人体运动非监督聚类分析*

王天树[1+], 郑南宁[1], 徐迎庆[2], 沈向洋[2]

[1](西安交通大学 人工智能与机器人研究所,陕西 西安 710049)

[2](微软亚洲研究院,北京 100080)

Unsupervised Clustering Analysis of Human Motion

WANG Tian-Shu[1+], ZHENG Nan-Ning[1], XU Ying-Qing[2], SHUM Heung-Yeung[2]

[1](Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China)

[2](Microsoft Research Asia, Beijing 100080, China)

+ Corresponding author: Phn: 86-10-62986677 ext 509, Fax: 86-10-82899633, E-mail: wangtsh@cn.ibm.com

http://www.aiar.xjtu.edu.cn

Received 2001-06-26; Accepted 2001-08-02

Wang TS, Zheng NN, Xu YQ, Shum HY. Unsupervised clustering analysis of human motion. *Journal of Software*, 2003,14(2):209~214.

Abstract: An unsupervised learning approach for analysis of human motion is proposed. In this approach, by learning a set of hidden Markov models under constrains of minimal descript length criterion, a continuous gestures sequence could be segmented and clustered, and thus the segments and labels of the original sequence are automatically extracted. The approach contains two steps. First continuous gestures are discretized and an original solution is found in discrete domain based on MDL criterion. Then coming back to continuous domain, a set of HMMs is learnt under constrains of MDL criterion, The HMMs exploit richer dynamics and thus generate better results. Experimental results by using real human gesture data demonstrate the effectiveness of the approach.

Key words: unsupervised learning; pattern recognition; artificial intelligence

摘 要: 提出了一种基于非监督学习的人体运动分析方法.该方法通过使用 MDL 准则约束下的 HMM 模型对连续运动序列进行分割和聚类,并实现对运动序列的自动分割和标记.该方法由两步组成,首先通过聚类将连续运动离散化,并按照最小描述长度准则在离散域得到初始解.在此基础上,返回到连续域训练 MDL 准则约束下的 HMM 模型.使用 HMM 模型可以进一步利用原始序列中的动态信息获得更精确的最终结果.通过对实际人体运动序列进行的实验验证了方法的有效性.

关键词: 非监督学习;模式识别;人工智能

中图法分类号: TP181 **文献标识码**: A

人体运动识别是计算机视觉研究工作的热点之一.它在人机接口、场景自动监控、计算机动画等领域有着广泛的应用[1].通常实现一个基于统计方法的运动识别系统需要 3 个步骤:(1) 收集并标记待识别运动的一些样

* Supported by the National Natural Science Foundation of China under Grant No.60024301 (国家自然科学基金)

第一作者简介: 王天树(1977－),男,博士,助教,主要研究领域为模式识别,计算机视觉.

本;(2) 对每类样本分别建立统计模型;(3) 利用最大似然方法识别观测到的未知运动序列.如上所述,这类系统的训练过程是一个典型的监督学习过程.其中事先收集训练样本数据是一个必要的环节.为了获取这些数据,需要手工标记和分割训练样本.但是,对人体运动的标记是一个易于出错且繁杂的工作,特别是在一些特殊应用场合,比如在场景自动监控和视频数据库索引中,我们无法事先得知可能会出现何种运动,也不可能事先收集到完全的训练数据集.针对传统方法存在的问题,本文给出了一个对人体运动非监督聚类方法.该方法不需要任何先验知识和标记的样本,只是利用观测到的人体运动所包含的上下文信息,就可以实现对连续的人体运动的自动分割和识别.

我们知道,在很多场合下,复杂的人体运动由一些基本动作构成.例如,舞蹈是由一些基本的舞蹈动作组成的.当我们观看了一段舞蹈表演之后,就会发现其中有一些重复出现的动作单元.同样,计算机也可以通过寻找连续人体运动中的重复出现的动作来自动学习构成动作的基本单元.而利用获取的基本动作单元就可以实现对连续动作的分割和标记.为了实现这一目的,需要解决以下两个问题:

(1) 人体运动不是精确重复的.同一个人的两次挥手,手在空间中的位置和速度会在一定范围内变化.给定的方法应当能够正确识别同一类的运动.

(2) 在非监督学习的过程中,我们事先既不知道可能存在多少种基本动作单元,也不知道基本动作单元是什么.因此在没有适当约束的情况下,能获得大量的可行解.例如有两种极端情况都是不可取的:(a) 整个运动序列被当作一个基本动作;(b)每一个采样时刻的姿态被当作一个基本动作.因此,我们需要合适的约束来避免不适当的解.

前人的相关工作[2~4]证明了利用隐马尔科夫模型(HMM)可以有效地识别人体运动.建立在统计学习方法基础上的 HMM 具有很好的鲁棒性.可以通过使用 HMM 来解决第 1 个问题.第 2 个问题实际上可以归结为选择一个表示数据的最优模型的模型选择问题.在没有先验知识的情况下,通过约束模型的复杂度,能够得到表示数据的最简洁模型.这就是在非监督学习中得到广泛应用最小描述长度(MDL)准则[5].实践经验证明,使用 MDL 往往可以得到符合人类认知的结果.如上所述,我们可以通过学习最小描述长度约束下的一组 HMM 模型,来实现对人体运动的非监督学习.本文给出了如何在 MDL 准则下学习 HMM 模型的学习算法.

1 算法概述

人体连续运动序列可以记为 $O=o_1o_2\ldots o_T$,其中 O_t 是 t 采样时刻人体的姿态,是一个表示人体关节位置的多维向量,T 是观测序列的长度.我们希望自动获取一组基本动作 $\{P_1,P_2,\ldots,P_n\}$.通过基本动作将原来的连续运动分割成 m 段,其中包含的 $m-1$ 个分割点记为 $s_k,k=1..m-1$,并将原始序列表示为基本动作单元的组合 $O=P_{l1}P_{l2}\ldots P_{lm}$,其中 $l_j\in[1\ldots n],j=1..m$ 为动作的标号.本文的问题,是在给定 O 的情况下,求取基本动作数目 n、对应基本动作 P_i 的统计模型 λ_i 及对 O 分割点 s_k 和标记 l_j.

我们的学习过程由两步组成,首先将对连续运动进行聚类,获取原始序列的离散化表示,然后通过基于 MDL 准则的算法提取符号串中的单词,获得候选词汇表.用从候选词汇表中选取的词汇初始化 HMM 模型.最后在 MDL 准则约束下通过 EM 算法训练 HMM 得到结果.系统流程如图 1 所示.

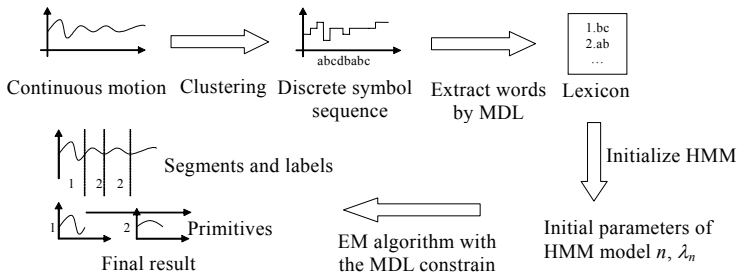

Fig.1 System framework

图 1 系统流程图

2 提取离散词汇

人体连续运动是由不同时刻人体姿态组成的序列.通过对人体姿态进行聚类,我们可以获得原连续运动的离散表示.本文采用了 CLUSTER 算法. CLUSTER 算法是一种在数据集上训练混合高斯模型(GMM)的算法.我们可以利用 CLUSTER 算法给出的结果,通过计算样本点与 GMM 中高斯函数均值的马氏距离来划分样本点的类别.CLUSTER 算法的优点是它利用 MDL 准则自动地确定类的个数[9].对得到的离散的符号串,我们合并相邻重复出现的字符.例如串“AAABBC”合并为“ABC”.

获得的符号串可被当作是一个去除了标点符号和空格的文本.从中寻找基本单元,类似于从文章中寻找单词.直接求取 MDL 约束下的最优解是一个 NP 难问题.本文中使用称为 COMPRESSIVE 算法的一种贪恋算法[10].

对一个长度为 M 的符号串,假定它在原始序列中出现了 N 次.如果将它作为单词,我们可以用替代符号在原序列中出现这个符号串的位置替换它.这里假定每个符号的描述长度都是 1.用长度为 1 的替代符号替换符号串,将使得总的描述长度减少 $M-1$.而该符号串出现 N 次,故描述长度减少 $N(M-1)$.但同时我们还要记录下这一条替换规则.这需要增加描述长度 $M+1$,其中用 M 来描述符号串的内容,1 描述替代符号.这样总的描述长度减少,

$$-\Delta DL = N(M-1)-(M+1) = M\cdot N-(M+N+1)\,. \tag{1}$$

COMPRESSIVE 算法的每一步均选出能够使序列描述长度减少最大的一个符号串作为单词,反复重复上述过程,得到一组单词,直到序列无法被进一步压缩为止.该算法可以使用前缀串(suffix array)数据结构来加速计算过程[11].前缀串是一个与原字符串同等长度的数组,数组中的每个单元为一个指向原字符串不同位置的指针,这些指针按照其指向字符串内容的字典序进行排序.在建立起前缀串数据结构后,只要顺序扫描一遍前缀串就可以得到使描述长度变化最大的单词.对于一个长度为 n 的符号串构造前缀串需要 $O(n\log n)$的计算复杂度,算法的计算复杂度近似为 $O(n^2\log n)$.图 2 给出了一个算法的具体示例.

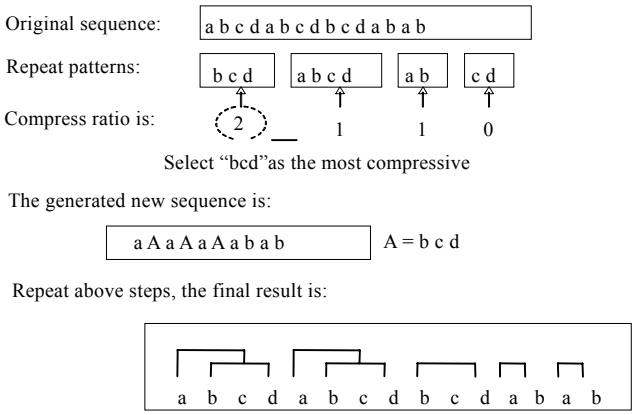

Fig 2 An example of COMPRESSIVE algorithm

图 2 COMPRESSIVE 算法示例

3 HMM 模型学习算法

利用在离散域提取的候选单词,可以对原始序列进行分割和标记.但由于这一过程中对原问题作了简化,因此得到的结果存在以下缺陷:离散化的过程本身对噪声比较敏感,同一种基本动作受噪声影响会导致不同的结果,产生不一样的单词;由于连续运动被离散化,实际得到的分割点不很准确.针对这些缺陷,我们使用在离散域得到的结果作为初值,通过训练 HMM 模型获得更好的结果.初始化的具体实现是:对离散域得到 N 个单词,利用每个单词在原连续序列中对应的部分分别训练 N 个 HMM 模型.

HMM 是一种统计意义下的有限状态机.它被广泛应用于对时间序列的识别过程中.HMM 模型的参数可以通过 EM 训练算法得到[12].我们使用 HMM 模型来表示连续运动中包含的基本动作.从连续运动中自动获取这

些基本动作,可以通过对 HMM 的训练来完成.这里我们给出了在 MDL 准则约束下训练多个 HMM 模型的方法.

对于一个连续观测序列,如果利用多个 HMM 模型来表示其中包含的基本动作,全部系统的统计描述长度(DL)可以如下计算[5]:

$$DL(K,\theta) = -\log p(O \mid \theta, K) + \frac{1}{2} LK \log(M) . \tag{2}$$

上式中,K 为 HMM 模型的个数,θ为模型的参数,L 为单个 HMM 模型包含参数的个数,M 为观测序列 O 中包含的数据总量,对一个维数为 N 的长度为 T 的运动序列,$M=N\cdot T$.

使用最小描述长度(MDL)准则估计模型参数的方法可以记为

$$(\hat{K},\hat{\theta}) = \arg\min_{k,\theta} DL(K,\theta) = \arg\min_{k,\theta}\left[-\log p(O \mid \theta, K) + \frac{1}{2} LK \log(M)\right]. \tag{3}$$

可以注意到,式(3)右端由两部分构成,分别为数据的似然度函数及模型的描述长度.如果固定模型的数目 K,则式(3)退化为通常使用的最大似然准则(ML).这里,估计模型参数和 K 采用如下方案:从一个较大的 K 出发,逐渐减小 K,而对固定的 K 使用最大似然准则更新模型参数.

$$(\hat{M},\hat{\theta}) = \arg\max_{\theta}\{\arg\max_{M} P(O,M,\theta)\} . \tag{4}$$

对于固定的 K 训练收敛后,我们使用 MDL 准则判断是否可以减少 K.具体过程是:首先选择两个最相近的 HMM.这里用 KL 距离准则来描述 HMM 模型之间的近似程度如下[8]:

$$D(\lambda_a,\lambda_b) = \frac{1}{T_a}\left|\log(O_a \mid \lambda_a) - \log(O_a \mid \lambda_b)\right| + \frac{1}{T_b}\left|\log(O_b \mid \lambda_b) - \log(O_b \mid \lambda_a)\right|, \tag{5}$$

其中 T_a 为观测 O_a 的序列长度,T_b 为观测 O_b 的序列长度.λ_a 和λ_b 分别为 HMM a 和 HMM b 的参数.

对这两个选定的最相近的 HMM a 和 HMM b 及相应的观测数据 O_a, O_b,如果我们合并这两个模型,即训练一个参数为λ_c 的 HMM c.来取代 a 和 b.DL 的变化为

$$\Delta DL = -\left[\log(O_a \mid \lambda_a) + \log(O_b \mid \lambda_b) - \log(O_{a,}O_b \mid \lambda_c)\right] + \frac{1}{2}\Delta L \log(T_{ab}N) , \tag{6}$$

其中 T_{ab} 为观测(O_a,O_b)的长度,ΔL 为一个 HMM 模型所包含的参数个数.

根据 MDL 准则,如果式(6)小于 0,则合并两个模型可以使 DL 减小,那么模型可以合并,且 K 变为 $K-1$.我们可以继续重复以上步骤.反之无法合并模型,可以认为算法收敛.

4 实验结果与结论

本文通过对手势的运动分析来验证算法的有效性.实验数据为摄像机拍摄的包含 3 种基本手势的视频运动序列.3 种基本运动如图 3 所示.从图中可以看到同类运动有一定范围的变化.这 3 种基本手势在一个连续手势序列中按照随机的次序反复出现多次.全部实验纪录的视频数据包含 4 000 帧(25 帧/秒).其中每个手势重复了 10 次以上.

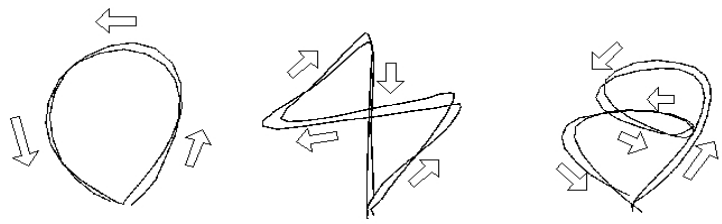

Fig.3 The samples of three primitive, which are O, X, B from left to right respectively. Two samples of each primitive are showed in the figure

图 3 3 种基本运动,从左至右分别为“O”,“X”,“B”.图中对每个动作都给出了两个样本

为了得到手在图像中的位置,我们使用了基于颜色特征的视频跟踪方法.该方法通过匹配特征区域的颜色

分布来获取在后续图像序列中特征的位置.实验中,该方法成功地跟踪了全部的视频序列,并且得到了较好的跟踪结果(在 640*480 图像中,最大误差小于 7 个像素).图 4 给出了几帧跟踪结果.

Fig.4 Tracking results

图 4 跟踪的结果

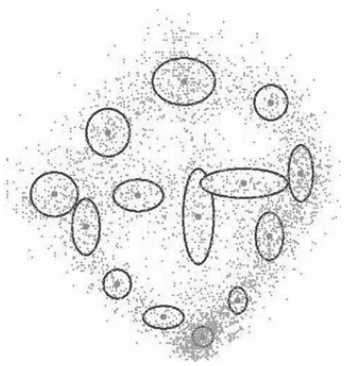

Fig.5 The samples of hand position and clustering results

图 5 手的位置所有样本点和聚类结果

通过跟踪过程,我们得到了待分析的人体运动序列,即手在图像中的二维位置.应用本文提出的算法计算过程如下:首先通过 CLUSTER 算法将全部样本聚为 14 类,如图 5 所示.其后在离散序列中使用 COMPRESSIVE 算法得到了 38 个基本单词.之后用这些词汇来初始化 HMM 模型,每个 HMM 模型转移矩阵设为全连接结构,而模型状态的个数由相应单词中符号的个数决定.最后用 EM 算法得到最终的结果.实验中我们发现,Segmental-Kmeans 算法的收敛速度很快,通常只需要 2~3 个循环.利用 MDL 准则合并相近的模型之后,最终的结果中包含 7 个基本动作,每个动作及对应的样本数据如图 6 所示.图中可见,我们得到了构成连续运动的基本组成部分,即基本动作.在获取基本动作的同时,运动序列被正确地分割和标记.全部计算在一台 PIII-500 计算机上利用 C 语言实现,计算过程共耗时 75 秒.

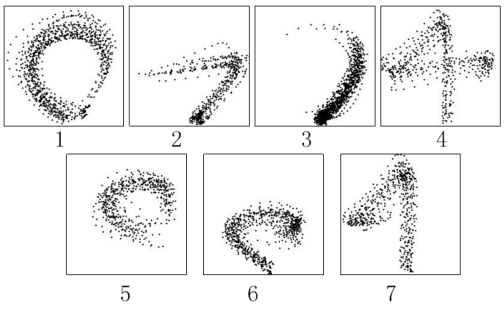

Fig.6 The extracted primitives: 1,3 compose O; 2,7 compose X; 3,4 compose X; 3,5,6 compose B

图 6 提取的基本运动:图案 1,3 组成“O”;图案;2,7 组成“X”;图案 3,4 组成“X”;图案 3,5,6 组成“B”

本文提出的是一种通用的非监督学习算法.我们没有利用与领域相关的特定知识,因此该方法也可以用于分析一般的时间序列.而在结合特定领域的相关知识后,该方法的性能可以得到进一步的提高.另外,本文只涉及如何提取运动序列中的基本单元,而没有涉及怎样得到这些基本单元之上的高层结构,即生成运动的统计语法规则.这将是我们的下一步的工作.

References:

[1] Gavria DM. The visual analysis of human movement: a suvey. Computer vision and Image Understanding, 1999,73(1):82~98.

[2] Bregler C. Learning and recognizing human dynamics in video sequences. In: Medioni G, ed. Proceedings of the IEEE Computer Vision and Pattern Recognition’97. Piscataway: IEEE Press, 1997. 568~574.

[3] Yang J, Xu Y, Chen CS. Human action learning via hidden Markov model. IEEE Transactions on Systems Man and Cybernetics, 1997,27(1):34~44.

[4] Starner T, Pentland A. Visual recognition of american language using hidden Markov models. In: Bichsel M, ed. Proceedings of the International Workshop on Automatic Face and Gesture Recognition. MultiMedia Laboratory, University of Zurich, 1995. 189~194.

[5] Rissanen J. A universal prior for integers and estimation by minimum description length. Annals of Statistics,1983,11:417~431.

[6] Clarkson B, Pentland A. Unsupervised clustering of ambulatory audio and video. In: Rodriquze J, ed. Proceedings of the ICASSP'99. Madison: Omini Press, 1999. 3037~3040

[7] Galata A, Johnson N, Hogg D. Learning structured behavior models using variable length hidden Markov models. In: Werner B, ed. IEEE International Workshop on Modeling people. Piscataway: IEEE Press, 1999. 95~101.

[8] Walter M, Psarrou A, Gong S. An incremental approach towards automatic model acquisition for human gesture recognition. In:Young DC, ed. Proceedings of the IEEE International Workshop on Human Motion. Bellingham: Applied Digital Imaging, 2000. 39~46.

[9] Bouman CA. CLUSTER: an unsupervised algorithm for modeling Gaussian mixtures. Soft Manual, 1995. http://dynamo.ecn.purdue.edu/~bouman/software/cluster/. 1995.

[10] Nevill-Manning C, Witten I. Online and offline heuristics for inferring hierarchies of repetitions in sequence. Proceedings of the IEEE, 2000,88(11):1745~1755.

[11] Sadakane K, Imai H. Constructing suffix arrays of large texts. In: Kesuregawa M, ed. CD-ROM Proceedings of the IEICE 9th Data Engineering Workshop. Tokyo:IEICE press. 1998. 96~101.

[12] Rabiner LR, Juang B. Fundamentals of Speech Recognition. New York: Prentice Hall, 1993. 321~387.

全国第 7 届计算语言学联合学术会议(JSCL 2003)

征 文 通 知

中国中文信息学会、中国计算机学会、中国人工智能学会和北京市语言学会于 2003 年 8 月 8 日~11 日在哈尔滨市与哈尔滨工业大学联合举办“全国第 7 届计算语言学联合学术会议(JSCL 2003)”。

一、征文范围

(1) 计算语言学的理论基础：知识表示、语义学、语用学、语料库语言学、记忆模型、机器学习、知识获取和推理技术；

(2) 现代汉语的句法分析和语义分析：汉语分析的策略、句法分析和语义分析中的计算问题、汉语分析的展望；

(3) 汉语语料库技术及系统；

(4) 汉语人机接口技术及系统；

(5) 机器翻译技术、系统及评测方法；

(6) 话语和篇章的分析与生成：话语的心理学和语言学模型、篇章分析、话语生成；

(7) 自然语言处理的应用系统：汉语自动分词系统、智能检索系统、自动文摘系统、自动校对系统、文本自动分类系统、信息抽取、信息过滤、智能搜索引擎、文本挖掘、智能拼音汉字转换等；

(8) 计算语言学的资源研究及建设：树库、语法词典、词汇语义分类体系和语义词典、汉语分词词表、概念词典、知识库等；

(9) 服务于计算语言学的支撑环境和软件技术。

二、来稿要求

全文不超过 8000 字，每篇论文均应有中英文两种文字标题、作者、姓名、单位和不超过 200 字的摘要。来稿全文一式 3 份，作者请自留底稿。会议概不退稿。大会录用的论文将收入有出版书号的会议论文集。

三、截稿日期

(1) 截稿日期：2003 年 4 月 1 日(以邮戳为准) 注：来稿请在首页上标明“JSCL 2003”。

(2) 录用通知发出日期：2003 年 5 月 1 日

(3) 作者提交的论文激光印刷版日期：2003 年 6 月 1 日(以到达日期为准)

四、联系方式

来稿邮寄地址：100084 北京市清华大学计算机科学与技术系 陈群秀 收 联系电话：010-62781479