基于多层动态贝叶斯网络的人的行为多尺度分析及识别方法

杜友田 [1] 陈 峰 [1] 徐文立 [1]

摘 要 人的行为识别是视频内容分析和计算机视觉领域中的一个重要问题. 在分析了人的行为包含多个尺度运动细节的基础上, 提出了一种分层且带驻留时间状态的动态贝叶斯网络 (Hierarchical durational-state dynamic Bayesian network, HDS-DBN). HDS-DBN 含有多层状态, 能够较好地表示人的行为包含的多尺度运动细节. 我们针对单人行为和两人交互行为进行了识别, 实验结果表明该方法具有较高的识别率, 并且在有噪声存在或信息缺失等不确定情况下均具有较好的鲁棒性. 实验结果表明 HDS-DBN 模型确实能够较好地表达行为中的多尺度运动细节.

关键词 人的行为识别, 计算机视觉, 视频监控, 动态贝叶斯网络
中图分类号 TP391

Approach to Human Activity Multi-scale Analysis and Recognition Based on Multi-layer Dynamic Bayesian Network

DU You-Tian[1] CHEN Feng[1] XU Wen-Li[1]

Abstract Human activity recognition is an important issue in the fields of video content analysis and computer vision. Based on analyzing multiple scales of motion details contained in the human activities, we propose a novel human activity recognition approach named hierarchical durational-state dynamic Bayesian network (HDS-DBN). The HDS-DBN contains multiple levels of states and represents multiple scales of motion details as well. Experiments are conducted on the recognition of individual activities and two-person interacting activities. Experimental results show that the HDS-DBN recognizes human activities with high rates and has good robustness to the noise and loss of information. In addition, experimental results demonstrate that the HDS-DBN can represent multiple scales of motion details correctly.

Key words Human activity recognition, computer vision, video surveillance, dynamic Bayesian network

在过去的十几年内, 基于计算机视觉的人的行为识别研究得到了迅速的发展[1−2]. 行为识别的目的是从视频中分析人的行为并产生一系列高层的描述. 人的行为识别在众多应用中都有着重要的作用, 譬如视频智能监控、人机交互以及基于内容的视频检索等.

对于不同的场合和目的, 行为识别采用不同层次的运动特征: 粗略层 (Gross)、中间层 (Intermediate) 和细微层 (Detailed)[3]. 在粗略层上, 人通常被建模成一个点或者矩形, 其运动轨迹是行为分析的重要特征; 在中间层上, 人的几个重要部位的运动需要被表达出来; 在细节层上, 人的某个肢体的运动需要被精确表示出来. 不同层次的特征反映了不同尺度上的运动. 譬如, 人的运动轨迹反映了大尺度上的运动, 其状态变化较慢; 人肢体的位置及速度等特征反映了较小尺度上的运动, 其运动比较细微而且状态变化较快.

实际上, 人的行为同时包含着多个尺度的运动细节, 不同尺度反映着行为的不同特性. 过去的绝大多数研究工作只在某个单一尺度上来研究行为识别, 仅有少量的工作同时采用多个尺度的运动特征来分析人的简单运动[4−5]. 图 1 (见下页) 给出了尺度的一个例子. 其中, 与轨迹相关的运动主要反映人与人以及人与外界环境之间的关系, 体现了行为在大尺度上的特性; 与人的姿态相关的运动反映了人每个时刻的姿态, 譬如弯腰、伸手等, 体现了中间尺度上的运动特性; 人的肢体的运动反映了行为在小尺度上的细微特性. 这三个尺度在下文中简称为大尺度、中尺度及小尺度. 不同尺度上的运动细节在行为识别中具有不同的角色, 如何恰当地将其结合起来对于解决行为识别问题有着重要的意义.

本文提出了一个利用多尺度运动细节进行行为识别的框架. 在该框架下我们提出了一种分层且带驻留时间状态的动态贝叶斯网络 (Hierarchical durational-state dynamic Bayesian network, HDS-DBN). 在 HDS-DBN 中, 每层状态反映了对应尺度

收稿日期 2007-09-04 收修改稿日期 2008-04-29
Received September 4, 2007; in revised form April 29, 2008
国家自然科学基金 (60772050) 资助
Supported by National Natural Science Foundation of China (60772050)
1. 清华大学自动化系 北京 100084
1. Department of Automation, Tsinghua University, Beijing 100084
DOI: 10.3724/SP.J.1004.2009.00225

上行为的运动特性, 不同层状态之间的联系反映了不同尺度间的关系. HDS-DBN 能够较好地表达行为中多个尺度的运动细节, 并能够将其有机地融合起来对人的行为进行建模和识别.

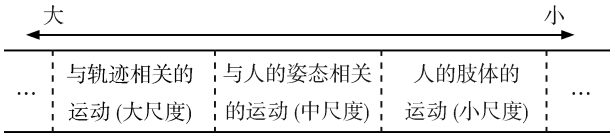


图 1 人的行为中的几个尺度

Fig. 1 Several scales in human activities

1 相关工作

在过去的十几年中涌现了很多人的行为识别方法, 如基于模板匹配的方法[6]、基于动态贝叶斯网络的方法[7−14] 以及基于文法技术的方法[15] 等. 其中基于动态贝叶斯网络的方法是被研究得最多的一类方法.

作为动态贝叶斯网络的一种, 隐马尔可夫模型 (Hidden Markov model, HMM) 被广泛地用于人的行为识别. 对于描述复杂的行为来说, HMM 的状态空间和参数空间都太大, 于是出现了很多改进的模型. 耦合隐马尔可夫模型 (Coupled hidden Markov model, CHMM)[7] 就是其中的一种, 它由多条马尔可夫链组成, 通常用来描述多人之间的交互行为. 另外还有一些其他的变形被用来分析交互行为, 如因子隐马尔可夫模型 (Factorial hidden Markov model, FHMM)[8]、观测分解隐马尔可夫模型 (Observation decomposed hidden Markov model, ODHMM)[9]、动态多链接隐马尔可夫模型 (Dynamically multi-linked hidden Markov model, DML-HMM)[10] 以及耦合半隐马尔可夫模型 (Coupled hidden semi Markov model, CHSMM)[11]. 分层隐马尔可夫模型 (Hierarchical hidden Markov model, HHMM)[12] 被用来分析时间序列里隐含的层次结构, 常被用于分析人的行为中蕴含的多层结构. 另外, 抽象隐马尔可夫模型 (Abstract hidden Markov model, AHMM)[13] 和层次隐马尔可夫模型 (Layered hidden Markov model, LHMM)[14] 也常常用来分析人的行为.

在以往的研究中, 针对特定的应用环境, 大多数方法是在某个单一的层次上来研究人的行为识别[3]. 目前只有少量的工作研究如何采用多个尺度上的特征来研究行为识别. Pers[4] 初步探讨了人行为中的多个尺度, 并采用简单的线性分类器对人的运动进行了分类, 结果表明采用的两个尺度的信息起到了互相补充的作用, 其识别效果好于单一的尺度. Fanti[5] 采用了人的位置及速度和人的表观两个层次的特征来对人的简单运动进行了识别, 其方法识别率较高, 模型的鲁棒性也较好.

2 HDS-DBN 模型

HDS-DBN 模型可由下面的多元组表示:

$$\mathcal{M}=(Q^{1:L}, D^{1:L-1}, O^{1:L}, \pi^{1:L}, A^{1:L}, B^{1:L}, P^{1:L-1}) \quad (1)$$

式中, $Q^{1:L}$、$O^{1:L}$ 和 $D^{1:L-1}$ 分别表示第 1 到 L 层的行为状态、观测及第 1 到 $L-1$ 层的驻留时间状态. HDS-DBN 模型是一种特殊的动态贝叶斯网络, 它的状态空间被分解为 L 个行为状态的集合 $\mathcal{Q}^l = \{1, 2, \cdots, |Q^l|\}$ 和 $L-1$ 个驻留时间状态的集合 $\mathcal{D}^h = \{d|d \in \mathbf{N} \cup \{0\}\}$, 其中 $\mathbf{N}$ 表示自然数集. HDS-DBN 模型的参数 Θ 包括初始概率 $\pi^l = \{\pi^l_{ij}\}$、转移概率 $A^l = \{a^l_{ijk}\}$、观测概率分布 $B^l = \{b^l_i(O_t)\}$ 以及行为状态的驻留时间分布 $P^h = \{p^h_i(d)\}$, 其中当 $l = 1$ 时, π^l_{ij} 和 a^l_{ijk} 退化为 π^1_j 和 a^1_{jk}. 即 HDS-DBN 模型的参数 Θ 可表示为 $\Theta = \{\pi^{1:L}, A^{1:L}, B^{1:L}, P^{1:L-1}\}$. 在文献 [16] 中定义了一个只含有 2 层状态的模型, 而本文将文献 [16] 中的模型推广到了任意层, 为描述具有多个尺度细节的视频内容建立了一个统一的模型. 图 2 显示了具有 3 层状态的 HDS-DBN 模型.

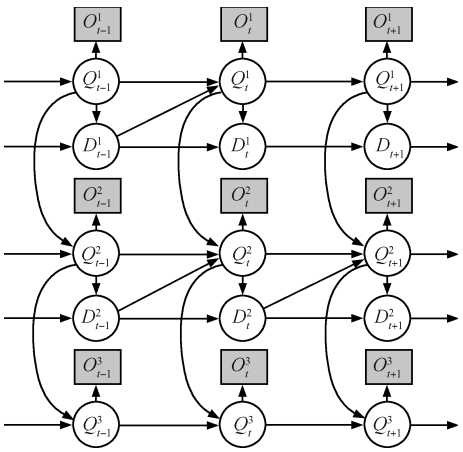


图 2 3 层 HDS-DBN 模型

Fig. 2 3-level HDS-DBN model

HDS-DBN 模型的条件概率分布 (Conditional probability distribution, CPD) 表示如下

$$P(Q^l_1 = j|Q^{l-1}_1 = i) = \begin{cases} \pi^1_j, & l = 1 \\ \pi^l_{ij}, & l = 2, \cdots, L \end{cases} \quad (2)$$

$$P(Q^l_t = k|Q^l_{t-1} = j,\, Q^{l-1}_t = i,\, D^l_{t-1} = d) = \begin{cases} \delta(j-k), & d > 0,\, l = 1, 2, \cdots, L-1 \\ a^1_{jk}, & d = 0,\, l = 1 \\ a^l_{ijk}, & d = 0,\, l = 2, 3, \cdots, L-1 \\ a^L_{ijk}, & l = L \end{cases} \quad (3)$$

$$P(D_t^l = d'|D_{t-1}^l = d, Q_t^l = i) = \begin{cases} p_i^l(d'), & d = 0 \\ \delta(d' - d + 1), & d > 0 \end{cases} \tag{4}$$

$$P(O_t^l|Q_t^l = i) = \sum_{k=1}^{M_i^l} w_{ik}^l N(O_t^l, \boldsymbol{\mu}_{ik}^l, \Sigma_{ik}^l) \tag{5}$$

其中, $p_i^l(d)$ 是行为状态 $Q^l = i$ 的驻留时间分布, $N(\cdot,\cdot,\cdot)$ 是高斯分布函数, $\delta(\cdot)$ 是 delta 函数, M_i^l, w_{ik}^l, $\boldsymbol{\mu}_{ik}^l$ 和 Σ_{ik}^l 分别是第 l 层观测的概率分布中的高斯分布个数、权重、均值向量和协方差矩阵.

2.1　模型结构分析

HDS-DBN 模型采用 L 层的行为状态来描述 L 个不同尺度上的运动细节. 每层行为状态中, 采用常用的马尔可夫性假设, 即 Q_t^l 依赖于 Q_{t-1}^l. 另外在不同的尺度间, 尤其是相邻的两个尺度间, 运动细节之间有很强的相关性. 通常在人的行为中, 大的尺度占有一定的主导性, 故我们只考虑相邻两个尺度间大尺度上的运动细节对小尺度上的影响, 在模型上则体现为 Q_t^{l-1} 对 Q_t^l 的影响, 如图 2. 另一方面, HDS-DBN 模型中不同层次的行为状态描述不同尺度大小的运动细节, 且 l 越大, 对应的尺度越小. 通常大尺度上的运动状态会持续较长的时间, 传统的 HMM 中状态驻留时间服从的几何分布不能准确反映实际情况[17]. 故在 HDS-DBN 的 $l = 1, 2, \cdots, L-1$ 层中加入驻留时间状态来改变行为状态的驻留时间分布. 由于第 L 层对应的尺度最小, 其行为状态转移的频率最高, 本身固有的几何分布基本符合实际情况, 故在第 L 层中不再加入驻留时间状态.

HDS-DBN 模型具有如下优点: 1) 它能够较好地表达人的行为中包含的多个尺度上的运动细节; 2) 它可以捕获不同尺度上的重要运动细节; 3) 和传统的 HMM 相比, HDS-DBN 模型通过分解的方式大大降低了参数空间的大小, 对于模型的训练有很大好处.

HDS-DBN 模型与 CHMM 不同, 后者一般由多条对称的马尔可夫链构成, 而 HDS-DBN 则是由描述不同尺度的多条不对称链构成, 它能够描述处于不同尺度的多个随机过程; 另外, 在描述人的行为的层次结构方面, HDS-DBN 模型与 HHMM 也不同, 后者是从单一的观测序列中来学习模型状态的意义和行为的结构的.

2.2　驻留时间分布

传统的 HMM 中, 状态的驻留时间长度 d 服从几何分布, 假设 a 是 HMM 中状态的自跳转概率, 则 d 服从 $P(d) = a^d(1-a)$. 但是对于很多场合, 尤其是人的行为识别来说, 几何分布是不合适的. 很多研究者采用了均匀分布[18]、Gamma 分布[19] 以及 Poisson 分布[20] 等, 并且取得了较好的效果. 本文采用离散 Coxian 分布, 该分布曾在文献 [21] 中得到了较好的应用. 离散 Coxian 分布是 Phase-type 分布的一种, 定义如下

$$DCox(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i=1}^{H_c} \mu_i S_i \tag{6}$$

其中, $DCox(\boldsymbol{\mu}, \boldsymbol{\lambda})$ 表示服从以 $\boldsymbol{\mu}$ 和 $\boldsymbol{\lambda}$ 为参数的离散 Coxian 分布的随机变量, $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_{H_c})^{\mathrm{T}}$, 且 $\sum_{i=1}^{H_c} \mu_i = 1$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_{H_c})^{\mathrm{T}}$, $S_i = X_i + \cdots + X_{H_c}$, X_i 是随机变量且服从几何分布, 即 $X_i \sim Geom(\lambda_i)$, H_c 为离散 Coxian 分布的阶段数. 实际上, 离散 Coxian 分布等价于一个 "从左到右" 型的马尔可夫链驻留时间的概率分布, 其参数 μ_i 和 λ_i 分别为状态 i 的初始概率和自跳转概率, 如图 3. 离散 Coxian 分布的一个突出优点是在 HDS-DBN 模型的计算中, 复杂度与离散 Coxian 分布的阶段数 H_c 有关, 而 H_c 通常远远小于状态的最长驻留时间. 而采用其他分布时, 计算复杂度与状态的最长驻留时间有关, 故采用离散 Coxian 分布可以显著地减小计算量.

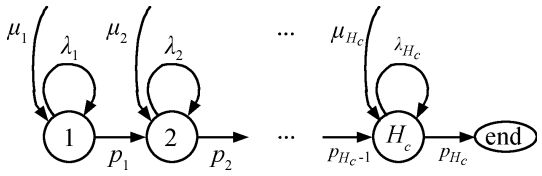


图 3　离散 Coxian 分布对应的 "从左到右" 马尔可夫链

Fig. 3　The left-to-right Markovian chain corresponding to discrete Coxian distribution

2.3　模型推理和参数估计

作为一种特殊的动态贝叶斯网络, 目前已有的动态贝叶斯网络推理算法都可以用于 HDS-DBN 的推理计算, 如团树算法等[22]. 令 $S_t = (Q_t^1, Q_t^2, \cdots, Q_t^L, D_t^1, D_t^2, \cdots, D_t^{L-1})$ 表示 t 时刻的状态, $S_{1:T} = (S_1, S_2, \cdots, S_T)$ 表示状态序列, $O_t = (O_t^1, O_t^2, \cdots, O_t^L)$ 表示 t 时刻的观测数据, $O_{1:T} = (O_1, O_2, \cdots, O_T)$ 表示观测序列. 在给定观测序列时, 可以根据推理算法求解模型的似然值 $P(O_{1:T}|\Theta)$, 以及状态平滑如 $P(S_t|O_{1:T}, \Theta)$ 和 $P(S_t, S_{t+1}|O_{1:T}, \Theta)$ 等. 本文采用团树算法对 HDS-DBN 模型进行推理.

模型学习的任务是根据给定的训练数据来估计模型的参数. 给定用于训练的观测序列 $O_{1:T}$, HDS-DBN 模型的参数 Θ 可以由最大似然方法估计, 即 $\hat{\Theta} = \arg\max_{\Theta} P(O_{1:T}|\Theta)$. 由于状态序列

$S_{1:T}$ 是不能被观测到的, 故需要通过 Expectation-maximization (EM) 算法来进行迭代求解

$$\Theta^{(n+1)}=\arg\max_{\Theta}\mathrm{E}\{\log P(O_{1:T},S_{1:T}|\Theta)|O_{1:T},\Theta^{(n)}\} \tag{7}$$

其中 $\Theta^{(n)}$ 是第 n 次的迭代结果. 通过式 (7) 的迭代计算, 参数可收敛到一个局部极值点. 在式 (7) 中, 观测序列 $O_{1:T}$ 和状态序列 $S_{1:T}$ 的联合概率密度如下式表示

$$P(O_{1:T},S_{1:T}|\Theta)=\prod_{t=1}^{T}\left(\prod_{l=1}^{L}P(Q_t^l|pa(Q_t^l))\times P(O_t^l|pa(O_t^l))\prod_{h=1}^{L-1}P(D_t^h|pa(D_t^h))\right) \tag{8}$$

其中 $pa(X)$ 表示节点 X 的父节点.

基于动态贝叶斯网络的人的行为识别实际上是一个推理过程. 给定 C 个已经训练好的 HDS-DBN 模型 $\mathcal{M}_1,\mathcal{M}_2,\cdots,\mathcal{M}_C$, 其中每个模型对应一类行为. 给定一个测试行为, 其观测序列为 $O_{1:T}$, 则我们选择 c 作为该测试行为的类别标号, 其中 c 由下式确定

$$c=\arg\max_i P(\mathcal{M}_i|O_{1:T})=\arg\max_i P(O_{1:T}|\mathcal{M}_i)P(\mathcal{M}_i) \tag{9}$$

式中 $P(\mathcal{M}_i)$ 是模型 $\mathcal{M}_i$ 的先验概率, 可以取为 $1/C$. 假设模型 $\mathcal{M}_i$ 的参数是 Θ_i, 则 $P(O_{1:T}|\mathcal{M}_i)=P(O_{1:T}|\Theta_i)$.

3 特征提取

本文采用图 1 中列出的三个尺度, 对应的特征分别称为大尺度特征、中尺度特征和小尺度特征.

3.1 大尺度特征

在大尺度上, 人的身体采用人的中心 (x^c,y^c) 表示. 对于多人之间的交互行为, 大尺度特征包括: 1) 第 i 个人的速度大小 $|\boldsymbol{v}_i|$; 2) 第 i 个和第 j 个人之间的距离 d_{ij}; 3) $\boldsymbol{v}_i$ 和 $\boldsymbol{r}_{ij}$ 之间的角度 φ_{ij}, 其中 $\boldsymbol{r}_{ij}$ 表示在视频序列开始时从第 i 个人的中心到第 j 个人的中心的向量. 对于单人的行为, 若场景中有明显的参照物体, 大尺度特征包括: 1) 人的速度大小 $|\boldsymbol{v}|$; 2) 人和参照物体间的距离 d; 3) $\boldsymbol{v}$ 和 $\boldsymbol{r}$ 之间的角度 φ, 其中 $\boldsymbol{r}$ 表示在视频序列开始时从人的位置到参照物体位置的向量. 若场景中没有明显的参照物体, 大尺度特征则只包含人的速度大小 $|\boldsymbol{v}|$.

3.2 中尺度特征

我们基于人的轮廓来提取行为的中尺度特征. 人的轮廓可以用上面的 K 个标记点 $p_1,p_2,\cdots,p_K$ 来描述, 故一个轮廓可以用一个 K 维的复向量来表示: $\boldsymbol{x}^p=(x_1^p+\mathrm{j}y_1^p,x_2^p+\mathrm{j}y_2^p,\cdots,x_K^p+\mathrm{j}y_K^p)^{\mathrm{T}}$, 其中 x_i^p 和 y_i^p 分别是点 p_i 的横坐标和纵坐标, $i=1,2,\cdots,K$. 为了使 $\boldsymbol{x}^p$ 对于轮廓的平移和大小具有不变性, 向量 $\boldsymbol{x}^p$ 需要进行归一化

$$\hat{\boldsymbol{x}}^p=\frac{U\boldsymbol{x}^p}{\|U\boldsymbol{x}^p\|},\quad U=I_K-\frac{1}{K}\mathbf{1}_K\mathbf{1}_K^{\mathrm{T}} \tag{10}$$

其中, I_K 是 $K\times K$ 的单位矩阵, $\mathbf{1}_K$ 是长度为 K 且元素全为 1 的列向量, $\|\cdot\|$ 表示向量范数. 若上式可以表示为 $\hat{\boldsymbol{x}}^p=(\hat{x}_1^p+\mathrm{j}\hat{y}_1^p,\hat{x}_2^p+\mathrm{j}\hat{y}_2^p,\cdots,\hat{x}_K^p+\mathrm{j}\hat{y}_K^p)^{\mathrm{T}}$, 则用实向量 $\hat{\boldsymbol{x}}'^p=(\hat{x}_1^p,\hat{y}_1^p,\hat{x}_2^p,\hat{y}_2^p,\cdots,\hat{x}_K^p,\hat{y'}_K^p)^{\mathrm{T}}$ 来表达人的轮廓. 向量 $\hat{\boldsymbol{x}}'^p$ 通常处于一个高维空间中, 而且其分布呈严重的非线性. 本文采用 LLE 方法[23] 对其进行非线性降维, 降维后得到的 V 维向量作为行为的中尺度特征. 在本文实验中, $V=4$.

3.3 小尺度特征

在小尺度上, 我们采用人的四肢和头的位置来刻画人的运动. 即提取人的头、两手及两脚的中心点位置 x_i^b,y_i^b 作为小尺度特征, $i=1,2,\cdots,5$. 这些关键部位的位置对于分析人的行为中更为微小、精细的动作具有重要意义. 为了使小尺度特征对于人的平移和大小具有不变性, 需对其进行归一化: $\tilde{x}_i^b=(x_i^b-x_c)/B$, $\tilde{y}_i^b=(y_i^b-y_c)/B$, 其中 $B=[\sum_{i=1}^{5}(x_i^b-x_c)^2+(y_i^b-y_c)^2]^{1/2}$. 归一化的 $\tilde{x}_i^b$, $\tilde{y}_i^b$ 排成一个向量即构成小尺度特征. 涉及多人时, 需要把每个人对应的特征排成一个向量.

4 实验结果和分析

本文实验中, 我们采用 2 层的 ($L=2$) 和 3 层的 ($L=3$) HDS-DBN 模型来对单人行为和两人交互行为进行识别. 2 层的 HDS-DBN 模型采用了两个尺度: 大尺度和中尺度; 3 层的 HDS-DBN 模型采用了三个尺度: 大尺度、中尺度和小尺度. 模型中的观测数据 O^1,O^2,O^3 分别对应于大尺度特征、中尺度特征和小尺度特征. 行为样本由多个人进行模拟并通过固定的单摄像头获取. 每类行为中的数据被分为两半, 一半用于训练, 另一半用于测试. 图 4 是一个行为样本中的几个关键帧及其运动跟踪的结果.

图 4 一个行为样本中的几个关键帧及跟踪结果

Fig. 4 Several key frames and tracking results of some activity sample

4.1　单人行为识别结果

单人行为数据包含以下五个类别: 1) Act 1: 人在场景中沿较为固定的方向行走; 2) Act 2: 人在场景中沿较为固定的方向慢跑; 3) Act 3: 人在场景中行走, 在行走期间从地上捡起一个物体, 然后继续向前走; 4) Act 4: 人在场景中行走, 在行走期间挥手一小段时间; 5) Act 5: 人在场景中行走, 在行走期间摔倒在地, 而后爬起继续前行. 该数据集的行为样本长度在 130 到 200 帧之间. 根据行为复杂性的不同, 2 层的 HDS-DBN 模型中的状态个数 $|Q^1|$ 取值为 3 或 4, $|Q^2|$ 也取值为 3 或 4. 状态数目是通过预先的实验确定的, 目的是使得识别效果最好. 表 1 列出了识别结果. 从结果来看, 识别效果较好, 尤其是能够较好地根据一些运动细节区分有些相近的行为, 譬如 Act 3 和 Act 5.

表 1　采用 2 层 HDS-DBN 模型时的单人行为识别结果

Table 1　The individual activity recognition results using 2-level HDS-DBN

行为类别	识别率 (%)	识别结果				
		Act 1	Act 2	Act 3	Act 4	Act 5
Act 1	95.5	21	0	0	1	0
Act 2	100	0	28	0	0	0
Act 3	93.3	1	0	28	0	1
Act 4	89.3	3	0	0	25	0
Act 5	92.3	0	0	2	0	24

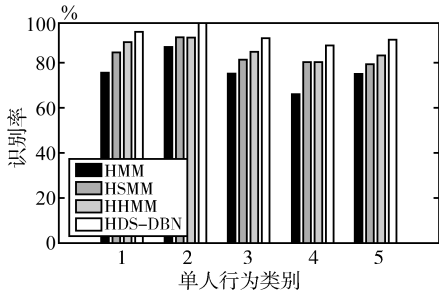
在采用 3 层的 HDS-DBN 模型时, $|Q^1|$ 取值为 3 或 4, $|Q^2|$ 取值为 3 或 4, $|Q^3|$ 也取值为 3 或 4. 表 2 列出了识别结果. 从表中可以看出, 3 层的 HDS-DBN 模型得到的识别率要稍高于 2 层模型. 其主要原因是增加了小尺度特征后, 数据集里有几类行为得到了更好的区分, 尤其是 Act 3 和 Act 5.

表 2　采用 3 层 HDS-DBN 模型时的单人行为识别结果

Table 2　The individual activity recognition results using 3-level HDS-DBN

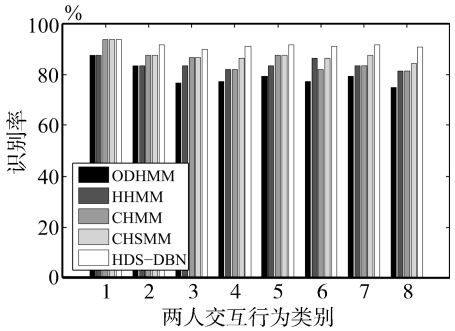
行为类别	识别率 (%)	识别结果				
		Act 1	Act 2	Act 3	Act 4	Act 5
Act 1	95.5	21	0	0	1	0
Act 2	100	0	28	0	0	0
Act 3	96.7	1	0	29	0	1
Act 4	92.9	2	0	0	26	0
Act 5	96.2	0	0	1	0	25

为了进一步评价 HDS-DBN 模型对于单人行为的识别效果, 本文同时对 HMM、HSMM 和 2 层 HHMM 在同样的数据集上进行了测试, 其结果如图 5 (a) 所示. 其中 HMM 和 HSMM 包含 4 到 6 个状态, HHMM 中每层包含 3 或 4 个状态, 状态数目的选择方法和 HDS-DBN 一样. 从图 5 (a) 中可以看出, 2 层的 HDS-DBN 识别效果优于其他几种模型.



(a) 单人行为识别

(a) Individual activity recognition



(b) 交互行为识别

(b) Interacting activity recognition

图 5　行为识别结果比较

Fig. 5　The comparison results of activity recognition

4.2　双人交互行为识别结果

双人交互行为主要包含如下八个类别: 1) Inter 1: 两人相向行走; 2) Inter 2: 两人相向跑动; 3) Inter 3: 两人步行靠近, 接近时两人交谈片刻, 然后沿着各自初始的方向离去; 4) Inter 4: 两人步行靠近, 接近时两人交谈片刻, 然后沿着各自的相反方向返回; 5) Inter 5: 两人步行靠近, 接近时两人交谈片刻, 然后其中一人沿着初始的方向离去, 另一人沿着另外的方向离去; 6) Inter 6: 两人步行靠近, 接近时两人握手片刻, 然后沿着各自初始的方向离去; 7) Inter 7: 两人步行靠近, 接近时两人握手片刻, 然后沿着各自的相反方向返回; 8) Inter 8: 一人在行进过程中置一物体于地上, 而后继续前行, 然后另一人迎面过来捡起物体并离去. 该数据集中的行为序列长度在 160 到 240 帧之间. 对于两人交互行为识别, 2 层的 HDS-DBN 模型中的状态个数 $|Q^1|$ 取值为 4、5 或 6, $|Q^2|$ 也取值为 4、5 或 6. 表 3 (见下页) 列出了两人交互行为的识别结果, 为了节省篇幅, 该表只列出了行为的识别率. 从表中可以看出, 2 层的 HDS-DBN 模型对两人交互行为的识别效果较好.

在采用 3 层的 HDS-DBN 模型时, $|Q^1|$ 取值为 4、5 或 6, $|Q^2|$ 取值为 4、5 或 6, $|Q^3|$ 取值为 5、6 或 7. 表 4 (见下页) 列出了采用 3 层 HDS-DBN 模型进行交互行为识别的结果. 从表 4 中可以看出, 对于

本文采用的两人交互行为数据集, 采用 3 层的 HDS-DBN 模型时, 识别效果和 2 层的模型相比没有较明显的提高. 这是因为 3 层的 HDS-DBN 模型增加了小尺度的特征, 如四肢和头的位置等, 而对本文的两人交互行为数据集的分类 (即识别) 来说, 这类特征起的作用较为有限. 另外, 3 层的 HDS-DBN 模型增加了模型的复杂度, 又采用了更精细的特征, 这使得在训练数据较为有限时, 模型容易出现过学习, 且识别结果易受噪声等干扰, 这也导致了在该测试集上模型的识别率提升的不明显. 相反, 如果在某些情况下小尺度的特征对于分类问题是必需的, 则可以预期 3 层 HDS-DBN 模型的识别效果会优于 2 层的模型. 所以, 模型并非是越复杂越好, 特征也并非是越细致越好, 而是应该与待解决的问题相适配.

表 3 采用 2 层 HDS-DBN 模型时的两人交互行为识别结果
Table 3 The interacting activity recognition results using 2-level HDS-DBN

	Inter 1	Inter 2	Inter 3	Inter 4	Inter 5	Inter 6	Inter 7	Inter 8
行为数目	16	24	30	22	24	22	24	32
识别率 (%)	93.8	91.7	90.0	90.9	91.7	90.9	91.7	90.6

表 4 采用 3 层 HDS-DBN 模型时的两人交互行为识别结果
Table 4 The interacting activity recognition results using 3-level HDS-DBN

	Inter 1	Inter 2	Inter 3	Inter 4	Inter 5	Inter 6	Inter 7	Inter 8
行为数目	16	24	30	22	24	22	24	32
识别率 (%)	93.8	91.7	86.7	95.5	91.7	90.9	91.7	93.8

为了进一步测试 HDS-DBN 模型对于两人交互行为的识别效果, 本文同时对 ODHMM, CHMM, CHSMM 和 2 层 HHMM 在同样的数据集上进行了测试, 其结果如图 5 (b). 其中 CHMM 和 CHSMM 中每条状态链包含 4 到 6 个状态, ODHMM 中包含 5 到 7 个状态, HHMM 中每层包含 4 到 6 个状态. 从图中可以看出, 2 层 HDS-DBN 模型的识别效果优于其他几种模型, 尤其是对于含有复杂运动细节的行为, 譬如 Inter 8 等.

4.3 模型鲁棒性分析

一般情况下, 观测数据会出现噪声或者数据缺失. 本节定量地分析 2 层的 HDS-DBN 模型在这些情况下的识别结果.

4.3.1 数据中存在噪声的情况

为了分析噪声对行为识别结果的影响, 我们在原始的观测数据中加入不同程度的高斯噪声. 在大尺度上, 轨迹受噪声污染情况表示如下

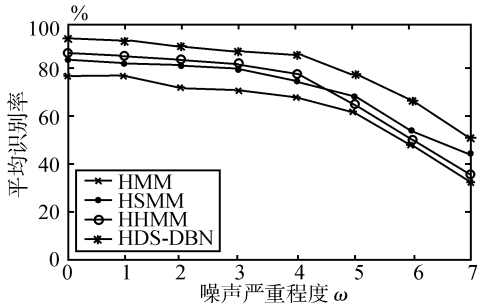
$$\begin{cases} \tilde{x}_t^c = x_t^c + \omega\xi_t, \quad \xi_t \sim N(0, \sigma_x^2) \\ \tilde{y}_t^c = y_t^c + \omega\eta_t, \quad \eta_t \sim N(0, \sigma_y^2) \end{cases} \tag{11}$$

其中, x_t^c 和 y_t^c 分别是 t 时刻人的中心的横、纵坐标, $\tilde{x}_t^c$ 和 $\tilde{y}_t^c$ 是对应的加噪声后的值, ξ_t 和 η_t 分别是标准差为 σ_x 和 σ_y 的零均值高斯噪声, 表示噪声对轨迹的污染程度. 轨迹中的噪声通常是跟踪不准确引起的, 这种不准确的程度通常较小, 在实验中令 $\sigma_x = 1$ (像素), $\sigma_y = 1$ (像素).

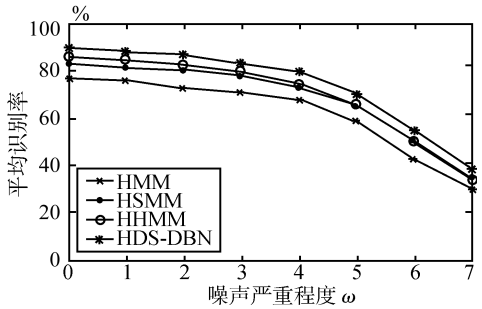
在中尺度上, 每一帧内人的轮廓受噪声污染情况表示如下

$$\begin{cases} \tilde{x}_i^p = x_i^p + \rho\upsilon_i, \quad \upsilon_i \sim N(0, \sigma_x^2) \\ \tilde{y}_i^p = y_i^p + \rho\zeta_i, \quad \zeta_i \sim N(0, \sigma_y^2) \end{cases} \tag{12}$$

其中, x_i^p 和 y_i^p 分别表示人轮廓上的第 i 个标志点的横、纵坐标, $\tilde{x}_i^p$ 和 $\tilde{y}_i^p$ 表示加噪声后的对应值, υ_i 和 ζ_i 分别是标准差为 σ_x 和 σ_y 的零均值高斯噪声, 表示噪声对人轮廓的污染程度. 我们同样采用和式 (10) 中相同的方差. 通常情况下人的轮廓受噪声干扰不是很大 (不考虑严重遮挡), 在此只考虑 $\rho = 0, 1$ 的情况. 图 6 是模型的单人行为的平均识别率随着噪声严重程度 ω 和 ρ 的变化曲线图. 从图中可以看出, 随着 ω 和 ρ 的增大, 识别率都有所下降, 在 $\omega \geq 5$ 时下降较为迅速. 另外 HDS-DBN 模型在不同噪声下的识别结果优于其他模型.



(a) $\rho = 0$



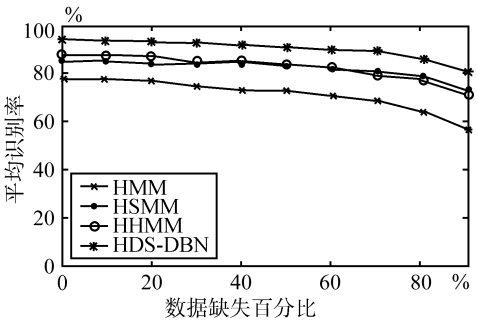
(b) $\rho = 1$

图 6 在不同程度噪声下的单人行为平均识别率变化
Fig. 6 Changes of average recognition rates with different levels of noise

4.3.2 数据部分缺失的情况

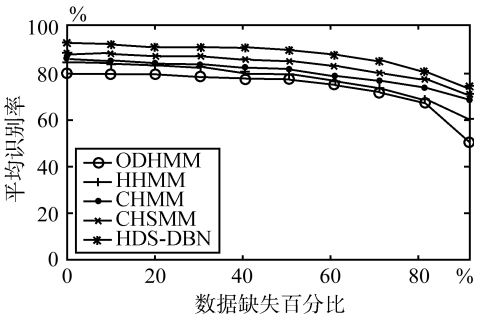
为了测试模型在数据发生部分缺失情况下的识

别结果, 我们去掉原始测试数据的部分帧, 用剩余的行为数据进行测试. 去掉帧的办法如下: 首先在 $1, 2, \cdots, 9$ 中随机产生 C 个数, 然后在原始的测试数据的每 10 帧中删除序号对应于 C 个数的那些帧. 由于 C 的取值范围从 1 到 9, 则删除的图像帧占整个行为序列长度的百分比在 10 % 到 90 % 之间. 图 7 是模型的平均识别率随删除帧的数量的变化曲线图. 从图 7 可以看出, 删除的帧数量在不大于 70 % 的情况下识别结果比较稳定, 并且和其他模型相比, HDS-DBN 的识别性能具有一定的优势. 另外和噪声的干扰相比, 模型在缺失信息时体现出了更好的鲁棒性. 这是由于连续的帧中含有大量的冗余信息, 缺失一部分帧对于行为的分类影响不大.



(a) 单人行为识别

(a) Individual activity recognition



(b) 交互行为识别

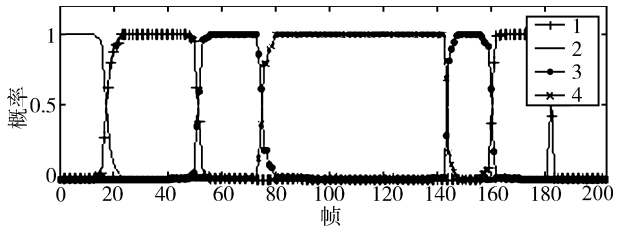
(b) Interacting activity recognition

图 7　在不同程度信息缺失时的平均识别率变化

Fig. 7　Changes of average recognition rates with different percentage information deleted

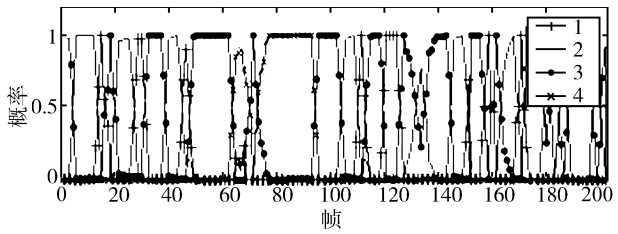
4.4　模型状态的后验概率及意义

在给定测试集上的识别率是反映模型性能的一个指标. 此外, 模型的结构还应该具有合理、清晰的物理意义, 这样模型在其他数据集上才具有好的扩展性. 本节分析了 2 层的 HDS-DBN 模型的内在结构. 给定一个行为序列, 其状态的后验概率表现出了行为的内在的特性. 图 8 反映了 Act 3 中的某一个测试序列对应的状态后验概率. 从图 8 可以看出, 状态 Q^1 的转移频率要大大小于状态 Q^2 的转移频率, 这反映了前者确实能够表现出大尺度上的运动细节, 而后者对于小尺度上的运动细节也能够准确地表示. 图 9 描述了 Q^2 的四个取值对应的人的轮廓, 可以看出该轮廓反映了 Act 3 中的几个重要姿态. 尤其是状态 $Q^2 = 4$ 对应的轮廓体现了人弯腰时的姿态, 该姿态在 Act 3 中是重要而又短暂的. 这说明 HDS-DBN 模型具有捕捉短暂而重要运动细节的能力.



(a) 状态 Q^1 的后验概率

(a) The posterior probability of state Q^1



(b) 状态 Q^2 的后验概率

(b) The posterior probability of state Q^2

图 8　Act 3 中某一个测试序列对应的状态后验概率

Fig. 8　The posterior probability of states of a test sequence in Act 3

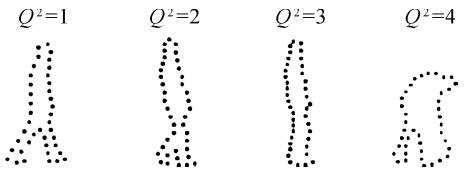


图 9　状态 Q^2 对应的几个姿态

Fig. 9　The poses corresponding to state Q^2

5　结论

本文提出了一种分层且带驻留时间状态的动态贝叶斯网络 HDS-DBN, 并且将其用于人的行为识别. 由于人的行为包含多个尺度的运动细节, 所以我们将人的运动分解为不同尺度上的多个随机过程, 用 HDS-DBN 模型来对其进行表示. HDS-DBN 模型能够较好地表示人的行为中包含的多个尺度的运动细节, 而且对运动的分解使得该模型的参数空间较小. 实验结果表明 HDS-DBN 模型具有很好的行为识别结果, 在具有噪声或者缺失信息的情况下其

识别性能仍然较好. 另外实验结果显示 HDS-DBN 模型能够捕捉重要而细微的运动细节.

References

1 Wang Liang, Hu Wei-Ming, Tan Tie-Niu. A survey of visual analysis of human motion. *Chinese Journal of Computers*, 2002, **25**(3): 225−237
(王亮, 胡卫明, 谭铁牛. 人运动的视觉分析综述. 计算机学报, 2002, **25**(3): 225−237)

2 Du You-Tian, Chen Feng, Xu Wen-Li, Li Yong-Bin. A survey on the vision-based human motion recognition. *Acta Electronica Sinica*, 2007, **35**(1): 84−90
(杜友田, 陈峰, 徐文立, 李永彬. 基于视觉的人的行为识别综述. 电子学报, 2007, **35**(1): 84−90)

3 Aggarwal J K, Park S. Human motion: modeling and recognition of actions and interactions. In: Proceedings of the 2nd International Symposium on 3D Data Processing, Visulization, and Transmission. Thessaloniki, Greece: IEEE, 2004. 640−647

4 Pers J, Vuckovic G, Dezman B, Kovacic S. Scale-based human motion representation for action recognition. In: Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis. Rome, Italy: IEEE, 2003. 668−673

5 Fanti C, Zelnik-Manor L, Perona P. Hybrid models for human motion recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 1166−1173

6 Bobick A F, Davis J W. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(3): 257−267

7 Brand M, Oliver N, Pentland A. Coupled hidden Markov models for complex action recognition. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Juan, Puerto Rico: IEEE, 1997. 994−999

8 Ghahramani Z, Jordan M I. Factorial hidden Markov models. *Machine Learning*, 1997, **29**(2-3): 245−273

9 Liu X H, Chua C S. Multi-agent activity recognition using observation decomposed hidden Markov models. *Image and Vision Computing*, 2006, **24**(2): 166−175

10 Gong S Q, Xiang T. Recognition of group activities using dynamic probabilistic networks. In: Proceedings of the 11th International Conference on Computer Vision. Beijing, China: IEEE, 2003. 742−749

11 Natarajan P, Nevatia R. Coupled hidden semi-Markov models for activity recognition. In: Proceedings of IEEE Workshop on Motion and Video Computing. Texas, USA: IEEE, 2007. 10

12 Fine S, Singer Y, Tishby N. The hierarchical hidden Markov model: analysis and applications. *Machine Learning*, 1998, **32**(1): 41−62

13 Bui H H, Venkatesh S, West G. Policy recognition in the abstract hidden Markov model. *Journal of Artificial Intelligence Research*, 2002, **17**: 451−499

14 Oliver N, Horvitz E, Garg A. Layered representations for human activity recognition. In: Proceedings of the 4th International Conference on Multimodal Interfaces. Pittsburgh, USA: IEEE, 2002. 3−8

15 Ivanov Y A, Bobick A F. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(8): 852−872

16 Du Y T, Chen F, Xu W L, Li Y B. Recognizing interaction activities using dynamic bayesian network. In: Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China: IEEE, 2006. 618−621

17 Bengio Y. Markovian models for sequential data. *Neural Computing Surveys*, 1999, **2**: 129−162

18 Gu H Y, Tseng C Y, Lee L S. Isolated-utterance speech recognition using hidden Markov models with bounded state durations. *IEEE Transactions on Signal Processing*, 1991, **39**(8): 1743−1752

19 Levinson S E. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1986, **1**(1): 29−45

20 Russell M J, Moore R K. Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Florida, USA: IEEE, 1985. 5−8

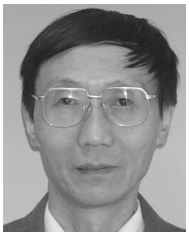
21 Duong T V, Bui H H, Phung D Q, Venkatesh S. Activity recognition and abnormality detection with the switching hidden semi-Markov model. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 838−845

22 Murphy K P. Dynamic Bayesian Network: Representation, Inference and Learning [Ph. D. dissertation], University of California, USA, 2002

23 Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000, **290**(5000): 2323−2326

杜友田 清华大学自动化系博士研究生. 2002 年获西安交通大学电气工程学院学士学位. 主要研究方向为计算机视觉和模式识别.
E-mail: dyt02@mails.tsinghua.edu.cn
(**DU You-Tian** Ph. D. candidate in the Department of Automation, Tsinghua University. He received his bachelor degree from Xi′an Jiaotong University in 2002. His research interest covers computer vision and pattern recognition.)

陈 峰 清华大学自动化系副教授. 2000 年获清华大学自动化系博士学位. 主要研究方向为计算机视觉和视频处理. 本文通信作者.
E-mail: chenfeng@mail.tsinghua.edu.cn
(**CHEN Feng** Associate professor at Tsinghua University. He received his Ph. D. degree from Tsinghua University in 2000. His research interest covers computer vision and video processing. Corresponding author of this paper.)



徐文立 清华大学自动化系教授. 1990 年获科罗拉多大学博士学位. 主要研究方向为视频处理、计算机视觉、机器人及自动控制.
E-mail: xuwl@mail.tsinghua.edu.cn
(**XU Wen-Li** Professor at Tsinghua University. He received his Ph. D. degree from University of Colorado at Boulder, Columbia, in 2000. His research interest covers video processing, computer vision, robotics, and automatic control.)