doi:10.11887/j.cn.201602012 http://journal.nudt.edu.cn

利用生成模型的人体行为识别*

王 军[1,2],夏利民[2],夏胜平[3]

(1. 电子科技大学中山学院 机电工程学院,广东 中山 528402;
2. 中南大学 信息科学与工程学院,湖南 长沙 410075;
3. 国防科技大学 自动目标识别国家重点实验室,湖南 长沙 410073)

摘 要:选取关键点轨迹的方向-大小描述符、轨迹形状描述符、外观描述符作为人体行为的特征;为了降低人体行为特征维数,利用信息瓶颈算法进行词表压缩;利用生成模型,结合已标记样本和未标记样本提出一种人体行为识别的半监督学习方法,解决了行为识别中的小样本问题。在YouTube数据库、中佛罗里达大学运动数据库上利用提出的方法与已有的方法进行对比实验,结果表明该方法具有更高的识别精度。

关键词:行为识别;词表压缩;信息瓶颈算法;生成模型

中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2016)02-068-07

Human behavior recognition using generative model

WANG Jun[1,2], *XIA Limin*[2], *XIA Shengping*[3]

(1. College of Mechanical and Electrical Engineering, University of Electronics Science and Technology, Zhongshan Institute, Zhongshan 528402, China;
2. School of Information Science and Engineering, Central South University, Changsha 410075, China;
3. National Key Laboratory of ATR, National University of Defense Technology, Changsha 410073, China)

Abstract: A novel method based on generative model was proposed for human behavior recognition. The behavior was represented by using a set of descriptors computed from key point trajectories, which included the orientation-magnitude descriptor, the trajectory shape descriptor and the appearance descriptor. In order to reduce feature dimensions, the agglomerative information bottleneck approach was used for vocabulary compression. The semi-supervised learning method for behavior recognition based on generative model was proposed to solve the problem of small sample in recognition, which made use of both the labeled and unlabeled samples. Compared with other state-of-the-art methods in both UCF sports database and YouTube database, results show that the proposed method has higher recognition accuracy.

Key words: behavior recognition; vocabulary compression; agglomerative information bottleneck; generative model

人体行为识别在视频监控、人机交互、运动与娱乐视频分析、虚拟现实等领域得到了越来越多的关注和应用,已成为计算机视觉领域的研究热点之一。但由于存在背景杂乱、遮挡、行为歧义性等问题,人体行为识别仍然是计算机视觉的难点。

人体行为识别的一个关键是行为特征的选择,直接影响到识别效果,常见特征包括全局特征和局部特征。其中,全局特征如空-时特征(Space-Time Volumes, STV)[1]、离散傅里叶系数(Discrete Fourier Transform, DFT)[2]对摄像机视角、噪声和遮挡较为敏感。而局部特征如尺度不变特征转换特征[3](Scale-Invariant Feature Transform, SIFT)、梯度直方图[4](Histogram of Oriented Gradient, HOG)、运动轨迹[5]等对图像平移、缩放和旋转具有不变性,且对噪声和光照变化的鲁棒性较强,但主要缺陷在于特征向量的维数过高。

人体行为识别的另一关键是识别方法,主要方法有统计法、句法方法和描述法。其中统计法是最常见的行为识别方法,如基于马尔可夫模型(Hidden Markov Models, HMMs)的行为识别[6-9]和基于动态贝叶斯网络模型(Dynamic Bayesian Networks, DBNs)的行为识别[10-13]。Natarajan等[9]将半隐马尔可夫模型和双层马尔可夫模型相结合,提出了基于双层半隐马尔可夫模型(Coupled Hidden Markov Models, CHMMs)的复杂行为识别方法;Bandouch[13]等将多层跟踪采样机制引入概率模型框架,利用贝叶斯模型进行行为

* 收稿日期:2015-05-31
基金项目:国家863计划资助项目(2009AA11Z205);国家自然科学基金资助项目(50808025)
作者简介:王军(1971—),男,山西应县人,讲师,博士研究生,E-Mail:106931289@qq.com

识别;另外 Hospedales[14] 等采用一种弱监督联合模型以及多类主题模型强监督联合主题模型(Weakly Supervised Joint Topic Mode, WSJTM)实现了少样本情况下行为的建模,实现了行为实时识别; Khoshhal[15] 等采用拉邦运动分析(Laban Movement Analysis, LMA)提取运动特征、采用二级概率模型进行建模识别人体行为,第一阶段利用贝叶斯网络估计人体运动参数,然后输入 HMM 中进行行为识别。然而,这些方法需要大量已知样本来训练概率模型,但在实际中很难得到足够的已知样本,这使得行为识别率下降。

为此,提出一种基于生成模型的人体行为识别方法。为了有效地表示人体行为,选取关键点轨迹的方向 - 大小描述符、轨迹形状描述符、外观描述符作为人体行为的特征;为了降低人体行为特征维数,利用信息瓶颈算法进行词表压缩;利用生成模型,结合标记样本和未标记样本提出人体行为半监督分类方法,解决了行为识别中的小样本问题,即利用已有的少量标记样本初始化模型,然后利用大量未标记样本对模型进行优化处理。在 YouTube 数据库、中佛罗里达大学(University of Central Florida, UCF)运动数据库上利用提出的方法与已有的方法进行对比实验。

1　人体行为特征

首先,利用金字塔卢卡斯奏托马西(Lucas - Kanade - Tomasi, KLT)跟踪器得到关键点轨迹[16]。在跟踪过程中,认为一条持续 5 帧图像的轨迹是"可靠的",短于 5 帧的轨迹自动被删除。当一条轨迹达到预定义的最大长度(25 帧)时,将自动分割同时生成一条新的轨迹。由于这些轨迹大部分是从背景区域中提取的,不是人体运动轨迹,因此,采用文献[17]中的轨迹修剪法,移除这些轨迹,同时保留描述人体行为的轨迹。在此基础上,提取下列人体行为特征。

1)方向 - 大小描述符。对于一条轨迹上的两个连续点:$\boldsymbol{p} = (x_l, y_l)$, $\boldsymbol{p}' = (x_{l+1}, y_{l+1})$,计算它们之间的位移向量 $\boldsymbol{d}_l = (x_{l+1} - x_l, y_{l+1} - y_l)$,对于长度为 L 的轨迹可计算出一组位移向量 $\boldsymbol{d} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{L-1}\}$,然后按下列方法对位移向量 $\boldsymbol{d}$ 的大小和方向进行量化。

对于位移向量大小的量化:首先,用同一轨迹中的最大位移量来归一化每个位移向量,然后,按 4 个均匀量化等级对位移向量大小进行量化。对位移向量方向的量化:将上、下半圆分为 8 个相等的扇形区,每个区域都为 22.5 度,如图 1 所示。根据大小、方向的量化,每个轨迹可由 32 位的直方图 O 表示,这种量化后描述符具有尺度不变性和方向不变性。

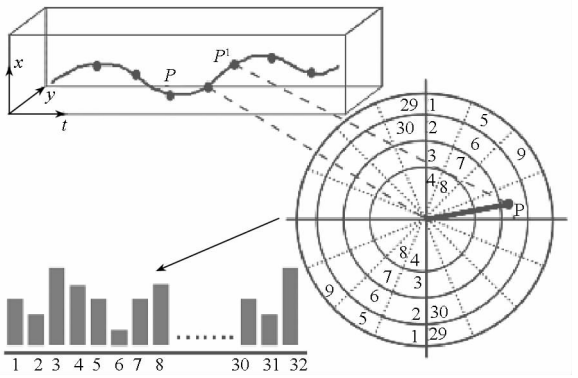

图 1　方向 - 大小描述符

Fig. 1　Orientation - Magnitude Descriptor

2)轨迹形状描述符。傅里叶描述符通常被用来表示物体的形状,以傅里叶描述符来描述行为轨迹形状。假设一个包含 L 个关键点 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_L, y_L)\}$ 的轨迹,可用由 N 个顶点 $\{z_i: i = 1, \cdots, N\}$ 组成的 2D 形状来表示轨迹。这 N 个顶点可由 N 个傅里叶变换系数 c_k 计算得到:

$$z_i = \sum_{k=\frac{-N}{2}+1}^{\frac{N}{2}} c_k e^{2\pi j \frac{ki}{N}} \tag{1}$$

傅里叶系数 c_k 表示轨迹的频率分量,其中,低频分量描述轨迹的近似形状,而高频分量则反映轨迹细节部分,因此,傅里叶系数提供了一个有效轨迹全局特征描述符。在 N 个傅里叶系数中,省略了 c_0,因为它表示一条轨迹的重心,且通过删除这个项,描述符具有平移不变性。此外,用 c_1 来归一化所有的傅里叶系数,使其具有缩放不变性。这样,每个轨迹可由 $N - 1$ 维的向量 F 表示了。

傅里叶描述符与方向 - 大小描述符不同,前者是全局形状描述符,反映一条轨迹中的全局行为信息,而后者是一种局部描述符,反映了轨迹的局部行为特征。两种描述符含有互补信息。

3)外观描述符。给定一个长度为 L 的轨迹,可提取 L 个关键点的 SIFT 特征 $S_i (i = 1, \cdots, L)$,则该轨迹的外观描述符 S 定义为 L 个关键点的 SIFT 特征的平均值:

$$S = \frac{1}{L} \sum_{i=1}^{L} S_i \tag{2}$$

4)轨迹表示。为了有效描述人体行为,采用词包(Bag Of Words, BOW)的方法将这三种互补的行为描述符结合在一起。对每条轨迹,将其描

述符 O, F 以及 S 归一化,并串联得到全局描述符 $G = [O,F,S]$。然后利用 BOW 法得到行为特征的 BOW 表示。具体步骤如下:

首先,利用 K 均值法生成 500 视觉词的码本来量表示全局描述符 G,并为每条轨迹分配一个码本。

其次,为了保留轨迹的时空信息,将视频中的一个兴趣区(Region Of Interest, ROI)的时空体积划分为 8 块,包括 4 个非重叠空间块和 2 个重叠的时间块(为 ROI 体积时间窗长度的 2/3)。随后每个时空块中的轨迹单独标记。这样可导致码本中有 $500 \times 8 = 4000$ 个视觉词来描述行为轨迹。

2 基于信息瓶颈算法的词表压缩

在上节建立了一个初始容量相对比较大的码本(4000 个视觉词)。为了减少特征维度,需要对词表进行压缩,这是一个典型的聚类问题,常见的聚类方法是 K 均值聚类法,但该方法很难确定视觉词的个数,并且这种聚类只考虑了视觉词之间的相似性,忽略了视觉词与行为类别间的关系,导致后续行为识别率下降,而信息瓶颈算法同时考虑了视觉词之间的相似性和视觉词与行为类别间的关系,因此可以得到一组有效的视觉词。

设离散随机变量 $A = \{a_1, a_2, \cdots, a_n\}$ 表示人体行为,其中 a_i 表示行为类别;随机变量 $W = \{w_1, w_2, \cdots, w_m\}$ 表示视觉词,而 w_i 表示第 i 个视觉词, $I(A,W)$ 表示 A, W 之间的互信息,而 $\tilde{W} = \{\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_k\}$ $(k < m)$ 表示 W 压缩后的视觉词,则由于视觉词压缩引起的互信息损失为:

$$D(\tilde{W}) = I(A,W) - I(A,\tilde{W}) \tag{3}$$

词表压缩就是寻找一个最佳的压缩词表 $\tilde{W}$,使得互信息损失最小。为此,采用凝聚信息瓶颈(Agglomerative Information Bottleneck, AIB)[18] 进行词表压缩,其基本思想是,迭代地将两类视觉词合并,这种合并引起互信息 $I(A,W)$ 减少是最小。

可以证明视觉词 w_i 和 w_j 合并造成互信息的损失为:

$$\begin{aligned} d(w_i,w_j) &= I(W_{\text{before}},Y) - I(W_{\text{after}},Y) \\ &= [p(w_i) + p(w_j)]JS[p(y|w_i), p(y|w_j)] \end{aligned} \tag{4}$$

其中, $JS[\cdot]$ 是 Jensen-Shannon 离散度,定义为:

$$JS_n[p_1,p_2,\cdots,p_M] = H[\sum_{i=1}^{M}\alpha_i p_i(x)] - \sum_{i=1}^{M}\alpha_i p_i(x) \tag{5}$$

式中, α_i 为权值, $H[p(x)]$ 是 Shannon 熵:

$$H[p(x)] = -\sum_{x} p(x)\lg p(x) \tag{6}$$

基于 AIB 的词表压缩算法如下:

1) 初始 $\tilde{W} = W$;

2) 对于 $\{w_i, w_j\} \in \tilde{W}, i < j$, 根据式(4)计算 d_{ij};

3) 合并:选择一对距离 $d(w_i,w_j)$ 最小的视觉词 $\{w_i,w_j\}$ 合并;

4) 重复步骤 2 和步骤 3,直到找到互信息最大(互信息损失最小)的 k 个视觉词为止。

3 基于生成模型的人体行为识别

在实际中,已标记的人体行为样本非常少,为了解决分类的小样本问题,采用基于生成模型的最大似然估计的半监督分类方法,首先,利用已有的少量已标记样本估计模型的参数,并以此作为模型参数的初始值;然后用大量未标记样本,通过递归计算方式对分类器参数进行优化处理,直到所有样本的似然函数收敛到局部极大值。对于待测样本,利用得到的分类器,计算其在各类别分布函数下的后验概率,以此进行分类。

3.1 概率生成模型

假设人体行为图像是由一个包含 c 类的混合模型生成的,且每个混合成分都满足一个特定的分布 $p(\boldsymbol{X}|\boldsymbol{\theta}_i)$,则数据的概率生成模型可表示为:

$$p(\boldsymbol{X}\mid\boldsymbol{\theta}) = \sum_{i=1}^{c} p(Y)p_i(\boldsymbol{X}\mid\boldsymbol{\theta}_i) \tag{7}$$

式中, $\boldsymbol{X}$ 为样本的特征向量, $p(Y)$ 代表该样本属于第 i 类的概率,或称先验概率; $\boldsymbol{\theta}_i$ 代表第 i 类样本的均值向量与协方差矩阵,也就是分类器训练过程中需要确定的参数, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_c\}$。假设每类行为近似符合高斯分布,用 $p(\boldsymbol{X}\mid\boldsymbol{\theta}_i)$ 表示,而整个样本集是由这些类别按比例混合生成的。

3.2 似然函数

样本集包括未标记样本和已标记样本,即 $D = L + U = \{(\boldsymbol{X}_1,Y_1),\cdots,(\boldsymbol{X}_l,Y_l),\boldsymbol{X}_{l+1},\cdots,\boldsymbol{X}_{l+u}\}$, $Y \in C = \{1,2,\cdots,c\}$, l 为已标记样本数, u 为未标记样本数。由于它们是由同一个混合模型生成的,所以其对数似然函数可写成下列形式:

$$\begin{aligned} &\ln L(\boldsymbol{\theta}\mid D) \\ &= \ln\Big\{\prod_{i=1}^{c}\prod_{k=1}^{l_i}[p_i(Y)p(\boldsymbol{X}_{ik}\mid\boldsymbol{\theta}_i)] \times \prod_{k=l+1}^{l+u} p(\boldsymbol{X}_k\mid\boldsymbol{\theta})\Big\} \\ &= \sum_{i=1}^{c}\sum_{k=1}^{l_i}[p_i(Y)p(\boldsymbol{X}_{ik}\mid\boldsymbol{\theta}_i)] + \sum_{k=l+1}^{l+u}\ln p(\boldsymbol{X}_k\mid\boldsymbol{\theta}) \end{aligned} \tag{8}$$

式中最后一个等式第一部分为监督分类部分,其仅涉及已标记的训练样本,$\boldsymbol{X}_{ik}$ 表示属于第 i 类的第 k 个已标记样本的特征向量,l_i 是属于第 i 类的已标记样本数目;而第二部分为无监督部分,其仅涉及未标记样本,$\boldsymbol{X}_k$ 表示未标记样本的特征向量。无监督部分可进一步写成:

$$\sum_{k=l+1}^{l+u} \ln p(\boldsymbol{X}_k \mid \boldsymbol{\theta}) = \sum_{k=l+1}^{l+u} \ln\Big(\sum_{i=1}^{c} p_i(Y) p(\boldsymbol{X}_k \mid \boldsymbol{\theta}_i)\Big) \tag{9}$$

将式(9)代入式(8)得到:

$$\ln L(\boldsymbol{\theta} \mid D) = \sum_{i=1}^{c} \sum_{k=1}^{l_i} (p_i(Y) p(\boldsymbol{X}_{ik} \mid \boldsymbol{\theta}_i)) + \sum_{i=l+1}^{l+u} \Big(\ln \sum_{i=1}^{c} p_i(Y) p(\boldsymbol{X}_k \mid \boldsymbol{\theta}_i)\Big) \tag{10}$$

与上述对数似然函数最大值对应的参数就是要估计的参数。

3.3　基于 EM 算法的分类参数估计

首先不考虑未标记样本的情况下,求式(10)最大值对应的参数,并作为模型参数的初始值,然后利用最大期望算法(Expectation Maximization algorithm, EM)算法来估计概率生成模型的参数。

E 步:应用对数似然函数(式(10))求未标记样本的概率值,即预测未标记样本的类别:

$$p_{jk}^{t} = p^{t}(Y = j \mid \boldsymbol{X}_k, \boldsymbol{\theta}_j) = \frac{p^{t-1}(Y = j) p(\boldsymbol{X}_k \mid Y = j, \boldsymbol{\theta}_j^{t-1})}{\sum_{j=1}^{c} p^{t-1}(Y = j) p(\boldsymbol{X}_k \mid Y = j, \boldsymbol{\theta}_j^{t-1})} \tag{11}$$

式中,p_{jk} 为当前参数分布下第 k 个未标记样本对应第 j 类的概率。$t-1$,t 表示迭代次数。

M 步:在已知当前未标记样本的预测类别之后,求似然函数取极大值时各参数的取值,即 $p(Y)$,$\boldsymbol{\mu}$(均值向量)和 $\boldsymbol{\Sigma}$(协方差矩阵):

$$p^{(t)}(Y = j) = \frac{\sum_{k=l+1}^{l+u} p_{jk}^{t-1} + l_j}{u + l} \tag{12}$$

$$\boldsymbol{\mu}_j^{t} = \frac{\sum_{k=l+1}^{l+u} p_{jk}^{t-1} \boldsymbol{X}_k + \sum_{k=1}^{u_j} \boldsymbol{X}'_{jk}}{u p^{t-1}(Y = j) + l_j} \tag{13}$$

$$\sum\nolimits_j^{(t)} = \frac{\sum_{k=l+1}^{l+u} p_{jk}^{t-1} \boldsymbol{Cov}_j(\boldsymbol{X}_k) + \sum_{k=1}^{u_j} \boldsymbol{Cov}_j(\boldsymbol{X}'_{jk})}{u p^{t-1}(Y = j) + l_j} \tag{14}$$

式中,$p(Y = j)$ 代表第 j 类的先验概率,$\boldsymbol{Cov}_j(\cdot)$ 表示协方差矩阵,u 和 l 分别是未标记样本和标记样本的数目,l_j 是属于第 j 类的已标记样本数目,而 $\boldsymbol{X}'_{jk}$ 表示属于第 j 类的第 k 个已标记样本。

不断重复 E 步和 M 步,直到收敛。其中收敛判别条件为:对数似然函数在相邻两次递归之间变化很小。

3.4　基于生成模型的人体行为识别

利用训练好的分类器,可识别人体行为,首先根据待识别行为的特征分别计算其在每个类别中的概率 $p(Y \mid \boldsymbol{X})$;然后,根据概率 $p(Y \mid \boldsymbol{X})$ 分类:若该样本在某类别分布函数下的后验概率 $p(Y \mid \boldsymbol{X})$ 最大,它便属于该类。

利用贝叶斯公式求得最大后验概率:

$$p(Y \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid Y) p(Y)}{\boldsymbol{\Sigma}_Y p(\boldsymbol{X} \mid Y) p(Y)} \tag{15}$$

由于采用的生成模型是假设高斯混合分布的,故式中 $p(\boldsymbol{X} \mid Y)$ 可由式(16)计算得到:

$$p(\boldsymbol{X} \mid Y) = \frac{1}{\sqrt{2\pi}\sqrt{|\boldsymbol{\Sigma}_Y|}} \exp\Big(\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_Y)^{\mathrm{T}} \boldsymbol{\Sigma}_Y^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_Y)\Big) \tag{16}$$

式中 $\boldsymbol{\mu}_Y$,$\boldsymbol{\Sigma}_Y$ 为属于类别 Y 的训练样本的均值向量和协方差矩阵,也就是分类器拟合过程中确定的参数向量 $\boldsymbol{\theta}$。

基于生成模型的行为识别步骤如下:

1)训练分类器。对于训练样本集 D,估计每个类别的先验概率 $p(Y)$;计算每个类别的均值向量和协方差矩阵,即估计参数 $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$。

2)行为识别。计算待识别行为对应各类别的后验概率,然后根据式(17)分类:

$$Y^* \rightarrow \boldsymbol{X} = \arg_Y \max p(Y \mid \boldsymbol{X}) \tag{17}$$

4　实验

为了验证本文方法的识别效率,分别在 YouTube 数据库、UCF 运动数据库中进行测试。测试环境是 Intel(R) Core(TM) i3 - 2310M CPU, 2.5 GHz 主频,2 G 内存的普通个人计算机(Personal Computer, PC),测试平台为 Windows XP 操作系统。实验中将文中方法与已有的一些识别方法进行了对比,对比方法包括:CHMMs 方法[9],DBNs 方法[13]和 WSJTM 方法[14]。

4.1　词表压缩实验

为了说明基于信息瓶颈算法的词表压缩方法

的有效性,分别采用K均值聚类法和信息瓶颈算法对词表进行压缩对比实验。词表中初始视觉词个数为4000个,利用信息瓶颈算法得到了750个最佳视觉词,为了比较效果,K均值聚类法也提取750个视觉词,在此基础上,采用文中的生成模型进行人体行为识别,实验结果如表1所示。从表1可看到,词表没有压缩时,行为识别率最低,识别时间最长,这是因为视觉词太多,计算量大,所以识别时间长,同时由于视觉词之间存在一定的相关性,使得识别率低;由于进行了词表压缩,K均值聚类法和信息瓶颈算法的识别时间明显减少(由于两者采用的视觉词个数相同,因此计算时间相同)。另外,从表1可看到信息瓶颈算法比K均值聚类法的识别率明显提高。

表1　不同方法的词表压缩结果

Tab.1　Vocabulary compression results of different methods

方法	没有压缩	K均值法	AIB法
识别率/%	80.3	82.1	91.6
识别时间/s	1.735	0.918	0.918

4.2　YouTube数据库

YouTube数据库包含11种动作类:投篮、骑自行车、潜水、高尔夫球挥杆、马术、足球运球、投球、网球发球、蹦床跳、排球扣球以及遛狗,图2为YouTube数据库的部分图像。该数据库视频存在大量的影响因素,如相机的运动;视角不同;目标外观以及姿势、尺寸不同;混杂的背景以及光照变化等。实验中,对投篮(a1)、骑自行车(a2)、跳水(a3)、高尔夫挥杆(a4)、骑马(a5)、足球运球(a6)、荡秋千(a7)、跳跃(a8)等行为进行了识别实验。图3为文中行为识别混淆矩阵表;表2为文中方法和其他方法在该数据库上的识别结果,从中可以看到,文中方法正确识别率达到了91.53%,比其他方法识别率都高,识别时间与其他方法接近。

注:从上至下依次为高尔夫球、跳水、骑马、自行车、网球、投篮行为。

图2　YouTube数据库部分图

Fig.2　YouTube database

	a1	a2	a3	a4	a5	a6	a7	a8
a1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
a2	0.00	0.98	0.00	0.00	0.02	0.00	0.00	0.00
a3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
a4	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
a5	0.00	0.02	0.00	0.00	0.97	0.00	0.00	0.01
a6	0.00	0.02	0.00	0.00	0.02	0.96	0.00	0.00
a7	0.00	0.00	0.01	0.00	0.00	0.00	0.87	0.12
a8	0.00	0.02	0.02	0.00	0.00	0.00	0.01	0.95

图3　在YouTube数据库上所提方法的混淆矩阵

Fig.3　Confusion matrix of the proposed method on YouTube database

表2　在YouTube数据库上不同方法的识别结果

Tab.2　Recognition results of different methods on YouTube database

方法	识别率/%	识别时间/s
CHMMs[9]	83.67	0.875
DBNs[13]	85.38	0.917
WSJTM[14]	86.37	0.926
所提方法	91.53	0.916

4.3　UCF运动数据库

UCF运动数据库包含多种运动视频,这些视频都是从电视直播频道收集得到的。该数据库包含了9种人体动作:跳水(16)、扣球(25)、举重(15)、骑马(14)、跑步(15)、溜冰(15)、投球(35)、高尔夫球挥杆(25)以及步行(22)等,每种行为后面括号里的数字表示数据库中含有该行为的相关视频数。文中选择了跳水(b1)、高尔夫挥杆(b2)、射门(b3)、举重(b4)、骑马(b5)、慢跑(b6)、溜冰(b7)、走(b8)等几个行为进行了实验。图4为该数据库中的部分图,图5为在该数据库上所提方法行为识别的混淆矩阵,表3为在该数据库上所提方法与其他方法的识别结果。从

表 3 可以看到所提方法的正确识别率达到了 91.72%,相比其他的方法都要高很多,识别时间与其他方法接近。

注:从上至下依次为高尔夫挥杆、射门、跳跃、举重、骑马、慢跑、滑轮、行走行为。

图 4　UCF 数据库的部分图

Fig. 4　UCF database

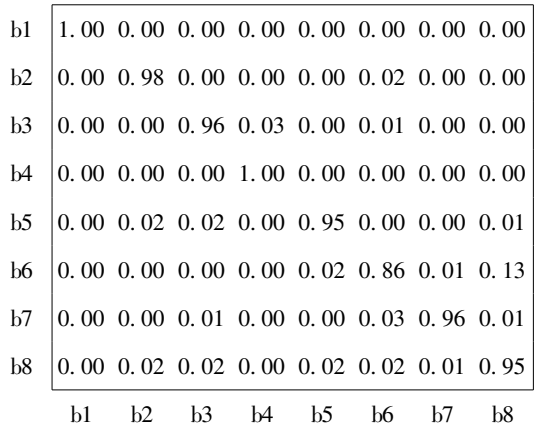

图 5　在 UCF 数据库上文中方法的混淆矩阵

Fig. 5　Confusion matrix of the proposed method on UCF database

表 3　在 UCF 数据库上不同方法的识别结果

Tab. 3　Recognition results of different methods on UCF database

方法	识别率/%	识别时间/s
CHMMs[9]	83.53	0.871
DBNs[13]	85.61	0.914
WSJTM[14]	86.43	0.921
所提方法	91.72	0.915

从表 2、表 3 可以看出,所提方法在相对复杂的 UCF 运动数据库和 YouTube 数据库中识别精度均有所提高。CHMMs 方法[9]采用一系列隐状态代表行为,通过计算各状态之间的转移概率识别行为,但需要足够多的训练样本,且无法识别时序结构复杂的行为,此外,大量的参数设置导致建模过程复杂;DBNs 方法[13]由于需要对识别系统进行实时更新,大大增加了系统难度,并且,需要较多的训练样本,在已知样本较少的情况下,系统识别不高,另外采用的是单一的人体关节点位置作为特征。因此上述两种方法在已知样本数量较少时识别效果不理想。WSJTM 方法[14]单一采用运动方向特征,利用弱监督主题模型进行人体行为识别,尽管所需要训练样本相对前两种方法少,但由于采用的是单一运动方向特征,很难有效描述复杂行为,因此识别效果也不太理想。所提方法由于采用了多个互补特征,能很好地描述人体行为,并且采用了半监督学习方法,利用少量已知样本就可得到准确的行为识别模型,所以文中方法比其他三种方法识别率要高。另外,尽管所提方法采用多特征描述行为,但由于采用了信息瓶颈算法对词表进行压缩,降低人体行为特征维数,所以识别速度与其他方法接近。

5　结论

提出一种基于生成模型的人体行为识别方法。主要工作如下:

1)提取关键点轨迹的方向 - 大小描述符、轨迹形状描述符、外观描述符作为人体行为的特征,由于这些特征具有互补性,能有效描述人体行为;

2)为了降低人体行为特征维数,利用信息瓶颈算法进行词表压缩;

3)利用生成模型,结合已标记样本和未标记样本提出人体行为半监督分类方法,从而解决行为识别中的小样本问题;

4)在 YouTube 数据库、UCF 数据库上利用所提方法与已有的方法进行对比实验,结果表明该方法具有更高的识别精度。

参考文献(References)

[1] Sun C, Junejo I N, Tappen M, et al. Exploring sparseness and self-similarity for action recognition[J]. IEEE Transactions on Image Processing, 2015, 24(8): 2488 - 2501.

[2] Shao L, Gao R Y, Liu Y, et al. Transform based spatio-temporal descriptors for human action recognition [J]. Neurocomputing, 2011, 74(6): 962 - 973.

[3] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition[C]// Proceedings of International Conference on Multimedia, Augsburg, Germany:

IEEE, 2007: 357 – 360.

[4] Kläser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients[C]// Proceedings of the British Machine Vision Conference, Bmvc, Leeds, UK: Springer, 2008: 213 – 222.

[5] Tu H B, Xia L M, Wang Z W. The complex action recognition via the correlated topic model[J]. Scientific World Journal, 2014, 2014(9): 983 – 990.

[6] Suk H, Jain A K, Lee S W. A network of dynamic probabilistic models for human interaction analysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2011, 21(7): 932 – 945.

[7] Xiang T, Gong S G. Beyond tracking: modelling activity and understanding behavior[J]. International Journal of Computer Vision, 2006, 67(1): 21 – 51.

[8] 韩磊, 李君峰, 贾云得. 基于时空单词的两人交互行为识别方法[J]. 计算机学报, 2010, 33(4): 776 – 783.
HAN Lei, LI Junfeng, JIA Yunde. Human interaction recognition using spatio-temporal words[J]. Journal of computers, 2010, 33(4): 776 – 783. (in Chinese)

[9] Natarajan P, Nevatia R. Coupled hidden semi-markov models for activity recognition[C]// Proceedings of IEEE Workshop on Motion and Video Computing, Austin, Texas, USA: IEEE, 2007.

[10] Zhang Y M, Zhang Y F, Swears E, et al. Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2468 – 2483.

[11] Damen D, Hogg D. Recognizing linked events: searching the space of feasible explanations[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA: IEEE, 2009: 927 – 934.

[12] Zhang Y M. Modeling temporal interactions with interval temperal Bayesian networks for complex activity recognition[C]// Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(10): 2468 – 2483.

[13] Bandouch J, Jenkins O C, Beetz M. A self-training approach for visual tracking and recognition of complex human activity patterns[J]. International Journal of Computer Vision, 2012, 99(2): 166 – 189.

[14] Hospedales T M, Li J, Gong S, et al. Identifying rare and subtle behaviors: a weakly supervised joint topic model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2451 – 2464.

[15] Khoshhal K, Aliakbarpour H, Mekhnacha K, et al. Lma-based human behaviour analysis using HMM[M]// Camarinha-Matos L M. Technological Innovation for Sustainability, Springer Berlin Heidelberg, 2011: 189 – 196.

[16] Matikainen P, Hebert M, Sukthankar R. Trajectons: action recognition through the motion analysis of tracked features[C]// Proceedings of the 12th IEEE International Conference on Computer Vision Workshops ICCV, Kyoto: Springer, 2009: 514 – 521.

[17] Bregonzio M, Li J, Gong S G, et al. Discrimina-tive topics modelling for action feature selection and recognition[C]// Proceedings of BMVC, Aberystwyth, UK: Springer, 2010: 1 – 11.

[18] Geiger B C, Petrov T, Kubin G, et al. Optimal kullback-leibler aggregation via information bottleneck[J]. IEEE Transactions on Automatic Control, 2015, 60(4): 1010 – 1022.