个卷积层的卷积核均为3×3×3）的密集连接网络，然后输入到一层卷积核为1×1×1的3D卷积层进行特征整合，最后将该3D密集连接残差结构的输入进行残差操作。3D密集连接残差网络的输入为一段连续视频，本文实验中采用的视频帧尺寸为112像素×112像素，视频帧长度为16帧，3D密集连接残差网络结构如图9所示。

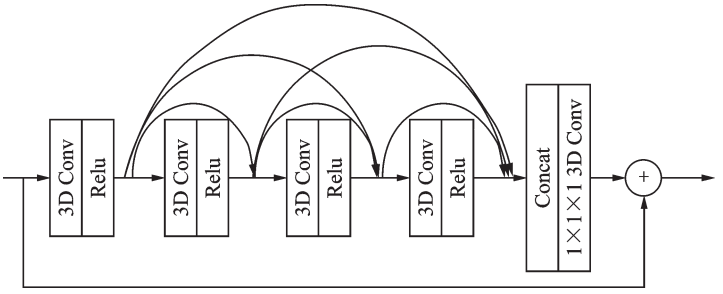

图9 3D密集连接残差网络架构

Fig.9 3D dense connection residual network architecture

4 实验结果与分析

为了验证不同特征的效果,在录制的数据库上,将关节点特征和HOG3D两种不同特征结合RBF-SVM作为识别方法,以及采用3D卷积网络和3D密集连接残差网络4种方法获取时空特征,反复5次求取平均识别率。实验结果如下。

4.1 基于关节点特征的实验结果与分析

利用式(8)提取关节点信息提取特征,采用SVM识别方法得到正视角的31个动作的总识别率为51.07%,侧视角31个动作的总识别率46.73%,动态视角31个动作的总识别率32.08%,混淆矩阵分别如图10—12所示。

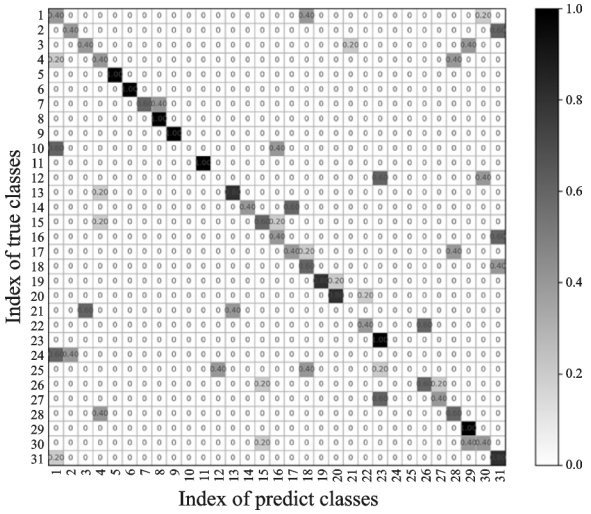

图10 正视角识别混淆矩阵

Fig.10 Confusion matrix of positive visual angle identification

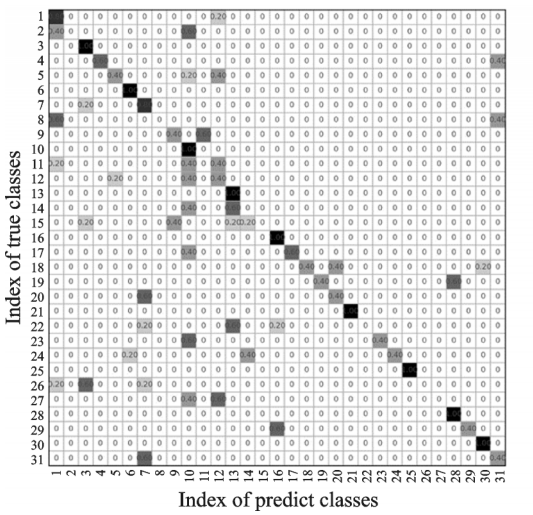

图11 侧视角识别混淆矩阵

Fig.11 Confusion matrix of side view angle identification

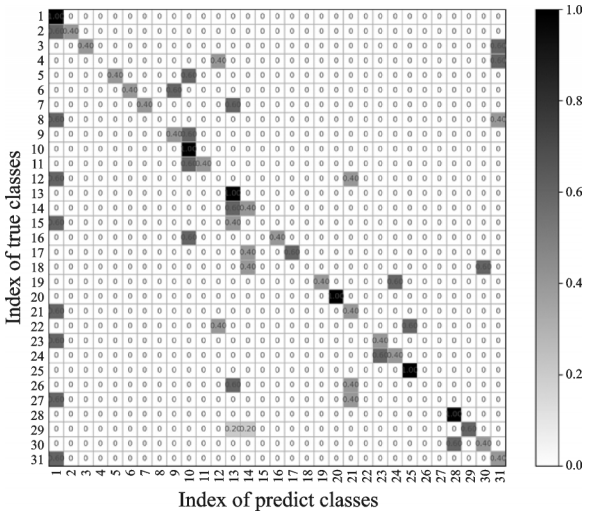

图12 动态视角识别混淆矩阵

Fig.12 Confusion matrix of dynamic perspective identification

在DMV action3D数据库基础上仅采用关节点作为特征向量获得的识别效果并不尽如人意。其原因是某些时刻关节点的运动具有极大的相似性，使得关节点特征的选取不具备代表性，另外关节点在运动中在部分图像中出现较大的跳动，也导致人体行为识别率较低。

4.2 基于HOG3D特征的实验结果与分析

在前景目标分割的基础上提取HOG3D特征，利用SVM进行分类的结果如图13—15所示，其中3个视角31类动作的总识别率分别为：正视角82.58%、侧视角85.15%和动态视角95.48%。利用RGB图像提取的HOG3D特征进行识别分类的结果总体要高于关节点特征提取方法，Kinect的RGB数据流可直接通过摄像头获取并保存，实时性和准确性都较高，Kinect

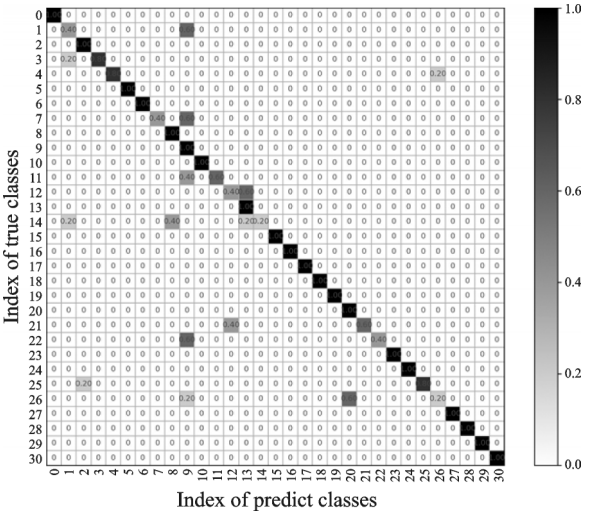

图13 基于HOG3D的正视角识别混淆矩阵

Fig.13 Confusion matrix of positive visual angle identification based on HOG3D

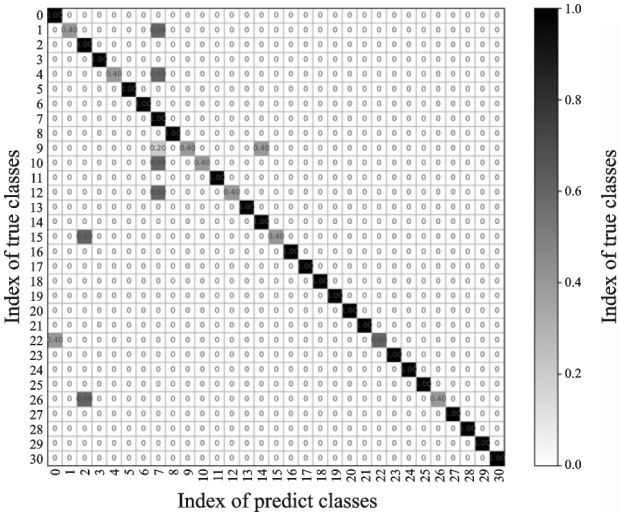

图14 基于HOG3D的侧视角识别混淆矩阵

Fig. 14 Confusion matrix of side view angle identification based on HOG3D

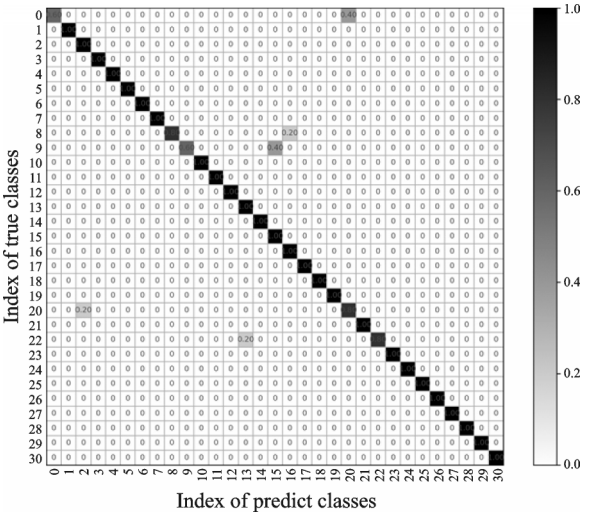

图15 基于HOG3D的动态视角识别混淆矩阵

Fig. 15 Confusion matrix of dynamic perspective identification based on HOG3D

关节点数据来源于对深度图像的计算，准确性较低，直接影响了基于关节点特征提取的准确率。

4.3 基于3D卷积神经网络的实验结果与分析

采用3D卷积网络提取人体行为时空特征，再经过softmax层实现分类，得到正视角的31个动作总识别率为36.29%，侧视角31个动作总识别率37.72%，动态视角31个动作总识别率38.26%，混淆矩阵分别如图16—18所示。

从图16—18可以看出：采用3D卷积网络提取人体行为时空特征时，获得的识别效果并不是很好。其主要原因可能是数据库中某些人体行为数据具有较大的跳变性，使得C3D提取连续帧间的信息不够充分，也可能是网络参数训练不充分，导致最终的人体行为识别率较低。

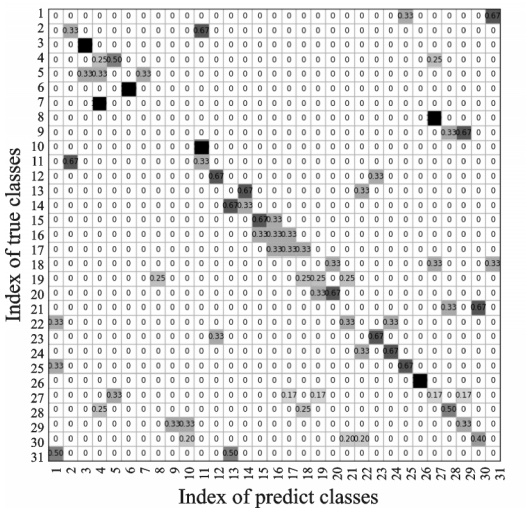

图16 基于3D卷积网络的正视角识别混淆矩阵

Fig. 16 Confusion matrix of positive visual angle identification based on 3D convolutional network

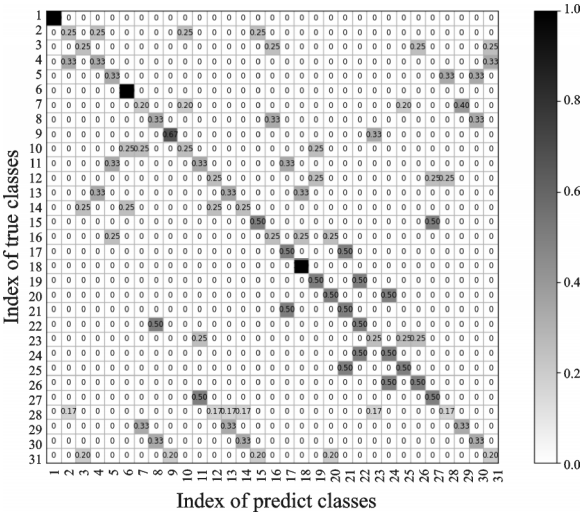

图17 基于3D卷积网络的侧视角识别混淆矩阵

Fig. 17 Confusion matrix recognition of side view angle based on 3D convolutional network

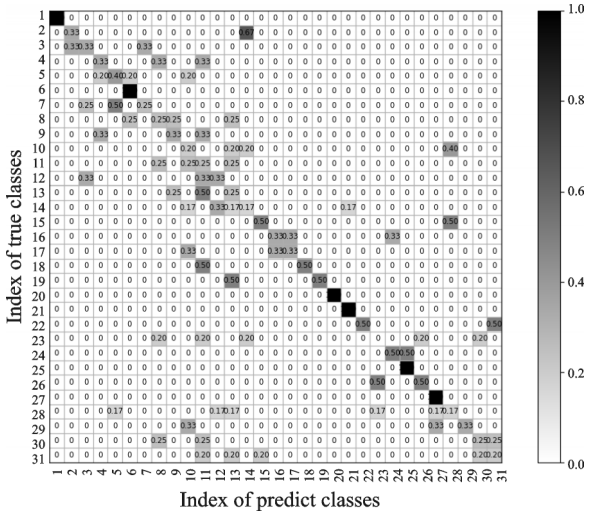

图18 基于3D卷积网络的动态视角识别混淆矩阵

Fig. 18 Confusion matrix of dynamic perspective identification based on 3D convolutional network

4.4 基于3D密集连接残差网络的实验结果与分析

最后采用基于3D密集连接残差网络的特征提取方法,数据库中31个动作的训练正确率为60.77%,31个动作的测试正确率为15.01%,由于该模型未很好地利用数据库交叉视角的优势所在,网络参数训练难以到达最优,导致最终的测试效果也不理想,结果如图19所示。

4.5 实验分析

多样化的人体行为识别数据库为验证人体行为识别方法的合理性和有效性提供了重要的基础和依据。在DMV action3D数据库基础上单独采用关节点、3D卷积网络或3D密集连接残差网络提取特

征时，当人体在被物体遮挡、动作幅度过大或视角信息较差的情况下，会出现抖动或产生数据偏差。例如在侧视角录制的关节点数据中，人的右腿膝关节在被遮挡的情况下容易出现抖动，此时提取的人体关键角和运动信息特征将会受到影响。

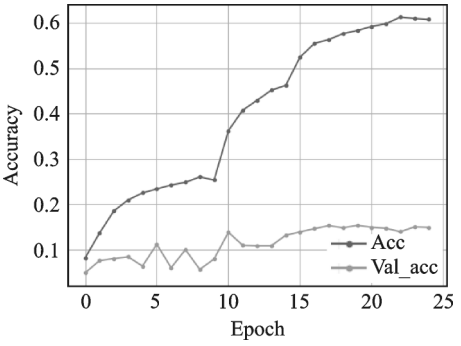

图19 3D密集连接残差网络实验结果

Fig. 19 Experimental result of 3D dense connection residual network

DMV action3D数据库相对于MSR Action3D和MSR DailyActivity数据库的难度更大。DMV action3D数据库相似性高的动作类别相对于MSR Action3D和MSR DailyActivity数据库更多，例如“自拍”和“看手表”，“读书”和“写字”，“擦汗”“打电话”和“戴眼镜”，“吃零食”和“喝水”，“搬箱子”和“搬凳子”，“摔倒”和“坐在地上”等相似动作，因此出现了大量数据识别错误的情况。另外，在动态视角中，人相对于Turtlebot的位置也是不断变化的。人在真实环境下保持不动，动态视角的机器人Turtlebot视频中的人是移动的，此时，人与机器人Turtlebot存在位移误差，使得关节点的位置特征、方向变化特征和关节点动能特征受到机器人运动的影响，因此，采用关节点作为特征时动态视角的识别率最低。

从图13—15的结果可以看出，CRFasRNN＋ HOG3D特征提取方法优于关节点、3D卷积网络以及3D密集连接残差网络特征提取的方法。Kinect的RGB数据流可直接通过摄像头获取并保存，实时性和准确性都较高。前景分割方法的优势在于，去除了大量的背景干扰，仅提取人体运动的有效特征。在现实情况中，静态视角下的背景是无规律变化的，也存在静态视角的动态背景例如水波晃动、喷泉和树木摇摆等，若不去除背景，HOG3D特征会将这些变化的部分同时作为目标提取特征。在DMV action3D静态视角中，正视角和侧视角都会把移动入镜的Turtlebot机器人与人体运动两者作为提取HOG3D特征的目标，而动态视角下的整个背景都是随着时间发生位移的，因此CRFasRNN＋HOG3D方法取得了良好的识别率。

从实验可以看出DMV action3D中的3大类行为数据，在不同的视角下，识别率有明显的差异，说明了视角的不同对识别率有较大的影响。在真实的人体行为识别中，视角差异会限制人体行为识别率的进一步提高，多视角的人体行为识别方法可能是一个有效的途径，为机器人研究不同视角下的人体行为识别和机器人寻找最佳视角提供了实验数据。

5 结束语

本文围绕人体行为识别和数据库两个紧密联系的部分，在录制的数据库上通过不同行为识别方法来验证录制数据库的合理性。文中利用移动机器人录制了动态视角的人体行为数据，现有31个不同的行为类，包括日常和交互和异常行为等3大类超过60万帧彩色深度图像。另外，所使用的几种人体行为识别的方法分别是：基于关节点信息特征、基于CNN和CRF结合的CRFasRNN方法提取的彩色图像HOG3D特征；基于3维卷积网络（C3D）和3D密集连接残差网络提取时空特征。4种不同特征的实验结果表明，DMV action3D人体行为数据库利用关节点、C3D卷积网络和3D密集连接残差网络提取特征的方法在不同视角下的识别率具有较大差异，效果较差；而基于RGB图像的HOG3D特征方法达到了良好的识别率。本数据库可提供给人体行为识别研究者和开发者，本文提供了人体行为数据库和部分实验代码的下载地址：ftp://10.10.108.138。今后将进一步补充和完善数据库样本。

参考文献：

[1] Koppula H S, Gupta R, Saxena A. Learning human activities and object affordances from RGB-D videos[M]. England: Sage

Publications Inc, 2013:951-970.
[2] Huang A S, Bachrach A, Henry P, et al. Visual odometry and mapping for autonomous flight using an RGB-D camera[M]. Switzerland Robotics Research: Springer International Publishing, 2017:235-252.
[3] 黄凯奇,任伟强,谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报, 2014, 37(6):1225-1240.
Huang Kaiqi, Ren Weiqiang, Tan Tieniu. A review on image object classification and detection[J]. Chinese Journal of Computers, 2014, 37(6): 1225-1240.
[4] Aggarwal J K, Ryoo M S. Human activity analysis: A review[M]. New York: ACM, 2011: 16.
[5] Gorelick L, Blank M, Shechtman E, et al. Actions as space-time shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(12):2247-2253.
[6] Rodriguez M D, Ahmed J, Shah M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition[C]// Computer Vision and Pattern Recognition 2008. [S.l.]: IEEE, 2008:1-8.
[7] Ryoo M S. Human activity prediction: Early recognition of ongoing activities from streaming videos[C]// IEEE Inter-national Conference on Computer Vision. [S.l.]: IEEE, 2012:1036-1043.
[8] Soomro K, Idrees H, Shah M. Action localization in videos through context walk[C]// IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015:3280-3288.
[9] Weinland D, Boyer E, Ronfard R. Action recognition from arbitrary views using 3D exemplars[J]. ICCV, 2010, 2380(7504):1-7.
[10] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points[C]// Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2010:9-14.
[11] Wu Y. Mining actionlet ensemble for action recognition with depth cameras[C]// IEEE Conference on Computer Visi-on and Pattern Recognition. [S.l.]: IEEE, 2012:1290-1297.
[12] Shahroudy A, Ng T T, Yang Q, et al. Multimodal multipart learning for action recognition in depth videos[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(10):2123-2129.
[13] Sung J, Ponce C, Selman B, et al. Unstructured human activity detection from RGBD images[J]. IEEE, 2012, 44(8):47-55.
[14] Sung J, Ponce C, Selman B, et al. Human activity detection from RGBD images[C]// IEEE International Conference on Robotics and Automation. [S.l.]: IEEE, 2011:842-849.
[15] Feifei L, Perona P. A Bayesian Hierarchical model for learning natural scene categories[J]. CVPR, 2005, 2: 524-531.
[16] Clevenger K A, Howe C A. Energy cost and enjoyment of active videogames in children and teens: Xbox 360 Kinect[J]. Games for Health Journal, 2015, 4(4):318-324.
[17] Kniss J, Jin K, Ivans R, et al. Robotics research with turtlebot 2016[D]. Boise: Boise State University, 2016.
[18] Wang Y, Li X, Ding X. Probabilistic framework of visual anomaly detection for unbalanced data[J]. Neurocomputing, 2016, 201:12-18.
[19] Razavian A S, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: An astounding baseline for recognition[C]// Computer Vision and Pattern Recognition Workshops. [S.l.]: IEEE, 2014:512-519.
[20] 李昕迪, 王云龙, 何艳,等. 基于Kinect的人体单关节点修复算法研究[J]. 自动化技术与应用, 2016, 35(4):96-98.
Li Xindi, Wang Yunlong, He Yan, et al. Research on human single-pass node repair algorithm based on Kinect [J]. Automation Technology and Application, 2016, 35(4):96-98.
[21] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks[C]// IEEE International Conference on Computer Vision.[S.l.]: IEEE, 2016:1529-1537.
[22] Wu C H, Tzeng G H, Goo Y J, et al. A real-valued genetic algorithm to optimize the parameters of support vector mac-hine for predicting bankruptcy[J]. Expert Systems with Applications, 2007, 32(2):397-408.
[23] Klaser A. A spatiotemporal descriptor based on 3D-gradients[C]// British Machine Vision Conference. Leed S, United Kingdom: British Machine Vision Association, 2008: 1-10.

作者简介:

王永雄(1970-),男,教授,研究方向:智能机器人、机器学习、视觉跟踪和人体行为识别,E-mail: wyxiong@ usst.edu.cn。

李璇(1993-),女,硕士研究生,研究方向:机器视觉,E-mail: Lydia_ia_LXH@126.com。

李梁华(1994-),男,硕士研究生,研究方向:机器视觉,E-mail: 1244094457@qq.com。

(编辑:刘彦东)