基于时空单词的两人交互行为识别方法

韩 磊 李君峰 贾云得

(北京理工大学计算机学院 智能信息技术北京市重点实验室 北京 100081)

摘 要 文中提出一种基于时空单词的两人交互行为识别方法,该方法从行为视频中提取丰富的时空兴趣点,基于人体剪影的连通性分析和时空兴趣点的历史信息,把时空兴趣点划分给不同的人体,并在兴趣点样本空间聚类生成时空码本(spatial-temporal codebook).对于给定的时空兴趣点集,通过投票得到表示单人原子行为的时空单词(spatial-temporal words).采用条件随机场模型建模单人原子行为,在两人交互行为的语义建模过程中,人工建立表示领域知识(domain knowledge)的一阶逻辑知识库,并训练马尔可夫逻辑网用以两人交互行为的推理.两人交互行为库上的实验结果证明了该方法的有效性.

关键词 交互行为分析;行为识别;时空特征;条件随机场;马尔可夫逻辑网

中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.00776

Human Interaction Recognition Using Spatio-Temporal Words

HAN Lei LI Jun-Feng JIA Yun-De

(*Beijing Laboratory of Intelligent Information Technology*, *School of Computer Science*, *Beijing Institute of Technology*, *Beijing* 100081)

Abstract This paper proposes a hierarchical approach for recognizing person-to-person interaction in indoor scenario from a single view, which is based on spatial-temporal feature extraction and representation. The dense space-time interest points detected from videos are divided into two sets exclusively according to the history information along the evolvement and the connectivity of the two human silhouettes. Then K-means clustering performs on points in the training set and learns the spatial-temporal codebook. For a given set of interest points, a spatial-temporal word is built by allowing each point to vote softly into the few centers nearest to it and accumulating the scores of all the points. The Conditional Random Field (CRF) whose inputs are the spatial-temporal words is used to modeling the primitive actions for each person, and common sense domain knowledge and first order logic production rules with weights are employed to learn the structure and the parameters of Markov Logic Network (MLN). The MLN can naturally integrate common sense reasoning with uncertain analysis, which is capable to deal with the uncertainty produced by CRF. Experiment results on the interaction dataset are provided to demonstrate the effectiveness and the robustness.

Keywords interaction analysis; action recognition; spatial-temporal feature; conditional random field; Markov logic network

---

收稿日期:2009-07-20;最终修改稿收到日期:2009-08-28.本课题得到国家自然科学基金(60905006,90920009)、国家"八六三"高技术研究发展计划项目基金(2009AA01Z323)资助.**韩 磊**,男,1982 年生,博士研究生,主要研究方向为基于视觉的人机交互、人体行为分析等.**李君峰**,男,1985 年生,硕士研究生,主要研究方向为基于视觉的人体行为识别.**贾云得**(通信作者),男,1962 年生,博士,教授,博士生导师,主要研究领域为计算机视觉、人工智能和人机交互等. E-mail: jiayunde@bit.edu.cn.

1 引　言

人体行为分析在智能视频监控、视频注解、虚拟现实、人机交互等领域中具有广阔的应用前景，已经成为计算机视觉和模式识别领域的研究热点[1]. 目前，人们对单人行为分析的研究工作很多，对两人和多人交互行为分析的研究工作较少. 两人交互行为在日常生活中非常普遍，比如握手、拥抱等，但如何有效地提取两人交互行为中的运动特征、建立多个目标之间的复杂交互模型是极具挑战性的问题.

通常，基于视觉的人体行为分析可以分为两个层次的任务：一是底层的特征提取和表示，二是高层的行为识别和建模. 从图像序列中提取出能够合理表示人体运动的特征，对行为识别和理解至关重要. 目前，基于边缘或形状的静态特征、基于光流或运动信息的动态特征以及基于时空体积数据的时空特征都得到了广泛的应用. 静态特征提取的准确性往往受到跟踪和姿态估计精度的影响，在运动物体较多或背景比较复杂的场景下，该类特征的鲁棒性面临严峻考验. 动态特征从相邻两帧图片中获取运动目标的运动信息，缺乏行为的全局分析. 时空特征方法把图像序列看作时空相关的三维体积数据，通过提取静态模式获得行为的时空表示，如 Niebles 等人[2]将视频序列表示为时空单词(spatial-temporal words)的集合，在背景移动或存在多个运动物体的情况下，他们的方法都取得了满意的识别效果.

本文通过检测行为视频中的时空兴趣点，提出一种两人交互行为的时空特征表示方法. 本文结合概率图模型和统计关系模型的各自优势，将两人交互行为识别分为两个层次的识别任务，即底层采用概率图模型建模单人的原子行为，高层采用马尔可夫网和一阶逻辑相结合的统计关系学习方法，实现两人交互行为建模. 概率图模型是当前最为流行的建模连续动态特征序列的工具，它已有了成熟的概率推理和优化方法，具有很好的理论基础，但其模型的拓扑结构依赖于行为内在的结构信息，随着行为复杂性的增加，如参与行为的人体个数的增加，需要大量的训练数据学习图模型的拓扑结构，因此基于概率图模型的方法更适合建模单人原子行为，而不是复杂的交互行为. 基于文法的方法其优势在于它能有效建模复杂行为的内部结构，但这类方法大多要求人工设定所有可能的产生式规则，需要大量繁杂的工作. 传统的基于知识或逻辑推理的方法只能进行知识的精确推理，对于输入数据的错误和不确定性无能为力，而在基于视觉的行为分析中，底层的视觉特征提取和中层的原子行为识别很可能存在误差或漏检的情况，如何在复杂行为建模的过程中，既能有效地利用先验知识，又能建模行为的不确定性是亟待解决的问题. 统计关系学习(Statistical Relational Learning，SRL)是一种将关系/逻辑表示、概率推理(即不确定性处理)、机器学习和数据挖掘综合在一起，以获取关系数据似然模型的机器学习方法，它非常适合复杂行为，尤其是交互行为的建模和识别.

2 算法框图

已有的研究工作中采用了多种推理模型建模交互行为. Oliver 等人[3]比较了马尔可夫模型(Hidden Markov Model，HMM)和舰合马尔可夫模型(Coupled HMM，CHMM)在两人交互行为分析中的性能，其结果表明在文中的监控场景下 CHMM 取得了比 HMM 更好的识别结果. Xiang 和 Gong[4]提出了一种状态间可动态联接的马尔可夫模型(Dynamically Multi-Linked HMM，DML-HMM)用于分析机场监控场景下地面运输和飞机之间的交互行为以及室内场景下两人之间的交互行为. Ivanov 和 Bobick[5]采用随机文法技术对多智能体的复杂行为事件和交互行为进行了检测和识别，其思想是将识别问题分成两层，底层采用 HMM 识别原子行为，其输出为高层上下文无关的句法分析机制服务. Park 和 Aggarwal[6]将交互行为识别分为 3 个层次：底层使用贝叶斯网络(Bayesian Network，BN)识别单个人体部分的姿态，并整合各个人体部分的姿态得到单个人体的姿态；中层采用动态贝叶斯网络(Dynamical Bayesian Network，DBN)建模单人行为，并将单人行为表示为动词居中的三元组结构(即“agent-motion-target”)；高层基于单人原子行为的时空约束创建描述交互行为的决策树用以交互行为的识别. Ryoo 和 Aggarwal[7]将交互行为分析分为 4 个层次，分别为人体部分提取层、姿态层、姿势层、单人动作和交互行为层，与 Park 的工作不同，他们采用 HMM 建模人体的姿势，并采用上下文无关文法建模人体的交互行为. Du 等人[8]将行为分解为多个交互的随机过程，每个随机过程对应一个尺度上的人体运动，提出层级周期状态 DBN(Hierarchical Durational-State Dynamic Bayesian Network，HDS-DBN)建模两人交互行为，该 DBN 不同层次的观测特征对应特征提取阶段提取的不同尺度的运

动细节. Hongeng 等人[9]提出一种面向室外监控场景中多人交互行为分析的层次化事件描述机制，定义了单线程事件(由单个人执行的行为)和多线程事件，采用包含时序和概率关系的时序逻辑网络将多个单线程事件组合起来，以表示复杂的多人交互事件. Hakeem 和 Shah[10]提出一种用于表示多人交互行为的事件层次架构 CASE，他们同样将高层的多人交互事件定义为底层事件的时序组合.

上述大部分工作都把交互行为识别分解为多个层次，本文借鉴了这种层级建模的思想，把两人交互行为分析划分为 3 个层次：

(1)时空特征提取和表示. 已有的工作主要采用静态特征或运动特征作为交互行为识别模型的观测特征，这些特征往往需要精确的运动跟踪和姿态估计的结果，因此其行为建模能力在很大程度上取决于特征提取的准确性. 我们通过提取交互行为视频中的时空特征表示人体运动，并结合静态特征(人体剪影)创建人体行为的时空特征表示(即时空单词，Spatial-Temporal Words). 本文的时空特征表示方法不需要运动跟踪和姿态估计的结果，可以有效地应用于具有复杂背景的视频序列.

(2)基于判别式模型的单人原子行为识别. 绝大多数交互行为分析方法都采用诸如 HMM、DBN 等产生式模型建模单人原子行为. 此类产生式模型存在严格的独立性假设，即假设每个观测都是彼此独立的，这并不符合实际情况，无法表现观测序列在长时间范围内的依赖关系. 与之不同的是，判别式模型直接对条件概率进行建模，本文采用条件随机场(Conditional Random Field, CRF)模型建模单人原子行为，该模型是一个直接计算给定输入节点情况下输出节点条件概率的无向图模型，具有很强的分类能力.

(3)基于统计关系学习模型的两人交互行为建模. 本文采用马尔可夫逻辑网(Markov Logic Network, MLN)建模两人交互行为，该模型将基于规则的表示和概率图模型相结合，不仅可以容易地引入人的领域知识，同时也具有处理不确定性的能力，可以很好地解决视觉特征提取阶段和单人原子行为识别引入的不确定性.

以上的 3 个层次包含了时空特征提取和行为识别两部分，图 1 和图 2 分别给出两部分的算法框图.

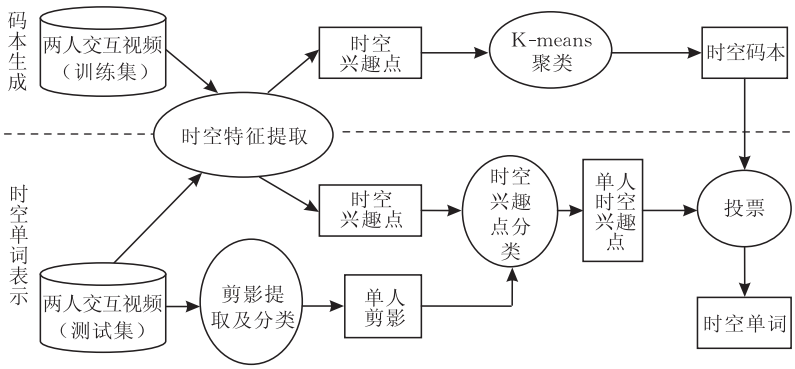


图 1 时空特征提取和表示框图

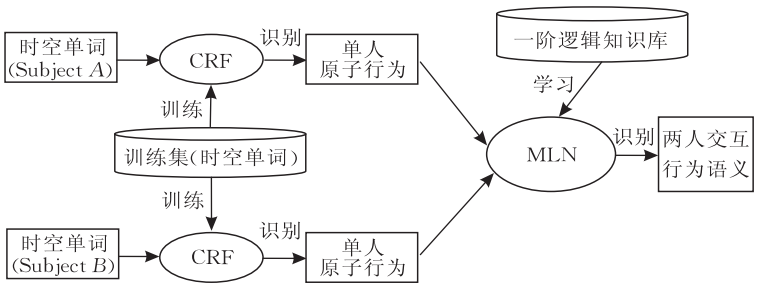


图 2 两人交互行为建模和识别框图

3 时空特征提取和表示

3.1 时空兴趣点

Dollar 等人[11]提出的时空兴趣点检测方法可以从视频序列中提取丰富的时空兴趣点，本文采用其中的线性滤波器检测图像序列中的时空兴趣点，该滤波器的响应函数为

$$R=(I\times g\times h_{ev})^2+(I\times g\times h_{od})^2 \quad (1)$$

其中 $g(x,y;\sigma)$ 是仅用于二维图像平滑的高斯核，

h_{ev}和h_{od}是一对正交的一维 Gabor 滤波器，仅用于时间维，定义为 $h_{ev}=(t;\tau,\omega)=-\cos(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2}$，$h_{od}(t;\tau,\omega)=-\sin(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2}$. 由于所有实验中均设定 $\omega=4/\tau$，式(1)中的参数减少到两个，即 σ 和 τ，他们分别控制检测器在空间和时间上的尺度. 出于计算效率的考虑，本文并没有在多个时空尺度上检测时空兴趣点，而仅在一个空间和时间尺度上进行检测(实验中设定 $\sigma=1$ 和 $\tau=2.5$). 图 3 分别显示了“握手”和“拳击”行为中部分图像中时空兴趣点的检测结果.

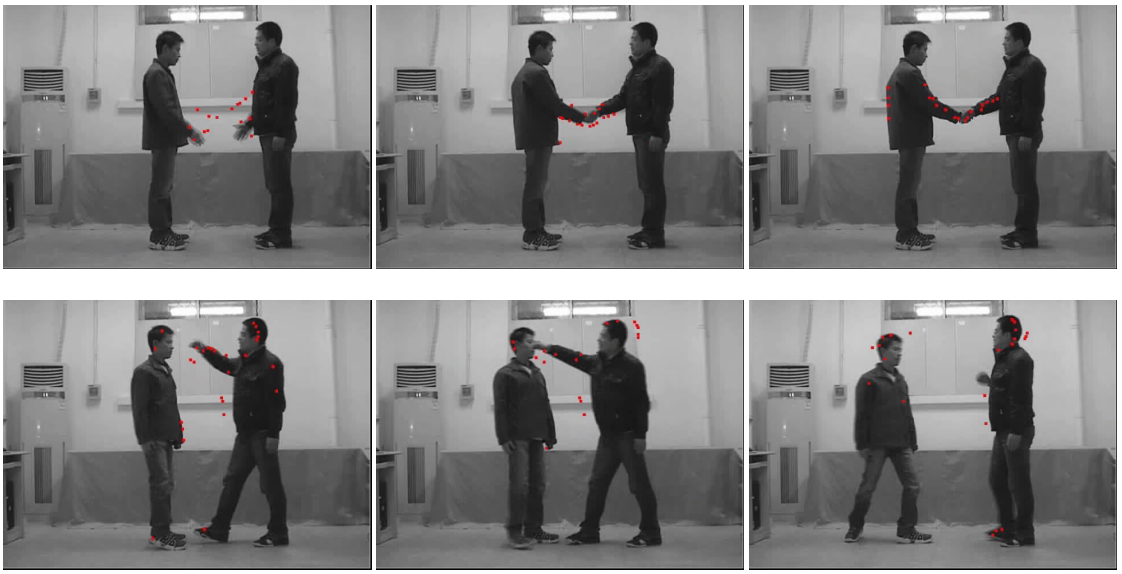


图 3　时空兴趣点检测结果

如图 3 所示，时空兴趣点可以正确地定位到视频序列中具有明显运动的区域. 值得注意的是，两人交互行为视频中的时空兴趣点是由两个不同的人产生，建模单人原子行为还需要按照不同的行为执行者对时空兴趣点进行分类. 剪影(silhouette)是基于视觉的人体行为分析中普遍使用的静态特征，从图像序列中鲁棒地提取人体剪影的技术已经比较成熟，本文我们基于两人剪影的连通性判断机制以及时空兴趣点的历史信息，提出一种可以动态划分时空兴趣点的方法，如算法 1 所示.

算法 1.　时空兴趣点分类算法.

定义：$p_{i,j}^k$为第 k 帧(i,j)位置的时空兴趣点，P^k为第 k 帧上的时空兴趣点集，S_k为第 k 帧的剪影图像；

$rad(p_{i,j}^k,r)=\{(x,y)\,|\,|x-i|=r \text{ or } |y-j|=r\}$，其中 r 为辐射半径；l 为时空兴趣点的分类标记且 $l\in\{l_1,l_2\}$，$l_{i,j}^k$表示对时空兴趣点 $p_{i,j}^k$的分类标记，则在以下两种情况下的分类算法如下：

1. 当 S_k不连通时

令 C_1，C_2为互不连通的两个剪影区域，对任意 $p_{i,j}^k$，按照下式计算

$$r_{\min}=\arg\min_r\{|C_1\cap rad(p_{i,j}^k,r)|\neq|C_2\cap rad(p_{i,j}^k,r)|\},$$

则按照如下准则对时空兴趣点分类：

当$|C_1\cap rad(p_{i,j}^r,r_{\min})|>|C_2\cap rad(p_{i,j}^k,r_{\min})|$时，$l_{i,j}^k=l_1$；

当$|C_1\cap rad(p_{i,j}^r,r_{\min})|<|C_2\cap rad(p_{i,j}^k,r_{\min})|$时，$l_{i,j}^k=l_2$.

2. 当 S_k连通时

令 $M=P^{k_{\max}}$，其中 $k_{\max}=\arg\max_{k=b,b+1,\cdots,e}\{|P^k|\}$，$b$，$e$ 为同时满足以下两个条件的值：

1) $S_b,S_{b+1},\cdots,S_e$均不连通并且 S_{b-1}连通；

2) $S_{e+1},S_{e+2},\cdots,S_{k-1}$均连通.

对任意 $p_{i,j}^k$，计算 $n_{\min}=\arg\min_n\{\|p_{i,j}^k-p_n\|\}$，其中$p_n\in M,n=1,2,\cdots,|M|$，假设 M 中第 $n_{\min}$个时空兴趣点对应的帧号和位置分别为 k^* 和(i^*,j^*)，则有分类准则 $l_{i,j}^k=l_{i^*,j^*}^{k^*}$.

注意到当第 1 帧连通时，不存在满足上述条件的 b，e，此时初始化 M 为 M_0，即

$$M_0=\{p_{i_1,j_{\min}}^1,p_{i_2,j_{\max}}^1\},$$

其中 $i_1\neq i_2$，$j_{\min}=\min\{j\,|\,p_{i,j}^1\in P^1\}$，$j_{\max}=\max\{j\,|\,p_{i,j}^1\in P^1\}$.

图 4 中给出了图 3 中图像的时空兴趣点的分类结果，结果表明本文的时空兴趣点分类算法可以有效地对时空兴趣点进行分类，可以为单人原子行为识别模型提供可靠的观测特征.

3.2　时空码本

行为建模过程中，特征向量的维数越高，需要的训练数据也就越多. 本文提出一种简洁有效的特征表示方式，在尽量保持原有信息关键成分不丢失的情况下，将原始特征数据从高维空间投影到低维

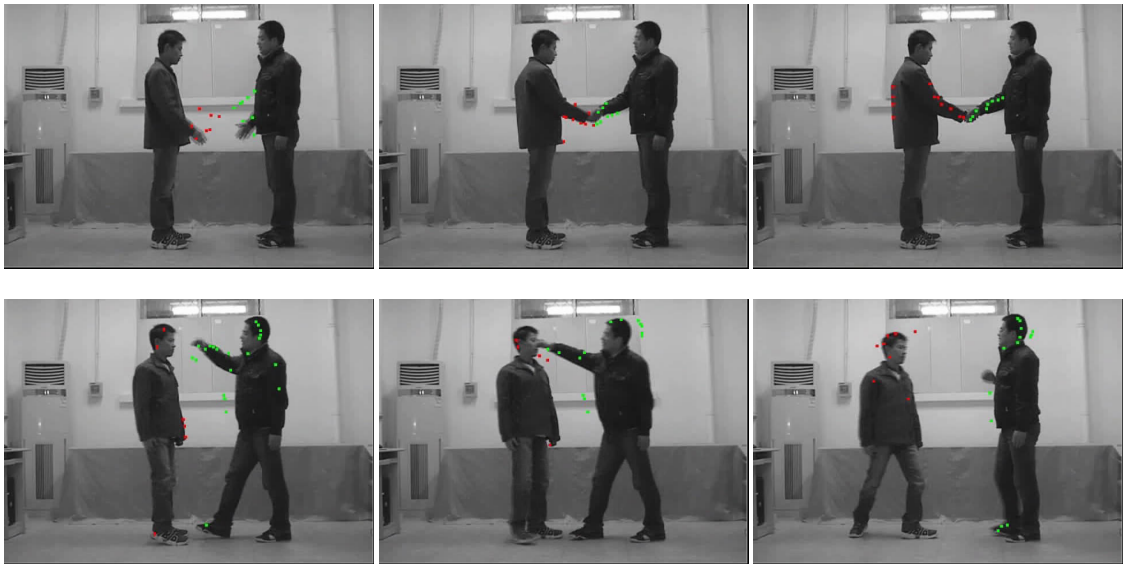
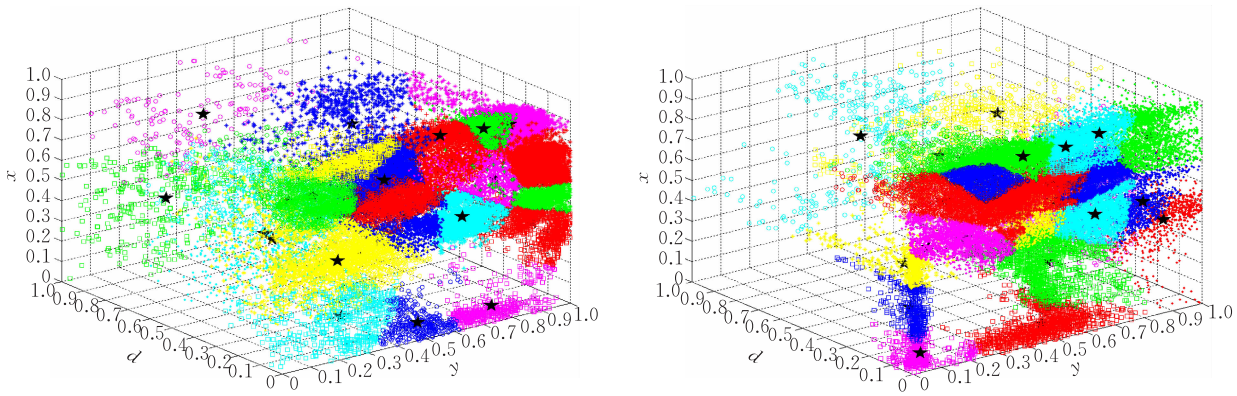


图 4 时空兴趣点分类结果

空间.

每个时空兴趣点都可以看作是三维空间((x,y,d),其中 d 是时空兴趣点的量值,x 和 y 是时空兴趣点在图像空间中的位置)中的一个点,因此每个人体行为可以看作是该三维空间中的一个时空兴趣点的集合.本文采用直方图量化技术将时空兴趣点集合量化为维数固定的直方图(即时空单词),时空码本采用 K-means 聚类算法生成(实验中选择时空码本的维数为 25,该维数在实验章节中进行了评估和说明).在聚类生成码本之前,每个时空兴趣点都进行了归一化,以保证其缩放和平移不变性.基于相似度的聚类算法中,如何评估样本之间的距离是个关键问题,由于我们对时空兴趣点进行了平移和缩放的不变性处理,我们的方法可以直接使用传统的欧式距离度量时空兴趣点的距离.图 5 给出了交叉验证实验的第一个训练集中两人时空兴趣点的分布以及聚类中心的选择,图中黑色五角星为聚类中心的标识,三维空间中部分聚类中心的标识被时空兴趣点遮挡未能在图中直观显示.



(a) 右侧人体 (b) 左侧人体

图 5 时空码本示例

对于一个给定的时空兴趣点集合,采用软投票(soft vote)机制生成相应的时空单词.投票时控制每个时空兴趣点向距离它最近的少数几个聚类中心投票.我们在每个聚类中心上放置一个高斯分布,以此计算每个时空兴趣点属于每个聚类中心的后验概率,该高斯分布的方差 σ 基于经验值设定(实验中 σ=0.2),以保证每个时空兴趣点都会明显地向 4～6 个聚类中心投票.

4 两人交互行为识别

4.1 单人原子行为识别

本文将两人交互行为建模分成分为单人原子行为建模和两人交互语义建模两个部分,前人的诸多

工作中，概率图模型都展现出了很强的原子行为建模能力. 自 Sminchisescu 等人[12]将 CRF 模型引入到人体行为分析领域后，CRF 及其改进模型在时序数据建模方面的卓越表现逐渐受到广大研究者的重视，本文采用线性链 CRF 模型建模两人交互行为中单人原子行为.

CRF 模型可以用具有单一状态链的图模型表示，假设 $S=\{s_t\}$ 是观测序列 $O=\{o_t\}$，$t=1,\cdots,T$ 对应的运动模式标签序列，$C=\{\{S_c,O_c\}\}$ 是图模型 G 中的团，则 CRF 建模给定观测序列下的状态序列的条件概率为

$$P_\theta(S|O)=\frac{1}{Z(O)}\prod_{c\in C}\Phi(S_c,O_c),$$
$$Z(O)=\sum_S\prod_{c\in C}\Phi(S_c,O_c) \tag{2}$$

其中 $Z(O)$ 是归一化因子，Φ 是团 C 上的势函数，定义为

$$\Phi(S_c,O_c)=\exp\Big(\sum_{t=1}^{T}\sum_n\lambda_n f_n(S_c,O_c,t)\Big) \tag{3}$$

$\{f_n\}$ 是特征函数集，可以是两种类型的特征函数：对应观测特征和标签转移的特征函数 $f_n(s_{t-1},s_t,O,t)$ 和对应观测特征和单个标签的特征函数 $g_n(s_t,O,t)$. 模型参数 $\theta=\{\lambda_n\}$ 是各个特征函数的实值权重，给定训练数据 $\{O^i,S^i\}_{i=1}^N$，模型参数可以通过优化以下的对数相似度函数得到

$$\Omega(\theta)=\sum_i\log p_\theta(S^i|O^i) \tag{4}$$

4.2　两人交互行为的语义建模

上述 CRF 模型为后续的两人交互语义表示提供了具有语义含义的输入数据，在视觉行为分析中，底层的视觉特征提取和中层的原子行为识别都可能存在误差和错误，MLN[13-14]将 Markov 网和一阶逻辑相结合，即保留了灵活的建模能力，又具有处理不确定性的能力.

MLN 中的每个逻辑公式 F_i 都有一个非负的实值权重 ω_i，其闭谓词(ground atom)集合 X 对应 Markov 网中的节点，假设闭谓词的子集 $x_{\{i\}}\in X$ 通过公式 F_i 相互联系，则 MLN 将第 i 个团上的特征定义为

$$f_i(x_{\{i\}})=\begin{cases}1, & F_i(x_{\{i\}})\text{ 为真}\\ 0, & \text{其它}\end{cases} \tag{5}$$

一阶逻辑知识库中的公式是创建 Markov 网的模版，Markov 网建模闭谓词的联合概率分布为

$$P(X=x)=\frac{1}{Z}\exp\Big(\sum_i\omega_i f_i(x_{\{i\}})\Big) \tag{6}$$

其中 Z 是归一化因子，$Z=\sum_{x\in X}\exp\big(\sum_i\omega_i f_i(x_{\{i\}})\big)$. 假设 $\phi_i(x_{\{i\}})$ 是定义在第 i 个团上的势函数，则 $\log(\phi_i(x_{\{i\}}))=\omega_i f_i(x_{\{i\}})$.

MLN 的网络结构确定后，可以采用概率推理学习模型参数. 由于模型的网络结构可能非常复杂(如可能有无向环)，精确的参数推理往往不能实现，通常采用 MCMC(Markov Chain Monte Carlo)方法，如 Gibbs 采样技术，进行近似的推理. MLN 中，给定闭谓词 X_i 的马尔可夫毯(Markov blanket) B_i，则该闭谓词为 x_i 的概率为

$$P(X_i=x_i|B_i=b_i)=$$
$$\Big(\exp\Big(\sum_{f_j\in F_i}\omega_j f_j(X_i=x_i,B_i=b_i)\Big)\Big)\Big/$$
$$\Big(\exp\Big(\sum_{f_j\in F_i}\omega_j f_j(X_i=0,B_i=b_i)\Big)+$$
$$\exp\Big(\sum_{f_j\in F_i}\omega_j f_j(X_i=1,B_i=b_i)\Big)\Big) \tag{7}$$

其中 F_i 是包含 X_i 的所有团的集合，f_j 采用式(5)计算.

一个完备的知识库(Knowledge Base，KB)对提高复杂交互行为的识别性能至关重要，本文将单人原子行为和两人交互行为语义均表示为一阶逻辑谓词(first order logic predicates). 知识库中引入了对两人交互行为分析的常识理解，并将它们定义为硬约束(hard constraints)，此类约束也包括可以从知识库中推理得到的谓词. 比如认为交互行为必须为两个不同人的原子行为的交互，则当两人握手的交互行为发生时，有如下硬约束：ShakeHands($p1$, $p2$)→!equal($p1$, $p2$). 单人原子行为和两人交互行为的逻辑关系通过软约束(soft constraints)，即具有权重的产生式规则建模. 比如本文采用以下产生式规则建模两个不同原子行为下的“握手”行为：action($p1$,*act_label*)^action($p2$,*act_label*)^!equal($p1$, $p2$)→shakeHands($p1$, $p2$). 软约束的初始权重通过 CRF 识别单人行为的性能设定，最终的权重由 MLN 从训练集中学习得到，MLN 的训练方法采用 Alchemy① 中提供的产生式学习方法实现.

5　实　验

5.1　实验设计

目前尚没有开放的两人交互行为数据库，为了

① Alchemy — Open Source AI. http://alchemy.cs.washington.edu/

评估本文的两人交互行为分析方法，我们在室内场景下采集两人交互行为视频建立两人交互行为数据库，所有视频序列都采用普通的数字视频设备在单一视角下获取. 目前，该行为库中共包含 5 种常见的两人交互行为，分别为“握手”、“击掌”、“拥抱”、“拳击”和“踢打”，每种交互行为均有 20 个行为样本.

整个行为数据库被平均分成 4 部分，实验采用交叉验证的方法评估本文的识别方法. 图 6 给出码本维数增加时，单人原子行为识别性能的变化趋势，当码本的维数为 25 时，原子行为的识别率最高，因此选择 K-means 聚类中心的个数为 25.

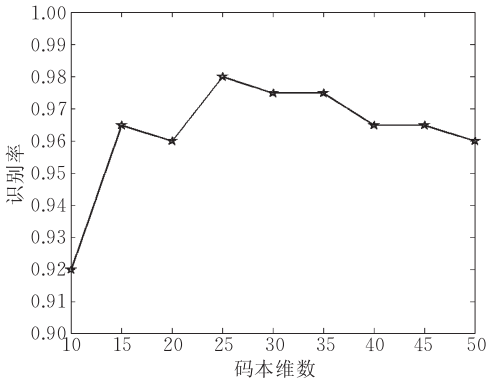


图 6 不同码本维数条件下单人原子行为的识别性能

5.2 实验结果及分析

本文的两人交互行为识别包括单人原子行为识别和两人交互语义分析两个层次，原子行为的识别结果在很大程度上影响最终两人交互行为的识别性能. 图 7 中给出了单人原子行为的识别率，实验中共定义了 9 个具有语义含义的单人原子行为，由于 CRF 模型在小样本集上具有很强的分类能力，单人原子行为的平均识别率达到了 98%. 两人交互行为识别的混淆矩阵如图 8 所示，得益于基于知识的方法在建模复杂交互行为上的灵活性，MLN 在底层原子行为识别错误的情况下也表现出了很强的纠错能力. 比如右侧人的“握手”(标签为 R1)行为错误地识别为“击掌”(标签 R2)行为时，由于 CRF 正确地识别了左侧人的“握手”(标签 L1)行为，最终 MLN 正确地识别两人“握手”的交互行为. 值得注意的是，当右侧人的“踢腿”(标签 R5)行为错误地识别为“握手”(标签 R1)行为时，尽管 CRF 正确地识别了左侧人的“避让”(标签 L4)行为，MLN 仍错误地识别为两人“握手”的行为，其原因在于当左侧人为“避让”行为时，MLN 建模此时右侧人可能的行为有两种，即“出拳”和“踢腿”，但右侧人为“握手”行为时，MLN 认为左侧人可能的行为只有一种，即“握手”，显然前者的不确定性要大于后者，因此 MLN 识别错误.

	L1	L2	L3	L4	R1	R2	R3	R4	R5
L1	1	0	0	0	0	0	0	0	0
L2	0	1	0	0	0	0	0	0	0
L3	0	0	1	0	0	0	0	0	0
L4	0	0	0.05	0.95	0	0	0	0	0
R1	0	0	0	0	0.95	0.05	0	0	0
R2	0	0	0	0	0	1	0	0	0
R3	0	0	0	0	0	0	1	0	0
R4	0	0	0	0	0	0	0	1	0
R5	0	0	0	0	0.05	0	0	0	0.95

图 7 单人原子行为识别的混淆矩阵(图中横纵标记为实验中定义的具有语义含义的单人原子行为)

	握手	击掌	拥抱	拳击	踢打
握手	1	0	0	0	0
击掌	0	1	0	0	0
拥抱	0	0	1	0	0
拳击	0	0	0	1	0
踢打	0.05	0	0	0	0.95

图 8 两人交互行为识别的混淆矩阵

6 结论及未来工作

本文针对室内场景中两人交互行为分析的任务提出一种基于时空特征的层次化交互行为建模方法. 该方法主要分为 3 个层次：时空特征提取和表示、单人原子行为识别以及交互语义表示和建模. 它充分利用了 CRF 模型在小样本集上强大的分类能力，为后续的交互语义建模提供了准确可信的输入语义. 它采用 MLN 建模高层的交互行为，既保持了基于逻辑推理方法的灵活建模能力，又为知识推理引入了不确定性处理的能力. 和传统的交互行为分析方法不同，它不是采用运动跟踪和姿态估计的结果作为特征提取的工具，而是通过提取运动视频的时空特征作为行为识别模型的观测特征，避免了传统行为识别系统对运动跟踪和姿态估计准确性的依赖. 此外，它并不局限于两人交互行为分析，可以扩展到多人交互行为识别. 在初步建立的两人交互行为库上的实验结果表明，该方法可以有效地建模两人交互行为并具有一定的鲁棒性.

本文实验中采用的两人交互行为库中的行为类别比较少，目前我们正在创建更大规模的两人交互行为库，在大规模数据库上验证本文的方法是下一步工作的重要内容. 另一方面，交互行为过程中人与人之间存在严重的遮挡和自遮挡，在多视角场景下

合理利用深度信息，可以有效提高特征提取的准确性，这也是我们未来的研究工作.

参　考　文　献

[1] Turaga P, Chellappa R, Subrahmanian V S, Udrea O. Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(11): 1473-1488

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

[3] Oliver N M, Rosario B, Pentland A P. A Bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 831-843

[4] Xiang T, Gong S. Beyond tracking: Modeling activity and understanding behavior. International Journal of Computer Vision, 2006, 67(1): 21-51

[5] 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

[6] Park S, Aggarwal J K. A hierarchical Bayesian network for event recognition of human action and interaction. ACM Journal of Multimedia Systems, Special Issue on Video Surveillance, 2004, 10(2): 164-179

[7] Ryoo M S, Aggarwal J K. Recognition of composite human activities through context-free grammar based representation//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. NY, USA, 2006: 1709-1719

[8] Du Y, Chen F, Xu W, Zhang W. Activity recognition through multi-scale motion detail analysis. Neurocomputing, 2008, 71(16-18): 3561-3574

[9] Hongeng S, Nevatia R, Bremond F. Video-based event recognition: Activity representation and probabilistic recognition methods. Computer Vision and Image Understanding, 2004, 96(2): 129-162

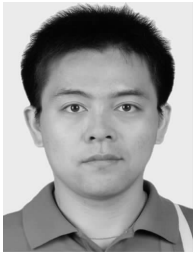
[10] Hakeem A, Shah M. Learning, detection and representation of multi-agent event in videos. Artificial Intelligence, 2007, 171(8-9): 586-605

[11] Dollar P, Rabaud V, Cotterll G, Belongie S. Behavior recognition via sparse spatio-temporal features//Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing, China, 2005: 65-72

[12] Sminchisescu C, Kanaujia A, Li Z, Metaxas D. Conditional random fields for contextual human motion recognition//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China, 2005: 1808-1815

[13] Richadson M, Domingos P. Markov logic networks. Machine Learning, 2006, 62(1-2): 107-136

[14] Tran S D, Davis L S. Event modeling and recognition using Markov logic networks//Proceedings of the 7th European Conference on Computer Vision. Marseille, France, 2008: 610-623



HAN Lei, born in 1982, Ph. D. candidate. His research interests include vision-based human-computer interaction and human action analysis.

LI Jun-Feng, born in 1985, M. S. candidate. His research interests focus on vision-based human action recognition.

JIA Yun-De, born in 1962, Ph. D., professor, Ph. D. supervisor. His research interests include computer vision, artificial intelligence and human-computer interaction.

Background

This work is supported by the National Natural Science Foundation of China under grant No. 60675021 and the National High Technology Research and Development Program (863 Program) of China under grant No. 2009AA01Z323.

Human interaction recognition is an important research topic in vision-based human action analysis. Generally, there are two key questions involved in this recognition task. One is to extract and represent the detail motion information from raw human interaction video data, and the other is to model the motions, which includes the primitive actions of a single person and the interactions among different persons.

Features used in the existing methods can be divided into three categories (Niebles et al.): static features based on edges and limb shapes, dynamic features based on optical flows, and spatial-temporal features based on space-time volume data. In particular, features from spatial-temporal interest points have shown to be useful in the human action recognition task, providing a rich description and powerful representation. Each video sequence is represented as a spatial-temporal word by extracting space-time interest points. The

dense space-time interest points detected from videos are divided into two sets exclusively according to the history information along the evolvement and the connectivity of the two human silhouettes. Then K-means clustering performs on points in the training set and learns the spatial-temporal codebook. For a given set of interest points, a spatial-temporal word is built by allowing each point to vote softly into the few centers nearest to it and accumulating the scores of all the points.

Since many methods have been presented for human interaction recognition. Inspired by the hierarchical recognition framework, the authors divided the task into two consecutive processes: primitive action recognition and interaction modeling. Temporal models such as HMMs have been widely used to model human primitive actions. However, a strong assumption of independence is usually made in such generative models. The authors use Conditional Random Field (CRF) to model the primitive actions of a single person. The inputs of CRF are the spatial-temporal words and common sense domain knowledge and first order logic production rules with weights are employed to learn the structure and the parameters of Markov Logic Network (MLN). The MLN can naturally integrate common sense reasoning with uncertain analysis, which is capable to deal with the uncertainty produced by CRF. Experiment results on the interaction dataset are provided to demonstrate the effectiveness and the robustness.