基于空-时域特征决策级融合的人体行为识别算法

李艳荻，徐熙平*

长春理工大学光电工程学院，吉林 长春 130022

摘要 提出一种基于空-时域特征决策级融合的人体行为识别算法。在空间域提取人体的形状上下文特征，用于同一时刻模板图像与测试图像的轮廓匹配；在时间域用变化的空间特征序列表征运动特征，联合稳健的空间特征进行有效的行为识别。识别阶段采用动态时间规划算法分别计算两种特征对于每种类别的后验概率，在决策级采用加权平均法对两种特征的后验概率进行融合，将最大概率对应的类别记为最终分类结果。针对动态时间规划算法提出一种基于椭圆边界约束的改进搜索策略，有效缩减最优路径的搜索空间，同时剔除视频中的噪声干扰。从计算复杂度和识别精度两方面对椭圆边界的约束性能进行分析，实验表明，椭圆边界约束性能优于平行四边形及菱形约束，并给出最佳边界尺寸范围。算法分别在 Weizmann、KTH 和 UCF101 行为数据集上进行测试，平均识别率分别优于 93.2%、92.7%和 81.2%，有效实现了室内智能监控系统的高效性及稳定性。

关键词 图像处理；行为识别；形状上下文；动态时间规划；决策级融合

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/AOS201838.0810001

Human Action Recognition by Decision-Making Level Fusion Based on Spatial-Temporal Features

Li Yandi, Xu Xiping*

College of Photoelectrical Engineering, Changchun University of Science and Technology, Changchun 130022, *China*

Abstract A human action recognition algorithm is proposed based on the decision-making level fusion with spatial and temporal features. Shape context feature of human body is extracted to match the contours of template images and test images in the spatial domain, while the motion feature is described by a changing spatial feature sequence in the time domain. Then, the motion feature is combined with the robust spatial feature for effective human action recognition. At the recognition stage, the dynamic time warping is applied to calculate the posterior probabilities of two kinds of features for each class. The weighted-average method is used to fuse the two posterior probabilities at the decision-making level, and the corresponding class with the maximum probability is recorded as the final classification result. Aiming at the dynamic time warping algorithm, we propose an improved searching strategy based on the elliptic boundary constraint, which can effectively reduce the space for searching for the optimal path, while eliminate the noise interference in the video sequence. The constraint performance of elliptical boundary is analyzed from two aspects of computational complexity and recognition accuracy. Experimental results show that the performance of elliptical boundary constraint is better than that of the parallelogram and diamond boundary constraints, and the optimal boundary size range is given. Experimental results on Weizmann, KTH and UCF101 datasets demonstrate that the average recognition rate of the proposed method is higher than 93.2%, 92.7% and 81.2%, respectively, indicating that the proposed method can effectively obtain the efficiency and stability of indoor intelligent monitoring system.

Key words image processing; action recognition; shape context; dynamic time warping; decision-making level fusion

OCIS codes 100.3008; 330.4150; 070.5010

收稿日期：2017-12-11；**修回日期**：2018-02-06；**录用日期**：2018-03-20

基金项目：吉林省科技发展计划(20160520018JH)

* **E-mail**：xxp@cust.edu.cn

1 引　　言

人体行为识别作为图像理解的一个重要分支，在视频监控、人机交互及虚拟现实等计算机视觉领域有着广泛的应用前景。本研究拟从视频或图像中提取和分析图像特征，考虑上下文环境信息和时序关系，估计、识别和重构人体姿态或行为[1]，是近年来计算机视觉领域的研究热点及难点。

根据现有的研究成果，人体行为识别方法可以分为3种：基于概率统计模型的方法[2-3]，基于模板的方法和基于深度学习的方法。

Yamato等[4]最先提出基于隐马尔科夫模型(HMM)的概率统计模型，输入人体运动区域块的特征信息，利用HMMs模型对人体动作进行识别。Peursum等[5]用层次HMM对动作在不同层次的信息进行描述，较好地实现了人体局部细节的行为识别。Natarajan等[6]提出了一种层次变量转变的HMM，对每个行为进行3次建模，并引入可变窗口，实时性效果显著。对于一般的HMM，模型参数通常会随着运动目标数目的增加成指数递增，复杂的计算量会增加模型应用的局限性；同时，该模型很难有效地融合特征信息，会导致序列间的特征出现重叠累积，影响识别精度。Lafferty等[2]提出条件随机场模型，利用大范围上下文信息进行参数学习和预测，相较于HMM具有更强的时序建模能力。Huang等[7]也证明，基于时空兴趣点和光流特征，隐条件随机场比HMM和支持向量机(SVM)方法更有效。但是，条件随机场模型的训练过程需要较多的人为标注数据，以获得空间特征随时间动态性所表现出的判别性能，复杂度相对较大，甚至会影响模型的稳定性。文献[8]通过提取视频中的时空兴趣点构造时空词袋模型，并结合潜在狄立克雷分配(LDA)主题模型和概率潜在语义(PLSA)主题模型进行行为识别，对动态背景下的行为识别稳健性较好；文献[9]将加速稳健特征(SURF)特征和稠密光流特征作为视频行为表征，利用随机抽样一致(RANSAC)算法完成特征点的精确匹配，该方法同样适用于相机运动的情况。

基于模板的方法主要包含模板匹配法和动态规划法。文献[10]在空间域使用Gabor滤波器提取图像的局部特征，然后对时间域的光流运动特征进行加权融合，最后利用SVM进行识别，取得了不错的效果。Liu等[11]用费德勒嵌入的方法将旋转图像和局部时空立方体嵌套到同一空间中，实现了单一视角和多视角下的行为识别。模板匹配[12-13]的优点在于易于实现、计算复杂度较低，但是对前景目标的提取精度要求较高，而且在时间序列时间长度不一致的情况下准确度会受到干扰，因此，不适用于时间尺度不固定的动作识别。动态规划方法[14]能够较好地解决人体行为在时间尺度上的不确定性，但是计算量会随着训练样本数量不断增加，易造成维数灾难。文献[15-16]将特征序列表示为几个人体状态的转移，通过确定输入特征序列与模板序列是否匹配得到行为类别，该方法引入了状态转移函数，有效提高了序列特征描述的稳健性。人体关节具有较大的自由度，对环境变化也十分敏感，因此对于基于模板的方法来讲，要确定一个稳健性较强的姿态描述是极具挑战的。

基于深度学习的识别方法是近年来新兴起的一种方法，基本思想都是将二维图像识别的神经网络框架扩展到三维(3D)视频中用于行为识别，在视频数据的时间维度和空间维度上进行特征计算[17-18]。Karpathy等[19]通过卷积神经网络学习局部时空特征，通过不同方式将行为视频描述成视频流形式的向量表征，最后用神经网络分类器进行行为识别。Baccouche等[20]利用3D卷积神经网络以类似的方式学习时空特征，利用长短时记忆网络获取视频片段在时间域上的联系，然后对提取的时空特征序列进行行为识别。目前行为识别方法所采用的数据库多是分割好的短视频片段，具有明确的行为类别，对时间尺度比较大且未做分割处理的视频效果并不好。有鉴于此，Shou等[21]提出了一种基于视频片段的3D卷积神经网络，对运动边界进行微调，提高了识别精度。传统深度学习的一个重要优势在于不需要手动提取特征(通常只须将整帧视频作为输入进行特征学习)，但是需要训练大量的网络参数，对样本的数量需求较大，特别是对于某些特定行为，很难收集到足够数量的有效样本。另外，3D卷积操作的计算量呈指数级增长，这些都将成为网络训练过程中的难点。

本文提出一种基于时间-空间域特征决策级融合的人体行为识别算法，在空间域提取人体的形状上下文特征，在时间域用变化的空间特征序列表征运动特征，然后联合稳健的空间特征进行有效的人体行为识别。通过动态时间规划算法分别计算出两种特征对于行为类别的后验概率，在决策级采用加权平均法进行融合，将最大概率对应的类别记为最终分类结果。针对动态规划方法易产生维数灾难的

问题，提出一种基于椭圆边界全局约束的搜索策略，从计算复杂度、识别精度两方面对其约束性能进行分析，最后分别在 Weizmann、KTH 和 UCF101 行为数据库进行测试及效率评估。

2　时间-空间域上人体行为特征序列匹配算法

由于人体行为在空间结构上可以描述成各个时刻的姿态集合，在时间序列上可以看作一段时间内姿态的演变过程；因此，可以用变化的空间特征序列表征运动特征，然后联合稳健的空间特征来共同描述人体行为。以形状上下文特征匹配算法为基础，在空间域用其来计算同一时刻模板图像与测试图像的轮廓相似度，在时间域用其来估计模板序列和测试序列中各自相邻两帧之间的形状变化，利用得到的两组由形状变化度组成的数据来计算两段视频序列在时间轴上的相似度。

2.1　形状上下文特征匹配

形状上下文特征具有良好的尺度不变性及旋转不变性，在目标发生微小几何形变及存在异常点的情况下，稳健性较好；因此，将其作为人体行为在空间域上的特征描述子[22]。形状上下文特征不是利用图像的某个区域或者轮廓来描述其形状特征，而是通过在对象轮廓上提取一些离散并且分布均匀或代表性较强，(如角点)的特征点集来表达，然后针对点集中的每一个特征点，计算其所对应的形状直方图，用来存储该特征点与轮廓上其他所有特征点全部的矢量关系。

形状上下文特征匹配算法的输入是一串二值人体轮廓序列，因此需要对视频图像进行目标轮廓提取。这一步骤的准确度直接影响特征的有效性，进而决定系统的识别性能。根据本课题组在文献[23]中提出的运动目标检测算法对每一帧图像提取前景，然后对其进行高斯平滑滤波及形态学处理，目的是消除噪声、空洞干扰。采用腐蚀运算消除细小目标、孤立的点或小区域，再利用膨胀处理填充目标内部的间隙和孔洞，以强化目标的空间相关性。对得到的连通域进行边缘检测，得到运动目标的轮廓，具体流程如图 1 所示。

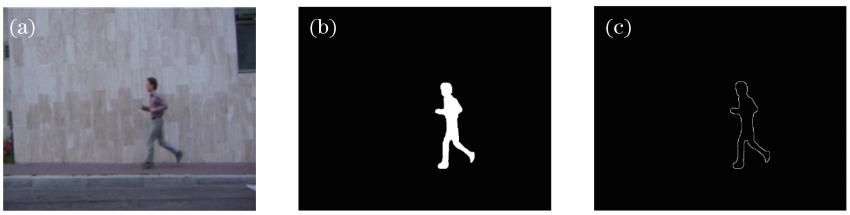

图 1　目标轮廓提取过程。(a)原始视频；(b)运动目标检测；(c)目标轮廓提取

Fig. 1　Target contour extraction process. (a) Original video; (b) motion detection; (c) target contour extraction

得到目标轮廓后，由于相邻轮廓点之间的信息是高度相关的，对轮廓点进行均匀下采样，记采样点数为 S。图 2 为 POSER 的 3D 人体仿真软件建立的人体站立姿态模板样本和测试样本，其中，模板样本中的采样点数小于测试样本中的采样点数。图 3

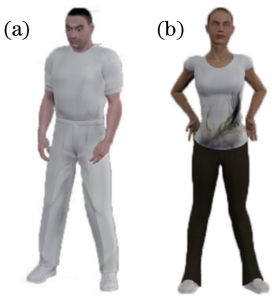

图 2　POSER 三维仿真的(a)模板样本和(b)测试样本

Fig. 2　(a) Template image and (b) test image of POSER 3D simulation samples

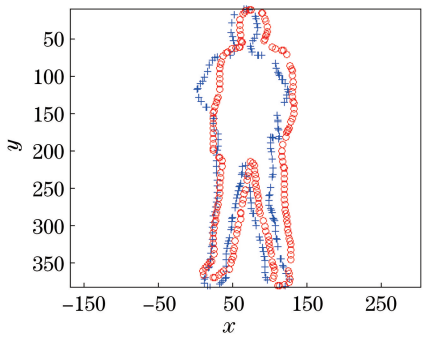

图 3　采样结果

Fig. 3　Sampling results

为两幅图像的采样结果，可以看出，模板图像的采样点数为 179，测试图像为 192。

轮廓上任意采样点都可以与其余 $S-1$ 个点建立向量关系，表征距离与方向两方面信息，对每个采样点建立极坐标系，按下式将笛卡尔坐标系下的采

样点映射到极坐标系：

$$\begin{cases} r=\sqrt{(x-x_0)^2+(y-y_0)^2} \\ \theta=\arctan[(y-y_0)/(x-x_0)] \end{cases}, \quad (1)$$

式中(x_0,y_0)为笛卡尔坐标系下的点，(x,y)是对应极坐标下的点。由于对数极坐标映射具有二维不变性，同时可简化尺度变换的计算量；因此，对图像进行对数极坐标转换。在对数极坐标系下分别将$\log r$和θ作A、B等分(通常情况下$A=5$，$B=12$)，按照距离和方向进行区块划分，构成5(半径)×12(角度)个区间，建立的对数极坐标系如图4所示。为保证尽可能多的点能够落在直方图区间，计算所有采样点之间的距离，用距离的平均值将log-polar直方图的径向距离归一化。

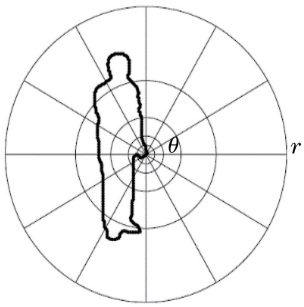

图4　极坐标下的轮廓点分布

Fig. 4　Distribution of contour points in polar coordinates

设轮廓采样点集$U=\{u_1,u_2,\cdots,u_S\}$，对点集中任意一点u_i，其特征可以用以u_i点为中心的极坐标系的60个区间中落入每个区间的离散点个数$h_i(k)$来表示，计算公式如下：

$$h_i(k)=\#\{v\neq u_i \& (v-u_i)\in \mathrm{bin}(k)\}, \quad (2)$$

式中：$k\in\{1,2,\cdots,K\}$，K为方向参数和距离参数的乘积(本文$K=60$)；操作#表示v_i落入第k个区间中不同于u_i点的轮廓上其余采样点的个数。这样就得到了一个含有60个分量的形状直方图。若模板图像与测试图像分别有m和n个采样点，则最终可得到m和n个形状直方图。如图5所示，图中区间内的数字表示落入该区间内的特征点个数。从图5(a)和(b)中可以看出，同一目标轮廓中不同轮廓点的形状直方图分布差异明显，而图5(a)和(c)中相似目标轮廓中相同位置点的形状直方图分布较为相似。

在对两组采样点集进行形状上下文描述子匹配时，采用文献[24]中提出的金字塔匹配核函数来直接评价两个集合的相似度，利用得到的匹配距离构造距离矩阵，匈牙利算法寻找最优匹配，以及下式最小化整个匹配代价。

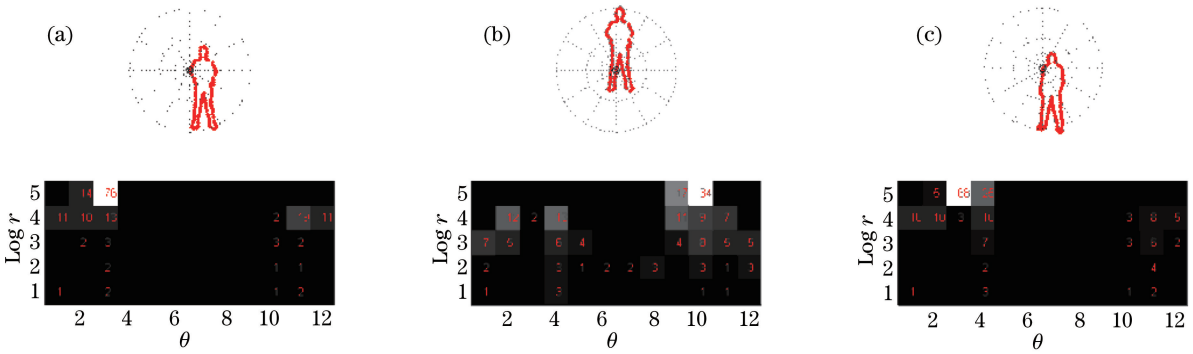

图5　不同轮廓采样点的形状直方图。(a)采样点1；(b)采样点2；(c)采样点3

Fig. 5　Shape histogram of different contour sampling points. (a) Sampling point 1; (b) sampling point 2; (c) sampling point 3

$$H(\pi)=\sum_i C[u_i,v_{\pi(i)}]。 \quad (3)$$

这样就可以得到两个形状之间相似采样点的对应关系。然后，利用薄板样条(TPS)插值函数运算来衡量两个形状之间的转变，最终的形状匹配距离用估计的变换来表示：

$$D_{SC}(U,V)=\frac{1}{m}\sum_{u\in U}\arg\min_{v\in V}\{C[u,T(v)]\}+\frac{1}{n}\sum_{v\in V}\arg\min_{u\in V}\{C[v,T(u)]\}, \quad (4)$$

式中$T(\cdot)$表示估计的TPS形状转换。$D_{SC}(U,V)$值越小，说明两幅图像形状的相似度越高。图6为对两个轮廓匹配的迭代结果，经过3次迭代，匹配结果基本稳定，最终匹配值为0.0438。

仿真过程中，为了能够得到最优匹配效果，对不同采样点的匹配结果进行测试。在Weizmann行为数据库的10类行为中，针对每类行为各手动截取5帧关键帧，采用“留一法”进行交叉验证实验，然后对所有行为的匹配结果进行统计。图7所示为模板样本和测试样本各提取不同采样点个数时得到的匹配结果平均值，其中，横轴表示模板样本的采样点数，

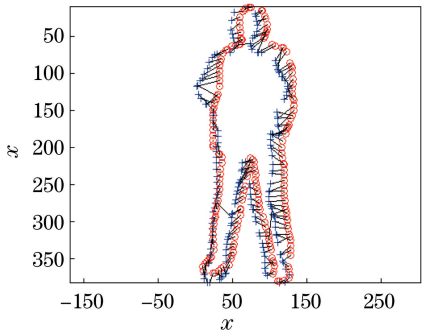

图 6　匹配结果

Fig. 6　Matching results

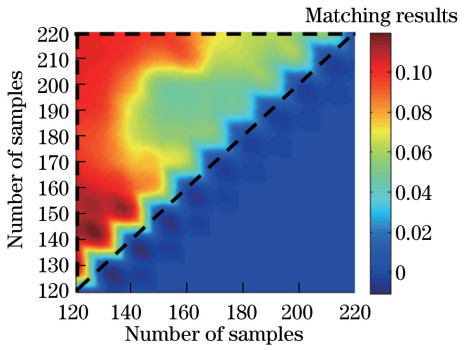

图 7　不同数量采样点的匹配结果

Fig. 7　Matching results of different sampling points

纵轴表示测试样本的当采样点数。为保证匹配过程中测试样本能够充分匹配到模板样本中的所有采样点，测试样本的采样点数量须大于模板样本中的采样点数量，图中黑色虚线框为有效匹配区域。可以看出，当 $S\in[150,200]$时，匹配结果最小且趋于稳定，说明此时特征的描述能力最强。

2.2　时间-空间域特征的视频序列匹配

图 8 所示为 4 种日常行为序列及对应的运动目标检测结果，提取序列上每一帧的形状上下文特征，用于计算测试行为序列与模板行为序列之间的匹配度。图 9 和图 10 分别描述了两段序列关于空间域的轮廓特征和时间域的运动特征的匹配过程。

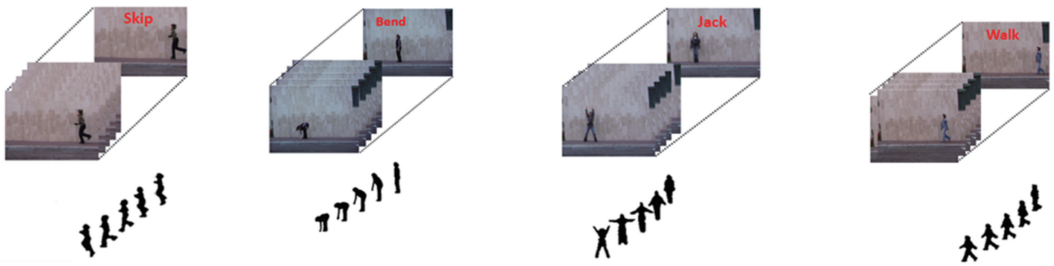

图 8　视频序列

Fig. 8　Video sequences

对于空间域上的轮廓特征匹配，构建匹配代价矩阵 $\boldsymbol{A}$，如图 9 所示，矩阵中的每一个元素为测试序列与模板序列上相对应两帧之间的匹配值。设测试序列长度为 N、模板序列长度为 M，则矩阵 $\boldsymbol{A}$ 的大小为 $N\times M$。

对于时间域上的运动特征匹配，首先利用形状上下文特征匹配算法计算得到模板序列上相邻两帧之间的轮廓特征匹配值，得到匹配值序列 S_1，同理得到测试序列上相邻帧之间的轮廓匹配值序列 S_2，S_1 和 S_2 可以分别表征模板序列和测试序列的运动特征，构建匹配代价矩阵 $\boldsymbol{B}$，如图 10 所示，矩阵中的每一个元素为这两串匹配度序列上对应元素之间的马氏距离，矩阵 $\boldsymbol{B}$ 的大小为 $(N-1)\times(M-1)$。

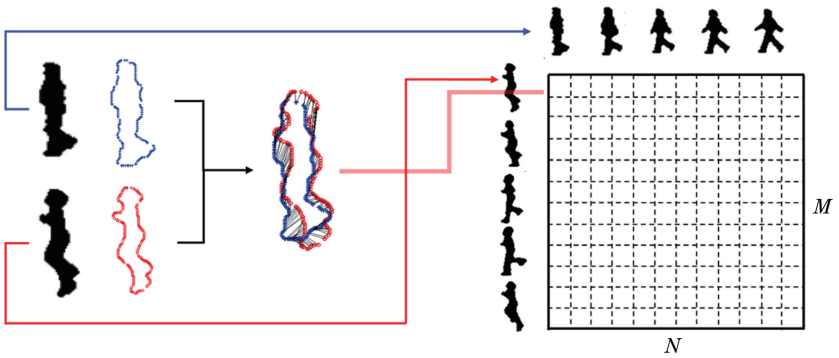

图 9　空间域特征序列匹配过程

Fig. 9　Matching process of feature sequence in spatial domain

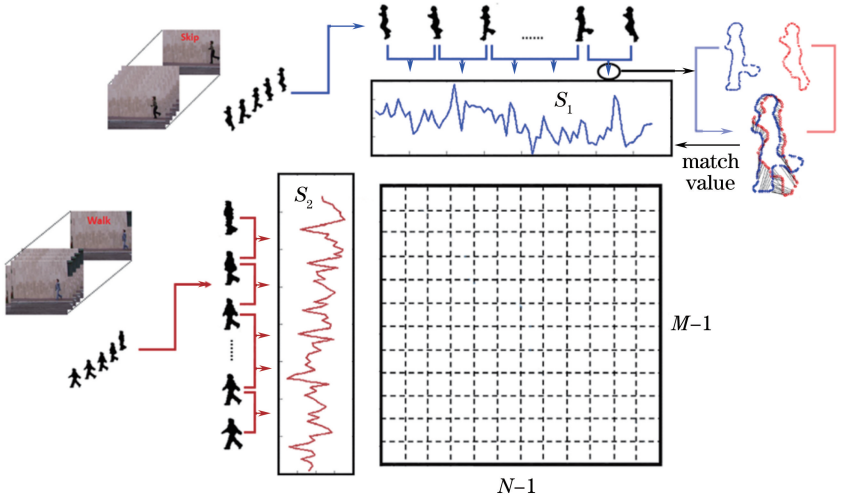

图 10　时间域特征序列匹配过程

Fig. 10　Match process of feature sequence in time domain

3　基于椭圆边界约束的动态时间规划识别算法

对匹配代价矩阵 ***A*** 和 ***B*** 分别利用动态时间规划(DTW)算法搜索一条能够使特征序列的总匹配值达到最大的路径。在原始 DTW 算法的全局约束方面进行改进，提出一种基于椭圆边界的搜索策略，以缩小 DTW 算法的搜索空间、提高识别效率。

3.1　动态时间规划算法

动态时间规划算法的实质是运用动态规划(DP)思想，利用局部最优搜索策略自动寻找一条路径(即时间规整函数)，使这条路径上每个元素对应的两个特征矢量之间的匹配距离最小，从而避免由于产生速度不同，以及始末点检测的差异而可能引入的误差。对 DTW 而言，即使测试序列与模板序列的时间尺度不能完全一致，只要时间次序约束存在，它仍能较好地完成两个序列间的模式匹配。

给出两个视频序列 $X=\{x_1,x_2,\cdots,x_N\}$ 和 $Y=\{y_1,y_2,\cdots,y_M\}$，首先将两个序列中各帧号分别在二维直角坐标系中标出，并通过这些帧号画出一系列纵横线构成一个矩阵网格，原点设置在左下角。网格中的每一个格点 $s\langle n,m\rangle$ 包含序列 X 中第 n 帧和 Y 中第 m 帧图像的相似度信息(也称失真度)。为了便于直观描述路径性质，引入时间轴 k，则：$v\langle n,m\rangle=v\langle i(k),j(k)\rangle,k=1,2,\cdots,K$。DTW 算法的示意图如 11 所示。

图 11 中路径 q 的搜索策略不是任意的，需要具备三个约束条件：

1) 端点约束。虽然行为发生的速度存在快慢变化，但每一帧姿态发生的先后顺序不能颠倒。因此，规划路径一定从左下角开始，至右上角结束。即起点 $i(1)=1,j(1)=1$；终点 $i(K)=N,j(K)=M$。

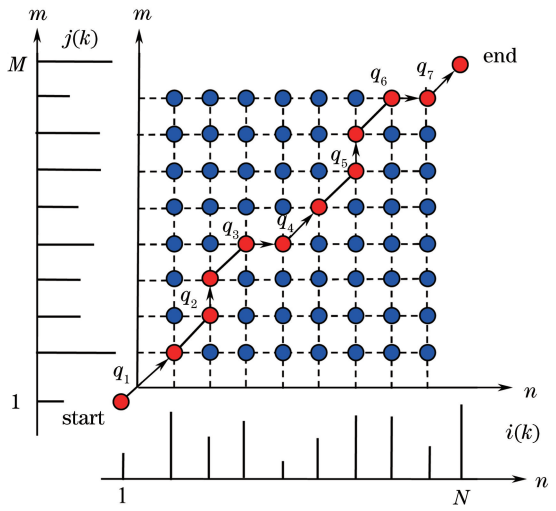

图 11　DTW 算法示意图

Fig. 11　Schematic diagram of DTW algorithm

2) 单调性约束。路径上每一点的搜索方向一定是向上或向右进行，反方向的搜索会出现无意义的循环。即 $i(k+1)\geqslant i(k),j(k+1)\geqslant j(k)$。

3) 局部连续性约束。该约束包括路径的全局走向和局部梯度两个要素。为避免路径过于平缓或陡峭，路径的斜率限制在[0.5,2]范围内，即若当前搜索点为 $c\langle i(k),j(k)\rangle$，则它的前一个搜索点只有 5 种情况——$c\langle i(k-1),c(k)\rangle$，$c\langle i(k-1),j(k-1)\rangle$，$c\langle i(k-1),j(k-2)\rangle$，$c\langle i(k),j(k-1)\rangle$，$c\langle i(k-2),j(k-1)\rangle$。图 12 所示为几种典型的局部路径。

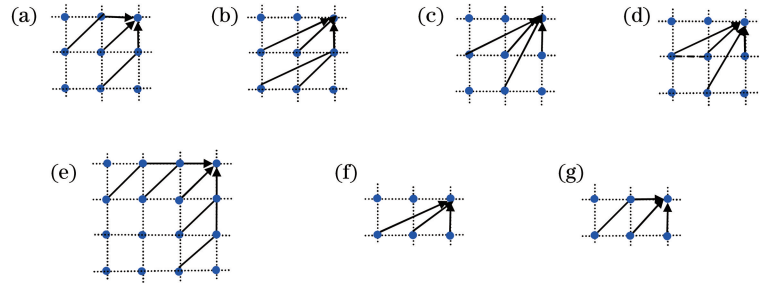

图 12　几种典型的局部路径约束示意图

Fig. 12　Schematic of typical local path constraint

基于上述三种约束条件，DTW 算法从"start"=(1,1)点开始，每经过一个格点，都会累加之前所有点的匹配距离，直到终点"end"=(M,N)，过程中寻找一条通过若干格点的最佳路径，使该路径上的格点所表示的累积距离之和最小。

设网格模型中任意一条有效路径为 $q=\{q_1,q_2,\cdots\}$，元素 q_i 代表路径 q 包含的所有子路径，设路径 q 上的格点依次为 $c_q(0),c_q(1),\cdots,c_q(r_q)$，其中 r_q 代表子路径的段数。根据局部连续性，有 $c_{q_{i+1}}(0)=c_{q_i}(r_{q_i})$。因此路径 q 的累积距离表示为

$$d(C_q)=\sum_{j=1}^{r_q} d\left[c_q(j)\right]w_{q,j}, \tag{5}$$

式中：C_q 代表路径 q 上所有格点的集合$\{c_q(0),c_q(1),\cdots,c_q(r_q)\}$；$w_{q,j}$ 代表每一段子路径的搜索权值，关于该权值的取值，文献[25]已做详细讨论。最优路径需要满足的失真度(即最小累加距离)表示为

$$D^*=\min_{q\in Q}D(q)=\rho\cdot\min_{q\in Q}\left[\sum_{i=1}^{L(q)-1} d(C_{q_i})\right], \tag{6}$$

式中：Q 为网格中所有有效路径的集合；归一化系数 ρ 将路径失真度归一化到测试序列的长度上，以消除规整路径中子路径个数的干扰，便于不同行为类别的比较。归一化系数 ρ 的表达式为

$$\rho(q)=\frac{N}{W(q)}, \tag{7}$$

式中 $W(q)=\sum_{i=1}^{L(q)-1}\sum_{j=1}^{r_{qi}} w_{q_i,j}$。由于人体行为过程满足马尔科夫性质，即每个动作的发生只与前一个动作有关，与过去无关；所以，$w_{q,j}=1$，$W(q)=N$。

3.2　基于椭圆约束的动态时间规划搜索策略改进

如图 13 所示，规整路径偏离了对角线附近的区域，形成"病态"扭曲路径，虽然符合路径搜索的几个约束条件，但是经 DTW 算法计算累积距离为 0，而且从直观上也可以看出两条时间序列并不相似；因此，对规整路径的搜索空间进行一定的几何约束，在计算之前剔除矩阵中无意义的元素点是十分必要的。

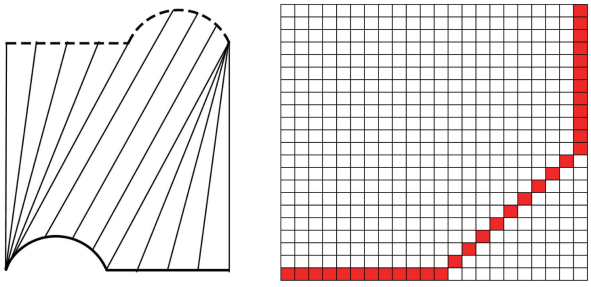

图 13　"病态"扭曲路径示意图

Fig. 13　Diagram of "Morbid" twisting path

根据端点约束及局部路径连续性约束可知，最优路径的搜索范围通常会集中在矩阵网格对角线附近的区域，设置一个形如椭圆边界的全局路径窗口对搜索范围进行约束，如图 14 中阴影区域指代代价矩阵中用于 DTW 算法处理的区域，最优路径只在该区域内搜索，当模板序列与测试序列内容越相似时，该约束的改进性能越明显。

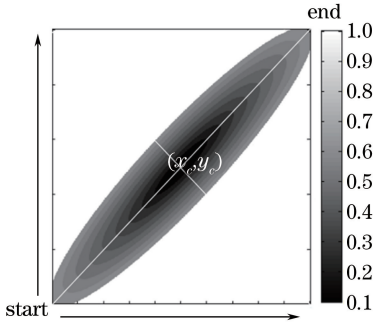

图 14　椭圆边界参数示意图

Fig. 14　Schematic of parameter of elliptic band

椭圆边界可由 5 个参数描述：中心点(x_C,y_C)，长轴 a，短轴 b，姿态角 $-\alpha\in(-\pi/2,\pi/2)$。在距离矩阵中，起始点"start"=(1,1)，终止点为"end"=(N,M)，中心点$(x_C,y_C)=(x_2/2,y_2/2)$，长轴 $a=\sqrt{|N|^2+|M|^2}$，短轴 $b=L$ 表征边界尺寸大小。参数示意图如图 14 所示。

以距离矩阵的左下角为坐标原点，椭圆公式可

表示为

$$\frac{(x-x_C)^2}{a^2}+\frac{(y-y_C)^2}{b^2}=1,b<a。\tag{8}$$

为了适应距离矩阵中规整路径的走向特点，对椭圆以圆心(x_C,y_C)为旋转中心进行一定角度的倾斜，设倾斜角为θ，旋转公式如下：

$$\begin{bmatrix}x\\y\end{bmatrix}=\begin{bmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}x^2\\y^2\end{bmatrix},\tag{9}$$

式中$P(x',y')$为旋转后椭圆边界上的点，$\theta=\arctan(M/N)$，设N为距离矩阵中纵向表示的序列、M为横向表示的序列。对于一般情况，椭圆边界内的元素(i,j)需满足：

$$\frac{(i\cdot\cos\theta-j\cdot\sin\theta-N/2)^2}{M^2+N^2}+\frac{(i\cdot\sin\theta+j\cdot\cos\theta-M/2)^2}{L^2}<1。\tag{10}$$

若两段时间序列长度相等，则$\theta=45°$。给定某一帧测试图像(y轴上的某一点)代入(10)式，得到两个x值，分别记为x_{max}、x_{min}。这样，在识别阶段y轴上的每一帧不需要与x轴上的所有帧进行比较，只须与x_{min}与x_{max}之间的数据帧比较即可，可显著地减少DTW算法进行匹配距离累积的计算量。

4　决策级融合

数据融合技术主要分为数据级、特征级和决策级三个融合层次。由于目标的运动特征和轮廓特征在维数上存在差异，若在特征级采用线性组合的方式直接对不同特征进行数据融合和降维，不仅没有考虑不同特征对识别结果贡献的差异性，而且也忽略了各个特征的受干扰程度，会导致识别误差较大；因此，采用决策级融合策略。决策级融合策略是一种高层次的融合方法，具有通信量小，抗干扰能力强，容错性好等优点。通过分类器获得每种特征关于类别的后验概率，然后按照一定策略融合各类别的度量信息，可以更有效地反映人体运动过程中各个侧面的不同类型信息，最后将融合结果中的最大值对应的类别作为最终识别结果。由于DTW算法的输出是沿着规整路径得到的累积匹配距离，因此需要将累积距离转换为概率形式输出，以满足后续计算需要。

根据模糊量度的相关理论，隶属函数可以将隶属度和距离联系起来，而隶属度可以理解为一种广义的概率，这样就可以将跟DTW相关的累积距离与决策级融合用到的后验概率联系起来。在没有确定的模糊度量的情况下，自定义一个应用于本研究的隶属度函数，参照模糊数学中模糊贴近度的概念，贴近度是表征模糊集接近程度的一种度量，某种程度上可以将贴近度看作是用来描述两类或者多类问题的隶属度。采用与搜索策略相关的距离来估计代价矩阵中对应两帧之间的贴近度，即：

$$\delta(m,n)=\exp[-D(m,n)],\tag{11}$$

式中$D(m,n)$对应匹配代价矩阵中的元素。对于空间域，$D(m,n)$代表测试序列第n帧和模板序列第m帧的形状上下文特征匹配值；对于时间域，$D(m,n)$代表测试序列上相邻两帧之间形状特征匹配值构成的序列上第n个数值与模板序列上相邻两帧之间形状特征匹配值构成的序列上第m个数值之间的匹配距离。计算最终的行为识别概率：

$$P=w_1\times\delta_s(m,n)+w_2\times\delta_t(m,n),\tag{12}$$

式中$w_i=p_i/\sum_{j=1}^{2}p_j$，$w_1$和$w_2$分别表示运动特征和轮廓特征的加权系数，即对于行为最终表征所占的比例。不同的比例分配对识别率会造成一定的影响，另外，对于不同的行为类别其比例分配也可能存在一定的差异，图15所示为采用不同分配比例的加权系数w_1和w_2，对不同行为类别识别结果的比较。从图中可以看出，形状特征和运动特征都会在一定程度上影响行为的整体表征，当$w_1=w_2=0.5$时，对于多数行为都能够达到最优识别率；因此，为简便计算，对于日常行为的识别，本文采用加权平均法进行决策融合。当然，对于特殊场合的特定行为识别，可以根据上述分析对两者的比例关系进行适当的调整。

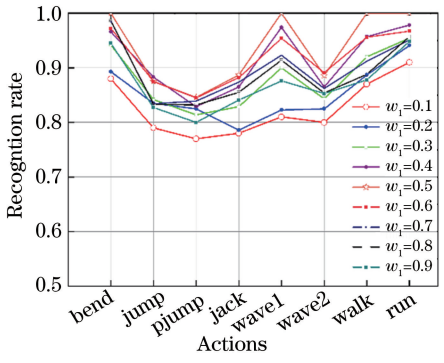

图15　权值分配比例对识别率的影响

Fig. 15　Recognition rate of different weight distributions

当测试序列进行融合决策时，匹配阈值可由多次实验获得，图16所示为不同匹配阈值对识别率的影响。

可以看出，当阈值处于0.85～0.95之间时，三

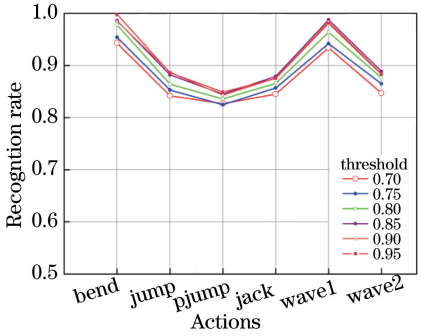

图 16　决策阈值对识别率的影响

Fig. 16　Influence of decision threshold values on recognition rate

条曲线几乎重合，说明识别效果区域稳定，且识别率能够取到最优效果，本文设置为 0.85。当两类特征同时属于某一类行为的融合后验概率高于此阈值时，则认定该类行为为最终识别结果。

5　实验结果及分析

实验环节主要分为两部分工作：1）针对利用椭圆边界约束来优化 DTW 的搜索效率问题进行相关实验，主要从计算复杂度、识别精度两方面对椭圆边界的约束性能进行分析；2）利用 Weizmann 和 KTH 公共数据库上提供的样本视频序列验证所提出的行为识别算法，首先通过实验验证所提的决策级融合形状特征和运动特征算法的有效性，然后分别展示所提行为识别算法在特定行为检测和行为分类识别中的结果，并与已有算法进行对比。

5.1　椭圆边界的约束性能分析

实验主要将椭圆边界与平行四边形和菱形两种全局边界进行对比。图 17 显示了相同边界尺寸下三种边界的形状示意图，其中，边界尺寸代表边界弯曲程度相同，即窗口沿对角线方向的长度占对角线的百分比，图中三种形状的边界尺寸均为 28%。窗口大小用窗口内包含的元素个数表征。

5.1.1　运行速度对比结果

首先，定义实验中对边界尺寸分别为 10%，50%，90%的约束边界在不同长度的时间序列下进行测试，分别用黑色、蓝色、红色表示平行四边形、椭圆、菱形边界。图 18 中横轴表示时间序列，长度范围设置为 10～6000 frame，纵轴表示运算时间。

从图 18 可以看出，当时间长度少于 1000 frame 时，所有的测试结果是相近的，当时间长度超过 4000 frame 时，基于椭圆边界的搜索策略在运算时间上要明显优于平行四边形边界。当边界尺寸为

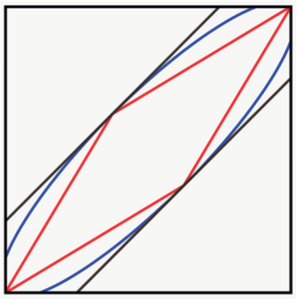

图 17　边界尺寸相同的三种常见的全局约束边界示意图

Fig. 17　Schematic of three global constraint boundaries with the same warping window size

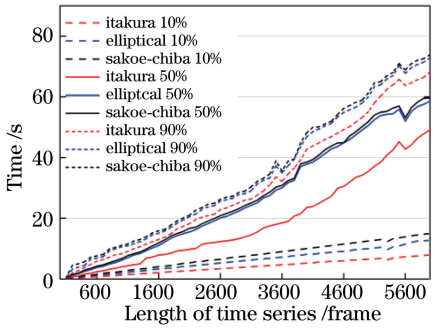

图 18　不同边界的搜索效率比较

Fig. 18　Comparison of searching efficiency of different bands

90%时，椭圆边界和平行四边形边界的搜索策略表现出相似的性能，这是由于在此条件下，二者内部包含的元素数量近乎相同。

由于识别系统多用于室内长时间监控，图 19 所示为长时间序列（5000 frame 以上）情况下不同边界形状及不同尺寸的搜索时间。

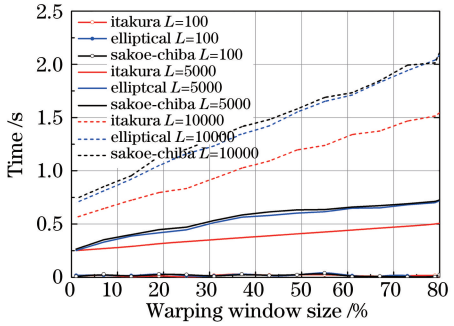

图 19　不同全局边界在较长时间序列下的搜索效率

Fig. 19　Searching efficiency of different global boundaries on large time series

可以看出，当对距离矩阵应用尺寸大小为 10% 的边界时，三者的运算时间均小于 1 s，当边界尺寸达到 50%时，运算时间均不超过 2 s。由图 18 和 19 的结果可以明显看出，基于椭圆边界约束的 DTW 搜索策略在运算时间上略优于常见的平行四边形约

束。至于菱形约束，虽然时间花销最小，但易遗漏有效匹配点，导致识别结果并不理想，因此菱形约束在运行时间上的略微优势竞争力较小。

5.1.2　识别精度对比结果

通常认为，全局边界尺寸越大，对识别精度的提升越有利。图 20 显示了在 Weizmann 行为数据库上应用不同边界尺寸对识别精度的影响，其中，实线代表模板序列与测试序列长度一致的情况（图例中用 Y 表示），虚线代表两序列长度不一致的情况（图例中用 N 表示）。图中纵轴表示精度值，横轴表示边界尺寸，范围为 1%～100%（因为 0 代表欧氏距离匹配，当测试序列与模板序列长度不一致时，是没有意义的），其中，100%代表没有约束的情况。

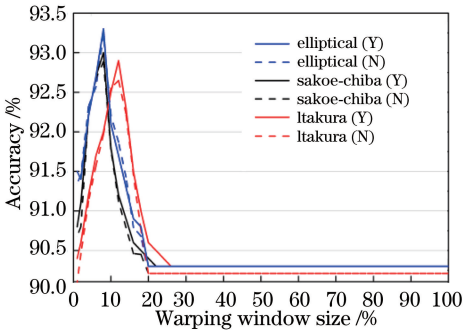

图 20　不同边界形状/尺寸对识别精度的影响

Fig. 20　Classification accuracies of different boundaries sharps and sizes

从图 20 可以看出，识别精度并不与边界尺寸呈正相关，三种边界的识别精度峰值均出现在边界尺寸较小的范围（5%～15%）内（椭圆和平行四边形边界的峰值出现在 5%～10%，菱形边界的峰值出现在 10%～15%），最高识别率可达到 93.3%。测试序列与模板序列长度是否一致并不影响这一结论，只是当两序列长度相同时，在识别精度方面更具有优势，在此区间内，三种边界的识别精度相差不超过 2%，但可以看出，椭圆边界的约束性能明显优于另两种。当边界尺寸处于 10%～18%时，精度呈下降趋势。这是由于边界扩大的同时，也增加了由人体行为随机性及肢体非刚性产生的近似轮廓的干扰。随着边界尺寸增加，边界对搜索策略的约束性能减弱，三种边界的精度近似，且均趋于稳定，但仍然可以看出，椭圆边界略优于平行四边形边界。

只考虑椭圆边界约束，利用不同间隔的下采样方式从序列中抽取一系列样本组成新的模板序列和测试序列，观察不同图像样本帧数下，边界尺寸对识别精度的影响，结果如图 21 所示。

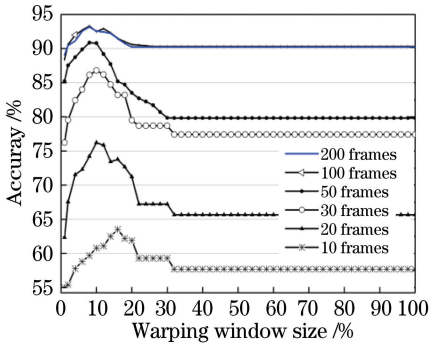

图 21　不同样本帧数下边界尺寸对识别精度的影响

Fig. 21　Classification accuracies of all warping window sizes with different frames

结果表明，调整窗口大小会影响准确性，而且效果也取决于匹配样本的数量，当样本数量减小时，识别精度会有所下降，而且精度峰值会对应更大的边界尺寸。比如，当只用 10 个样本时，精度只有 63%，对应的边界尺寸约 15%，而当对视频序列的连续帧进行测试时，精度可以达到 93%，对应的边界尺寸在 8%左右。当序列中的图像帧数超过 100 时，识别精度趋于稳定。

5.2　行为分类识别

5.2.1　KTH 行为数据库

KTH 行为数据库包含 6 种行为类别："walking""jogging""running""boxing""hand waving""hand clapping"，这 6 种行为分别由 25 人完成，并拍摄于 4 种场景（室外场景、室外拍摄距离变化场景、室外人物服饰变化场景和室内场景）。该数据库包含 2391 段行为视频，分辨率为 160 pixel×120 pixel。由于本文算法是基于少样本的行为识别方法，无法对类内差异极大的行为进行识别。因此，该实验设置不同于以往的交叉验证法或者基于拆分法，在实际实验中需要将每种场景的视频数据作为一个单独的行为数据库，随机地从每类行为中选择 3～5 个行为视频作为训练样本，对其他所有行为进行分类识别。根据研究范围，实验主要针对室内场景进行测试。同时，为了验证形状特征描述子和运动特征描述子各自的有效性和互补性，分别单独对其分类性能进行测试，识别结果的混淆矩阵如图 22 所示。

结果显示，本文算法在样本数量较少（3～5 个样本行为序列）的情况下，能够实现对特定行为的检测，其中，"jogging"和"running"两类动作由于个人在执行的速率上存在差异，"hand clapping"和"hand waving"两种行为之间存在一定的相似性，导致误

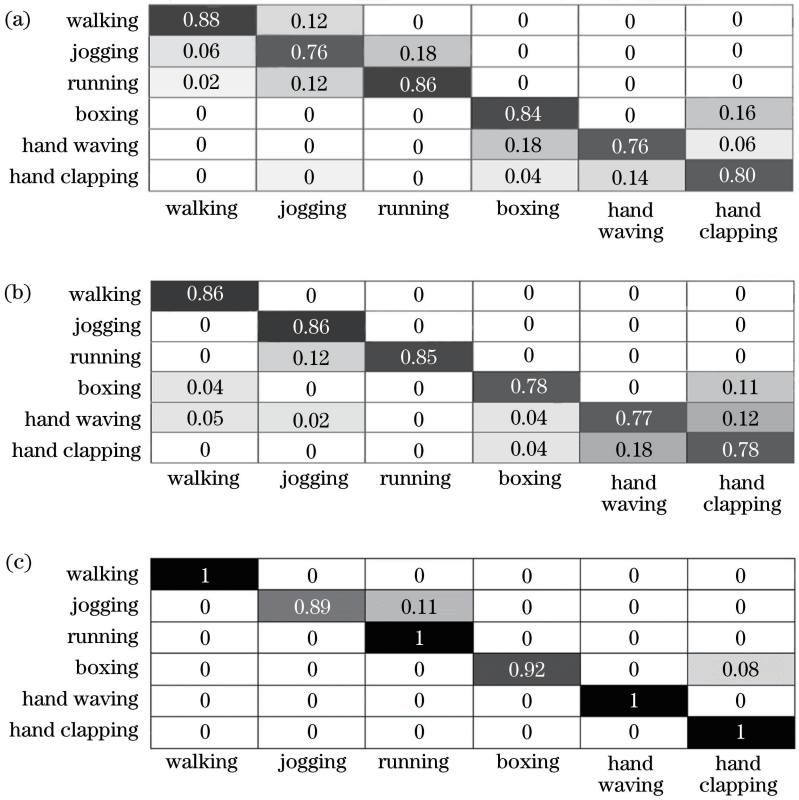

图 22　KTH 数据库上分类结果的混淆矩阵。(a)形状特征;(b)运动特征;(c)融合特征

Fig. 22　Confusion matrix of classification results on KTH dataset. (a) Shape feature; (b) motion feature; (c) fusion feature

分。表 1 为本文算法与其他方法在 KTH 数据库上的实验结果比较。

表 1　不同算法在 KTH 行为数据库上的平均识别率与运行时间比较

Table 1　Comparison ofaccuracy and computation time of different algorithms on KTH dataset

Algorithm	Average accuracy /%	Computation time /ms
Method in Ref.[7]	89.70	23.9
Method in Ref.[10]	85.67	30.3
Proposed method (motion)	84.89	18.9
Proposed method (shape)	81.11	19.3
Proposed method (fusion)	92.70	21.7

5.2.2　Weizmann 行为数据库

理论上,本文算法能够实现对不同行为的分类识别,为验证这一设想,在 Weizmann 行为数据库上进行多类行为识别实验。该数据库包含 93 个视频,由 9 人执行 10 种动作,包括“弯腰(bend)”、“向上跳跃(jump)”、“双腿跳着走(pjump)”、“开合跳(jack)”、“单腿跳着走(skip)”、“侧边走(side)”、“跑步(run)”、“行走(walk)”、“单手挥手(wave1)”、“双手挥手(wave2)”,视频分辨率为 144 pixel × 180 pixel,这些行为均采自摄像机固定条件下,环境背景单一,不受外界因素侵扰,可以用来模拟室内场景。不同于传统的对数据库样本进行交叉验证的实验方法,本文方法不需要模板训练过程,实验中随机地选取一个人的行为视频作为模板样本序列,其他 8 个人的行为视频序列作为测试样本序列,根据时间域和空间域上各自搜索到的最优状态匹配路径得到对应的行为分类,经过加权平均融合得到最终分类结果,并与实际情况比较,行为分类识别的混淆矩阵如图 23 所示。

从显示结果可以看出,对于样本库提供的几种行为,总体上识别结果比较理想,“jump”和“pjump”,以及“jack”和“wave2”这两对行为的识别结果略有偏差。这是由于人在运动过程中会产生一些非刚性形变,加之每个人的行为习惯不同,导致这些行为中会存在类似的动作,对识别结果产生一定的影响。表 2 为分别只用文中的形状特征和运动特征来识别,以及本文方法和其他方法在 Weizmann 数据库上的实验结果比较。可以看出,决策级融合方法稳健性更强,性能更优。

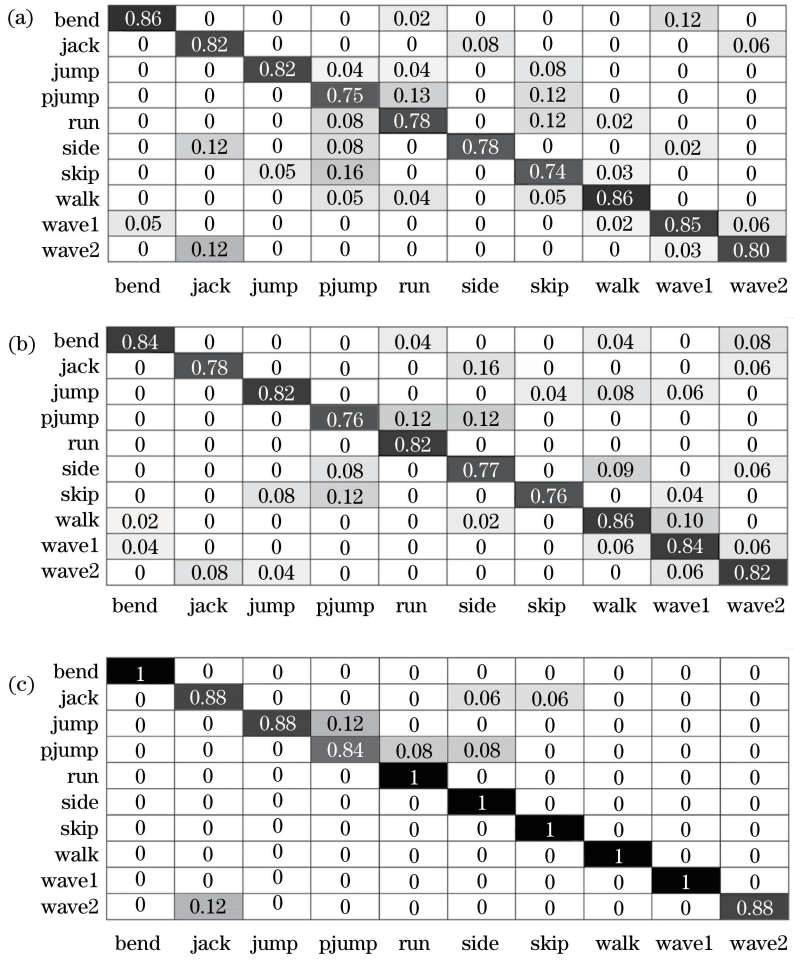

图 23　Weizmann 数据库上分类结果的混淆矩阵。(a)形状特征;(b)运动特征;(c)融合特征

Fig. 23　Confusion matrix of classification results on Weizmann dataset. (a) Shape feature; (b) motion feature; (c) fusion feature

5.2.3　UCF101 数据库

为证明本文算法的普适性和稳健性,进一步在 UCF101 数据库上进行实验。UCF101 数据动作库包含 101 类共计 13320 个动作视频,视频来源于网络和电视节目,分辨率均为 320 pixel×240 pixel。选择“跳高(HighJump)”“太极拳(TaiChi)”,“骑马(HorseRiding)”“跳远(LongJump)”“打保龄球(Bowing)”“举重(Lift)”6种常见运动类别进行测试,每种运动分别选取 100 段视频,采用 three-train/test split 交叉验证,分别选取训练样本和测试样本,并与当下主流方法进行识别率对比,结果如表 3 所示。

表 2　不同算法在 Weizmann 行为数据库上的平均识别率和运行时间比较

Table 2　Comparison of average accuracy and computation time of different algorithms on Weizmann dataset

Algorithm	Average accuracy /%	Computation time /ms
Method in Ref.[11]	89.26	35.7
Method in Ref.[8]	90.00	26.8
Proposed method (motion)	83.69	17.9
Proposed method (shape)	82.80	19.4
Proposed method (fusion)	93.20	20.9

表 3　不同算法在 UCF101 行为数据库上的平均识别率比较

Table 3　Comparison of average accuracy of different algorithms on UCF101 dataset

Algorithm	Average accuracy /%
Method in Ref.[9]	78.67
Method in Ref.[19]	81.60
Proposed method (motion)	77.90
Proposed method (shape)	76.50
Proposed method (fusion)	81.20

可以看出,本文算法在 UCF101 行为库上的识别率表现稍逊色于 KTH 和 Weizmann 数据库,这是因为 UCF101 数据库中的运动类别更加多变,场景也较复杂。另外,UCF101 数据库的视频长度相对于另两者更长,包含了除行为之外的环境信息,对

识别效果也会产生一定影响。文献[19]的效果具有一定的竞争力，该方法主要在时序动作检测的基础上，基于三维卷积(C3D)设计了一个卷积逆卷积网络，通过输入一小段视频，即可输出每一帧的动作类别概率。该方法的优势主要在于对时序动作检测的动作边界进行微调，对于时间尺度较大的视频更具优势，动作边界更加准确，也提高了检测精度。基于神经网络的识别方法，即便层数不多，训练过程也较本文方法复杂。因此，针对利用场景简单，特别是基于少样本的室内日常行为的精准识别，本文算法更加简便快捷。

经过在 KTH、Weizmann 和 UCF101 三种行为数据库上的测试，可以看出，单一的形状特征描述子在行为表征上略优于单一运动特征描述子。由于本文提取的动作特征是从形状特征在时间域的规律中学习得到的，能够很好地描述不同行为因运动速率不同而存在的类别差异，如跑、走等，然而对空间域上的特征变化匹配并不十分严格，因此对一些局部肢体动作较多的行为，如挥手、击掌等，会导致误判。总体来说，两者均具有良好的动作时空特征表征效果。将两种特征融合后，表征效果明显提高，由此也证明了二者的互补性。

6　结　论

提出一种基于时间-空间域特征在决策级融合的人体行为识别算法。首先从视频中提取有效的前景信息，然后从空间域和时间域两个角度出发，分别提取人体的形状特征和运动特征，其中，运动特征用随时间变化的形状特征表征。识别阶段利用动态时间规划算法分别计算两种特征在时间序列上的累积距离，并根据模糊数学的相关理论建立隶属度函数，将累积距离转换成后验概率，在决策级采用加权平均融合策略进行行为分类。另外，针对 DTW 算法提出一种基于椭圆边界约束的改进搜索策略，大大缩减了最优路径的搜索空间，有效剔除了视频中的噪声干扰。实验从计算复杂度、识别精度两方面对椭圆边界的约束性能进行分析，表明椭圆边界约束的改进性能均优于其余两种，并给出了识别精度达到最高时的边界尺寸范围。最后，分别在 Weizmann、KTH 及 UCF101 行为数据库对算法进行测试，平均识别率分别达到 93.2%、92.7% 和 81.2%。由于本文算法不需要训练大量样本，因此，利用单一样本对特定行为进行快速检测即可达到很好的效果，相比基于神经网络的识别方法更简便快捷，但是针对多分类行为的识别效果还有待进一步提升。另外，由于本文利用的形状上下文特征属于人体行为的全局表征，对遮挡因素比较敏感，因此，局部特征与全局特征的有效结合将作为下一步工作的重点。

参　考　文　献

[1] Li R F, Wang L L, Wang K. A survey of human body action recognition[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(1): 35-48.
李瑞峰, 王亮亮, 王珂. 人体动作行为识别研究综述[J]. 模式识别与人工智能, 2014, 27(1): 35-48.

[2] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic model for segmenting and labeling Sequence data[C] // Proceedings of the 18th International Conference on Machine Learning, 2001: 282-289.

[3] Wang J, Liu P, She M, *et al*. Human action categorization using conditional random field[C]. IEEE Workshop on Robotic Intelligence in Informationally Structured Space (RiiSS), 2011: 131-135.

[4] Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequential images using hidden markov model[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992: 379-385.

[5] Peursum P, Venkatesh S, West G. Tracking-as-recognition for articulated full-body human motion analysis[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2007: 1-8.

[6] Natarajan P, Nevatia R. Online, real-time tracking and recognition of human actions[C]. IEEE Workshop on Motion and video Computing, 2008: 1-8.

[7] Huang K, Zhang Y, Tan T. A discriminative model of motion and cross ratio for view-invariant action recognition[J]. IEEE Transactions on Image Processing, 2012, 21(4): 2187-2197.

[8] Niebles J C, Wang H, Li F. Unsupervised learning of human action categories using spatial-temporal words[J]. International Journal of Computer Vision, 2008, 79(3): 299-318.

[9] Wang H, Schmid C. Action recognition with improved trajectories[C]. IEEE International Conference on Computer Vision, 2013: 3551-3558.

[10] Schindler K, Van Gool L. Action snippets: how many frames does human action recognition require?[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[11] Liu J, Ali S, Shah M. Recognizing human actions using multiple features[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[12] Wang X Y, Zhang Y Z, Chen D Y. Face detection based on MB-LBP and eye tracking[J]. Chinese Journal of Scientific Instrument, 2014 (12): 2739-2745.
王小玉, 张亚洲, 陈德运. 基于多块局部二值模式特征和人眼定位的人脸检测[J]. 仪器仪表学报, 2014 (12): 2739-2745.

[13] Ando H, Fujiyoshi H. Human-area segmentation by selecting similar silhouette images based on weak-classifier response[C]. 20th International Conference on Pattern Recognition, 2010: 3444-3447.

[14] Fu Y, Guo J Y. Dynamic time warping-based human action recognition [J]. Eletronic Measurement Technology, 2014 (3): 69-72.
傅颖, 郭晶云. 基于动态时间规整的人体动作识别方法[J]. 电子测量技术, 2014(3): 69-72.

[15] Zhang J, Gao W, Liu A A, *et al*. Modeling approach of the video semantic events based on motion trajectories[J]. Eletronic Measurement Technology, 2013(9): 31-36.
张静, 高伟, 刘安安, 等. 基于运动轨迹的视频语义事件建模方法[J]. 电子测量技术, 2013(9): 31-36.

[16] An D, Rong C Q, Yang D, *et al*. Speaker recognition method basd on PSOA clustering and KMP algorithm[J]. Chinese Journal of Scientific Instrument, 2013(6): 107-112.
安冬, 荣超群, 杨丹, 等. 基于 PSOA 聚类和 KMP 算法的说话人识别方法[J]. 仪器仪表学报, 2013 (6): 107-112.

[17] Li Q H, Li A H, Wang T, *et al*. Two-stream networks with sequence opticsl flow image for action recognition[J]. Acta Optica Sinica, 2018, 38(6): 0615002.
李庆辉, 李艾华, 王涛, 等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报, 2018, 38(6): 0615002.

[18] Ijjina E P, Mohan C K. Human action recognition based on motion capture information using fuzzy convolution neural networks[C]. Eighth International Conference on Advances in Pattern Recognition, 2015: 1-6.

[19] Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks[C] // Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014: 1725-1732.

[20] Baccouche M, Mamalet F, Wolf C, *et al*. Sequential deep learning for human action recognition [C]. International Workshop on Human Behavior Understanding, 2011: 29-39.

[21] Shou Z, Chan J, Zareian A, *et al*. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1417-1426.

[22] Zhao S, Wang B, Tang C Y. Arm vein feature extraction and matching based on chain code[J]. Acta Optica Sinica, 2016, 36(5): 0515003.
赵珊, 王彪, 唐超颖. 基于链码表示的手臂静脉特征提取与匹配[J]. 光学学报, 2016, 36(5): 0515003.

[23] Li Y D, Xu X P, Chen J, *et al*. Background updating based on dynamic characteristic block matching in the application of the motion detection [J]. Chinese Journal of Scientific Instrument, 2017 (2): 445-453.
李艳荻, 徐熙平, 陈江, 等. 动态特征块匹配的背景更新在运动检测的应用[J]. 仪器仪表学报, 2017 (2): 445-453.

[24] Li Y D, Xu X P, Wang J Q. Feature extraction based on pyramid match kernel algorithm with adaptive partitioning[J]. Acta Photonica Sinica 2017, 46(12): 1210001.
李艳荻, 徐熙平, 王佳琪. 基于自适应分块金字塔匹配核的特征提取算法[J]. 光子学报, 2017, 46(12): 1210001.

[25] Rabiner L R, Juang B H. Fundamentals of speech recognition[M]. Englewood Cliffs: PTR Prentice Hall, 1993.