可视媒体中的人体运动分析

赵 旭 张 鸣 刘允才
上海交通大学

关键词： 可视媒体 人体姿态重建 动作识别

前言

基于视觉媒体信息的人体运动分析是机器视觉研究领域中的一项前沿课题，其主要目的是使计算机能够从图像中检测和发现人并重建其运动行为，最终达到在语义层面上对人体运动的感知和理解。

人是人类环境中最重要的元素，各种各样的人体运动构成了可视媒体的主要内容，这些运动携带了大量有关人类社会交互的重要信息。因此，基于视觉信息的人体运动分析具有广阔的应用前景。

人体运动的视觉分析的研究任务如图1所示。对于给定的图像序列，人体运动分析算法期望得到的输出结果包括：图像中是否有人，有多少人，这些人的位置在哪里，他们的姿态怎样等等。此外，还希望据此推断出其它有用的语义信息，例如：这些人携带了什么物体，他们在做什么等等。有关的课题研究衍生的内容包括：高级人机交互、安全监控、虚拟现实、体育运动分析、医学运动分析、基于内容的图像检索、超低带宽的视频压缩、老电影中的演员更换以及老年人看护等等。在人工智能研究中，基于视觉技术的人体运动分析扮演着十分重要的角色。未来智能机器人必须具有从周围环境提取视觉信息的能力，而最重要的信息之一就是有关周围环境中人的信息，这需要人体运动的视觉分析研究作为基础。

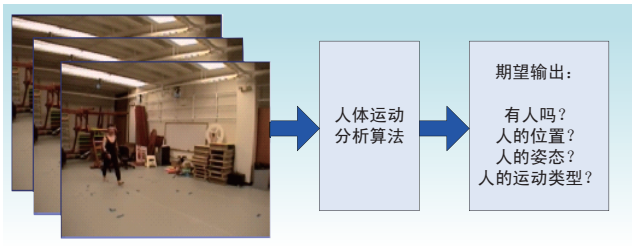

图1 人体运动视觉分析的任务

由于人体运动内在的复杂性、平面视觉信息的模糊性、噪声的干扰以及深度信息的缺失，还有图像中人体的自遮挡、互遮挡和二义性等因素，都使得研究困难重重。尽管如此，技术上的挑战和潜在的应用前景仍使这项研究在近年来十分受到关注。人体运动的研究内容不断得到丰富，其理论和方法日渐扩展，已涉及到包括计算机视觉、计算机图形学、机器学习、模式识别、人体运动学和心理学等在内的多个学科领域。

近年来，人们在人体运动分析方面开展了大量的研究工作，本文基于视觉认知层次按照从低到高的顺序对人体运动的视觉分析研究进行系统的梳理和回顾。这些层次包括：人体检测和跟踪、人体姿态重建以及人体运动、动作的语义识别。人体姿态重建将是介绍的重点。

虽然这方面已有一些综述文章[1~3]，但它们在总结人体运动重建方法时，一般都以有无先验人体模型为准则对各种重建方法进行划分。这种划分有其合理性，但是人体模型的使用只体现了具体的重建手段，并不反映人体运动重建方法的本质。为此，本文从人体运动重建的基本求解框架出发，介绍人体运动中

生成式重建、诊断式重建以及几何重建这三种不同方法。

人体检测与跟踪

人体检测与跟踪[1]是人体运动重建的基础，其目的是从图像中发现人并以特定的图像特征进行表示。这两部分研究既有区别又有交叉，检测是跟踪的基础，同时跟踪又是检测的手段。

通过检测和跟踪分割出来的人体通常用图像特征进行表示，以便进一步利用相关信息。如果目标仅为跟踪运动的人，不估计姿态参数，则人体可以用简单的矩形框、像素块等来标识；如果需要进行姿态重建，则要提取人体轮廓、边缘、轮廓线和颜色等视觉特征；如果涉及动作识别，则要用分类的特征作表示方式。下面重点介绍人体检测。

人体检测方法主要包括：基于时域信息的检测、基于空域信息的检测以及基于学习的检测。根据人体自身结构和外观等特点，检测方法还有基于模型的检测和基于皮肤颜色的检测等。

基于时域信息的检测是利用不同时刻图像间的差别进行的，主要包括光流法和减除法。

光流法是利用图像灰度在时间上的变化与场景中运动体的结构与运动的关系来检测的。达拉尔（N. Dalal）为了处理背景运动和相机运动引入了光流描述子，用支持向量机（Support Vector Machine，SVM）对像素进行分类以确定图像属于人还是背景，训练和分类时都要进行光流计算[4]。顾（H. Gu）等人则利用光流约束跟踪边缘特征[5]。利用光流法也可以把检测和姿态估计放到一个框架内解决。布雷格勒（C. Bregler）利用逆运动学和人体肢体像素的光流约束方程组，求解得到以旋转平移表示的三维运动参数[6]。光流法检测缺点是计算复杂、鲁棒性差，因此在使用时结合其它方法会有更好的效果，例如文献[4]同时使用了外观特征（形状、强度和颜色等）。

减除法是通过从当前图像中逐像素地减除掉背景图像，来检测人体的方法。减除的信息可以是灰度值或梯度值。在减除法中，一般需要自适应地估计出背景，然后再执行减除。最简单的做法是先把不同光照条件下的静态背景拍下来，然后以平均后的图像作为减除对象来检测前景。有很多改进的减除方法，例如文献[7]提出可以先通过一定图像帧数的训练建立背景模型，然后实时周期性地更新背景模型，以适应光照变化和场景变化。文献[8]则采用对每个像素以高斯混合模型建模，然后迭代更新这个模型。背景减除法的应用比较广泛，但在处理复杂动态场景时有较大局限性。

基于空域信息的检测是依据图像像素间灰度与颜色的差别进行的。如果图像主体的灰度或颜色与背景不同，则通过设置合适的阈值可以检测出前景主体。常用的利用空域信息检测人体的方法是统计建模法。其基本思想是根据训练图像对每个像素或表征特定对象的像素群进行统计建模，据此将当前图像中的像素分为前景或背景，通过这种统计建模松弛了减除法中的前景假定。例如，雷恩（C. R. Wren）等人[9]的方法是先对由聚类像素块表示的人肢体建立统计模型，然后将当前图像的每个像素根据其颜色、空间特性以概率的方式分配到不同的块中或背景中。

基于学习的检测是在样本集的基础上依靠有效的学习算法得到分割人与背景的决策边界。这类方法包括样本集选取、分类机制和描述子等几个关键点。描述子一般包括外观描述子和运动描述子，是区分人与背景的主要依据。文献[10]用径向基函数为核函数的支持向

[1] 人体跟踪主要关注图像序列中人体在图像帧间的对应信息。人体检测在文中有较详细的介绍。

量机作为分类器，以光流为描述子，得到判断人与背景的分类函数的相关参数，然后通过位置、尺度可变的探测窗口在图像上滑动检测人体；达拉尔等人[11]用线性支持向量机分类器和有向梯度直方图描述子的方法检测直立行走的人体。最近几年基于学习的方法成为研究热点，通过纳入学习框架，人体检测被转化为一个分类问题。

人体姿态重建

如果以高维姿态参数空间内的点表征人体姿态，则姿态重建就是要找到图像中人体在姿态参数空间内的对应点。姿态重建使图像中人体的姿态得到描述和还原，其结果既可作为运动和行为识别的基础，也可直接面向运动分析和虚拟现实等具体应用。

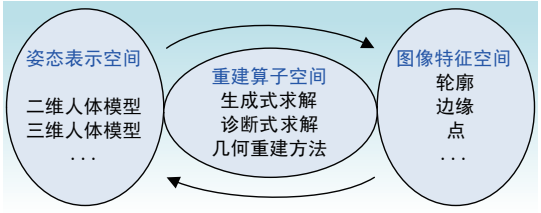

图2 人体姿态重建的三个部分

人体姿态重建可以分解为姿态表示空间、图像特征空间和重建算子空间三个部分（图2）。重建算子空间集合了姿态重建的各种方法，是人体姿态重建的关键部分，其中的每种方法都提供了从图像特征到姿态参数的某种映射机制；姿态表示空间是各种人体姿态描述方法的集合，例如三维人体模型就是最常见的姿态表示载体；图像空间则包括了表达图像人体的各种图像特征，比如轮廓、边缘和关节点等，人体检测和跟踪的结果都可作为其中的元素。

姿态表示空间的划分是人体姿态重建方法重要的分类依据。莫斯伦德（T. B. Moeslund）[1]据此把姿态估计问题总结为：无模型方法、非直接模型使用方法以及基于模型直接使用方法。此外，阿加尔沃（J. K. Aggarwal）[2]和加瓦里拉（D. M. Gavrila）[3]也使用了基于模型、不基于模型的分类方法。但对于姿态重建来说，核心是具体的重建框架，因此本文将从重建算子空间的分解出发，对人体姿态重建进行总结。如果把要重建的人体姿态看作原因，把图像看作结果，则目前见到的大多数人体姿态重建方法可以纳入两种框架：一种是从原因到结果，即生成式（Generative）求解；另一种是从结果直接到原因，即诊断式（Discriminative）求解。此外，还有建立在传统机器视觉方法基础上的几何重建算法。

人体姿态的生成式重建

生成式重建是自顶向下的求解方法（典型的图模型见图3）。其实质是在状态先验和动态模型的指导下，生成预测的姿态参数，再根据图像信息修正预测，是一个预测—匹配—更新的过程，也称为AbS（Analysis-by-Synthesis，综合分析）方法[1]，包括状态预测和匹配更新两个重要方面。

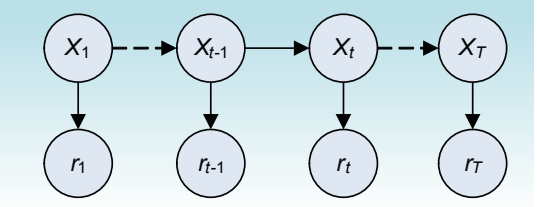

图3 生成式重建图模型表示

图3中的x_{t-1}是姿态参数时间序列，r_{t-1}是图像观测序列，生成式重建根据状态先验$p(x)$以及时域动态模型 $\boldsymbol{p}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots)$ 生成本时刻的预测状态，再根据 $\boldsymbol{p}(\boldsymbol{x} \mid \boldsymbol{r}) \propto \boldsymbol{p}(\boldsymbol{x}) \cdot \boldsymbol{p}(\boldsymbol{r} \mid \boldsymbol{x})$ 通过匹配更新修正预测以得到最佳的后验状态。

状态预测的目的是削减状态空间。这种预测建立在人体姿态表示方式的基础上。最常见的描述方式是显式的人体模型。它直观地呈现了人体姿态，同时也丰富了姿态重建的方法。