ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.7, July 2008, pp.1623−1634
DOI: 10.3724/SP.J.1001.2008.01623

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

视角无关的动作识别*

黄飞跃+, 徐光祐

(普适计算教育部重点实验室,清华信息科学与技术国家实验室(筹),清华大学计算机科学与技术系,北京 100084)

Viewpoint Independent Action Recognition

HUANG Fei-Yue+, XU Guang-You

(Key Laboratory of Pervasive Computing of the Ministry of Education, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: E-mail: huangfeiyue@gmail.com

Huang FY, Xu GY. Viewpoint independent action recognition. *Journal of Software*, 2008,19(7):1623−1634.
http://www.jos.org.cn/1000-9825/19/1623.htm

Abstract: Action recognition is a popular and important research topic in computer vision. However, it is challenging when facing viewpoint variance. So far, most researches in action recognition remain rooted in view-dependent representations. Some view invariance approaches have been proposed, but most of them suffer from some weaknesses, such as lack of abundant information for recognition, dependency on robust meaningful feature detection or point correspondence. To perform viewpoint and subject independent action recognition, this paper proposes a representation called "Envelop Shape" which is viewpoint insensitive. "Envelop Shape" is easy to acquire from silhouettes using two orthogonal cameras. It makes full use of two cameras' silhouettes to dispel influence caused by human body's vertical rotation, which is often the primary viewpoint variance. With the help of "Envelop Shape" representation and Hidden Markov Model, the inspiring results on action recognition independent of subject and viewpoint are obtained. Results indicate that "Envelop Shape" representation contains enough discriminating features for action recognition. Extension of "Envelop Shape" is also proposed to make it run under fewer restrictions of camera configurations, which increases its application value effectively.

Key words: action recognition; viewpoint independent; envelop shape; hidden Markov model

摘 要: 人体动作识别是计算机视觉中一个流行而且重要的研究课题.当观察视角发生变化时,动作识别变得格外困难.至今为止,关于动作识别和手势识别的大多数研究工作都是围绕着视角相关的表达展开的.有一小部分利用了视角不变的表示开展研究,可是它们大多数存在一些缺陷,比如缺少用于识别的足够信息,依赖鲁棒的语义特征点的检测或者是点对应.为了解决这个问题,实现视角无关、动作人无关的动作识别,提出了"包容形状"的表示,这种表示不依赖于特定视角.在人体动作识别中,人的身体旋转通常是引起视角变化的主要原因.包容形状充分利用了两个正交摄像机拍摄的轮廓信息以去除由人的身体旋转产生的影响.从来自两个正交的摄像机拍摄的外轮廓,可以很容易计算得到包容形状.利用包容形状的体态表示和隐马尔可夫模型,取得了非特定人、任意视角下动作识别的很好的实验结果.这些实验结果也表明了包容形状包含有足够区分度的信息.同时提出

* Supported by the National Natural Science Foundation of China under Grant No.60673189 (国家自然科学基金)
Received 2007-09-22; Accepted 2008-01-21

了包容形状的扩展表示,以便在两个摄像机并不完全正交的更为一般的摄像机配置条件下也可以应用,这极大地加强了其实用价值.

关键词: 动作识别;视角无关;包容形状;隐马尔可夫模型

中图法分类号: TP391 **文献标识码**: A

人体动作识别是计算机视觉里一个活跃的研究方向,在这个方向上有不少综述力图把以前的相关研究方法进行总结和分类,比如文献[1–4].本文重点研究面向动作识别的、与视角无关的体态表示,同时提出了一个通用的非特定人任意视角下的动作识别方法.

在我们提出的动作识别系统中,有 3 个组成模块:预处理、姿态估计和识别.预处理模块包括人体检测和跟踪,它为姿态估计提供底层的人体表示信息.姿态估计模块用来确定和表示人的躯干和肢体在一个静态帧里的对应配置状况.而识别模块则利用对每一帧进行姿态估计得到的结果来对动作进行分类和识别.

我们定义体态(posture)为某个时刻的人体姿态.它可用相应时刻视频帧中的人体姿态特征向量来表示.比如,人体轮廓的水平和垂直直方图[5];从人体边缘像素点到人体质心距离构成的向量[6]等等.这个描述人体姿态的特征向量是进一步处理的基础,因此是人体动作识别的关键.姿态估计模块的功能是根据帧图像来完成当前时刻人体特定体态的描述和表达.经过姿态估计模块的处理,序列中的每一帧可以计算得到一个特征向量,从而图像序列可以转换成特征向量序列用于下一步的动作识别.我们认为,动作可以看成是有实际语义的一段体态序列,比如举手、步行、奔跑等等.动作识别就是针对体态序列进行分类和识别.所以体态表示是人体动作识别系统中最为基础和关键的一个部分.

如文献[7]中所述,一个用于分类的良好表示方法应该具有如下度量特性:对于不同分类下这些度量应该有较大的区分度,而在同一个分类中的物体这些度量应该相似.根据这个观点我们需要寻找既具有高区分度,同时对外在条件变化具有容忍能力的特征.在非特定人任意视角动作识别情景下,观察视角的变化、识别人变化、背景变化、光照条件变化这些都是常见的条件变化.我们希望找到一个好的特征表示可以容忍上述这些变化的干扰.在这些变化中,最需要考虑的是视角的变化.其他变化带来的干扰可以根据应用场景进行特定的训练;然而为了进行自然的人体动作识别,我们无法限制动作人的身体进行移动和旋转,而人体的运动则会不可避免地导致观察视角的变化.

寻找可以用于动作识别、能够容忍视角变化的体态表示,即视角不变的体态表示是一项具有挑战性的工作.当前已经提出了一些视角不变的动作识别方法.比如Campbell等人提出的基于立体视觉数据的三维手势识别系统[8].Seitz 和 Dyer描述了一种用于检测周期运动的仿射不变的方法[9].Cen Rao提出了多种用于人体动作识别的视角不变的分析方法[10,11].他使用人手质心的轨迹来描述由一只手完成的动作,并利用轨迹的仿射不变量开发了可以自动运作的识别系统.Vasu Parameswaran重点研究了视角不变的人体动作识别方法[12,13].他选择了人体的 6 个连接点,并且计算每一个体态中它们的三维不变量.这样,每个体态可以被约束在三维不变量空间中的一个参数曲面上.Daniel等人提出使用历史运动体(motion history volume)作为动作识别的一个视角无关的表示,这种表示需要多个标定的摄像机[14].Ogale等人则通过为每个姿态保存各种不同视角下的多个表观模型来进行视角不变的动作分析[15].

虽然在现阶段已经开展了不少的视角不变的动作识别的研究工作,可仍有很多问题亟待解决.比如,大多数的方法依赖鲁棒的语义特征点检测或者是点对应,而这些实现起来是比较困难的.另一方面,视角不变的方法通常会存在信息的损耗,导致用于识别不同动作的区分信息的缺失.如何使得体态的表示可以容忍视角的变化同时又保留足够的可以用于动作识别的可区分信息,这是一个关键问题.

出于这个考虑,我们提出了一种叫做“包容形状(envelop shape)”的表示.在仿射摄像机投影模型的假设下,从理论和实验两方面证明了这种表达对于视角的变化具有不敏感性.包容形状很容易从底层信息中提取,可以直接通过两个正交摄像机拍摄的动作人的轮廓计算获得.它比以前的视角不变的表示包含了更多的信息,而且不需要依靠任何较难实现并且对误差很敏感的语义点检测和点对应过程.在表 1 中,我们把常见的视角不变的

动作识别方法和包容形状进行了比较和评价,分别列举了各自的优缺点.利用这种包容形状的体态表示和隐马尔可夫模型,我们开发了自己的动作识别系统.实验结果表明,我们的系统对于非特定人、任意视角下的动作有着很理想的识别能力.同时,为了在两个摄像机并不完全正交的更为一般的摄像机配置条件下应用本方法,本文又提出了包容形状的扩展表示,这极大地加强了它的实用价值.

Table 1 Comparison of some view independent action recognition methods

表 1 常见的视角不变的识别方法的比较与评价

Methods	Used features	Advantages	Disadvantages
Cen Rao[10,11]	Trajectory of hand	Simple, one camera	Just fit for classify simple actions
Vasu[12,13]	Articulation	One camera	Need articulation points
Daniel[14]	Silhouette	Better distinguishability	Need multiple calibrated cameras
Ogale[15]	Appearance	Also classify view-point	Need data from all kinds of views
Envelop shape	Silhouette	Easy to apply	Need a pair of cameras

本文第 1 节陈述具有不依赖于特定视角的体态表示.第 2 节描述实现的动作识别系统并且给出实验结果.第 3 节给出更为普遍的扩展表达.第 4 节给出结论.

1 视角无关的体态表示

在人体动作识别领域中,体态的表示一直是一个基本而关键的问题.大多数的人体动作识别方法通常要求人体朝向相对摄像机固定,比如正面或者侧面.这时视角固定,从而可以基于视角相关的体态表示进行识别.然而,为了实现实际环境下的人体自然动作的识别,例如,在识别讲台上教师的动作时,教师需要在讲台上自由地走动,为此我们无法限制动作人身体的自由移动和旋转.这些人体运动会不可避免地导致观察视角的变化.一个好的动作识别系统应该能够有良好的视角不变特性,因此很自然的想法就是去发掘视角无关的体态表示,即在各种不同的视角下基本保持一致的体态不变量表示.

1.1 动作识别中的视角变化

本文中,视角是观察视角的简称,是指摄像机观察主体(人体)的方向.在人体动作识别中,观察视角的变化可以由两种原因——人体运动或者摄像机运动引起.一般的动作识别中采用固定的摄像机,本文也只考虑摄像机固定的情况,这时只需考虑由人体运动引起的视角变化即可.

视角变化可以分成两个部分:平移和旋转.在动作识别中,几乎所有的体态表示都已经有了平移的不变特性,所以我们只需要考虑旋转不变性.图 1 显示了我们的体系中采用的坐标系.在这个坐标系中,Y 轴是竖直向上的.在坐标系中,一共存在 3 种旋转分量用来表示旋转,它们分别是:转动(roll)、倾斜(pitch)和偏转(yaw).这 3 个分量分别描述了绕着 Z 轴(α)、X 轴(β)和 Y 轴(γ)各自的旋转.

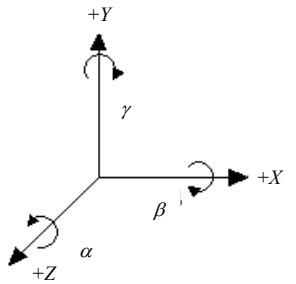

Fig.1 The coordinate system

图 1 系统中采用的坐标系

通常人体在做出一个有含义的动作时,会自然地伴随着人体朝向的变化.这里说的“朝向”是特指人体围绕竖直轴的旋转.比如黑板前的教师,写板书是面向黑板的,而讲解则是面向学生的.当动作者在一个固定的摄像机面前完成某些自然动作的时候,人体朝向是任意的,可以是正面,也可以是侧面,然而它们还是同一个动作.不同的人体朝向之间仅仅对应了人体的偏转变换.出于这方面的考虑,如果只有偏转分量变化,我们应该把这些体态归为同一个分类中.而如果还存在另外两种旋转分量(转动或倾斜),我们会把它们归为不同的分类中.例如,一个人直立着和躺在地面上,这时候存在转动或者倾斜两种旋转的分量,我们会把它们看成是两种不同的体态;而当一个人仅仅是把他的身体旋转到另一个朝向,我们则可以认为他的体态还是同一种.通过如上讨论,我们可以有如下结论:研究视角无关的动作识别中只需考虑偏转情况下的不变量表示.

1.2 包容形状的体态表示

在人体动作识别的实际应用场景之中,由于人体本身的深度变化相对于人体到摄像机的距离通常较小,我们可以采用仿射摄像机模型.为了得到动作识别中的视角不变量的表示,我们提出了一个双摄像机配置方案,如图 2 所示.这两个摄像机的成像平面都和竖直轴 Y 平行,它们的光轴是正交的,同时它们像平面坐标系中的 V 轴那样都和 Y 轴平行.让我们来考虑人体的一个水平截面 H,在这个截面上的所有点到像平面 1 上的投影都在直线 L1 上,而在这个截面上的所有点到像平面 2 上的投影都在直线 L2 上(即直线 L1 是点 p2 的外极线,而直线 L2 则是点 p1 的外极线).这样,人体偏转就相当于人体所有的水平截面在自身对应的二维平面内做了一个旋转.为了发掘偏转不变量,我们可以分析人体在二维水平截面 H 上的切面形状在旋转时的投影变化情况.

如图 3 所示,由于两个摄像机光轴正交,所以 U1 轴和 U2 轴的夹角是 90°.假设在水平截面 H 上人体外轮廓对应形状 S,它在原始的 U1 和 U2 轴中的投影线段是 AB 和 BC,那么 S 在矩形 ABCD 里面.在另外一个旋转了某个角度 θ 的 U1′和 U2′轴中,它的投影在线段 EF 和 FG 中.这里我们定义原始投影线段的长度为 x 和 y,而新的投影线段的长度则是 x'和 y'.那么,我们可以得到如下关系式:

$$x' \le x\cos\theta + y\sin\theta \quad y' \le y\cos\theta + x\sin\theta \tag{1}$$

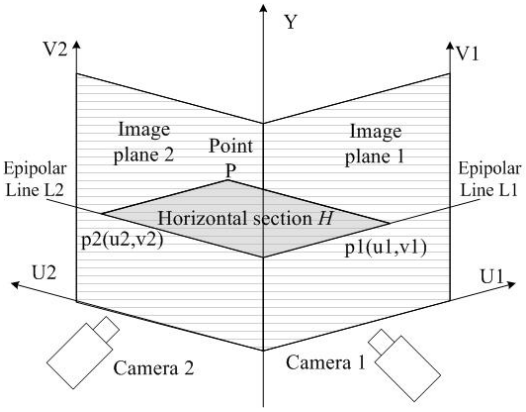

Fig.2 Two cameras configuration

图 2 双摄像机配置方案

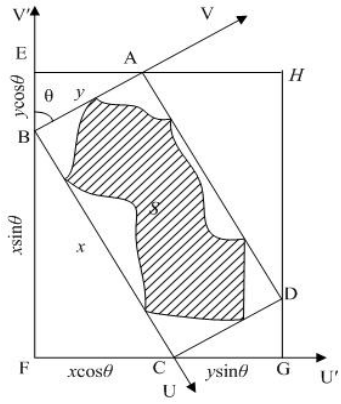

Fig.3 2D shape projection on different rotations

图 3 二维形状在不同旋转下的投影

定义 r:

$$r = \sqrt{x^2 + y^2} \tag{2}$$

那么旋转后的 r'为

$$r' = \sqrt{x'^2 + y'^2} \le \sqrt{x^2 + y^2 + 2xy\sin 2\theta} \le \sqrt{2}r \tag{3}$$

取r_0是所有旋转对应的各个r中的最小值,那么在任何旋转情况下,相应的r值都会满足如下取值区间:

$$r_0 \le r \le \sqrt{2}r_0 \tag{4}$$

与原始投影值 x'和 x 或者 y'和 y 之间比值的无限范围区间相比,这是一个相当小的取值区间,也就是说,我们找到了一种视角不敏感的人体表示.对于每一个水平截平面,我们利用公式(2)来计算一个 r 值.这样,对于每一个静态帧的人体体态,我们可以得到一个 r 值向量.由于这个向量构成的形状可以把人体的轮廓包围在内部,因而我们把这个 r 值向量称为“包容形状(envelop shape)”.我们给出一些在不同视角下合成人体模型的包容形状图像.如图 4 中(a),(b)两组图像分别表示两种体态围绕着竖直轴(Y 轴)旋转了 8 个不同角度时的情况.每组前两行是两个正交摄像机拍摄的外轮廓图像,而第 3 行则是包容形状图像.从图中我们可以看到,在视角发生变化时,包容形状的变化很小.

(a) Posture A at different viewpoints

(a) 不同视角下的体态 A

(b) Posture B at different viewpoints

(b) 不同视角下的体态 B

Fig.4

图 4

下面给出计算每帧对应的包容形状表示的算法:

(1) 从两个摄像机相应图像帧中提取人体外轮廓.

(2) 对外轮廓进行尺度正规化,使所有人体外轮廓都是同一个高度.

(3) 对于摄像机的一对对应帧,在归一化外轮廓的每一个高度上取对应的水平截面,根据公式(2)来计算 r 值,其中 x 和 y 是两帧中外轮廓在当前高度的宽度.

包容形状具有如下优点:

(1) 包含了两个自由度的信息:垂直轴和水平截面.与诸如轨迹投影[10,11]这样的视角不变量相比,它包含了更多的信息,这些视角不变量实际上只有一维的信息.因此,包容形状在可以容忍视角变化的同时有着更好的区分度.

(2) 很容易获得,只需要外轮廓作为输入.提取外轮廓通常比语义点检测、跟踪或者是点对应要容易得多.

虽然我们提出这两个摄像机的摆放需要保证成像平面和竖直轴平行同时光轴正交,但是,实际上并不需要严格的摄像机来标定.我们知道,精确的摄像机标定是比较困难的.在我们的这种方案中,摄像机的摆放大致满足这样的要求即可.这意味着我们无须在摄像机的精确摆放和标定上花费大量的时间.正如本文前面提出的,这种表示仅仅是视角不敏感,它的取值被限制在一个较小的范围内变化.所以,大致的摆放即可.我们将在第 2 节中展示我们所进行的实验.相关的视频数据都是利用两个大致摆放并未精确校准的摄像机采集的,而我们可以看到实验结果依然非常理想.

2 识别系统和实验

2.1 动作识别系统和HMM模型

基于这种包容形状的体态表示,我们可以把动作的每一帧转换成特征向量,这样,每一个动作就是一个特征向量序列.动作识别系统只需分析特征向量序列即可.我们在一个智能教室中实现和部署了任意视角下的动作识别系统.图 5 展示了系统流程图.我们首先采用PFinder算法[16]来提取人体外轮廓.然后,由两个摄像机的视频序列作为原始输入,我们为每一帧时刻生成了相应的包容形状向量,然后利用主分量分析(principal component analysis,简称PCA)来进行降维.

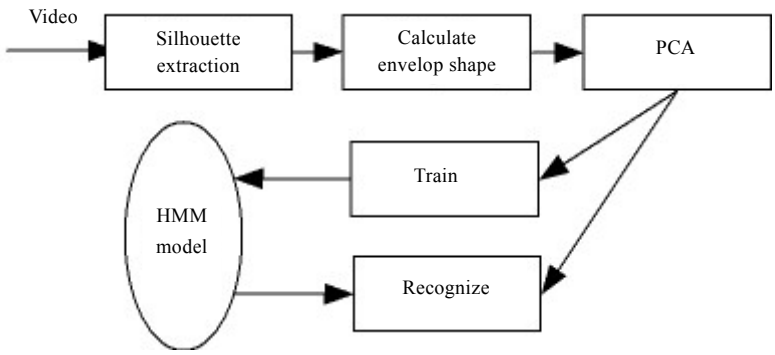

Fig.5 Action recognition system flow diagram

图 5 动作识别系统的流程图

对于每一个视频段,经过预处理和体态表示模块,我们可以得到对应的时序特征向量序列.有很多种方法可以对时序特征向量序列进行分类和识别,例如动态时间归整、隐马尔可夫模型[17]、耦合马尔可夫模型[18]或者是概率解析[19]等方法.这里,我们采用连续隐马尔可夫模型来进行动作的训练和识别.

隐马尔可夫模型(hidden Markov model,简称HMM),作为一种基于随机过程和概率观测的动态统计模型,最初被应用在语音识别中并且获得了很大的成功[20].随后被广泛地应用在包括语音信号处理、语义理解、手写体识别、唇读识别、动作识别等多种包括状态转换的动态过程系统的识别中.目前,在基于视觉的人体动作识别中,隐马尔可夫模型及其改进模型已经是应用最为广泛的一种动态模型.HMM是在马尔可夫链的基础上发展起来的,马尔可夫链是状态和时间参数都离散的马尔可夫过程,从数学上,我们可以给出如下定义[20]:

随机序列X_n,任一时刻n,它可以处在的状态为$\theta_1,\theta_2,\ldots,\theta_N$.已知$m$时刻的状态$q_m$,在$m+k$时刻所处的状态$q_{m+k}$的概率只与$q_m$相关,而与$m$时刻以前它所处的历史状态条件相独立,即有:

$$P(X_{m+k}=q_{m+k}\mid X_m=q_m,X_{m-1}=q_{m-1},\ldots,X_1=q_1)=P(X_{m+k}=q_{m+k}\mid X_m=q_m) \tag{5}$$

其中 $q_1,q_2,\ldots,q_m,\ldots,q_{m+k}\in(\theta_1,\ldots,\theta_N)$,那么称X_n为Markov链,而且称

$$P_{i,j}(m,m+k)=P(X_{m+k}=q_{m+k}\mid X_m=q_m),1\le i,j\le N,\ m,k \text{ 为正整数} \tag{6}$$

为k阶转移概率.当$P_{i,j}(m,m+k)$与m无关时,称X_n为齐次马尔可夫链.

HMM 由状态和观测两部分组成,状态表示当前时刻的本质属性,是隐含而无法观测的,只能通过观测来推理.而观测是状态的外在表现,是可以测量或者计算得到的.目前通用的隐马尔可夫模型的状态序列都是基于一阶齐次马尔可夫链,其意义是系统当前所处状态的概率只与前一个时刻状态有关系,与其他历史状态条件独立.图 6 以图的形式表现了隐马尔可夫模型.其中,θ_t 表示 t 时刻的状态,O_t 表示 t 时刻的状态 θ_t 的观测值.状态 θ_t 是隐藏而无法观测的,因此只有通过 O_t 表现出来.由于 HMM 模型假定状态序列是一阶齐次马尔可夫链,即

$$P(\theta_t\mid\theta_1,\ldots,\theta_t)=P(\theta_t\mid\theta_{t-1}) \tag{7}$$

其中, $P(\theta_t\mid\theta_{t-1})$ 称为状态转移概率,那么一个 HMM 可以由以下参数描述:

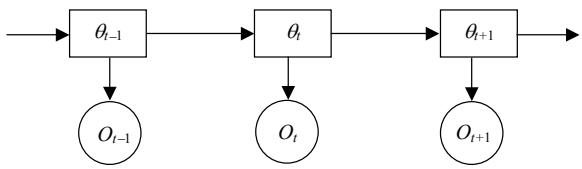

Fig.6 Hidden Markov model

图 6 隐马尔可夫模型

N: HMM中马尔可夫链中的状态数目.记N个状态为 $\theta_1,\ldots,\theta_N$,记t时刻马尔可夫链所处状态为q_t,

$q_t \in (\theta_1,...,\theta_N)$.

M: 每个状态对应的可能的观测值的数目.记M个观测值为$V_1,...,V_M$,记t时刻观测到的观测值为O_t,$O_t \in (V_1,...,V_M)$.

π: 初始状态概率向量,$\pi = (\pi_1,...,\pi_N)$,其中$\pi_i = P(q_1 = \theta_i)$,$1 \le i \le N$.

A: 状态转移概率矩阵,$A = (a_{ij})_{N\times N}$,其中$a_{ij} = P(q_{t+1} = \theta_j \mid q_t = \theta_i)$,$1 \le i, j \le N$.

B: 观测值概率矩阵,$B = (b_{jk})_{N\times M}$,其中$b_{jk} = P(O_t = V_k \mid q_t = \theta_j)$, $1 \le j \le N$, $1 \le k \le M$.

这样,可以记一个 HMM 为$\lambda = (N, M, \pi, A, B)$,或者简写为$\lambda = (\pi, A, B)$.

上述HMM的观测值是M个离散的值,在实际应用中通常需要把观测值利用向量量化的方法离散化,这样通常会丢失部分信息.所以,我们采用了它的改进模型:连续HMM.在连续HMM中的B不再是矩阵,而采用一组观测值概率密度函数来表示,即$B=\{b_j(X),j=1,\ldots,M\}$,这里,$b_j(X)$通常取高斯概率密度函数:

$$b_j(X) = \sum_{k=1}^{K} c_{jk} b_{jk}(X) = \sum_{k=1}^{K} c_{jk} N(X, \mu_{jk}, \Sigma_{jk}), \quad 1 \le j \le M \tag{8}$$

其中,$N(X, \mu_{jk}, \Sigma_{jk})$为多维高斯概率密度函数.这里的$K$表示混合模型的高斯函数个数,这时候的$M$也不再表示离散观测值的总数,而是观测值向量的维度.

由于包容形状对应的特征向量的维度较高,并且各维之间包含了较多的冗余信息,所以我们首先把包容形状向量利用 PCA 进行降维处理后得到观测向量.本系统中,我们降低维度为 8 维.这样,在我们的动作识别系统中,HMM 模型的参数包括:

N: HMM 模型的状态数.本系统中取 N=5.

M: 观测向量的维度.本系统中取 M=8.

K: 混合模型的高斯函数个数.本系统中取 K=10.

隐马尔可夫模型作为一种成熟的动态系统分析工具,应用在动作识别中可以屏蔽动态系统分析的细节,从而让我们更加专注于动作识别中的动作表示和姿态估计.

2.2 识别实验

为了进行动作识别的实验,我们采集了相应的视频段并建立了自己的视频数据库.在这个数据库中,一共包含有 7 个不同的动作者.每个动作者表演了 9 种自然动作.这些动作分别是“指向”、“举手”、“挥手”、 “摸头”、“交流”、“鞠躬”、“捡起(东西)”、“踢腿”和“步行”等.这些动作涵盖了人手、人头、躯干、腿部的运动,具有一定的典型性.动作者会在 3 个任意视角下面执行每一个动作 3 遍,也就是每个动作者共重复一个动作 9 遍.图 7 和图 8 显示了我们实验数据的一些示例.每个图包含有 5 行,前两行是两个摄像机的图像,接下来的两行是利用 PFinder 方法提取得到的人体轮廓,最后一行则是包容形状向量生成的图像(每一个实际的动作包含大约 30 帧左右的图像,图例中仅显示了部分采样).动作序列都是在任意视角下面采集的,这意味着我们的实验是视角无关的.

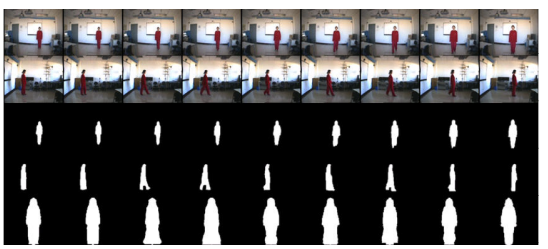

Fig.7 One group of “Walk” sequences

图 7 一组“步行”的动作序列

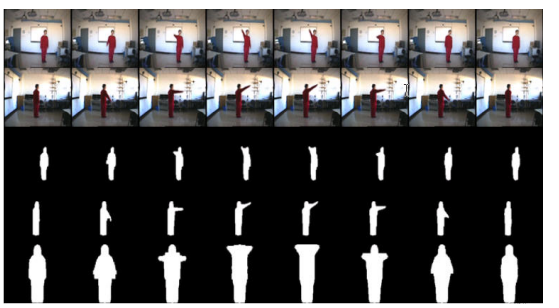

Fig.8 One group of “Point to” sequences

图 8 一组“指向”的动作序列

现在的动作识别实验系统仅仅针对已经分割好的动作段进行训练和识别,对于连续动作,采用人工分割的方法.实验参数如下:每个输入向量(也就是包容形状向量经过 PCA 降维后的输出向量)的维度是 8,每个 HMM 模型的状态数是 5,而混合观测值模型中包含的高斯模型个数是 10.通过我们的视频数据库,我们分别进行了特定人和非特定人的实验.

在特定人动作识别这种情况下,我们为每一个动作者训练他们个人的 HMM 动作模型,然后使用他自己的 HMM 模型进行动作识别.对每一个动作,我们用 6 个序列作为训练集,同时把另外 3 个序列作为测试集.表 2 显示了实验结果.前 14 列中的值是每个动作者动作中识别正确的个数.对于每个动作者下面的两列,第 1 列是训练集(共 6 遍动作)的识别结果,第 2 列是测试集(共 3 遍动作)的识别结果.最后两列的值是每个动作所有动作者的平均识别率.

Table 2 Subject dependent action recognition results

表 2 特定人的动作识别结果

	Actor 1		Actor 2		Actor 3		Actor 4		Actor 5		Actor 6		Actor 7		Average (%)	
Point to	6	3	6	3	6	3	6	3	6	3	6	3	6	3	100	100
Raise hand	6	3	6	3	6	3	6	3	6	3	6	3	6	3	100	95.2
Wave	6	3	6	3	6	3	5	3	6	3	6	3	6	3	97.6	100
Touch head	6	2	6	3	6	3	6	3	6	3	6	3	6	3	100	95.2
Communication	6	2	5	3	6	3	6	3	5	2	6	3	6	3	95.2	90.5
Bow	6	3	6	3	6	3	6	3	6	3	6	3	6	3	100	100
Pick up	6	3	6	3	6	3	6	3	6	3	6	3	6	3	100	100
Kick	6	3	6	3	6	3	6	3	6	3	6	3	6	3	100	100
Walk	6	2	6	3	6	3	6	3	6	3	6	3	6	3	100	95.2

对于非特定人动作识别这种情况,我们采用训练集中所有人的动作为每一个动作类型训练对应动作的 HMM 模型.对于每一种动作,我们用 5 个动作者的视频数据作为训练集,另外两个动作者的数据作为测试集.这样,对于每一种动作,训练集中一共有 45 个视频段,测试集中一共有 18 个视频段.表 3 列出了识别的正确率.由于在实验数据中,不同演员甚至同一个演员的同一次的动作都不完全规范、一致.因此,从实验定性的角度而言,动作识别的实验结果,尤其是非特定人的识别实验结果,表明了系统对于演员动作偏差存在较好的容忍度和鲁棒性.

为了证明包容形状用于视角无关的动作识别上的有效性,我们尝试了一些对比实验.在表 4 中给出了与利用视角相关的体态表示进行动作识别的结果比较.表的第 1 行显示了我们采用的 3 种方法.第 1 种是已有的基于包容形状的方法;在第 2 种方法中,我们采用外轮廓宽度在 Y 轴的投影向量(即公式(2)中的 x)作为输入特征,这是和视角相关的表示;在第 3 种方法中,我们参考文献[21],采用了一种同样是视角相关的运动特征作为输入.表格的每种方法下面包含两列,第 1 列是任意视角的动作识别,第 2 列则是特定视角的动作识别.从比较中我们可以看到,在任意视角的场景下,只有基于包容形状的方法才取得了很好的识别效果,而另外两种则受限于视角相关的表示,无法取得令人满意的结果.

Table 3 Subject independent action recognition results

表 3 非特定人的动作识别结果

	Train set (%)	Test set (%)
Point to	97.8	100
Raise hand	100	100
Wave	95.6	88.9
Touch head	95.6	94.4
Communication	88.9	83.3
Bow	100	100
Pick up	100	94.4
Kick	100	100
Walk	100	94.4

Table 4 Comparison with view dependent action recognition methods

表 4 与视角相关的识别方法进行比较

	Envelop shape (%)		Silhouette projection on Y (%)		Motion feature[21] (%)	
Point to	100	94.4	44.4	88.9	38.8	100
Raise hand	100	100	61.1	94.4	33.3	94.4
Wave	88.9	94.4	33.3	61.1	55.5	94.4
Touch head	94.4	88.9	38.8	83.3	22.2	100
Communication	83.3	83.3	27.7	61.1	27.7.	88.9
Bow	100	100	61.1	94.4	38.8	94.4
Pick up	94.4	100	38.8	72.2	33.3	83.3
Kick	100	94.4	55.5	83.3	38.8	88.9
Walk	94.4	94.4	33.3	72.2	44.4	100

3 包容形状的扩展

在实际应用中,由于受到应用场景的限制,比如为了兼顾较大的观察视野,有时无法保证双摄像机光轴夹角是正交的.在这种情况下,我们可以扩展包容形状的表示,以便可以在更多的场景下得到应用.除了摄像机光轴不必要求正交以外,其余的配置要求和图 2 所示一样,即要求两个摄像机的像平面都和竖直轴 Y 平行.

记双摄像机光轴的夹角是$\pi-\alpha$,由立体视觉的知识可以推导出两个像平面夹角为α,即投影轴 U1 和 U2 的夹角为α.当α为锐角时,模仿图 3 画出水平截面图如图 9 所示.

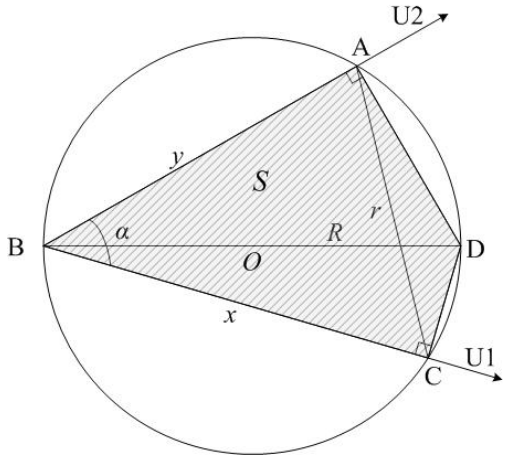

Fig.9 Section projection whose optical axis angle is $\pi-\alpha$

图 9 光轴夹角是$\pi-\alpha$时的水平截面图

对于二维截面中的形状 S,它在两个摄像机的成像平面上投影轴分别为 U1 和 U2,U1 和 U2 夹角是α.S 在 U1 和 U2 轴的投影线段分别是 AB 和 BC.记 AB 为 x,BC 为 y,AC 为 r,同时记三角形 ABC 的外接圆为圆 O,直径为 R.

当$\alpha>\pi/2$时,U1 和 U2 夹角为钝角.由于投影只考虑方向角,不计正负方向,所以我们令其中一个投影轴反向即可,同样可以按照图 9 来分析.于是我们令β为

$$\begin{cases}\beta=\alpha, & \text{当}\alpha<\pi/2\\ \beta=\pi-\alpha, & \text{当}\alpha>\pi/2\end{cases} \tag{9}$$

我们由公式(10)定义扩展的包容形状为

$$r=\sqrt{x^2+y^2-2xy\cos\beta} \tag{10}$$

由几何关系我们可以得到:

$$R\sin\beta=r \tag{11}$$

当人体偏转θ角时,相当于形状 S 旋转了角θ.记新的投影线段的长度是 x'和 y'.由于形状 S 一定在外接圆 O 的内部,所以

$$x'\le R,\ y'\le R \tag{12}$$

由式(9)~式(12)可以推导得到 r'的取值区间:

$$\begin{cases}r'\le\dfrac{r}{\sin\alpha}, & \text{当}0<\alpha<\pi/3\text{或者}2\pi/3<\alpha<\pi\\ r'\le\dfrac{r}{\cos\dfrac{\alpha}{2}}, & \text{当}\pi/3\le\alpha\le\pi/2\\ r'\le\dfrac{r}{\sin\dfrac{\alpha}{2}}, & \text{当}\pi/2<\alpha\le2\pi/3\end{cases} \tag{13}$$

这样,取r_0是所有旋转对应的各个r中的最小值.那么在任意旋转下,相应的$\frac{r}{r_0}$值都会满足如下取值区间:

$$\begin{cases} 1 \le \dfrac{r}{r_0} \le \dfrac{1}{\sin\alpha}, & \text{当} 0 < \alpha < \pi/3 \text{或} 2\pi/3 < \alpha < \pi \\ 1 \le \dfrac{r}{r_0} \le \dfrac{1}{\cos\dfrac{\alpha}{2}}, & \text{当} \pi/3 \le \alpha \le \pi/2 \\ 1 \le \dfrac{r}{r_0} \le \dfrac{1}{\sin\dfrac{\alpha}{2}}, & \text{当} \pi/2 < \alpha \le 2\pi/3 \end{cases} \tag{14}$$

$\frac{r}{r_0}$的最大取值是 α 的一个函数,记作 $f(\alpha)$,它的取值对应了包容形状的视角不变性.我们画出 $f(\alpha)$的对应函数曲线如图 10 所示(考虑到显示效果,$f(\alpha)$>5 的部分已被截除不显示).

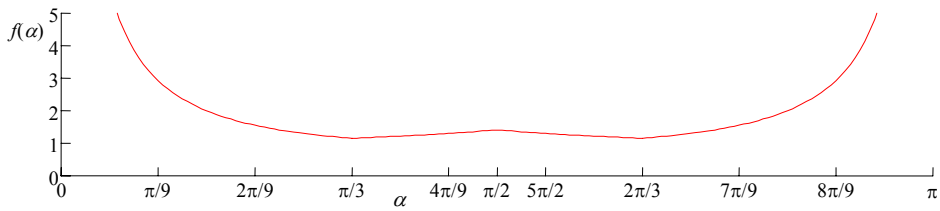

Fig.10 Viewpoint invariant of envelop shape: $f(\alpha)$ curve

图 10 包容形状的视角不变性:$f(\alpha)$函数曲线

当双摄像机光轴的夹角是$\pi-\alpha$,即两个成像平面夹角是α时,我们根据公式(10)来定义扩展的包容形状 r.当$\alpha=\pi/2$ 时,公式(10)就变成了公式(2).

$f(\alpha)$表示了包容形状在不同视角下的变化区间,$f(\alpha)$越小表示包容形状 r 的视角不变性越强.表 5 罗列了不同α角度下 $f(\alpha)$的值.当α在[π/4,3π/4]区间时,$f(\alpha)$的取值在 1.15~1.414 之间,数值不大,此时我们可以把包容形状看成视角变化不敏感的体态表示.图 11 列出了合成人体模型对应图 4(a)中第 1 种姿态在不同α时对应的包容形状图像.每一行对应一个α值,每行包括了该体态围绕着竖直轴(Y 轴)旋转了 8 个不同角度时的情况.其中每一行α的取值和表 5 中一致,分别为π/6,π/4,π/3,π/2,2π/3,3π/4,5π/6.从图中我们可以看到,在视角变化时,包容形状的变化不大.α取值π/6,π/4,3π/4 和 5π/6 时,包容形状的每一行值都比较小,这是由于两个摄像机主轴夹角较小,方向较一致,具有抵消作用,导致包含的可区分信息量减少.

Table 5 $f(\alpha)$ value at different α angle

表 5 不同α角度下$f(\alpha)$的值

α	π/6	π/4	π/3	π/2	2π/3	3π/4	5π/6
$f(\alpha)$	2	1.414	1.15	1.414	1.15	1.414	2

由函数曲线来看,当$\alpha=\pi/3$ 或$\alpha=2\pi/3$ 时,$f(\alpha)$取值最小,也就是说此时视角不变特性最好.不过,作为动作识别所用的体态表示而言,cosβ越大,根据公式(10)计算而来的 r 值会越多地由于 $2xy\cos\beta$项的抵消而减少了原始信息量,从而降低了可区分性,所以当$\alpha=\pi/2$ 时可区分性最好.在不同的角度α下,我们进行了和第 3 节方法类似的非特定人动作识别实验,表 6 给出了实验结果(其中的α采用近似测量).实验表明,当α取值在π/3~2π/3 以内,利用包容形状表示进行动作识别都可以得到令人满意的结果.而α取值在π/2 附近,则是视角不变性和可区分性两者之间一个较好的平衡点.

Table 6 Recognition results of different α (%)

表 6 不同α角度下动作识别结果(%)

α	π/4	π/3	π/2	2π/3
Point to	72.2	94.4	100	100
Raise hand	83.3	100	100	83.3
Wave	77.8	94.4	88.9	94.4
Touch head	88.9	100	94.4	94.4
Communication	77.8	83.3	83.3	88.9
Bow	94.4	100	100	100
Pick up	89.9	83.3	94.4	88.9
Kick	77.8	100	100	94.4
Walk	89.9	94.4	94.4	100

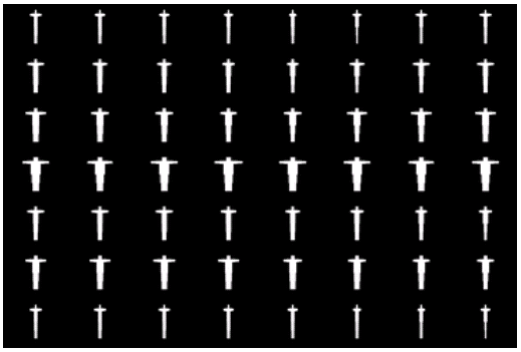

Fig.11 Envelop shape of different α

图 11 不同α角度不同视角下的包容形状

4 小 结

利用包容形状的表示,我们构建了一个视角不变的动作识别系统,通过包容形状的表示和隐马尔可夫模型的应用,我们的实验系统对于任意视角下的动作都有较高的识别正确率.实验结果表明,包容形状的表示包含了足够的可区分性信息可以用于任意视角下的非特定人动作识别.在两个摄像机非正交时,我们提出了包容形状的扩展,这极大地加强了本方法的实用价值.

包容形状的表示对视角变化不敏感.与以前的方法相比,它很容易获得而且有着更丰富的信息.这种表示不依赖于较难获取而且对于误差敏感的语义特征检测或者点对应.然而作为一种视角不敏感的表示,它还是损失了一些视角变化的信息,这些信息有时候对于某些特定的动作还是很重要的.比如,仅仅利用这一种表示,我们无法区分出是左手在运动还是右手在运动.一些视角相关的信息可能对于解决这种类型的问题有所帮助.如何把视角不变的表示和视角相关的表示相结合和利用,这是下一步需要完成的研究工作.另一方面,现在的动作识别系统针对的是已经人工分割的动作段.如何针对自然连贯动作进行自动分割然后进行识别,这也是我们下一步要重点研究的内容.

References:

[1] Cedras C, Shah M. Motion-Based recognition: A survey. Image and Vision Computing, 1995,13(2):129−155.

[2] Aggarwal JK, Cai Q. Human motion analysis: A review. Computer Vision and Image Understanding, 1999,73(3):428−440.

[3] Moeslund TB, Granum E. A survey of computer vision-based human motion capture. Computer Vision and Image Understanding, 2001,81(3):231−268.

[4] Wang L, Hu WM, Tan TN. Recent developments in human motion analysis. Pattern Recognition, 2003,36(3):585−601.

[5] Leo M, D'Orazio T, Spagnolo P. Human activity recognition for automatic visual surveillance of wide areas. In: Int'l Multimedia Conf., Proc. of the ACM 2nd Int'l Workshop on Video Surveillance & Sensor Networks. 2004. 124−130. http://portal.acm.org/citation.cfm?id=1026799.1026820

[6] Wang L, Tan TN, Ning HZ, Weiming Hu. Silhouette analysis-based gait recognition for human identification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(12):1505−1518.

[7] Duda RO, Hart PE, Stock DG. Pattern Classification. New York: John Wiley & Sons, 2001. 11.

[8] Campbell LW, Becker DA, Azarbayejani A, Bobick AF, Pentland A. Invariant features for 3D gesture recognition. In: Proc. of the Int'l Conf. on Automatic Face and Gesture Recognition. 1996. 157−162. http://doi.ieeecomputersociety.org/10.1109/AFGR.1996.557258

[9] Seitz1 SM, Dyer1 CR. View-Invariant analysis of cyclic motion. Int'l Journal of Computer Vision, 1997,25(3):231−251.

[10] Rao C, Yilmaz A, Shah M. View-Invariant representation and recognition of actions. Int'l Journal of Computer Vision, 2002,50(2):203−226.

[11] Rao C, Shah M, Mahmood TS. Action recognition based on view invariant spatio-temporal analysis. ACM Multimedia, 2003, 518−527.

[12] Parameswaran V, Chellappa R. Using 2D projective invariance for human action recognition., Int'l Journal of Computer Vision, 2006,66(1):83−10.

[13] Parameswaran V, Chellappa R. Human action recognition using mutual invariants. Computer Vision and Image Understanding, 2005,98(2):294−324.

[14] Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding, 2006,104(2):249−257.

[15] Ogale AS, Karapurkar A, Aloimonos Y. View-Invariant modeling and recognition of human actions using grammars. In: Proc. of the Int'l Conf. on Computer Vision (ICCV), Workshop on Dynamical Vision. 2005. http://www.cs.umd.edu/~karapurk/nsfhsd/media_files/ActionsPCFG.pdf

[16] Wren C, Azarbayejani A, Darrell T, Pentland A. Pfinder: Real-Time tracking of the human body. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997,19(7):780−785.

[17] Yamato J, Ohya J, Ishii K. Recognizing human action in time-sequetial images using hidden Markov model. In: Proc. of the 1992 IEEE Conf. on Computer Vision and Pattern Rec. IEEEE Press, 1992. 379−385. http://ieeexplore.ieee.org/xpls/abs_all.jsp?&arnumber=223161

[18] Brand M, Oliver N, Pentland A. Coupled hidden Markov models for complex action recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition. Puerto Rico, 1997. http://citeseer.ist.psu.edu/118270.html

[19] Ivanov YA, Bobick AF. Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8):852−872.

[20] Xie JH. Hidden Markov Model (HMM) and its Application on Speech Processing. Wuhan: Huazhong University of Science and Technology Press,1995 (in Chinese).

[21] Masoud O, Papanikolopoulos N. A method for human action recognition. Image and Vision Computing, 2003,21(8):729−743.

附中文参考文献:

[20] 谢锦辉.隐 Markov 模型(HMM)及其在语音处理中的应用.武汉:华中理工大学出版社,1995.

黄飞跃(1979－),男,江苏南通人,博士生,主要研究领域为计算机视觉,数字图像处理,模式识别.

徐光祐(1940－),男,教授,博士生导师,CCF 高级会员,主要研究领域为计算机视觉,移动机器人视觉导航,多媒体技术,自然的人机交互,普适计算.