莫斯科国立罗蒙诺索夫大学

关于手稿的权利

UDC 004.932.4

叶夫根尼-瓦迪莫维奇-沙尔诺夫

研究和开发在视频序列中为人及其身体部位伴奏的方法

特产 05.13.11 -

"计算机、综合体和计算机网络的数学和软件"

物理学和数学的博士论文

导师：物理和数学博士

，副教授

Anton Sergeyevich Konushin

莫斯科 - 2018年

目录

页。

简介

在当今世界，视频监控系统正在成为城市和企业基础设施的一个重要组成部分。视频监控系统被定义为一套软件和硬件工具，用于获取和分析视频以帮助人类决策。目前，在大多数情况下，视频监控系统被用来捕捉视频事件，以便日后由人类操作员进行分析，例如，在发生紧急情况后。操作员需要回答的关键问题是 "谁出现在视频中？"和 "发生了什么事件？

目前计算机视觉算法的发展水平使得这些问题可以在一些重要的实际场景中实现自动化。车辆检测和车牌识别方面的进展使人们能够开发出一个自动捕捉交通违规行为的系统，迫使司机跟随他们。在过去的几年里，高效的算法已经被开发出来，用于视频面部识别和面部识别。这些算法已被用于创建带有面部识别的门禁系统，在检查站或申请贷款时使用生物识别护照进行身份验证的自动化，等等。

然而，视频监控的潜力要大得多。 识别视频中脸部被隐藏或图像分辨率低的人的问题仍有待解决。在图片中

0.1介绍了拍摄的非法活动的例子。尽管有可能利用所获得的数据重建事件的时间顺序，但在许多情况下，识别录像中的人需要手工劳动，因为事件的参与者可能会隐藏他们的脸。在这一点上，重要的是

图0.1 - 非法活动的闭路电视录像的例子。第一行是对钢琴和汽车的纵火。第二行是闯入商店和盗窃自行车。

发展的主要重点是根据人的面部表情和行为，特别是他们的步态来识别人。一个人在摄像头或多摄像头系统视野中的轨迹信息对识别一个人也很重要。这可能使我们有可能确定感兴趣的人来自哪里，去了哪里，或者找到他或她的脸还没有被掩盖或遮盖的时间点。例如，在许多情况下，由于身材或衣服颜色的相似性，很难在人群中分辨出护送的目标。如果被追捕的人故意要把目标人物甩开，这项任务就变得更加困难。在这方面，有必要为所有在场的人使用护送者。

视频序列。即使不能在人群中识别出感兴趣的人，这种方法也能确定附近或与感兴趣的人相似的所有人员的轨迹，这大大降低了搜索活动的复杂性。

跟踪视频序列中所有人员的任务还有其他应用。它可以通过分析人员和车辆的数量和路线来简化城市规划。例如，根据行业准则ODM 218.6.003-2011和GOST 52289-2004，根据交通和行人流量调查的结果决定是否需要设计交通灯设施。这些文件规定了建议使用交通灯调节的流量密度。因此，使用自动人车统计工具可以及时跟踪交通流的变化，并做出城市规划决策。

然而，现代算法在多人伴奏的质量上明显不如人类[1]。在这方面，它们在解决实际任务方面的应用是非常有限的。另一个重要的限制是许多视频分析算法的高计算复杂性，这不允许它们在目前的技术水平上实际应用。视频摄像机的广泛使用和计算机网络的发展使我们能够创建拥有超过数十万台摄像机的视频监控系统。然而，即使是初级分析算法，如检测感兴趣的物体（人、车等），也不允许在中央处理器上处理超过几个视频流，或为昂贵的图形加速器设计。

解决CCTV数据处理结果计算复杂度高、质量差的问题的一个可能办法是

[1] 最好的现代伴奏算法的结果可以在以下网站找到
MOTChallenge https://motchallenge.net/

是使用关于所用相机的位置和属性的信息，即其校准参数。这些信息限制了帧中感兴趣的物体的可能位置，这既可以用来减少分析的图像区域的数量，也可以用来检测检测算法中的假阳性。不幸的是，现有的获取相机信息的算法要么需要与用户和校准模式互动，要么只能应用于小范围的可能的相机位置，限制了它们的适用性。

视频监控系统的发展需要开发在准确性和质量方面优于现有算法的分析算法。在我的工作中，我考虑了一个基本的视频监控场景，涉及到一个单一的静态摄像机。在这种情况下，[1]中描述的视频监控数据分析的标准方法是解决以下子问题：

1. 相机校准（在世界坐标系和图像坐标系之间建立一个映射）；
2. 检测和跟踪视频中感兴趣的对象（如人）；
3. 行为分析（包括通过行为自动识别类型和识别异常行为）。

本文的**目的**是制定方法来提高质量
通过使用关于摄像机校准和场景中人的运动的信息，在用静态摄像机拍摄的视频序列中对人进行定位、跟踪和姿势检测。

为了实现这一目标，必须解决以下问题

任务：

1. 开发并实现一种算法，根据人类检测的结果确定摄像机在场景中的位置和方向，允许倾斜角度从0到π。

2. 设计并实现一种算法，利用摄像机校准信息和场景入口区域跟踪视频序列中的每个人，以提高轨迹的准确性。
3. 开发并实现一种以去序列化的方式确定人类姿势的算法，该算法基于身体关节的位置和速度的联合模型，与以前的方法相比，允许提高解决问题的准确性。
4. 在提出的算法的基础上，开发一个软件工具，用于在视频序列上构建人及其肢体的轨迹，使我们能够解决这个问题，并允许使用不同的算法来定位人和视觉跟踪，替换个别模块。

提出了辩护的主要内容：

1. 提出了一种基于人的检测确定静态摄像机在场景中的位置和方向的原创方法，该方法仅基于合成视频监控数据的显示训练。
2. 对于静态摄像机拍摄的视频序列，已经开发了一种人类追踪算法，利用摄像机的位置和方向来过滤掉检测器的假阳性。
3. 提出了一种在视频序列中估计人类姿态的算法，该算法同时考虑了视频序列帧中人体每个关节的位置和速度。
4. 在所提出的算法的基础上，开发了一个用于自动跟踪和确定视频序列中人类姿势的软件包和一个用于在每一帧中构建人类姿势的专家标记的自动化软件工具。

科学的新颖性：

1. 首次提出了一种基于视频录像中人的检测来确定静态摄像机在场景中的位置和方向的算法，该算法基于机器学习，可以只对合成数据进行调整。研究表明，与真实CCTV数据分析中的类似方法相比，所提出的算法的准确性不会随着摄像机倾斜角度从0到90度的增加而降低。
2. 首次提出了一种基于机器学习并可在合成数据上进行调整的算法，将静态摄像机图像中的人的检测分为可信的和不可接受的。研究表明，所提出的算法的应用提高了图像中人的检测速度和平均精度。
3. 首次提出了一个人体骨骼模型，将视频序列中每个人体关节的位置和运动同时描述为一个线性动态系统。研究表明，以前存在的模型是所提模型的特例。在该模型的基础上，提出了一种新的算法，通过搜索目标函数的局部最优来确定视频中每一帧的人体骨架（姿势）。与基于先前模型的算法相比，所提出的算法显示出更高的姿势检测精度。

实用性 录像发展的领域之一

数据处理的第一阶段（特别是物体检测）是利用摄像机本身的资源进行的。鉴于相机上可用的计算资源有限，本文提出的用于自动校准相机和检测场景中人物的算法具有很大的实际意义。这些算法扩展了各种基本的物体检测算法，能够处理

在给定的运行时间约束下的图像，即允许使用更先进的检测器，这通常需要更多的计算资源。

如果方向和焦距的变化被执行器量化，所提出的利用校准信息检测人的算法也可以应用于PTZ摄像机。

所提出的用于视频中人的姿势检测的算法允许构建一个与部分专家标记相对应的解决方案。基于这个想法，我们创建了一个软件工具来构建视频中人的姿势的参考样本，包括两个重复的步骤：

– 应用一种算法来寻找视频中的最佳人体姿态，其中包括部分专家的标记；
– 部分专家标记的扩展，以纠正当前解决方案中的错误。

所提出的工具的价值在于大大减少了对视频序列进行标记的手工工作。这种标记的数据是出现新的、改进的视频中人类姿势估计算法的关键因素。

所提出的算法已经以软件工具（SW）的形式实现。所开发的用于在视频序列中构建人及其肢体轨迹的软件具有模块化结构，其中每个模块都解决了分析输入数据的个别任务。模块的替换提供了利用新算法提高解决既定任务的质量的可能性。

该工作已被批准。该工作的主要成果在以下会议上作了介绍：

– 该研讨会在M.R. Shura-Bura研讨会上举行。M.R. Shura-Bura在M.M. Gorbunov-Posadov的指导下；
– 在R.L. Smeliansky的指导下，莫斯科国立大学高等数学和控制论系的ASVC和SCI的研究生研讨会；

– CMC MSU-Huawei国际研讨会 "多媒体图像处理和分析中的选定主题"，俄罗斯，莫斯科，2016年8月31日；
– 第五届图像分析国际研讨会（第五届图像挖掘国际研讨会。理论与应用），德国柏林，2015；
– 第11届国际会议模式识别与分析图像俄罗斯，萨马拉，2013年；
– 第26届计算机图形、图像处理和机器视觉、可视化和虚拟环境国际会议GraphiCon 2016，俄罗斯下诺夫哥罗德，2016年9月19-23日；
– 第25届国际计算机图形、教育会议GraphiCon 2015， GraphiCon 2015， GraphiCon 2015， 俄罗斯Protvino，2015年9月22-25日；
– 第24届国际计算机图形学、教育会议GraphiCon 2014是世界领先的成像和机器视觉、可视化和虚拟环境系统，俄罗斯顿河畔罗斯托夫，2014年9月30日至10月3日；
– 微软博士生暑期学校（微软研究院博士生暑期学校）。学校），英国剑桥，2014年。

个人贡献。作者的个人贡献包括论文工作中所描述的大部分理论和实验研究，包括理论模型的开发、方法论以及算法的开发和实施，分析和设计结果的出版物和科学报告的形式。

在已发表的作品中，A.S. Konushin定义了该问题并讨论了其解决方案的结果。A.S. Konushin的贡献是

构建一个可视化方法的概述，并讨论结果。

出版物。该论文的主要成果如下

在5个出版物中，其中4个是在高级鉴定委员会推荐的期刊上发表的。

工作的范围和结构。本论文由导言、六个

章节和结论。该论文共115页，包括22个图和7个表。参考文献列表包含65个参考文献。

第一章：文献回顾

1.1 确定摄像机的位置和方向

在计算机视觉中，确定摄像机在场景中的位置和方向的问题，也称为摄像机姿态，已经被研究了很长时间[2-9]。其解决方案是构建世界坐标系到与摄像机相关的坐标系的映射的方法。构建这种映射的输入数据是由摄像机获得的帧。

在[8]中提出了一种在PTZ摄像机移动时提取其信息的方法。作者在摄像机旋转和变焦过程中使用了对关键点的跟踪。这允许估计摄像机的焦点位置和世界坐标系轴的方向。同时，大量的CCTV摄像机是静态的，即不随时间改变它们在场景中的位置和方向。

在我的工作中，我考虑的是静止摄像机的情况。对于它，我们可以区分两种方法来解决摄像机的姿势估计问题。第一种方法的算法分析了场景图像中直线方向分布的特殊性，以重建世界坐标系的方向。归属于第二种方法的方法使用观察到的已知物体的大小分布，如人或车，在图像的不同部分。

1.1.1 分析图像中线条方向的分布情况

第一种方法假设场景代表的是所谓的 "曼哈顿世界"。这个定义描述了由以下方面创建的场景

在人类的作品中，有三个正交方向的直线占主导地位：两条水平线和一条垂直线。通常建议选择这些线的方向作为作品中世界坐标系的轴。透视变换导致的事实是，这些线的图像在三个相应的消失点相交。因此，地平线的消失点位于地平线上，而垂直线的交点形成了天顶或天底。在 "曼哈顿世界 "的假设下，最大数量的线条相交于所述的三个消失点。第一种方法的工作目的是在图像中定位这些消失点。为了简洁起见，以下我将把对应于场景中正交线的消失点称为正交点。在[2]中，提出了三个正交消失点的位置和相机的焦距之间的关系。它作为以下摄像机姿势检测算法的基础。在[3]中，提出了从建筑物等物体的图像中提取正交直线。然而，所提出的方法不能应用于缺少这种结构或不能找到所有必要的消失点的场景。因此，在[9]中，提出利用高速公路上汽车的明显方向来提取图像中的水平线。作者[4；5]利用人们的运动方向和他们图像的方向来寻找地平线和垂直消失点。在这种方法中，场景中的人是由垂直线描述的。当拍摄方向与水平方向不同时，这个模型的准确性会下降。因此，在我的工作中，我没有使用人的图像的方向信息来估计摄像机的姿势。

1.1.2 分析图像中物体的大小

第二种方法的算法是分析场景图像中已知物体的大小分布。这些方法的经典假设是，有一个单一的地平面，所有的物体都位于这个平面上。这种方法最有名的算法是在[10]中提出的。作者建立了一个概率图形模型，描述了摄像机位置与场景中人和车的大小之间的关系。在[7]中，构建了一个取决于图像中心的人的大小和地平线位置的摄像机姿势的函数。提出的算法有一些明显的局限性。作者假设限制人类图像的矩形边界与场景中人的上下两点的位置重合。当摄像机的拍摄方向接近水平时，这一条件得到满足，而当摄像机的方向垂直向下时，这一条件就完全错误了。所描述的工作没有考虑到错误物体检测的可能性及其对摄像机姿态估计结果的影响。此外，作者还不得不受限于没有摄像机滚动的情况下，建立物体大小与摄像机位置映射的分析公式。

在我的论文中，我提出了一种基于以下因素评估相机姿态的算法

该算法评估了场景中人们头部的大小。与之前的方法不同，该算法在0,π 的范围内估计摄像机的倾斜度，并估计摄像机的滚动。

在[- ,ππ]范围内。在我的工作中，我允许出现假阳性现象
所提出的方法是通过估计预处理中的误差来适应它们。

摄像机的定位是基于观察到的数据。

1.2 对象的定位

在图像中构建一个物体检测器的任务一直是计算机视觉研究者感兴趣的。现代物体检测算法是根据滑动窗口原理操作的，它将处理周期分成两个阶段：1）为物体在图像中的位置构建一组假设，称为窗口；2）对窗口内的图像进行分类。通常情况下，开发的算法需要尽可能快的处理速度和尽可能少的假阳性数量。这些约束是相互矛盾的。提高检测的准确性往往也会导致计算的复杂性增加。数据处理的速度是实际视频监控应用的一个关键参数。因此，大量的研究致力于在保持质量的同时降低物体检测的计算复杂性。这一领域的工作有两个主要方向：建立快速窗口分类器和减少窗口的数量。

1.2.1 构建快速分类器

从历史上看，第一个关于加速检测的工作集中在应用分类器的加速上。[11]的作者提出使用简单分类器的级联来检测图 像 中的人脸。级联的第一阶段抛弃了大量不包含人脸的 "简单 "分类窗口，减少了总的分类时间。提出的想法非常有效，以至于级联

检测器甚至已经在数码相机中使用。这种方法的一个重要缺点是不可能改变已经构建的分类器的准确性/完整性比率。论文[12]通过改变级联的结构克服了这个限制。作者提出了一个所谓的 "软 "级联，其中级联的简单分类器的构建阶段和分离正负例子的边界的选择是分开的。这使得调整所获得的级联以满足精度要求成为可能，而无需重新训练分类器。在[13]中，他们通过仅在图像的稀疏金字塔上计算符号来实现分类器的加速。在中间层，作者提出用插值法重建特征。

在我的工作中，我提出了一种降低检测器计算复杂性的算法，这种算法与所使用的窗口分类器的类型无关，因此它可以与快速分类器一起使用。

1.2.2 减少窗口的数量

加快物体检测的另一个趋势是减少考虑的窗口数量。14]的作者使用相邻窗口中分类器响应的相关性来选择图像中可能存在物体的区域。为此，在处理的第一阶段，只对一组稀疏的窗口进行分类。之后，只对获得正面分类器响应的区域进行详细分析。由于神经网络算法在图像分类方面取得了相当大的成功[15-18]，卷积神经网络也被用于图像中物体检测的任务。通常情况下，神经网络分类器需要大量的

计算资源。因此在[19]中提出，只对图像中选择的一小部分窗口应用神经网络分类器。在[20；21]中，我们发展了以前的想法，提出将窗口的神经网络分类器分成选择感兴趣的窗口区域和细化物体位置两个阶段。这增加了窗口的大小，减少了它们的数量。

我在论文中提出的算法可以与任何一种提议的方法相结合，以减少处理窗口的数量。它对感兴趣的区域的位置有一个先验的估计，也就是说，它在应用检测器之前限制了处理的图像区域。

1.3 设施支持

物体跟踪包括在视频序列中绘制其运动轨迹。在标准的表述中，运动是在图像坐标系中考虑的，而不是在观察的场景中。这个问题有两种方法：视觉跟踪和通过检测跟踪。第一种方法用于生成视频中的物体轨迹，其中被追踪物体的类型是未知的。这种方法只用于那些可以被检测器定位的物体类别。

1.3.1 视觉支持

当物体的类型不为人知时，视觉跟踪被用来定位物体。与这种方法有关的算法同样是

适用于构建人脸或轮胎图像的运动路径，也适用于跟踪视频中更抽象的图像区域的运动。因此，为了建立一个物体的内部表示（模板），这种算法需要在第一帧中指定它的位置。在随后的帧中，将搜索与指定区域最相似的图像区域（按某种标准）。

视觉跟踪算法在所使用的物体表征和确定物体在后续帧中的位置的方式上有所不同。对后续帧进行搜索的最简单方法之一是模式的交叉相关[22]。在这种情况下，物体的代表是它在第一帧上的图像。这种方法实现了高速度，但它对被跟踪物体的变化是不稳定的。特别是，由于不同身体部位的相对运动，它不适合构建一个行走者的轨迹。因此，基于交叉相关模式的物体跟踪被用于跟踪人体的各个部分，或者用于处理物体图像没有发生重大变化的短视频片段。

被称为粒子过滤器的跟踪算法被广泛使用[23]。它的特点是搜索物体在下一帧的可能位置的方法。与之前的方法相比，该算法[23]不是通过单点来描述物体在画面上的位置，而是通过定义在二维空间上的随机变量来描述。分布密度与算法的置信度相对应，物体的预测位置是这个随机变量的数学期望值。该算法将分布的离散近似值构建为一组加权的 "粒子"。使用位置分布而不是它的模式，就有可能找到被跟踪的物体，即使在之前的几个帧中，物体的位置是错误的。

对于高质量的视觉物体追踪来说， 建 立 一个描述被追踪物体特征的表示是很重要的。其中一个使用的表示类型是一组物体图像的局部特征--关键点。在这种情况下，对后续帧的搜索不是针对整个物体，而是只针对这些点。在[24]中显示，选择物体纹理的角落作为局部特征可以在跟踪过程中实现对物体损失的高稳定性。这些结果被工作[25]所证实，它提出了使用几个这样的点以及在视频中一起追踪它们的程序。

视觉跟踪算法的主要缺点是需要对物体位置进行良好的初始化。在视频序列中跟踪人的任务中，专门的检测算法被用来解决这个问题，但这导致了错误的检测轨迹。另外，视觉跟踪方法没有考虑到视频序列中存在多个伴随类的物体。这导致了两种类型的错误：将跟踪切换到另一个类似的物体，以及为一个物体建立多个轨迹。

1.3.2 通过检测提供支持

通过检测进行跟踪的方法被用来建立只针对预定的一类物体的轨迹，例如汽车或人。在这种情况下，数据处理被分为检测视频关键帧中的物体和将检测集合分为对应于不同物体观测的组等阶段。

这种方法中算法的质量取决于用于检测图像中物体的算法以及将检测结果组合成轨迹的方式。错误检测器触发的增加会增加建立不对应于感兴趣物体的假阳性轨迹的概率。在[26]中，假设这种轨迹对应于背景图像区域的常规探测器触发，因此是静止的。为了对付假阳性轨迹，作者使用了图像中物体运动的统计数据。如果一个物体在任何关键帧中都没有被检测到，在所考虑的方法中就不能为其绘制轨迹。为了提高对人的检测的完整性，包括那些图像部分重叠的人，在某些情况下，人的检测器被身体部位检测器所取代[27]。

轨迹构建的问题通常被简化为将一组探测结果分成若干组（轨迹）的离散问题。在[26-29]中，使用动态贝叶斯网络对轨迹进行建模并找到最大后验概率。这种网络的马尔可夫特性使人们可以通过物体在前一个时间点的状态来确定其状态。在作品[30；31]中，作者将构建轨迹的问题简化为搜索图中最小成本流的问题。

追踪的另一种方法是构建与检测到的物体位置一致的连续曲线（轨迹）。在[32；33]中，有人提出用样条来模拟人类的轨迹。作者将任务划分为将探测器的探测结果与物体联系起来和构建物体运动轨迹两个阶段。轨迹的构建包括在一般函数最小化的框架内循环地重复这些阶段。

大多数轨迹构建方法可以用最小化函数问题来表示，其关键因素决定了一组检测结果对应于一个物体的可能性有多大。这个问题

因素描述了一组内检测的相似性，并以解决验证问题为基础。在护送人员时，在许多情况下，不可能应用方法从面部图像中验证一个人，因为它可能是隐藏的或分辨率低。因此，人们在画面中的位置和他们的速度信息被广泛使用。视觉跟踪被用于通过观察算法来估计物体在被检测到的框架附近的速度[26；27]。这一信息允许对两个检测结果进行比较，即使它们对应的是原始序列中遥远的（时间上）帧。作者[33]提议考虑轨迹的曲率和与视频序列中其他轨迹的距离作为正则化。34]的作者利用关于场景的语义信息预测视频中人的运动方向。在[29；30]中，人们在群体中的运动被考虑到了。作者估计了整个群体和其中每个人的运动轨迹。在[35；36]中，建议在一个单一的优化问题的框架内，将一帧物体的检测和它们在视频序列中的伴奏结合起来。

本文提出的方法是指通过检测护送的方法。本文提出在检测人时使用摄像机校准信息，在构建轨迹时考虑到场景的输入/输出区域。

1.4 确定一个人的姿态

有几种方法可以在图像中定义一个人的姿势。在第一批作品中，姿势被定义为人体许多部位的位置，如小腿、前臂、躯干、头部等。每个身体部位的位置都由一个矩形来描述，矩形的两边不能有空格。

是与坐标轴平行的。在这种表述中，问题的解决变成了困难，因为图像中身体部位的尺寸会因人的姿势不同而有很大变化。因此，在[37]中，作者将姿态定义为图像中连接人体部位的关节的位置。目前，这种姿态被用来确定姿态。

1.4.1 识别图像中人的姿势

确定一个人的姿势的任务是找到人体图像中一组固定的点的位置。这些点中的一些对应于人体的关节，并形成一个虚拟骨架。姿势估计与检测任务不同，它预测的是一个结构化的输出，即图像中不同关节的位置是相互依赖的。这种依赖性是由人体的物理尺寸决定的。确定图像中人的姿势有两种基本方法：使用一组可变形部件的模型和关节位置的回归。

由一组可变形的部件组成的模型

第一种方法结合了定位单个骨骼关节和在整体最小化问题中选择最可能的人的姿势配置这两个阶段。这种方法的显著特点是可以估计任何姿势的可信度。

在有关的图像中。换句话说，一组脱胎换骨的部件的模型定义了图像中人类姿势的概率分布。这是我在工作中广泛使用的一个特点。

一组可变形部件的模型是标准的定位方法在由几个部件组成的物体上的延伸。这种方法的特点是可以考虑到物体各部分相对位置的可能变化。这种方法首先被应用于检测场景中的物体[38]，但后来它被应用于确定图像中人的姿势[37]。

该对象由马尔可夫网络建模，其顶点对应于所寻求的关节，而边则设定了对其相互安排的约束。在[37；39]中，作者限制自己只考虑边上给定的成对效力。为了使图形模型的推理准确有效，作者限制自己只考虑树状模型。

一般来说，一组部件的模型将一个人的姿势定义为一个mi

由两类电位（因素）描述的能量函数：

$$E(\square) = \sum_{i \in V} \phi_i(p_i, \square) + \sum_{(\square, \square) \in E} \psi_{(\square, \square)}(\square, \square, \square) \text{、} \qquad (1.1)$$

其中，j_i 是联合$\square_i$ 的单项势，ψ^s $_{(\square,}$ 设定联合的配对势

的p_i 和p_j 在图像中，而s是一个人类大小的离散参数。在[37；39]中，假定成对势 ψ_i ，与输入图像无关。通过使用全局潜势参数

s人体的大小，避免了在找到的姿势中某些部位比其他部位不成比例地大的情况。

在[37]中提出的可变形部件集模型的缺点之一是关节之间的依赖关系呈树状结构。例如，膝关节的位置没有直接关系，而是通过

躯干关节的位置。这导致该算法能够在一条腿的图像中定位一个人的两条腿的关节。为了解决这个问题，[40]提出用一组*poslet*（英文版本为*poselet*）来扩展人体模型，这些*poslet*约束了人体关节的一个子集的相对位置。尽管由此产生的图形模型不再是一棵树，但推理算法仍然是有效的，因为它允许在一个小的poslet状态集上列举。

基本模型[37]假设人体的所有关节都在图像中可见。在人体在图像中部分可见的情况下，由于重叠和部分重合，这就成为一个严重的问题。在[41；42]中，作者确定一个关节是否是另一个人的重叠图像。

在[43]中，指出了对人的外表进行建模以更准确地估计其姿势的重要性。作者用图像颜色的直方图扩展了人的姿势模型，并提出了一种联合参数估计的方法，这提高了姿势估计问题的准确性。

在[37；39-41]中，作者使用启发式特征来描述图像。图形模型的参数是用结构参考矢量方法[44]训练的。在[45]中，作者表明，模型中来自一组可变形部件的输出可以表示为前向传播卷积神经网络。这使得随后的工作[42；46；47]能够与图形模型的参数一起训练图像特征。

在[48]中，作者表明，来自一组可变形部件的模型不仅可以找到一个最小化$E(□)$的姿势，而且还可以搜索其他位置不同的函数的最小值。

至少有一个关节与之前的关节相同。这种最小值的集合描述了图像中最佳人体姿势的假设集合。当所需的假设数量远远小于可能的假设数量时，就会出现以下情况

构建一组这样的假说的计算复杂性不超过构建人类姿势的最佳假说的两倍。

在我的论文工作中，我使用了一种算法，根据一组可变形部件的模型来确定图像中的姿势。构建图像中人的姿势的假设和扩展最小化函数的能力使我能够应用它来确定视频序列中人的姿势。

关节位置的回归

确定图像中人的姿势的另一种方法是来自图像的关节位置回归法。与可变形部件集模型不同，它不允许估计图像中任意姿势的质量，但可以隐含地考虑关节组的位置。

在[49]中，作者构建了一个输入图像到每个关节坐标的映射。他们提出的算法是两个神经网络的级联，依次细化图像中每个关节的位置。作者表明，第一个调节器表示整个身体的关节的大致位置，而第二个调节器则指定单个关节的位置。在[50]中，作者指出，用同样的方法预测关节位置的热图会产生更好的结果。

在[51]中，作者扩展了以前的方法。为了明确每个关节的定位，他们建议也使用其他关节位置的热图。这种方法最接近于集合模型

这种方法最难用的情况是，同一图像中有几个人，他们的图像相互重叠。这种方法最难用的情况是当同一图像中有几个人，彼此的图像重叠。

1.4.2 识别视频序列中一个人的姿势

一组可变形部件的模型允许扩展到一连串帧的情况。为了做到这一点，我们引入了一个运动模型来描述帧之间的姿势变化。最简单的变体是为每个关节指定独立的运动模型：

$$\square(\{\square\}_{\square=}) = \sum_{\square=}^{T} E(\square) + \sum_{i\in V} \sum_{\square=}^{\square-1} \Psi^{s}(p^{t+1}, p^{t}, s)^{t} \qquad (1.2)$$

其中$E(\square)$是图像中人类姿势的模型（1.1）。

在[48]中，提出了一个简单的姿态变化模型，假设帧之间的变化很弱。帧之间的关节位置变化的二次函数被选为成对的潜力。

寻找函数（1.2）的最小值被证明是一项困难的任务，因为相应的图形模型包含循环。由于其高计算复杂性，精确推断被证明是不可能的。因此，作者[48]使用构建图像中人类姿势的最佳假设来减少可接受姿势的数量，并减少确定马尔可夫链中最佳状态的问题。

在这篇论文中，我提出了一个考虑到人体骨架各关节速度的最小化函数的一般化。我还提出了一种在关节位置集和速度参数上寻找其局部最小值的算法。

第2章：确定摄像机的姿势

在这一章中，我提出了一种基于对被观察场景中感兴趣的物体的图像大小的分析来确定摄像机姿势的方法。建议的方法有两个关键的优点：

1. 它允许对感兴趣的物体进行错误检测，并考虑到存在的物体的定位误差；
2. 它预测摄像机在范围内的位置和方向。
   的区域，其倾角（0、 $\frac{\pi}{2}$）

摄像机的姿势是指其相对于被拍摄场景的位置和方向。因此，摄像机姿势检测的问题是找到世界坐标系到与摄像机相关的坐标系的转换。为了使问题陈述正规化，在下一节中，我提出了一个观察数据和使用摄像机的数学模型。

2.1 观察数据的数学模型

在本文中，我提出了一个由平面静态场景和其中的人组成的观察数据的模型。这个近似值适用于描述大多数监控场景。下面我们提出对其三个组成部分的描述：场景、摄像机和人。

2.1.1 模型现场

在这项工作中，我指的是由相机成像的静止物体。因此，一个场景可以由道路、建筑、树木、长椅等组成。然而，所提出的方法并不使用关于场景物体的语义信息，所以在这项工作中，我考虑了一个简单的场景模型，它由一个单一的水平平面--地平面组成。

为了设置摄像机的姿势，我们需要选择一个与场景相关的NAT世界坐标系。让x_w，y_w 和 z_w 表示其基向量。选择世界坐标系是为了使地平面与平面z=0重合，而向量z_w ，与场景中的向上方向重合。我选择了马球的投影作为世界坐标系的原点。

摄像机与地平面的距离。在本文中，我假设矢量

x_w 与摄像机方向矢量对地平面的投影相吻合。所描述的约束条件毫不含糊地定义了观察到的场景中的世界坐标系。

2.1.2 相机型号

一个坐标系也与相机相关。在本文中，我使用x_c，y_c 和 z_c 来表示其基础向量。摄像机的光学中心是原点，x_c ，与图像中向右的方向吻合，y_c ，与向下的方向吻合。矢量z_c 表示摄像机的方向。

我使用透视相机投影模型，它由相机的焦距f_c 。所用相机的物理特性由几个参数给出：相机传感器的尺寸

$_c w$, h_c , 原理点的位置（ x_c , y_c ），像素的大小（ w_p , h_p ）和其斜角α_c 。我使用关于像素的方形形状的假设（ $w_p = y_p$, $\alpha_c = 0$）和原理点在图像中心的位置（ $x_c = \frac{w}{2}$, $y_c = \frac{h}{2}$ ）。

在给定的约束条件下，模型的透视投影是完全是由相机的内部校准矩阵决定的，其内容如下：

$$K = \begin{bmatrix} f & 0 & \frac{w_I}{2} \\ 0 & f & \frac{h_I}{2} \\ 0 & 0 & 1 \end{bmatrix}, \quad (2.1)$$

其中通过$f = f_c$是以像素尺寸计算的焦距，以及 (w_I , h_I) - 图像大小。

尽管它很简单，但所使用的摄像机模型也足以描述更普遍的情况。例如，如果某些摄像机的失真和针尖位置的参数是已知的，那么它可以通过应用修改输入视频序列的每一帧的确定性程序而与所述模型相一致。

2.1.3 人体模型

场景中唯一移动的物体是人。我使用了[52]中提出的人类模型。这个模型是将姿势和身材映射到三维人体模型的顶点位置。

2.2 摄像机的姿势

所选择的场景模型毫不含糊地定义了世界坐标系和摄像机坐标系的相互位置。摄影机的姿势l_c，由三个参数明确地确定：

– 摄像机在地平面上的高度h；

– 摄像机的角度；

– 摄像机的滚动角度r。

摄像机离地平面的高度h决定了摄像机光学中心在□轴上的位置。倾斜角和滚动角是摄像机的t和r的角度

摄像机是一个世界坐标系和自身旋转的坐标系统。

从形式上看，从图像中确定摄像机姿势的任务如下：

输入： - 一系列的$\{□_t\}^T$图像，通过静态的方式获得。
摄像机；

- f相机的焦距；

退出： 场景中的相机姿势参数：$l_c = (h, t, □)$.

所开发的算法可以应用于包含彩色和单色图像的序列。对观察到的数据施加以下限制：

– 至少有三张未定位的人的图片
两者在同一条直线上；

– 人们头部的图像至少是16×16像素的大小；

– 摄像机离地面的高度不超过20米；

– 摄像机的滚动角□在（- ,ππ）之内；

– 摄像机的角度是在)0,π/2(；

– f相机的焦距被限制在5000像素的尺寸。

表1 - 合成样本中的姿势参数和相机焦距的分布。

搭配米	名称	最小值	最大值
		价值	价值
h	高度（米）	0	20
t	斜率(rad)	0	$\frac{\pi}{2}$
r	滚动 (rad)	$-\frac{\pi}{6}$	$\frac{\pi}{6}$
f	焦距 (像素)	0	5000

输入的图像序列可以是任何长度，特别是包含一个单一的图像。

2.3 建议的方法

我开发了一种确定摄像机姿势的方法，基于对感兴趣的物体的图像大小的分析。我选择人头作为感兴趣的对象，因为它们在监控场景中不太可能重叠，不像整个人体。

我使用机器学习技术与教师一起构建输入图像与摄像机参数的映射。由于缺乏已知物体和摄像机在场景中的位置的大型标记集合，训练是在合成采样上进行的。

2.3.1 建立一个培训样本

训练样本由合成图像序列组成。我使用观察数据的模型来构建它、

2.1节中描述的。相机的姿势参数从表1所示的范围内统一选择。

所提出的摄像机姿势估计算法只使用人的检测结果，即人的头部在图像中的位置和大小。因此，合成样本并没有模拟现实世界中人们的各种姿势，所有的人都是一个标准的姿势。我选择了1.75米作为场景中人的高度，这是欧洲成年人的平均高度。在构建合成样本时，图像被丢弃了、不满足观测数据模型的约束。合成样本由100373个序列组成，每个序列至少包含200张图像。一幅图像在图像的任何位置都只捕捉到一个人。没有重叠的情况下，检测器可以检测到一个人，并检测出假阳性。

2.3.2 选择一个特征描述

合成图像序列在视觉上与真实数据不同（图2.1）。因此，为了在数据上训练回归算法，有必要选择一个对所使用的样本不变的特征描述。

我使用检测图像中人的头部的结果作为描述。这样一来，图像中的每个人都描述了

三个数字（x_h，y_h，s_h）对应于他的头部中心位置。
和它的线性尺寸。当然，人体模型[52]可以确定合成图像中人头的真实位置。然而，这样的方法不能应用于真实数据图像。使用不同的方法来估计人的头部在真实数据中的位置

图2.1 - 观察到的视频序列帧和合成帧的比较。专家标记的头部区域（对于真实帧）和检测器检测到的头部区域（对于合成数据）用红色标出。

和合成数据会降低摄像机姿势检测算法的通用性。因此，人头检测器甚至适用于合成数据。此外，这种方法允许学习过程适应检测算法的定位和头部尺寸误差。这种方法还允许图像中存在三维模型和检测算法的任何物体被用于摄像机姿势检测。

选择了算法[53]，并使用该算法的优化版本[1]来定位图像中的人头。使用这种算法是由于它的高速度。这个因素对于建立一个大型的合成图像集合非常重要。另外，这种算法对人体图像中没有纹理的情况不敏感。

[1] https://bitbucket.org/13e_sha/fasterhog

为了建立一个训练姿势回归算法的先例，我使用了场景中64个人的信息。

2.3.3 摄像机姿态回归

所考虑的问题是一个回归问题，用于将发现的人的尺寸映射到一对摄像机的位置。

在回归训练中，关键是选择优化的损失函数。误差函数的标准选择是目标变量的预测值和真实值之间的欧几里德距离。这同样 "惩罚 "了所有情况下对正确答案的偏离。然而，随着相机在地平面以上的高度增加，其预测的准确性可能会下降。例如，如果摄像机安装在高处，以相同的数量改变摄像机的高度可能不会对观察到的数据造成任何明显的变化；如果摄像机靠近地平面，则可能会显著改变图像中人们头部的大小分布。因此，预测的准确性必须取决于摄像机的姿态。换句话说，不同的选择序列可能有不同的学习难度。

我将这一假设纳入最小化的损失函数中。它代表了四次方偏差的加权和，其中目标变量的值和它对损失函数的贡献权重都是由回归的算法预测的。

从形式上看，所构建的算法使用了正态的假设。

摄像机姿态分布 $l =_c$ $(h, t, \square)$受观察到的的先例，并预测了这个分布的数学期望值： r_c，方差Σ_c

$$_c p(l_c \mid x, \Theta) = \square(\square(x, \Theta), \Sigma_c (x, \Theta) \qquad (2.2)$$

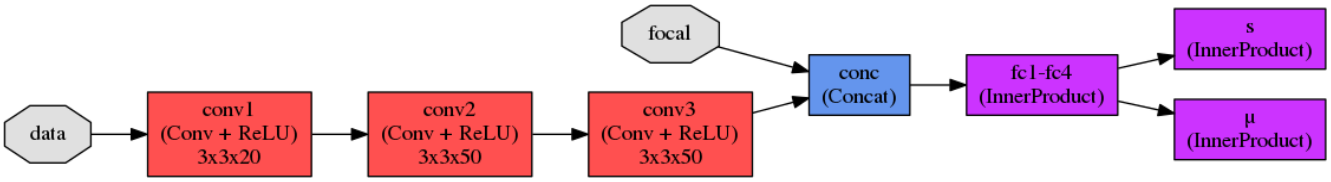


图2.2 - 预测摄像机位置和方向参数的神经网络示意图。

其中Θ为可教参数。

协方差矩阵Σ_c必须是正向确定的。我教

我在损失函数中使用这一约束，只考虑以下形式的对角线协方差矩阵：

$$\Sigma_c(x, \Theta) = diiag(\square e^{s(x,\Theta)}) + \varepsilon I \tag{2.3}$$

其中$\square(\square, \Theta) = (s_h(x, \Theta), s_t(x, \Theta), s_r(x, \Theta)$- 是神经网络预测的协方差矩阵参数的向量。为了防止协方差矩阵在训练过程中变质，在其上添加一个等于10的正则化参数$^{-6}$。

协方差矩阵的行列式Σ_c描述了摄像机姿势预测的不确定性区域的大小。因此，$\lambda_c(x, \Theta) = \Sigma^{-1}(x, \Theta)$的值可以解释为算法对预测的信心。

训练映射$\square \tilde{} \square(\square, \Theta)$和$\Sigma_c(x, \Theta)$是用最大似然法完成。因此，观察数据的可能性的负对数被用作损失函数：

$$L^{(\square)} = -\sum_{\square} \log N(\square l_i \square \tilde{l}^i, diag(\square e^{s^{i)\square}} + \varepsilon I)\square \tag{2.4}$$

所用的损失函数的导数可以用类比法计算：

$$\frac{\partial \square}{\partial \tilde{l}^j} = \frac{l_c^j - lc}{esj + \varepsilon} \tag{2.5}$$

$$\frac{\partial \square}{\partial s^j} = \frac{1}{2}\frac{e^{s^j}}{esj + \varepsilon}\left(1 - \frac{(\tilde{l}_c^j - lc\square)^2}{esj + \varepsilon}\right)\square \tag{2.6}$$

表达式（2.4）、（2.5）和（2.6）允许损失函数在现代图形加速器上有效实现。

我使用一个前向传播卷积神经网络，通过
图2.2显示了函数□˜□(□, Θ)和Σ$_c$ (x, Θ)。输入
的神经网络是一个3×8×8的张量，描述了
和检测到的64个头中的每个头的大小。为了避免算法不得不适应前述物体的各种排列组合，它们被按大小升序排列。网络的另一个输入是所用相机的焦距f，以尺寸计算
图像的像素。

构建的神经网络由3个卷积层和一个非线性ReLu函数组成。每个卷积层的大小为3×3。这种方法允许卷积层1）使用尺寸明显不同的远处物体的信息；2）通过使用尺寸相近的物体来适应数据中的噪音。

在应用卷积层后，结果与摄像机的焦距f相结合，并反馈给全链路层的输入。上述损失函数允许神经网络被训练来预测摄像机位置和预测的信心。

2.3.4 结合先例的结果

构建的神经网络利用场景中不超过64个人的位置来预测摄像机的姿势。在真实的监控场景中，场景中的人的数量可能大大超过这个值。因此，下面我考虑采用一种方法来结合不同数据集上的结果。

使用天真贝叶斯方法将算法在K不同的人类检测子集上的结果结合起来：

$$\overline{l_c} = \overline{\Sigma_c}^{-1} \left(\sum_{k=1}^{K} \Sigma_{l_c,k}^{-1} l_{c,k} \right) \tag{2.7}$$

$$\overline{\Sigma_c} = \left(\sum_{k=1}^{K} \Sigma_{l_c,k}^{-1} \right)^{-1}, \tag{2.8}$$

其中$\overline{l_c}$是预测的相机姿势。

需要注意的是，这种方法假定在不同的数据子集上预测的摄像机姿势之间没有依赖性。为了实现这一点，可以在稀疏的帧子集上使用非重叠的物体检测子集。

2.4 学习和体验式评估

本节介绍了用于训练神经网络参数的方法以及对合成和真实数据的测试结果。

2.4.1 培训

神经网络是在一个只包含正确头部检测的合成样本上训练的。只使用这种 "干净 "的数据并不允许构建的算法在含有错误的真实数据上概括出检测器的结果。

为了解决这个问题，我提出了两种方法来模拟真实数据中的噪声：1）场景中的错误头部检测；2）重复的

在一个先例中检测。如果被观察的人在一段时间内没有改变他/她在场景中的位置，第二种类型的噪声会在真实数据中发生。为了模拟噪声数据，采用了以下先例构建算法：

1. n元素的子集（$0 < n \leq 64$）从属于同一序列的检测集合中选出；
2. 在选定的检测中 m $0 < m <$ 任意替换 探测器在图像中的位置和大小是随机定位的，以防止出现假阳性；
3. 通过选择64个有重复的检测，从构建的集合中产生一个先例。

这允许模拟人类检测算法中的假阳性和不同帧中的重复检测。在所提出的算法的帮助下，为每个合成检测序列构建了三个先例。

所提出的卷积神经网络只有67393个参数，隐藏层的神经元数量相对较少。这使得在训练中可以批量处理32768个先例。我使用亚当的方法进行训练[54]。其速度是

学习按照参数γ=0.95的幂律下降，每
1500次迭代。训练需要200000次迭代。在实验中，我使用80%的合成样本序列进行训练，20%用于验证。

2.4.2 对一个合成样品的实验评估

实验评估包含了将所提算法应用于合成和真实CCTV数据的结果。

表2 - 训练和验证样本上的归一化偏差对全连接层的数量和训练样本的

大小的依赖。

尺寸 样品	数量 全胶合层	平均每小时误差 培训	平均每小时误差 验证
20448	3	0.1061	0.121
20448	4	0.0956	0.1167
20448	5	0.07	0.12
30179	4	**0.09015**	0.117
51366	4	0.09836	0.1064
80298	5	0.09764	**0.1009**

第一个实验显示了训练样本的大小和神经网络的隐藏层数对构建的回归器的质量的影响。我测试了具有不同全连接层数的神经网络（见表2）。验证样本上的回归器的均方根误差似乎是衡量算法质量的一个无信息量，因为相机姿势参数的尺寸和取值范围是不同的

预测值和真实值之间的距离。所以我用□ $_1$ 预测值和真实值之间的距离

通过对训练样本中的相机姿势参数值进行归一化处理（见表1）。比较结果表明，增加训练样本的大小和全连接层的数量可以提高摄像机姿势检测的准确性。在80298个场景的训练样本上训练一个包含5个半连接层的神经网络时，获得了最好的结果。另外，测试表明，在选定的参数下没有过度训练，因为训练和验证样本的误差很接近。

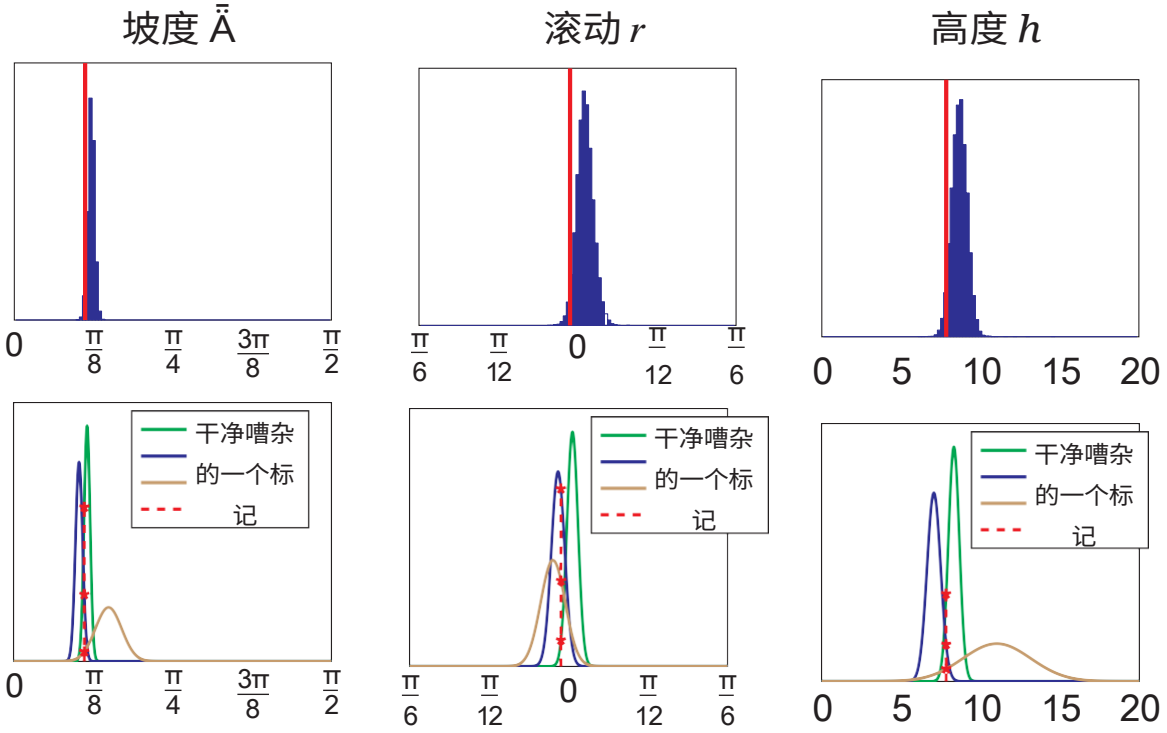


图2.3--TownCentre样本的相机姿势检测结果。第一行显示了样本中不同的正确人类检测子集上预测的摄像机姿势参数的直方图（蓝色）。第二行描述了相机姿势参数的预测分布，这些预测结果结合了a）不含噪音的 "干净 "数据（绿色曲线）；b）含有人类检测的假阳性的 "噪音 "数据（蓝色曲线）；c）含有一个人类检测的64个副本的先例（棕色曲线）。专家标记中提出的相机姿势参数用红色标记。各栏对应的是摄像机在地平面以上的倾斜角、旋转角和高度。

2.4.3 对真实数据的实验评估

为了评估算法的通用性，还有必要在现有的真实CCTV数据样本上提出测试结果。实验包括对TownCentre样本的质量评估[26]，因为它包含了专家对人们头部位置的标记和摄像机的校准。

表3 - TownCentre样本中摄像机姿势检测的结果

	倾斜度	转向 r	高度 h
专家打标	0.3497	-0.0251	7.84
"清洁 "数据	0.3634±0.0149	0.0130 ± 0.0182	8.3290 ± 0.3453
"噪音 "数据	0.3237±0.0176	-0.0345±0.0219	7.0696 ± 0.4294
只有探测	0.4694±0.0649	-0.0513±0.0402	11.0312±2.1441

在测试过程中，检测fastHOG图像中人的头部的算法被应用于序列帧[53]。如果这些检测与专家标记的重合度在IoU指标中超过0.5，则被视为正确。在这些条件下，该算法的准确率为48%，检测结果包含4501帧中的19061个头部检测。图2.3的第一行显示了不同先例上的反应分布。可以看出，这些分布有一个接近正确参数值的模式。

为了结合算法在不同先例上的结果，采用了第2.3.4节所述的方法。随机选择20个检测集不重叠的先例。这确保了算法对不同先例的结果的独立性。当预测结果被合并时，两个参数都被估计：目标变量的平均值和它的协方差矩阵。此外，协方差矩阵也是对角线的。因此，在不同先例上组合预测的结果可以表示为摄像机姿势的各个组成部分的密度函数（见图2.3，第二行，绿色曲线）。

图2.4显示了在预测的摄像机姿态下，合成的人在地面上的可视化情况。 可以看出

图2.4 - 预测地平面上合成的人的可视化。

真实的人和合成的人的尺寸是相似的，这证实了预测的相机姿势的真实性。

在接下来的实验中，使用TownCentre样本中的所有检测结果，包括检测器的假阳性结果，重复所提出的姿势检测方法。结果显示在图2.3的第二行（蓝色曲线）。值得注意的是，样本中探测器的误报比例明显高于训练样本中的模拟误报比例（52% vs 10%）。尽管如此，结果显示，即使在大量的错误探测器报警的情况下，预测的超过摄像机也接近于专家的标记。

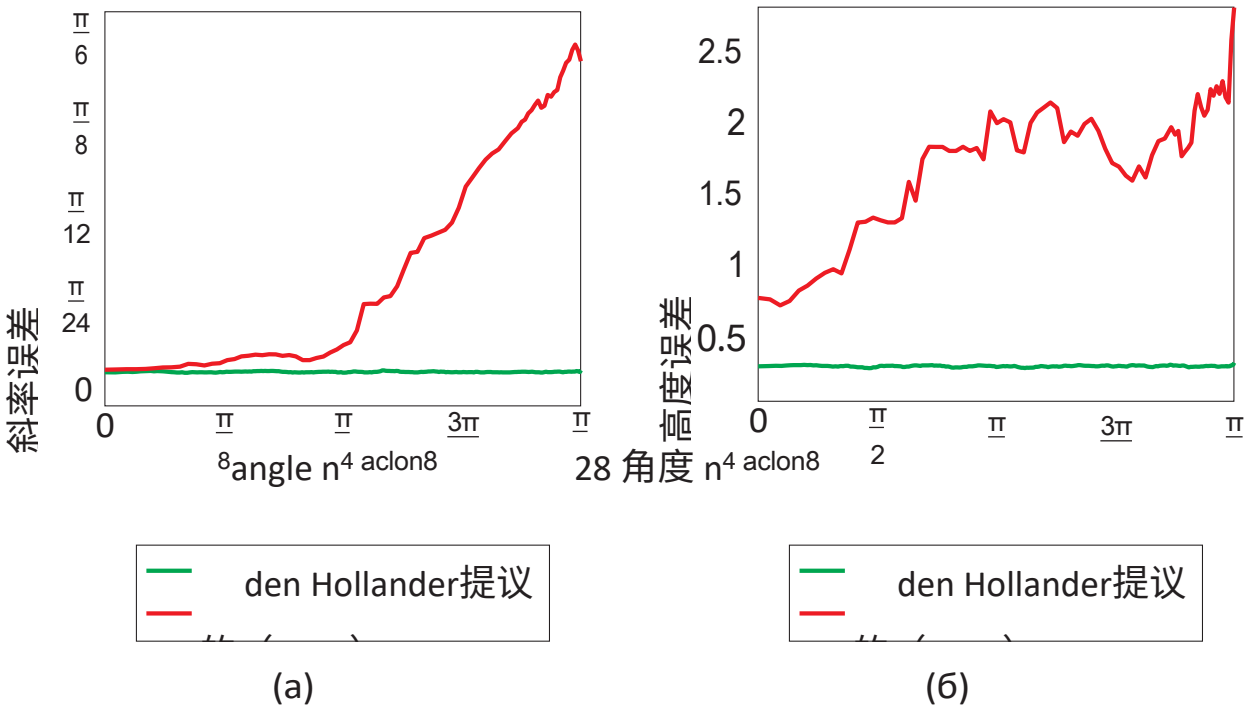
用TownCentre样本进行的最后一个实验考察了在一个先例中重复检测的极端情况。在这种情况下，先例由64份随机选择的对场景中人的头部的检测组成。这种情况可能出现在场景中只有一个人，而这个人在很长一段时间内不改变他的位置。该算法不能预测摄像机的位置，因为唯一的检测只设定了摄像机到地平面某一点的距离。在图2.3的第二行

表4- PETS 2006样本中预测的相机姿态参数。该表包含预测的相机姿态参数及其标准偏差。最后一行显示合成验证样本的平均误差。

序列		倾斜	转弯	高度
PETS 1	加价 被评估的	0.104 0.445 ± 0.061	-0.017 -0.001 ± 0.035	1.878 4.595 ± 0.878
PETS 2	加价 被评估的	-0.037 0.16 ± 0.071	-0.095 -0.037 ± 0.042	4.609 9.37 ± 1.081
PETS 3	加价 被评估的	0.289 0.357 ± 0.024	-0.03 -0.024 ± 0.022	5.501 5.823 ± 0.298
PETS 4	加价 被评估的	0.458 0.453 ± 0.024	-0.109 0.059 ± 0.023	6.567 5.367 ± 0.326
合成的	中型 错误	0.084	0.093	0.818

(棕色曲线）呈现了对摄像机姿态的估计。可以看出，当进入这样一个复杂的先例时，分布的方差明显增加。因此，预测的algo节奏准确性参数λ_c，使我们能够确定复杂的先例进行分析。对TownCentre样本的所有实验结果见表3。

所提出的方法还在更复杂的PETS 2006样本的四个序列上进行了测试[55]。值得注意的是，该样本的第一和第二序列违反了所使用的假设之一。在这些序列中，人们位于几个楼层，因此不可能识别一个单一的地平面。尽管如此，我还是将建议的方法应用于该样本的所有序列，并使用所有的头部检测来构建先例。测试结果（见表4）显示，建议的方法正确地估计了第三和第四个序列的相机姿势，而在第五个序列中，建议的方法正确地估计出相机的姿势。



(a) (б)

图2.5 - 摄像机的倾斜度（a）和高度（b）的预测误差与倾斜角的真实值的关系。

算法[7]的平均误差显示为红色，拟议算法的平均误差显示为绿色。

测试是在一个只包含正确探测器触发器的合成测试样本上进行的。

对于前两个序列是错误的。然而，对于这些复杂的序列，该算法的置信度明显降低。在分析算法的过程中，我们还研究了摄像机姿势检测的准确性对摄像机倾斜角的依赖性。为此，将摄像机倾斜角值的区间0，π，分成若干段，使每段对应于500个测试样本。在每个区段内，通过两种算法计算摄像机倾斜度和高度值的平均误差：建议的方法和[7]的方法（图2.5）。由于工作[7]假设检测器没有误报，对比的合成样本只包含正确的人类检测。分析表明，拟议算法的摄像机定位精度不取决于摄像机的倾斜角度。误差函数的这种行为的原因在于训练方法

数据集。在训练过程中，神经网络必须适应场景中各种摄像机的位置和方向。同时，用方法[7]对场景中摄像机位置估计的平均误差随着摄像机倾斜角度的增加而增加。造成这种行为的原因是假设边界矩形的边界包含检测到的物体的上下两点。这一假设被以下事实所违反

摄像机角度越接近π，就越大。 2

2.5 结语

在本章中，提出了一种算法，用于确定固定的CCTV摄像机的位置和方向，使用人的图像作为校准对象。该算法的一个显著特点是应用机器学习来构建人的检测结果与摄像机姿势参数值的映射。由于这种方法，所提出的算法有两个重要特性：

1. 抵抗摄像机角度在$0), (\pi$之间的变化。
2. 预测摄像机位置误差的能力。 2

该章的结果发表在[56]。

第三章：图像中的人物定位

本章讨论的是在已知校准参数的相机所拍摄的图像中检测人的问题。摄像机的内部属性，如焦距和摄像机的姿态，对图像中物体的可能大小和位置都有限制。例如，在视频监控系统中，人的图像不能完全高于地平线。与基本检测方法相比，利用这种限制提高了数据处理的准确性和速度。

在这一章中，我提出了一种方法，在已知校准参数的相机所拍摄的图像上建立人类检测的算法。所提出的方法考虑了摄像机在场景中的位置限制，以提高人头检测的基本算法的准确性和性能。它还可以推广到检测画面中其他类别的感兴趣物体的情况，如汽车、行人等。

从形式上看，有关问题的定义如下：输入： 1.图像；

2. 姿势参数l_c 相机（见第2章）；
3. f相机的焦距。

输出：一组限定感兴趣的物体图像的矩形。

3.1 建议的方法

所提出的算法是$g(A_b, A_f)$基本人类头部图像检测算法A_b及其结果分类器A_f的叠加。拟议的叠加法对算法A_b的选择没有任何限制，并将其作为一个黑箱。拟议的映射$g(A_b, A_f)$的输出是只有那些被A_b算法的检测的集合，这些检测在给定的相机校准参数下被A_f分类为 "可信的"。

提出的叠加法有几个重要的特点。首先，通过检测特定场景的 "不可能 "的检测器触发器，与基本算法A_b相比，检测精度得到了提高。此外，如果A_b算法是基于滑动的

窗口，那么分类器A_f，允许你识别那些不含

即使在对图像进行分析之前，也能获得 "可信的 "检测结果。这一特性可以提高视频监控系统中使用静止摄像机的叠加速度$g(A_b, A_f)$。

开发检测算法$g(A_b, A_f)$的任务归结为构建一个分类器A_f。为此，我采用了机器学习的方法，应用于合成视频监控数据。因此，有必要

有三项任务需要解决：

1. 构建一个观察数据的合成样本；
2. 构建对合成和真实数 据 不变的特征；
3. 训练分类器 A_f 算法结果

   物体不受相同条件的制约。

3.1.1 建立一个培训样本

由于缺乏具有已知人的位置和摄像机参数的大样本标记的监控数据，很难在真实数据上训练分类器。这就是为什么使用2.3.1节中描述的合成样本。(1)检测到的头部区域和(2)使用的摄像机的参数被用作检测结果的特征描述。因此，分类算法的输入是一个由以下内容组成的向量：

– 探测结果$o = (x_o, y_o, s_o)$对象的基本算法。

– ；

– l_c = (ℎ, 𝑡, □)相机姿势；

– f相机的焦距。

从形式上看，考虑到构建样本的特殊性，检测器触发分类的问题属于异常搜索问题的范畴。这类问题的明显特征是训练样本中没有关键类（检测器错误）的例子。在我的工作中，我提出了一种为这类对象建模的方法。为了构建它们，我结合了两种策略：1）使用针对具有不同相机校准参数的训练样本序列的检测；2）使用具有任意位置和大小的图像区域。第一种策略允许训练一个对相机校准参数有辨别力的分类器。第二种策略可以将正确的检测与背景图像上的基本检测器A_b的错误警报区分开来。在构建负面例子时
训练样本的使用比例为1:9。

3.1.2 建立一个分类器

为了解决分类问题，我使用了一个由5个隐藏层组成的全链路神经网络，以ReLu为激活函数。每个隐藏层有20个神经元。隐蔽网络层的数量和它们的神经元是从验证抽样中选择的。最后一层有1个输出，对其应用Logistic函数。神经网络的输出被解释为属于 "可信 "检测类别的概率。

3.2 学习和体验式评估

3.2.1 培训

分类器的训练是使用Adam方法[54]和Logistic损失函数进行的。训练是在使用caffe库[57]的24143个场景的合成样本上进行的。训练率每1500次迭代降低，功率率γ=0.95。在训练过程中，准确度/完整性曲线下的面积值在验证样本上达到了0.93的数值。

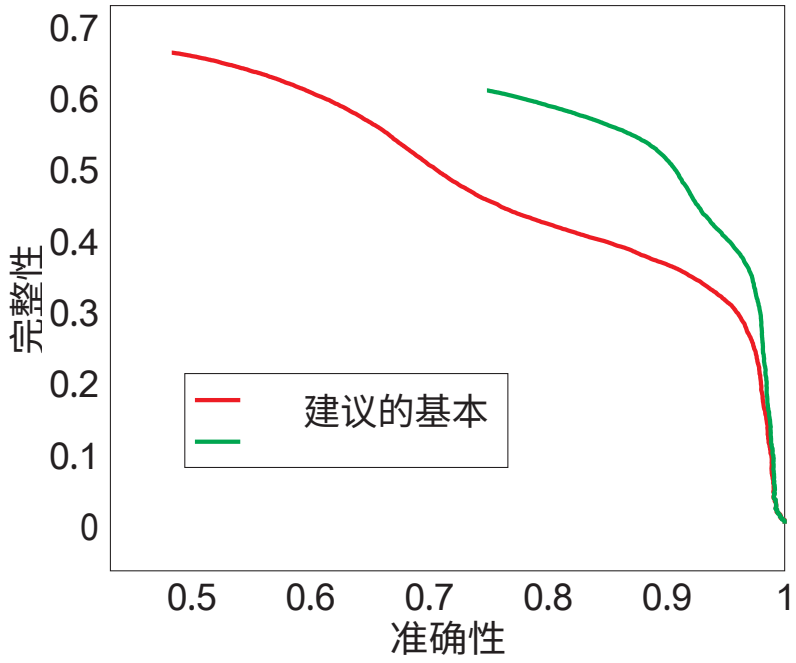


图3.1 - 将经过训练的分类器应用于真实世界的视频监控数据时，检测质量的变化。

3.2.2 对真实数据的实验评估

为了对真实数据进行专家判断，有必要知道样本的每一帧的相机位置。这使得使用标示图像的标准数据库变得困难。

因此，TownCentre样本[26]被用于测试。该算法A_f，将其置信度超过0.25的检测归为 "可信 "检测。如果一个检测与标记中的人的头部区域相交，则被认为是正确的，至少需要

其整合度的25%。原始检测器和所产生的检测器的质量比较（见图3.1）表明，所提出的过滤算法提高了18%的准确性，同时减少了2.1%的完整性。

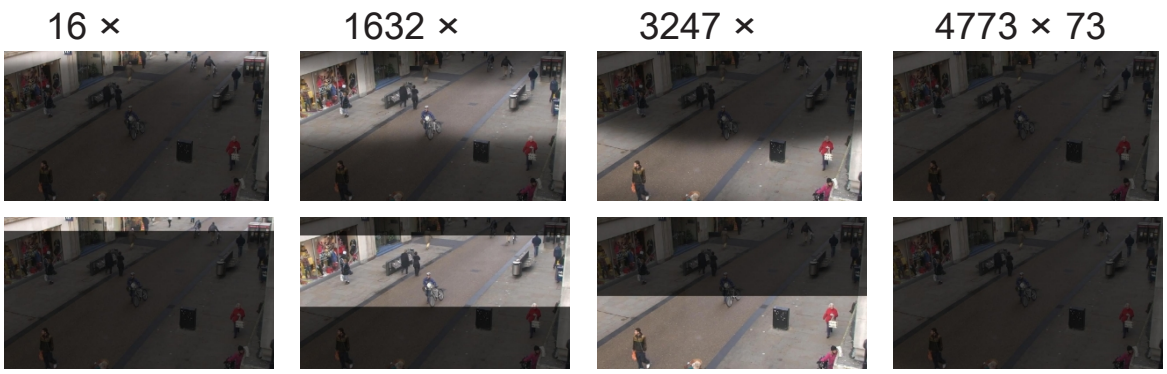


图3.2 - 检测分类器结果。

亮度表示由分类器（a）预测的、由检测器（b）使用的特定大小的头部的观察可能性。

3.2.3 与检测算法的整合

提出的分类算法A_f与基本检测器A_b相比，提高了$g(A_b, A_f)$的生产率。它对光影图像在画面上的允许位置设置了静态约束。图3.2（第一行）显示了概率

与TownCentre样本相对应的相机的不同尺寸的 "可接受 "检测。对于每个头部图像的尺寸，分类器预测了 "可接受 "检测所在区域的掩码。

在静态摄像机的情况下，所描述的区域的位置不会随时间而改变。这一信息被用来放大

与基本算法相比，组成性能$g(A_b, A_f)$
检测A_b。如果A_b使用一个图像金字塔，则只需要对其每一级的分类器预测的小区域进行处理（图3.2）。对于产生集合回归的算法

分类器为每个图像窗口预先确定矩形（"锚"），并预测哪些 "锚 "可能不对应于感兴趣物体的正确位置。

在本论文中，使用了[53]中提出的算法作为基本检测器。在TownCentre样本上测试时，训练好的分类器A_f，表明所有金字塔层上只有21.44%的窗口需要被处理。使用一个任意的

处理窗口的掩码降低了图像处理的效率。因此，在拟议的实现中，算法在金字塔的每一级处理矩形图像区域（见图3.2第二行）。这相当于处理了所有窗口的24.03%。这种方法可以提高人头定位算法的性能。

在TownCentre样本上，每秒20.03至34.36帧。

3.3 总结

在这一章中，提出了一种在已知校准参数的相机所拍摄的图像中检测人的算法。提出的算法是人类检测的基本算法（A_b，A_f）A_b及其分类器A_f的叠加。使用叠加法，任何用于检测图像中人的算法都可以被选作基础。分类器A_f是在一个合成的CCTV样本上使用机器学习构建的。实验评估表明，构建的叠加法提高了图像中人员检测的准确性和速度。

该章的结果发表于[58]。

第4章。在视频序列中的陪同人员

本章讨论了在视频序列中为人伴奏的任务。陪伴是指绘制场景中每个人的运动路径。陪伴任务可以被认为是人的检测问题的延伸，它需要指出每个人的位置，不是在一个，而是在视频中他出现的所有帧中（图4.1第二行）。因此，许多视频中的人物追踪算法将图像中的人物检测作为第一步处理。

从形式上看，人类跟踪算法的输入是一个视频序列$\{I_t\}$，输出是一组人类运动的轨迹$\{T_j\}$。每条轨迹T都由一连串的约束条件描述
视频段是一个矩形$\{\square^t\}$，是视频段中一个人的图像
$t \in [t_b, t_e]$，他在那里出席：

$$T = \{b^t\}_{t=}$$

4.1 基本算法

如第1.3节所述，通过检测进行护送的方法被用来护送人。它包括依次解决在视频序列的帧中检测人的任务，并将检测结果组合成对应于一个人的组--轨迹（图4.1）。

在许多用于追踪视频中的人的算法中，算法[26]被选为基本方法。它似乎是最有希望用于实际的，因为它允许人们建立

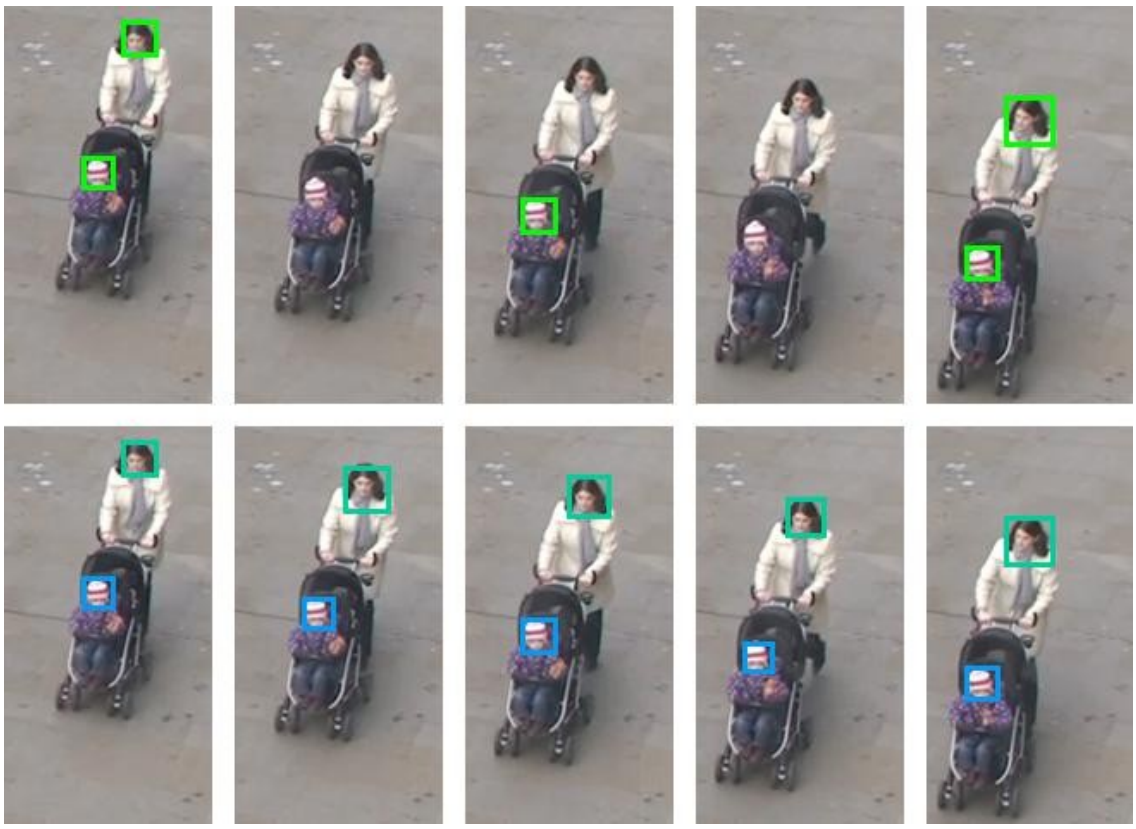


图4.1 - 护送-检查-检测方法的可视化。在第一行中，矩形显示了在一组稀疏的关键帧上正确检测人的结果：第一、第三和第五。
第二行显示的是护送期间接待的人员的位置。

检测算法甚至可以应用于数量有限的帧（以下称为关键帧），这大大增加了数据处理的速度。

基本算法包括四个步骤：

1. 检测视频片段中的人（他们的头）的图像；
2. 构建多条轨迹D--时间中的轨迹片段
   该人已被检测到在框架的附近；
3. 将小轨道合并为轨迹；
4. 在没有找到人的地方，恢复人们在干部中的地位。

本论文的第三章专门讨论了在视频序列中检测人的任务。在下面的小节中，将详细描述基本算法跟踪问题的剩余解决步骤和所提出的修改意见。

4.1.1 构建小轨道

对于每个在时间t检测到的头部图像，构建一个包含人在帧$_t$ 的I时间范围内的位置信息的追踪子。在这个阶段，人的运动被描述出来。特别是，它允许确定他的运动方向，这对进一步构建轨迹很重要。

小轨的构建是通过对检测到的人类头部区域进行视觉复制来完成的。基本算法[26]提出使用一些关键点[24]，用KLT算法[59]伴随头部图像。

4.1.2 将小轨道组合成轨迹

这一步解决了将小轨道的集合D分割成相同轨迹的小轨道组的问题。为简单起见，每一组这样的小轨道也被称为一个轨迹。基本算法的作者认为，同一帧的不同轨迹不能对应于同一个人。因此，一条轨迹在时间上不能包含一个以上的小轨道。同时，每个tracklet都对应着一个特定的轨迹。轨迹集D的任何划分都是以轨迹的形式进行的。

$\{T_j\}$满足这个约束条件的，称为假设。因此

因此，轨迹构建问题是寻找最佳假设□* 的问题。

对于所描述的问题，一个概率分布在多个

最有可能的假说代表了将一组小轨道划分为轨迹的最佳方案：

$$* □= \arg\max_{□} □(□|□) = \arg\max_{□} □(□, □) = \arg\max_{□} □(□)□(□|□) \quad (4.1)$$

因此，需要对轨迹算法进行描述：

– 设置假设集P（$D|H$）的概率分布；

– 定义一种寻找最佳假设的方法。

运动模式

运动模型定义了一个似然函数P（$D|H$），描述了将小轨道划分为轨迹的过程。需要注意的是，一些小轨迹可以由错误的探测器检测构建，因此不应归属于任何人类轨迹。为了解决基本方法中的这个问题，每个轨迹包含一个离散的二进制变量，即类标签，取两个值之一：$_{cped}$ - "人类"，c_{fp} - "虚假检测"。

论文[26]使用的假设是人们在场景中独立移动：

$$□ (□|□) = □(□_1 , □_2 , ..., □_□ | □) = \prod_{j\ T\in H} □(□_□ |□)_□ \quad (4.2)$$

并根据图中的动态贝叶斯网络对每条轨迹的可能性进行因子化。4.2:

$$p(T = \{d_0 , d_1 , d_2 , \ldots , d_{n-1} , d_n \} \mid \text{D450}) = p(d_0 \mid \text{D450}) \prod_{□=1}^{□} p(d_i |d_{i-1} ， □) 、 \quad (4.3)$$

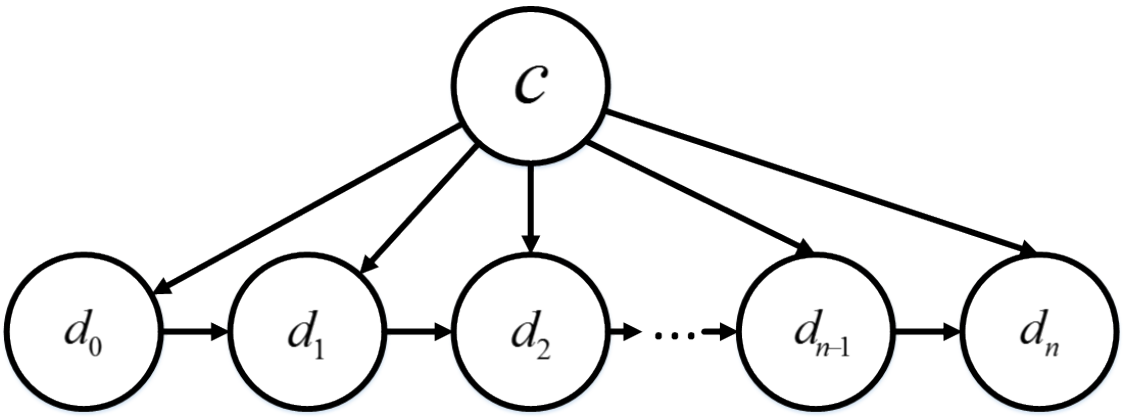


图4.2 - 轨迹对应的图形模型

$\{\square\}\{d\}_i^n$ $\square=$ - 是有关轨迹的跟踪小集，c是物体的类型：人类或虚假检测。

其中，$p(d_0 |\square)$是轨迹的第一个小轨道的观察概率，以及
$p(d_i |d_{i-1}, \square)$是一个小轨道到下一个的过渡概率。在基线方法中，这些概率描述了小轨道的特征，如检测的位置x，其大小$\square$，速度分布m
并估计一个小轨道内人类运动的速度Y:

$$\square(\square_0 |\text{D450}) = \square(\square_0)\square(\square_0 |\text{D450})\,\square(\square_0 |\text{D450}) \tag{4.4}$$

$$p(d_i |d_{i-1}, \square) = p(s_i |s_{i-1})\square(x_i |x_{i-1}, Y_{i-1}, c)\,p(m_i |\square) \tag{4.5}$$

为了描述所提出的方法与基线方法之间的差异，只应考虑限制人类图像在关键帧中的位置的因素$\square(x_0 |c = \square\square\square\square)$和$\square(x_i |x_{i-1}, Y_{i-1}, c = \square\square\square\square)$。

基本算法使用的假设是在轨迹开始时人的位置均匀分布，而不考虑轨迹的类型：

$$\square(x_0 |\square) \propto \frac{1}{\alpha}, \tag{4.6}$$

其中α是图像的尺寸，单位为像素。

$\square(x_i |x_{i-1}, c = \square\square\square\square)$这一因素对于寻找属于同一轨迹的三个单元最为相关。它描述了位置的变化

帧之间的人，并使用轨迹中包含的运动信息。这个因素被定义为检测到的人与人

之间的距离。

(a) (б)

图4.3 - 追踪器位置相似性系数的可视化。绿色矩形显示了关键帧的头部检测器的性能，绿色曲线显示了构建的小轨道，红色矩形显示了估计的人类位置。在另一个关键帧上。(a) □(x_i |x_{i-1} , c = □□□□)估计在一个关键帧上的位置。

(b) □(x_{i-1} |x_i , c = □□□□)估计在下一个关键帧的位置。前一个关键帧。

人像的位置并进行预测（图4.3(a)）。为了预测人像的位置，使用了所有对人像速度的估计，也就是说，使用了t帧′ 和keyframe ⍰ 之间的平均速度：

$$\square(x_i \mid x_{i-1}, Y_{i-1}, \square\square\square\square) = \frac{\delta_t}{\lvert Y \rvert_{i-1}} \sum_{y \in Y} \square(\square_y, \Sigma_y + 2\Sigma_d) + (1 - \alpha\ \square)\, N(\square_p, \Sigma_p + 2\Sigma_d), \quad (4.7)$$

其中$|Y_{i-1}|$是由小轨道d_{i-1}获得的速度估计值的集合。详细来说

预测位置的计算方法x^{i-1} 和x^{-}_{1} 人体图像

下面的关键帧和预测的置信度Σ_y，Σ_d，Σ_p在[26]中描述。

4.1.3 寻找最佳假说的算法

为了构建人类在场景中的运动轨迹，必须找到最佳假设H^*：

$$H^* = \arg\max_{H} P(Z|H)\,P(H) \tag{4.8}$$

基本算法指出，没有有效的算法来获得分布函数的全局最大值的假设。因此，作者提出使用一种近似推理方法，即MCMC DA算法。这种算法包括使用马尔科夫链方案从$p(H, \square)$的分布中抽取假设。然后，具有最高后验概率的假说被接受为最佳假说的近似值。

4.1.4 恢复原状

基本算法的最后一步是确定每个人在视频序列的每一帧中的位置。这是通过在检测器检测到的关键帧之间对人的位置进行线性插值来完成的。

表5 - 镇中心样本的维护结果

测试结果	MOTA	非物质文化遗产
Benfold等人。[26]	0.454	0.508
模式的交叉关联性	0.507	0.524
"—包圆点"。	0.519	0.525
"包点 "+出入境区域	0.54	0.522

4.2 复杂的算法

为了提高轨迹的准确性，我们提出了一种新的跟踪算法，该算法以[26]为基础，但在几个关键参数上与它不同。

轨迹构建的质量取决于确定哪些小轨迹对应于一个人的准确性。这项任务的基本算法是评估检测到的图像位置的一致性

基于因素$p(d_0 \mid \square)$和$p(d_i \mid d_{i-1}, \square)$ 的人类轨迹。
因此，所提出的修改意见旨在提高小轨道结构的准确性并修改所描述的因素。本文提出了基本算法的几处修改：

1. 第三章提出的方法用于检测视频序列关键帧中的人头；

2. 建议使用一种可靠的视觉引导方法来提高小轨道的准确性；

3. 不仅对下一次检测，而且对前一次检测在时间上也使用位置一致性；

4. 增加了对场景入口区域的估计和对轨迹的第一个小轨道的位置的先验偏

好。

图4.4 - 角的位置（用红叉子标记），在75帧期间用KLT算法（左）和Swarm of Dots算法（右）获得。

4.2.1 构建小轨道

在将人的检测结果结合到追踪器轨迹中时，追踪器内人的定位的准确性起着重要作用。可以区分两种类型的跟踪错误：1）在某一帧中对人的定位不准确；2）丢失一个陪同人员。第二种类型的错误出现在对人在下一个关键帧上的位置预测不正确时。值得注意的是，这种类型的错误是最关键的，因为它可能会把对应于不同人的轨迹合并到一个轨迹中。一个小轨道所对应的速度估计值在预测一个人在下一个关键帧的位置方面起着至关重要的作用。

基本的方法是使用视觉追踪算法，即KLT，它在人的头部图像上追踪多个角。这种算法没有对追踪点的相对位置进行约束，这可能会导 致 在构建追踪点时出现误差（图4.4）。因此，建议使用羊群算法来生成更可靠的速度估计值

点"。[25].与KLT相反，"点群 "算法可以检测并重新初始化在当前帧中位置被错误检测的角。另外，当头部图像的纹理很差时，它无法找到足够的角，无法被"KLT "和 "点群 "算法可靠地引导。这种情况在头部图像较小或人背对着摄像机时很典型。在这种情况下，基于交叉相关模式的跟踪[22]可以得到更可靠的结果。

4.2.2 评估一个人的立场是否一致

在基线工作中，归属于一个人的检测结果的一致性由因子（4.7）描述。这个系数只考虑了为前一个关键帧$I_{t\,\square i-1}$ 构建��小轨速度估计值d 。然而，据观察，不仅使用小轨道速度估计值$\square_{i-1}$ ，而且使用小轨道速度估计值d_i ，可以提高跟踪的质量。因此，建议将小轨道的位置分布d_i ，定义如下：

$_{i-1}\square(x_i \mid x_{i-1}, Y_i, Y_{i-1}, \square\square\square\square) = \beta\square(x_i \mid x_{i-1}, Y_i, \square\square\square\square) + (1-\beta)\square(x_i \mid x_{i-1}, Y\ \square\square\square\square)$, (4.9) 其中 Y_i 和 Y_{i-1} 是小轨道速度估计集 $\square_{\ i-1}$ 和 d_i

分别。系数$\square(x' \mid x, Y, \square\square\square\square)$根据（4.7）确定。参数β表示哪种估计被赋予更多的优先权，取决于小轨长度$d_{\ i}$ 和 $\square_{i-1}$：

$$\beta = \frac{|Y_{i-1}| + 1}{|Y_{i-1}| + |Y_i| + 2} \tag{4.10}$$

与基本算法不同的是，本文提议使用所有的小轨道速度估计值，只使用这些速度估计值、

这些数据来自于被护送者在被发现时的位置，以及他们在对应于追踪器的钥匙卡之间的位置。

d_i 和 d_{i-1} 。这确保了，首先，只有最可靠的
其次，它只假设相邻检测之间的人类运动是一致的。

4.2.3 限制第一次轨迹检测的位置

使用图像中人类位置的先验偏好需要关于被观察场景的语义信息：道路、天空、房屋等的位置。提取这种语义信息需要复杂的场景分析和大量的计算资源[60]。因此，对于轨迹中人的第一次检测，建议只使用与场景输入区域的接近程度：

$$_{e}\square(x_0) = \square(\rho(x_0, ent\square\square)|0, \sigma^2) \quad (4.11)$$

其中，$\rho(x_0, ent\square\square)$是第一个小轨道$x_0$ 的位置到最近的场景入口区域的距离。因此，轨迹的第一个小轨道的先验分布出现为分布（4.11）和（4.6）的混合物：

$$\square\tilde{}(x)_0) = \frac{\square_{\square}(\square_0) + \square(\square_0)}{2}。 \quad (4.12)$$

设置场景输入区域的最简单方法是选择图像边界作为它。然而，当地平线在图像上或场景中有障碍物时，例如建筑物的墙壁，这可能会导致算法的不正确操作（图4.5）。因此，在工作中，我们提出要确定场景中的入口区域，以抵制。

在本论文中，假定有关区域是图像中的一个凸形多边形（图4.5（a））。那么

(a) (б)

图4.5 - 搜索场景入口区域。(a)找到的场景入口区域；(b)专家标记轨迹的可视化

到场景入口区域的距离可以定义为到该多边形边界的距离。可信轨迹中包含的第一个和最后一个人类探测的位置的凸壳被选为这样一个多边形。场景入口区域的这种表示方法使它在每次小轨道与轨迹相关联时都能被更新。

这种方法的缺点之一是它不可持续。例如，如果一个场景的入口区域被错误地扩大了，它就再也无法被缩小。正因为如此，只有可靠的轨迹应该被考虑在内。在提议的算法中，包含至少5个人类轨迹的轨迹被认为是可靠的。可以看出，构建的输入区域与包含专家标记轨迹的区域边界相对应（图4.5(b)）。

所提出的方法的另一个缺点是，场景边界的障碍物被当作入口区域来考虑。图4.5（a）所示的建筑墙就是这样一个障碍物的例子。

4.3 故事评估

对所提算法的实验评估是在一个室内TownCentre数据库[26]上进行的。它包含一个由静态摄像机拍摄的高分辨率视频序列（1920x1080/25fps）。还有一个专家标记，包含71500个标记的人的头部位置。平均而言，每一帧有16个人。

MOTA和MOTP标准被用来评估跟踪的质量[61]。MOTA描述了人员轨迹构建的质量，而MOTP描述了视频序列帧上人员定位的准确性。在这方面，最有参考价值的标准是MOTA。MOTA不超过1，而MOTP为非负值。MOTA值越高，说明绘制的人的轨迹质量越好，而MOTP值越低，说明检测人的准确性越高。

测试结果显示在表5中。表中第二行和第三行介绍的算法在构建跟踪子的方式上有所不同。在TownCentre测试平台上，"点群 "算法给出了最好的结果，但该算法的追踪质量取决于人的头部图像的质量。在纹理较弱的图像上，并不总是能够找到足够数量的角。在这种情况下，基于交叉相关的模式匹配算法对人的头部的损失更为稳健。

最后一行显示了使用 "点群 "算法并考虑到场景的输入区域时提出的算法的结果。

4.3.1 算法

如上所述，拟议的维护方法是一个依次应用算法的管道。因此，一个重要的问题是这个管道的哪一部分是该方法最薄弱的部分。回答这个问题的最有效方法之一是基于依次将管道的第一步骤从算法中排除的方法。

这种方法包括用专家标记的结果连续替换管道初始阶段获得的结果。这样就有可能估计出每个步骤对整体误差的贡献有多大。

分析结果列于表6。第一列显示了管道阶段，该阶段已被专家标记所取代。唯一的例外是第一行，它显示的是整个管道的结果。第二列和第三列分别显示了MOTA和MOTP标准的数值。分析结果显示，改进寻路算法可以使跟踪可靠性得到最大的提高（约0.28，从0.629到0.916）。另外，分析表明，检测之间人员的线性假设是一个非常粗略的近似，因为用专家标记上的位置代替重建中间帧上人员位置的算法可以使跟踪可靠性增加0.07以上。这可以从使用从专家标记中得到的轨迹的MOTP的增加中得到证明。这种增加意味着在非线性运动的情况下，轨迹算法的错误最多。

分析表明，用于找人的算法能够对人进行可靠的跟踪，但定位精度

表6 - 所提出的支持算法的质量分析

	MOTA	非物质文化遗产
整个传送带	0.54	0.522

是低的。值得注意的是，为什么即使流水线的最后一步被专家标记结果取代，也不能获得理想的结果。在所提出的算法中，每秒钟对物体进行5次搜索，也就是说，只对五帧中的一帧进行搜索。同样，对于被护送者的位置，没有在第一次和最后一次检测之间的片段之外进行推断。这导致了这样一个事实：专家标记中护送人的时间在很多情况下超过了所提出的算法的护送时间。图像中人的定位的准确性受到人的头部图像所包围的矩形的宽度和高度相等的假设的影响。这一假设的贡献反映在最后一行的MOTP标准值中。

4.4 挑剔的人

在这一章中，提出了一种在去序列中跟踪一组人的算法，由于使用了一种可靠的方法来估计不同帧中人的检测的相似性，其轨迹准确性优于基线算法。可靠性的提高受到以下因素的影响

提出了两个因素：一个是跟踪时检测位置的相似性因素，一个是惩罚脱离场景入口区域的轨迹的因素。

该章的结果已经发表在[62；63]。

第5章：在视频序列中确定一个人的姿势

本章讨论的是以去序列化的观点来确定人的姿势问题。如第1.4节所述，目前图像中人的姿势被定义为其关节的位置。因此，一个人在四分之三时间的姿势是以下的序列

检查图像中的$P^t \in \{\{p^t\}^K \mid p^t \in \mathrm{R}\}$。[2]

在我的论文中，我研究了确定一个人的姿态的算法。作为伴奏后视频数据处理的下一个阶段。因此，我对视频序列中姿势检测任务的表述如下：

输入：
- 视频序列 $I = \{I\}_t^N$；
- 人的轨迹 $T = \{b\}^{tN}$

退出：人类在每一帧的姿势 $P = \{P\}^{tN}$。

一个人的已知运动轨迹限制了确定其姿态的图像区域。

5.1 观察数据的数学模型

我认为视频中的人体姿势检测是能量函数$E(P, \Theta|I, \square)$的最小化问题，其中Θ是隐藏模型参数。参数Θ既可以包括单帧中人体姿势的隐藏参数（如尺寸参数），也可以包括人体模型的全局参数（颜色模型）。我们可以假设能量函数$E(P, \Theta|I, T)$将视频序列中的非归一化姿势似然函数设定为$\square\tilde{}(P, \Theta|I, T) = \exp(-E(P, \Theta|I, T)$。）此后，为了简化

我将隐含地假设能量函数对原始视频序列和每一帧中人的位置的依赖性，即 $E(P, \theta) = E(P, \theta|I, T)$ 。

所使用的观察数据模型是图像中的人体姿势模型在视频序列情况下的一般化。为此，基本模型通过假设人的姿势对不同帧的依赖性而得到扩展。归纳基本模型的标准方法对应于以下能量函数：

$$E(P, \theta) = \sum_{t=}^{T} E_I(P^t, \theta) + \sum_{t=}^{T-1} E_T(P^{t+1}, P^t, \theta) \quad (5.1)$$

其中$\square_I(P^t, \Theta)$是一帧中人的姿势模型，$E_T(P^{t+1}, P^t)$是一帧间姿势变化模型。这个模型可以被看作是一个马尔科夫式的

一个一阶链，每个时间点的状态都是多个

是一个同质的维度，描述了人类的姿势。它由两部分组成：1）图像中人类姿势的基本模型$\square_I$（P^t, Θ）；2）她的关节运动模型E_T（P^{t+1}, P^t, Θ）。

5.1.1 对图像中人的姿势进行建模

根据对现有方法的审查，科学论文提出了两种主要的人体姿势模型：一种是来自一组可变形部件的模型，另一种是姿势的回归模型。

尽管在撰写这篇论文时，回归模型为图像中的姿势检测提供了较好的结果，但将其推广到视频序列的情况是困难的。输入图像I_t到图像上的人的姿势P^t的映射不是

允许使用邻近帧的先验信息。

因此，我们选择了一组可变形部件的模型作为图像中人类姿势的基础模型。相应的马尔科夫

该网络定义了一个非正常化的似然函数 $\square\tilde{}(P^t, \Theta^t) =$ $\exp(-\square_I(P^t, \Theta))$。这一属性允许将其作为一个更大的图形模型的一部分来整合，使用以前的框架信息来构建

先验的限制。

一组零件的模型用一个能量函数来描述

$_I E(P^t, \Theta^t)$，其最小值被定义为图像I_t 中的人体姿势：

$$_I E(P^t, \Theta) = \sum_{i} \varphi_{i}(p^t, s^t) + \sum_{(i,j)} \psi^{s}_{(i,j)}(p_i, p_j, s_i) \quad (5.2)$$

单项潜力$j_i(p^t, s^t)$可以看作是人类关节检测算法在特定图像比例下的反应。成对的

势$\psi^s_{(i,j)}(p^t, p^t, s^t)$是一个二次函数形式，取决于

关节之间的位移。能量函数取决于当前图像中的人体尺寸参数s^t，而不取决于模型的其他隐藏参数。

在实践中，关节的位置参数p^t 是一个离散值

尺寸在图像中以一些增量来定义。尺寸参数

s^t，也是不连续的，当关节暴露时，对应的是不同的尺度。

5.1.2 运动模式

建立姿势变化模型的最简单方法是假设关节运动的独立性：

$$ {}_{T}E(P^{t+1}, P^{t}, \Theta) = \sum_{i=} \psi_{i}^{t}(p_{i}^{t+1}, p_{i}^{t}, \Theta) \tag{5.3} $$

由于不同关节的运动模式相似，只看其中一个就足够了。为了简化本小节中的符号，我省略了有关关节的索引。例如，使用p^t。

用来表示有关关节在帧t的状况。

这种将图像中的人的姿势模型扩展到视频序列的情况也被用于以前的工作中。例如，[48]提出了一个运动模型，假设人的姿势在帧之间有微弱的变化：

$$\psi^t(p^{t+1}, p^t, \Theta) = \frac{1}{2\pi\sqrt{2}}(p^{t+1} - \square^t)^{T-1}(\Sigma_p)^{-1}(\square^{t+1} - \square^t)$$

因此，当视频中人的姿势不变时，这种运动模式的最佳值就达到了。在移动时改变姿势，结果是

"可接受的噪音"。

在本文中，我扩展了这个运动模型。我使用一个线性动力系统来描述人体关节的运动。为此，该模型的潜在状态Θ被每个关节的运动特征所扩展。在本文中，我考虑的是关节运动的线性模型，即每个关节的状态由其在框架上的位置p^t和瞬时运动速度$v^t \in \mathrm{R}^2$描述。通过与人类姿势的类比，我用V表示视频序列中所有关节的速度。如果我们用$h^t = [p^t, v^t]$表示人类姿势中被考虑的关节在帧中的状态，那么所提出的运动模型的形式是：

$$\psi^t(p^{t+1}, p^t, \Theta) = \frac{1}{2\pi\sqrt{2}}(h^{t+1} - \square h)^{tT}\Sigma^{-1}(h^{t+1} - \square h)^t$$

$$\square = \begin{bmatrix} 0 & 1 & 0 & \\ & 0 & 1 & 0 & 1 \\ & & 0 & 0 & 1 & 0 \\ & & & & & \\ & & & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.4}$$

线性运动模型的可接受偏差由一个对称的正定义矩阵$\Sigma_p \in S_+$给出。为了减少参数的数量，我只考虑以下形式的对角线矩阵Σ_p：

$$\Sigma_p = \begin{bmatrix} \Sigma_{\square} & \Theta \\ \Theta & \Sigma_{\square}^{1} \end{bmatrix}$$

$$\Sigma_p = \alpha\text{-}1\square_{2\times 2} \tag{5.5}$$

$$\Sigma_v = \alpha\text{-}1\square_{2\times 2}$$

$$\alpha_p > 0，\ \alpha_v > 0、$$

矩阵$\Sigma_{\square}$描述了关节$\square^{t+1}$与其位置的可允许的偏差线性预测$p^t + v^t$, a Σ^v是关节速度的允许变化。

人员配置表中的工作人员数量。

如果没有额外的正则化，运动模型（5.4）允许关节速度的数值大得令人难以置信，因为它只限制其变化。为了解决这个问题，增加了一个对第一帧中关节速度的先验偏好的因素：

$$\psi^0(v^1, \Theta) = \square^{-1}\square(\square\Sigma_v^{1}) \tag{5.6}$$

$$\square^{-1}\Sigma_p = \alpha\square_1\square_{2\times 2}$$

此外，我的模型还限制了人的尺寸在各帧之间的难以置信的变化：

$$\eta^t(s^{t+1}, s^t) = \frac{1}{2}\frac{(\square\square\square^{+1} - s^t)^2}{\square\square\sigma_s} \tag{5.7}$$

因此，拟议的运动模型有以下形式：

$$\sum_{\square=1}^{T-1}\psi(P^{t+1}, P^t, \theta) = \sum_{\square=1}^{\square}(\square\psi^0(v^1) + \sum_{\square=1}^{T-1}\psi^t(h^{t+1}, h^t, \Theta) + \sum_{\square=1}^{T-1}\eta^t(s^{t+1}, s^t) \tag{5.8}$$

5.1.3 个别案例

提出的模型有两个有趣的特例，描述了以前关于姿势定义的工作。

考虑$\alpha_v \to +\infty$，$\alpha_v 1 \to +\infty$和$\sigma_s \to +\infty$的情况。这种情况描述了当任何关节的速度值与0不同时，能量的显著增加：

$$\begin{aligned} &\text{肢体}_{\substack{\alpha\square 1\\ \to+\infty}} \arg\min \psi^0(v^1) = 0 \\ &\text{肢体}_{\alpha_v \to+\infty} \arg\min_0 \psi^t(h^{t+1}, h^t, \Theta) = \\ &\text{肢体}_{\substack{\sigma_s\\ \to+\infty}} \eta^t(s^{t+1}, s^t) = 0 \end{aligned} \tag{5.9}$$

也就是说，最优解是所有速度参数为0的解。在这个条件下，关节运动模型的形式是：：

$$\begin{aligned} &\psi_{(hi)}^{\square+1}, hi, \Theta)|_v \quad \frac{1}{2\square\square 2}(\square^{\square+1} - p)^T \sum p)^{-1} (\square^{\square+1} - p) \\ &\psi(\square^1)|_{v\,1=0} = 0 \end{aligned} \tag{5.10}$$

也就是说，所提出的模型变得等同于[48]中描述的运动模型。

考虑另一种特殊情况。如果我们放松对帧间关节速度微小变化的约束，那么模型（5.8）描述了每一帧中人类姿势的独立确定。事实上

$$\begin{aligned} &\text{肢体}_{\alpha_v \to 0+0} \psi^t(h^{t+1}, h^t, \Theta) = 0 \\ &\text{肢体}_{\substack{\alpha\square 1\\ \to 0+0}} \psi^0(v^1) = \\ &\qquad 0 \\ &\text{肢体}_{\substack{\sigma_s\\ \to+\infty}} \eta^t(s^{t+1}, s^t) = 0 \end{aligned} \tag{5.11}$$

也就是说，图形模型被分解成独立的、连贯的组件，与每一帧中的人体姿势相

对应。

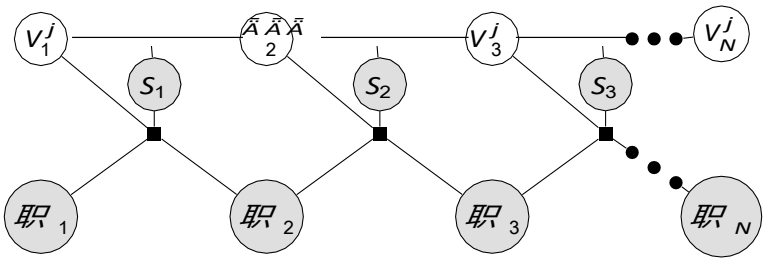


图5.1 - 与速度定义问题相对应的 因素图。

观察到的变量用灰色标记

5.2 优化方法

5.2.1 模型分析

为了描述所提出的优化算法，有必要考虑与所提出的能量函数有关的两个问题：

1. 确定视频中已知人体姿势和高度的关节速度 - $\arg\min_V E(P, \Theta)$；

2.在其余部分已知的情况下，确定画面中人的姿势和尺寸 t

模型参数为 $\arg\min_{P^t,s^t} E(P, \Theta)$.

确定速度

让我们考虑确定视频中人在已知姿势和高度下的关节速度问题$V(P, \Theta_{\setminus V}) = \arg\min_V E(P, \Theta)$。图像中的人体姿势模型$E_I(P^t, \Theta)$和尺寸变化模型$\eta^t(s^{t+1}, s^t)$与速度参数无关。因此，所考虑的问题在优化问题上是等价的：

$$V = \arg\min E(V|P, \Theta_{\setminus V}) = \arg\min \sum_{t=1}^{T-1} \psi^t(h^{t+1}, h^t, \Theta) + \psi^0(v^1), \quad (5.12)$$

□ □ □

其中$\Theta_{\setminus V}$，表示不包含速度参数□的隐藏参数集。

从这个表格中你可以看到，不同关节的速度从不包含在同一个总和中，这意味着可以独立优化每个关节的速度：

$$_{i}\square = \arg\min_{\square_{\square}} \psi^{0}(v^{1}) + \sum_{\square=1}^{T-1} \psi_{\square}^{t}(h_{\square}^{t+1}, h_{\square}^{t}, \quad V = \cup_{\square=1}^{K} \square_{\square} \tag{5.13}$$

考虑寻找单个关节的速度的最佳值。为了简化计算，我将省略当前关节的指数。 能量函数$E(V \mid P, \Theta_{\setminus V})$对应于图5.1中的因子图。可以看出，相应的图形模型是一个一阶马尔可夫链。鉴于（5.4）和（5.5），优化后的能量$E(V \mid P, \Theta_{\setminus V})$可以被重写为单数和对数的势：

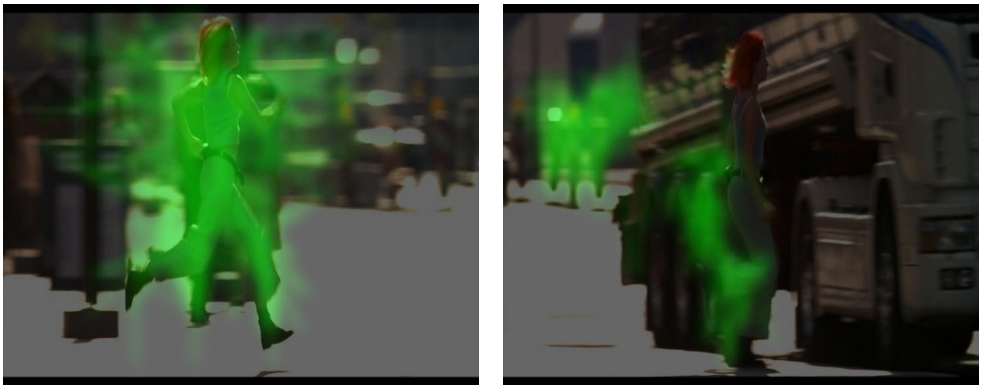
$$E(V \mid P, \Theta_{\setminus V}) = \psi^{0}(v^{1}) + \sum_{\square=1}^{T-1} \psi^{t,u}(v^{t} \mid p^{t+1}, p^{t}, s^{t}) + \psi^{t,p}(v^{t+1}, v^{t} \mid s)^{t}$$

$$\psi^{t,u}(v^{t} \mid p^{t+1}, p^{t}, s^{t}) = \frac{(\Delta p^{t} - A\, v)^{utT}\, \Sigma^{p-1}\, (\Delta p^{t} - A\, v)^{ut}}{2\square\square 2}$$

$$\psi\square^{,\square}(\square\square+1, {}^{vt}|st) = -\frac{(\square\square+1 \_ Apvt)T\, \Sigma v\text{-}1(\square\square+1 - \square\square\square\square)}{2\square\square 2} \tag{5.14}$$

$$\Delta p^{t} = p^{t+1} - A_{\square} p^{ut}$$

$$A = \left[\begin{array}{c|c} \square\square & \square\square \\ \square & \\ \hline \Theta & \square\square \end{array}\right] \quad Ap, A^{\square\ \square}, A \in \mathrm{R}^{2\times 2}$$



(a) (б)

图5.2 - 画面上人类姿势的最佳假设的可视化。绿色高亮部分的强度与构建的人体姿势假设的数量相对应。在画面（a）中，大多数检测到的假设都接近于正确的人体姿势。在(b)帧中，检测器无法找到一套好的人体姿势假说。

B 这个 阐述 的任务是 要找到 最小 职能能量

$E(V \mid P, \Theta_{\setminus V})$与确定线性动态系统（LDS）的状态问题相吻合：

$$\square = \arg\max_{\square} p(\square V \mid \{\Delta p^t\}_{\square=1}^{T-1}, \{s^t\}_{\square=1}^{T})\square$$

$$\Delta p^t \sim N(A^u v^t, (s)^{t2p} \Sigma_{\square})$$

$$^{t+1}\square \sim N(A v^{pt}, (\square_{\square}^{t})^{2v} \Sigma)$$

$$^{1}v \sim \square(\Theta, (s \quad \Sigma_p^{\square})^{v2})$$

$$vT = ApvT\text{-}1 \tag{5.15}$$

对最可能的配置V的搜索是利用卡尔曼滤波和RTS方程进行的。

确定框架上的姿势和尺寸

考虑在已知其他模型参数的情况下，确定某个帧中人的姿势P^t和尺寸s^t的问题 $\arg\min_{P^t, s^t} E(P, \Theta)$：

$$E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s}) = E(P, \Theta) = \sum_{(p,q)\in E} \psi_{(p,q)}(p^t, p^t, s^t) + \sum_{p=1}^{T} \psi_p(p^t, \Theta) + \sum_{t=1}^{T-1} \eta^t(p^{t+1}, p^t, \Theta) \quad (5.16)$$

视频中的人体姿势模型是一阶马尔可夫链，因此有关能量的形式是：

$$E(P^{t-1}, P^t, P^{t+1} | P^{\setminus t}, \Theta_{\setminus s}) = E_T(P^t, \Theta) + E_T(P^t, P^{t-1}, \Theta) + E_T(P^{t+1}, P^t, \Theta) + C, \quad (5.17)$$

其中，C是一个独立于所考虑参数的常数。为了不单独考虑边界情况，我们可以进一步将时间时刻$t = 0$和$t = T + 1$的模型状态分别定义为时间时刻$t = 1$和$t = T$的值。

给出（5.2）和（5.4），有关的能量是

$$E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s}) = \sum_{i=1}^{n} \psi_i(p^t, s^t) + \sum_{(i,j)\in E} \psi_{(i,j)}(p^t, p^t, s^t) + \psi_s(p^t, p^{t-1}, s^t) + \eta^t(s^{t+1}, s^t) + \eta^t(s^t, s^{t-1}) + C \quad (5.18)$$

因此，条件人类模型$E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s})$对框架$I_t$对于每个规模$s^t$，是单数$\psi_i(p^t, I^t)$和配对的$\psi_{(i,j)}(p^t, p^t, s^t)$电位与人体姿势的关节位置有关。

ka.由于运动模型没有在能量中加入成对的势能

$E(P^t, s^t \mid P^{\setminus t}, \Theta_{\setminus s}$ □)，那么这个能量是按照与姿势模型$E_I(P^t, \Theta)$ 相同的图形模型来分解的。因子$\eta^t(s^{t+1}, s^t)$和$\eta^t(s^t, s^{t-1})$对姿势大小参数的值设定了一个后验的偏好。

因此，人体姿势的条件模型$E(P^t, s^t \mid P^{\setminus t}, \Theta_{\setminus s}$ □)是图像中人体姿势模型$E_I(P^t, \Theta)$ 的概括。由于在优化模型□$_I(P^t, \Theta)$的过程中，算法列举了不同的人体尺寸参数□，那么对于条件人体姿势模型来说

世纪的特性也得到了满足：

1. 全局最优搜索算法的计算复杂度等于□(□□)，其中K是所考虑的人体骨架的关节数，M是图像中一个关节的允许位置数；
2. 构建图像中人的姿势的最佳假设集N的算法，这些假设至少相差一定量，其复杂度等于□（□□□）。

图5.2显示了一组最佳姿态假设的例子

在检测器正确定位关节而不能正确识别人的姿势的情况下，图像上的几个世纪。

寻找全局最优的难度

描述人体姿势模型的每个能量函数都有一个马尔可夫网络。寻找模型最优值的算法的复杂性取决于其属性。在一般情况下，寻找图形模型的全局最优值是由信心传播算法进行的。其计算复杂性取决于将图形模型分为两部分的顶点最小组的可接受状态的数量。

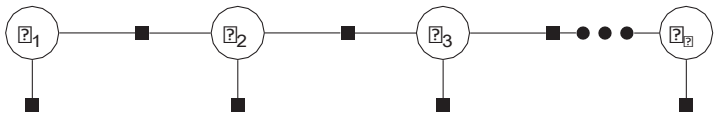


图5.3 - 基本姿势模型的因子图。

如果一个图形模型是一棵树，那么算法的复杂性就取决于每个顶点允许的状态数量的四倍。这种图形模型的一个例子是给定一个已知比例参数的图像中人的姿势模型。一种特殊的成对电位，使得图形模型中输出的复杂性降低到对图像中关节的可接受位置的数量的线性依赖。

然而，当这个模型被扩展时，马尔科夫网络中出现了循环，因此寻找最优解的复杂性大大增加。图5.3显示了对应于模型[48]的图形模型。它可以被看作是一个一阶马尔可夫链，其中每个状态都由一个框架的顶点集合来描述。

考虑一下这样一个马尔科夫链的全局最小搜索算法的计算复杂性。如果帧中每个关节有M个可能的位置，那么置信度传播算法的复杂度就等于$\mathcal{O}(M^K$ □)。事实上，运动模型（5.8）是一个二次函数形式，为了计算网络中的信息可以使用距离变换方法。因此，寻找最佳姿势集的复杂性线性地取决于图像中可能的人体姿势的数量。由于参数M的值
可以超过10个[4] ，而关节的数量可以达到几十个。
事实证明，寻找精确解的节奏在实践中并不适用。通过减少图像帧上的可接受姿势的数量，作者[48]构建了一种算法
复杂度为$\mathcal{O}(N^2$ □)的局部最优搜索节奏，其中N为数字
在一帧中允许的姿势的数量。

本文提出的模型是[48]的扩展，并将其作为一个特例。似然传播算法不能用于寻找所提模型的全局最优，因为它需要太多的计算资源。另外

潜在状态Θ还包含连续速度参数

这里描述的模型是离散的、不连续的。这不允许直接使用算法[48]。我提出了两种算法来寻找构建的能量函数的最优值。第一种算法使用了[48]中提出的减少框架中人类姿势的可接受假设的数量的想法来寻找局部的optii mum，第二种算法使用从分布中抽样的方法来完善结果。

5.2.2 确定性的算法

考虑一下所提出的算法中的第一个。它的伪代码显示在清单1中。

该算法的想法是基于两个任务的顺序解决：1）根据条件姿势模型，构建人在画面中的姿势假设

$E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s}$ □); 2) 计算速度参数的最佳值、

根据模型$E(V | P, \Theta_{\setminus V})$。解决这些问题的方法将在第5.2.1节中描述。

为了构建初始化，我使用了[48]中提出的方法，即最小化模型的能量，条件是

关节的速度为零 $E(P, \Theta)$ $|_{\square=0}$.对于构建在

通过最小化$E(V | P, \Theta_{\setminus V})$估计关节的运动速度。

数据: □

数据: □

结果: □

$r \leftarrow \max(I.width(), I.height())$

$P \leftarrow$ □_□□□□(E(P, | □=0, □)

$V \leftarrow \arg\min_V E(V \mid P, \Theta_{\setminus V})$

虽然□>1**做**

  $E \leftarrow$ □(□, $_0$

  为 □ = 1,□ **do**

    $\leftarrow$ □_ $best\ E(P, \mid P, \Theta_{\setminus st}), N, \square)$

    为□ = 1,□**做** ;

      $\bar{P}_{\square} \leftarrow P_{\setminus t} P$

      $\bar{\Theta} \leftarrow \{\Theta, \bar{s}^t_{\square}$

      $\bar{V} \leftarrow \arg\min_V E(V \mid \bar{P}_{k,\Theta k,\setminus V})$

    结束

    $(P, \Theta) \leftarrow \arg\min \bar{P}_{\square}\bar{\Theta}$ □(□), $\bar{\Theta}_{\square})$;

  结束

  如果 $E(P, \Theta) \setminus =$

    $r \leftarrow \frac{r}{2}$

  结束

算法1：在视频中构建人体姿态的迭代算法。

之后，反复进行直到收敛，所提出的算法选择视频序列中的一个任意帧，并改进与该帧相关的姿势和隐藏参数的估计。根据第5.2.1节，如果其他模型参数是已知的，可以确定任意一帧上的最佳姿势和大小参数。然而，对于选定的帧I_t，在上一步中得到的速度估计可能不是最优的。因此，拟议的算法建立了一组

N人体姿势的假设。对于每一种假设，都要确定速度参数的值，并选择导致能量函数减少最多的假设。

[48]中提出的在图像中构建人体姿势假设的方法取决于两个超参数：1）半径□，它决定了构建的假设之间的最小差异；2）要构建的假设数量N。半径r的数值越大，就越能得到差别很大的姿势假说，适合于工作的需要

在速度值确定不正确的情况下。小半径值

r允许在以前的解决方案附近对姿势进行局部搜索。因此，在拟议的算法中，这个参数随着时间的推移而减少。

由于能量函数$E(P, \Theta)$可以包含几个位置

在局部最小值的情况下，局部优化算法不能保证找到一个最优。确定性的算法1找到了一个局部的

能量函数□（P，Θ）的最小值。该算法的关键问题之一是

是逐帧更新一个人的姿势。例如，该算法不能离开一个局部最小值，即人的姿势大小的参数， 因而他的姿势在某个时间段的定义是不正确的。为了解决这个问题，我们开发了一种随机算法来完善结果。

5.2.3 随机算法

数据: □

数据 □

数据： $\Theta\tilde{}$ = (□ ~,

□ ~)

数据: □

结果: □

$H_0 \leftarrow$ (□ ~, □ ~)

对于□ = 1,□**做**

$i \sim$ sample(□step():)□

□□□ ←最小 $\frac{\square\tilde{}_{\square}(H_{\square})\ p}{\square\tilde{}(H)p_{Tr}}$, 1 ;

如果*rand*()< □□□□ **那么**

$_i H \leftarrow$ □;

否则

结束

$_i(P, \square) \leftarrow \arg_{\min Hi=(Pi,Si)} E(H$ _00

算法2：在视频中构建人体姿态的采样算法。

一般描述

考虑函数$E_{\backslash V}(P, \square) = \max_V E(P, \Theta)$，它只取决于人类的姿势和尺寸参数。计算这个最大值的方法在第5.2.1小节中描述。需要注意的是，函数$E_{\backslash V}(P, \square)$

是在一个离散的人类姿势和可接受的尺寸值的集合上定义的。因此，这个函数指定了在下列情况下的非标准化概率

视频中人的可能姿势集 □~(P, □) = exp(-E_V(P, □)。

所提出的算法使用该表示法，使用马尔科夫链方案（MCMC）从分布□~（P, □）中构建样本。采样构造使我们能够在长时间的采样过程中获得先例

从不同的分布模式□~（P, □）。因此，即使使用ini

从局部最优的角度来看，所提出的算法可以找到视频中人的姿势的最佳假设。具有最低能量值的假设被选为算法的结果。

此外，在本小节中，我所说的假设是指视频中人的姿势和尺寸参数的当前值：

□ = (□, □) 、 (5.19)

并以V（□）表示速度参数的相应值：

$$V(\square) = \arg\max_{\Theta)} E(P, \mid_{(\square,\square)=\square} \quad (5.20)$$

有许多方法可以从分布中构造出一个样本。最常见的是Gibbs和Metropolis-Hastings算法。由于Metropolis-Hastings算法允许更新一组模型状态参数（例如，一个人在几个帧中的姿势同时被更新），所以它被选作样本构建。算法2中介绍了所提出的在视频中寻找最佳人体姿势的随机方法的高层表示。

为了用Metropolis-Hastings算法建立一个样本，还必须定义一个过渡模型$p_{T\,r}$(H|□)。它描述了一种从先前的□构建新的姿势H假设的方法。根据Metropo

的算法

新假设被接受的概率为$c(H|\square)$：

$$\alpha(H|\square) = \min\left(1, \frac{\tilde{p}(H)p_T(H|\square)}{\tilde{p}(H)p_{Tr}(H|\square)}\right) \tag{5.21}$$

值得注意的是，Metropolis-Hastings算法并不要求计算
分布归一化常数□~。此外，为了构建
只要有一种方法可以计算出单个假设的概率值就足够了。这使得抽样方法也可以用来优化视频中人的姿势的更复杂的模型，包括全局外观参数，如衣服的颜色和人的身材。

过渡模式

为了通过Metropolis-Hastings算法从分布中构建样本，必须选择一个过渡模型$p_{T,r}(H|H)$。它描述了如何改变关节位置和尺度参数值以建立新的假设。

过渡模型的选择是构建能量优化算法$E_{\backslash V}$（P, □）的关键步骤。最简单的过渡模型是一个在可接受的低水平的集合上的均匀分布。
论文，与当前状态?无关。然后，根据（5.21），接受新假说的概率与这些假说的概率比成正比
□~（□）模型中的论文。然后，如果当前的假设H与局部最优匹配，Metropolis-Hastings算法就会拒绝大多数的假设H。

当过渡模型构建具有相似或更高概率的假说时，这个问题的影响就会减少。因此，拟议的过渡模型在每一步随机选择以下类型的假设变化之一：

1. 一个人的姿势的关节在某个时间段内的意外位移，在lo
   目前的位置是在他们当前位置的附近；

2. 用时间-1的值更新时间t的人的姿势和尺寸参数；

3. 使用姿态假设之一，根据条件模型

$E(P^t, s^t \mid P^{\setminus t}, \Theta_{\setminus s} \square)$

4. 在各时间点之间对人体姿势进行线性插值

$[\square_1, \square_2]$。

所提出的假设过渡模型可以在某个时间点t或者在某个片段$[t_1, t_2]$更新人体姿态。为了加快收敛速度，过渡模型更多的是选择时间时刻 $\square$，其中有最大的偏离拟议模型$E_{\setminus V}$（ $\square$ ），即在来选择速度参数的最佳值：

$$V(\square) = \arg\max_{\square} E(P, \mid_{(\square,\square)=\square}$$

$$\xi(t \mid \square) = (E) \quad (P^t, \theta) + \frac{1}{2}\psi)(P^t, P^{t-1}, \theta) + \psi(P^{t+1}, P^t, \theta) \Big(\Big)\square \Big|_{\substack{(\square,\square)=\square \\ \square(\square)}} \tag{5.22}$$

$$p(t) \propto \text{最大}_{\tau} \xi(\tau) - \xi(t)$$

$$t \sim p(t)$$

对段$[t_1, t_2]$的选择等同于对两个时间极限时刻t_1 和 t_2 的选择。

前三种类型的过渡模拟了局部搜索。第一种过渡类型在选定的时间点上产生关节位置P^t 和姿势大小s^t 的小偏离。偏差的步骤是根据以下规律选择的：

$$\begin{aligned} {}^t\acute{p}_{\square} &= p^t + \beta s^t \quad ; \\ {}^t s' &= s^t + \gamma \\ \beta &\sim \square(0, \beta \square \\ &)^{-1}{}_{2\times 2} \\ \gamma &\sim \square(0, \gamma_{\square})^{-1} \end{aligned} \tag{5.23}$$

第二种类型的转换是在考虑到前一帧姿势的当前假设的情况下，更新人在I_t 的姿势，同时增加一个小的

根据（5.23），正态分布噪声。这种类型的过渡相当于

倾向于尽量减少这些框架之间的运动模式：

$$(\underline{P}t,\ s^{t}) = \arg\min_{\square,\square} \left\{ \Psi(P)\square, \square \square_{1}^{-} (P^{t-1}, s^{t-1}) \in H \right. \tag{5.24}$$

之后，对结果应用第一种类型的过渡。

第三种类型的转换包括选择一个随机的姿势假设，根据模型$E(P^{t}, s^{t} | P^{\backslash t}, \Theta_{\backslash s}$ □)。5.2.1小节描述了构建合适的假设的算法。 可以注意到，只应用这种类型的转换并选择画面中人的姿势的最合理的假设，相当于算法1。然而，对于这一步，假设的选择是随机进行的，这使我们能够研究不同的分布模式

$(P, \square)$。

第四种过渡方式是在区间$[t_1, t_2]$上更新当前假设。为此，对关节位置和尺寸参数进行了线性插值。由于所提出的运动模型$E_T(P, \Theta)$描述了关节的直线均匀运动，这种假设的更新

可以减少配对势对能量函数值的贡献。

只有第一种类型的转换是可逆的，即在构建

从□~(□)分布中的一个样本，返回到先前状态的概率

$p_{Tr}(H|\square)$对于其他类型的过渡应该等于零。然而，由于目前的任务是找到视频中的最佳人体姿势，我使用的假设是：对于任何类型的过渡，$\underline{p}_{Tr}(H|\square) = p_{Tr}(H|\square)$。

所提出的模型的设置使选择第一种假设变化的概率平均比选择其他类型过渡的概率高50倍。因此，在视频中寻找最佳人体姿势的算法过程可以分为两个重复的步骤：

图5.4 - 测试序列的示例帧。从左到右分别是步行、投球、Lola1、Lola2序列的帧。

1. 从□˜(□)分布中构建一个样本（应用第一通类型）；
2. 选择取样的初始假设（应用不同类型的过渡）；

5.3 实验评估

5.3.1 抽样调查

在[48]中提出的样本上，对所提出的算法与基线进行了数值评估。图5.4显示了该样本的帧的例子。该样本由4个视频序列组成，分别是**Pitching**、**Lola1**、**Lola2**和**Walking**，它们的复杂性各不相同。**Pitching**、**Lola1**和**Lola2**视频序列包含摄像机的运动。**投球的**视频序列包括焦距的变化。大多数序列包含一个人在画面中，但在**Lola2**中，有些画面包含几个人。

(a) (б)

图5.5 - 姿势可视化为一组部分（片段）。(a)专家对帧上姿势的标记，(b)建议方法的结果。

专家的标记以蓝色显示。根据PCP标准，绿色部分的标记是正确的，红色部分的标记是错误的。

5.3.2 比较结果

使用[64]中提出的PCP（正确确定部分的比例）标准来比较这些算法。然而，这个标准有一个重要的缺点。它将一个姿势解释为一组部件或片段（图5.5a），并独立地评估每个部件的正确性。如果一个部分的两端到其正确位置的距离相差不超过一半，则认为该部分的位置定义正确。

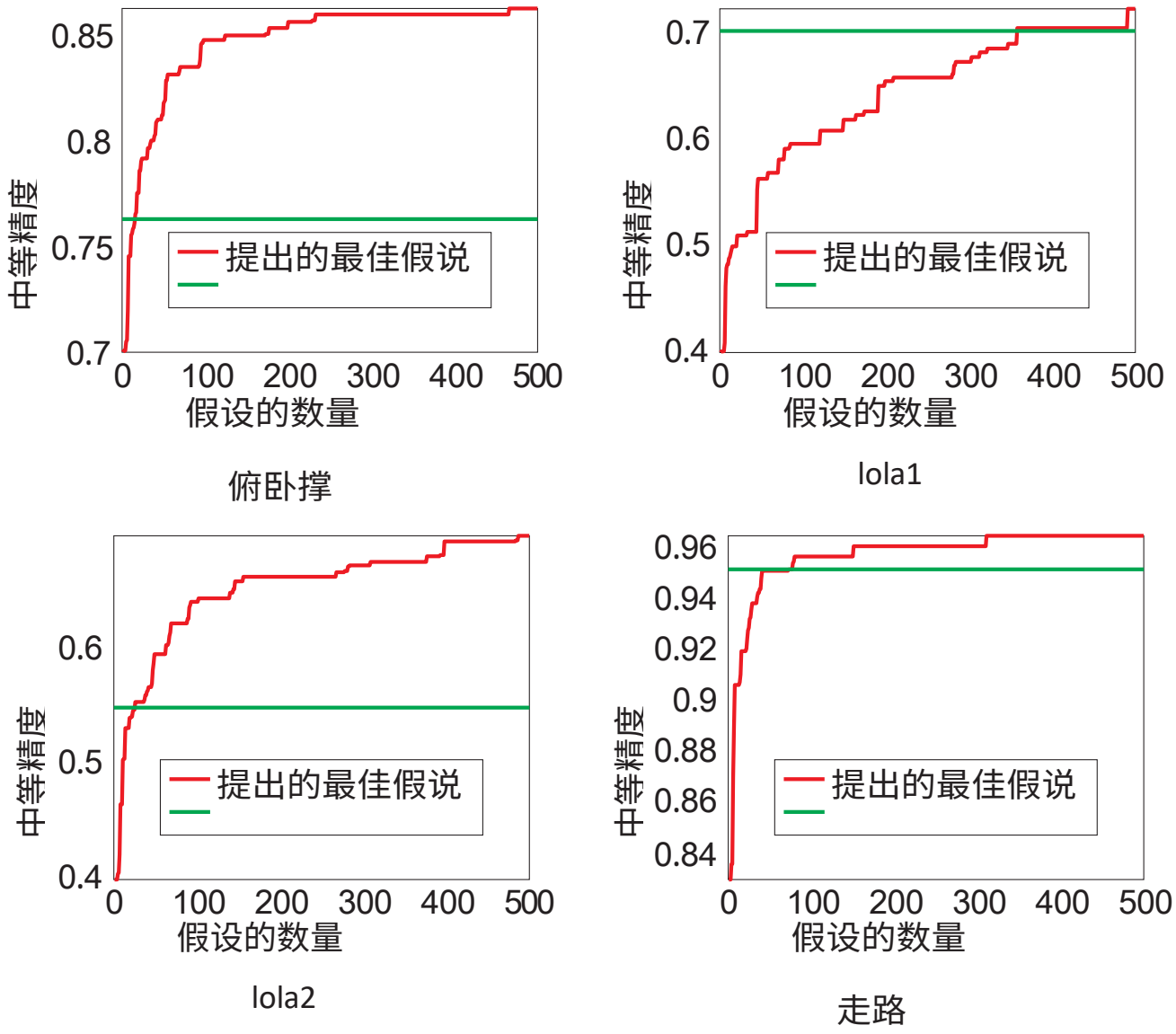


图5.6--算法性能的最高估计值（红色曲线）。

[48]作为每帧假设数量的函数。拟议算法的结果显示为绿色。

在专家的标记中，该段的长度。由于两部分的尺寸不同，可以认为关节位置在一部分中被正确确定，而在另一部分中则不正确（图5.5 b）。然而，由于PCP标准仍然被用来评估方法的质量，我用它来进行比较。

为了进行公平的比较，我不使用伴奏的结果在初始化，即在初始时间，人的姿势在图像中的位置被假定为是均匀的。此外，算法参数α_p ，α_v 和 σ_s 在整个采样序列中是固定的。结果

比较结果见表7。用于比较的参考方法是[48]。

表7 - 所提方法与基础方法的比较结果。比较是根据正确定位的零件（PCP）的平均数量进行的。基础方法的结果取自[48]。

算法	走路	俯卧撑	lola1	lola2
基本的	**0.950**	**0.797**	0.670	0.500
	0.950	0.763	**0.685**	**0.545**

在最困难的序列**Lola1**和**Lola2**上，所提出的方法优于基线方法。该算法通过使用方向和速度信息，能够解决在**Lola2**的画面中有几个人的不确定性。

对于算法来说，最容易确定视频中人的姿势的是**步行**视频序列。它呈现的是一个人均匀地行走。基本算法和提议的算法在这个序列上的结果没有差别，这与框架上使用的人类姿势模型的限制有关[37]。

当对**投球**序列进行测试时，与基线相比，提议的方法显示出较低的PCP标准值。这个例子被证明是对拟议算法最具挑战性的，因为所介绍的运动员的运动模式与关节的线性运动模式不一致。

为了估计所使用的人类模型的可接受能力[48]，我估计了所构建的人类姿势假设的质量上限对其数量的依赖性（图5.6）。为了估计每一帧的上限，根据PCP标准在构建的假说中选择最佳假说。结果显示，在**lola1**和**行走**序列上，所提出的方法选择的姿势接近于最佳姿势。对于**投球**和**lola2的**序列，由于复杂的人体肢体运动模式，获得的解决方案与构建的上界有很大的不同。

摄像机的运动。**步行**和**Iola1**序列包含与监控场景相对应的运动类型。

[48]中考虑的许多测试序列对所开发的算法来说是困难的，因为由于摄像机的 移 动 和变焦，图像中人体关节运动的均匀性假设被违反。因此，在静态摄像机的视频监控场景中，也对人体姿势检测的准确性进行了评估。使用第六章中描述的自动工具，对TownCentre序列的一个子样本进行了标记[26]。由此产生的分区包含了视频中2000个人的姿势。另外，在图像中不可见的人体骨架的关节，在标记中也被标记为不可见。在评估该算法的质量时，只对专家标记中两个关节都被标记为可见的身体部位的子集进行了比较。在构建的样本上，提议的算法
在PCP指标上，节奏的质量为0.673，而基线的质量为0.586。
姿势检测的质量通过使用关于人体骨骼关节运动速度的信息得到了改善。事实上，在所考虑的场景中，大多数人的运动都接近于均匀。因此，所提出的模型可以更好地预测过渡到下一帧时人体关节的位置，甚至在人体关节重叠的情况下也能恢复其位置。

5.4 总结

在这一章中，提出了一种在视频序列中估计人体姿势的算法，该算法同时考虑了帧中人体每个关节的位置和速度。事实表明，所提出的基于一组零件模型的姿势估计算法是

是所提算法的一个特殊情况。通过使用联合速度信息，所提出的算法提高了静止摄像机场景下的姿势检测的准确性。

该章的结果发表在[65]。

第6章。软件实施

6.1 一般描述

第2、3、4和5章中提出的算法被作为独立的模块来实现（图6.1）。在其基础上，我开发并实施了两个软件工具。第一个解决了自动检测和跟踪所有行人以及他们在视频序列中的姿势变化的问题。第二个软件工具提供了一个在视频序列中检测人的姿势的自动工具。

6.2 陪同人们并确定他们在视频中的姿态

第2、3、4和5章中提出的算法是开发的软件工具的基础，用于引导人们并识别他们在视频中的姿势。图6.2显示了所开发的应用程序的各组成部分之间的互动示意图。

输入是一个帧序列，输出是一组人的姿势，每个人都有一个路径标识符。我假设输入帧序列的频率至少为每秒25帧。

在第一阶段，图像由头部检测器进行处理。所用的检测器有两种操作模式。在初始化模式下，视频序列中的第一帧被处理。假设摄像机的位置是未知的，检测器与参照物相匹配，也就是说，在金字塔的每一级都处理整个帧。在检测器被应用于20帧之后，估计的摄像机位置被认为是可靠的、

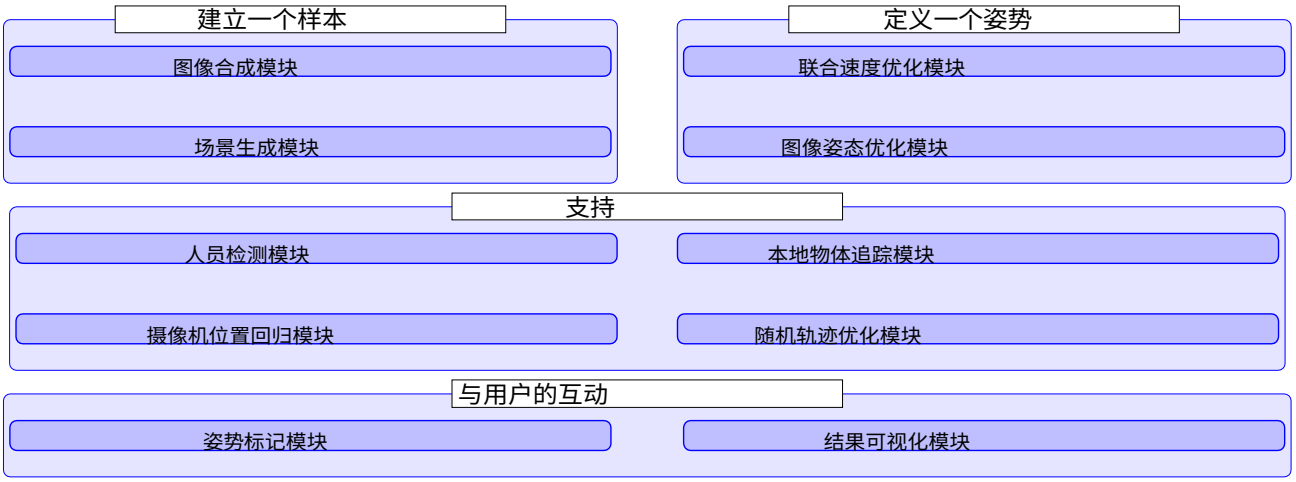


图6.1 - 论文工作中设计和实施的模块

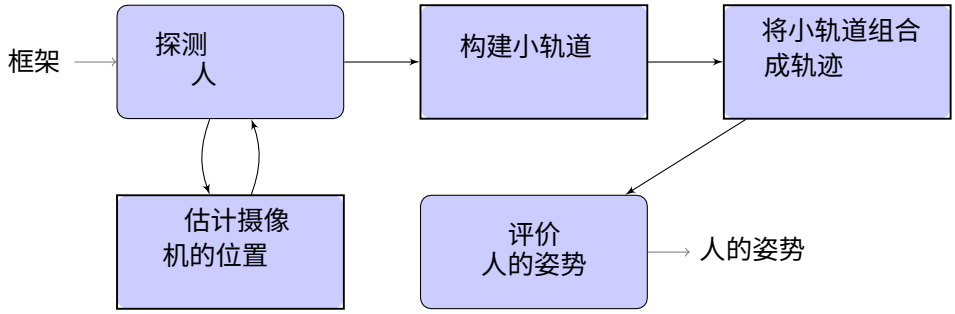


图6.2 - 构成人体跟踪和姿势检测软件工具的各部分之间的互动示意图

并使用第三章中提出的算法。为了加快数据处理速度，检测器被应用于每5帧，即每秒5次。然而，一个不正确的摄像机位置将无法在输入的视频序列中找到人。因此，基本的检测器被应用到视频序列中，每秒一次，其结果被用来完善摄像机的位置。检测结果和视频序列中的帧被送到下一阶段，用于构建跟踪单元。

追踪器的构建和将追踪器合并为轨迹的工作如第四章所述。轨迹是由一个包含最后200帧（8秒）的小轨道的临时窗口建立的，检测的标签（轨道ID）是为中心的

这个时间窗口的帧。在构建轨迹时，每一帧都要进行10,000次的随机优化迭代。

视频序列姿势检测模块接收一个视频序列帧和对该帧的陪衬结果。由于跟踪的结果与4秒前处理的帧相对应，该模块对最后收到的帧进行缓冲。对于每个人的轨迹，其图像从处理过的帧中切出，并通过包含它的整个视频片段进行姿势搜索。模块操作的结果是被处理帧上的人的姿势和相应轨迹的标识符。

用来实现模块的语言是C++。为了加快工作进度
该实现是用CUDA C++编写的，使用CUDA流对连续的图像进行流水线处理。姿势检测器中使用了特征库来实现矩阵操作。

这个软件工具在视频分析技术有限公司被用来计算通过警报线的人。通过使用摄像机校准信息，它允许你增加人流检测器应用于输入视频帧的频率。这使得在视频流处理环境中提高人员计数的准确性成为可能。

6.3 自动构建人类姿势的专家标记

第5章描述了在视频序列中自动检测人体姿势的拟议算法。然而，如结果所示，它不能准确地确定每一帧中关节的位置。为了开发视频中人类姿势检测的算法，有必要解决

构建包含每一帧人的姿势标记的视频序列样本的自动化任务。为了解决这个问题，基于第五章提出的姿势检测算法，构建了一个在每一帧中标记人体姿势的自动化软件工具。

拟议的工具允许用户指定两种类型的附加信息：

1. 在当前图像中，正在使用的骨架的关节是否可见。
2. 在当前帧中限制关节位置的矩形的位置

然后，这些额外的信息被用作帧中允许的人类姿势集的额外约束。第一类信息可以用一个二进制变量来描述

^{t} 只有当第i个关节在图像中可见时，其值为1。

$_{t}$。第二类信息是由矩形b^{t} 上的位置描述的。

帧$_{t}$，里面是一个关节。与前一章一样，M^{t} 和B^{t} 表示人类姿势的所有关节在帧$_{t}$ 中的可见性和位置的限制矢量。我进一步认为，用户没有指定额外信息的关节是可见的，它们的约束是

下面是一个例子，说明如何在视频序列中搜索一个人的姿势，同时考虑到建议的约束条件。下面你将看到如何在视频序列中搜索一个人的姿势，同时考虑到建议的约束条件。

视频序列中的姿势检测问题的解决方案被简化为函数（5.1）的无条件优化。考虑到额外的信息，优化问题变得如下：

$$\min \sum_{=1}^{} {}_{I} E(P^{t}, \theta, M^{t}) + \sum_{=1}^{} {}_{T} E(P^{t+1}, P^{t}, \Theta) \rightarrow \quad (6.1)$$

$$^{t}p \in b^{t}$$

需要注意的是，引入的约束条件并不影响使用的运动模型E_T。下面我将说明所述的约束条件是如何改变模型的优化函数和找到最佳状态的方式的。

关于关节在图像帧中是否可见的信息只影响优化函数的分量E_I。 如上所述，这

组成部分是两个因素的总和：$\phi_i(p^t, s^t)$ - 探测器反应

的图像和ψ^s (p^t, p^t, s^t) - 对相互定位的限制。

关节。如果在处理过的框架中看不到第i个关节，那么单数的

公式（5.2）中的潜力j_i (p^t, s^t) 不能包含关于polo

使用这个因素也会导致不正确的姿态检测。此外，使用这个因素会导致对人类姿势的不正确判断，因为它在优化函数中引入了额外的噪音。因此，我使用了因子 $^t_i\ \phi m(p^t, s^t)$, 如果接头处被标记为可见，其视觉就等于原始视觉

和0，然后开始。这样就可以避免根据图像中检测器的反应来调整人的姿势，而只根据相邻关节的位置来选择关节位置。

对关节在给定矩形内的位置的限制 b^t

需要使用约束性优化方法来确定人体姿势。然而，优化问题（6.1）可以使用拉格朗日不确定集方法以等效形式重写：

$$\underset{P\text{s.t.}}{\sum_{t=1}^{T}} (E_I(P^t, \Theta, M^t) + E_L(P^t, B^t) + \sum_{t=1}^{T-1} E_T(P, {}^{t+1}t, \Theta)) \to min_{P,b}$$

$$_L(P^t, B^t) = \sum_i \delta(p_i^t, b^t) \qquad (6.2)$$

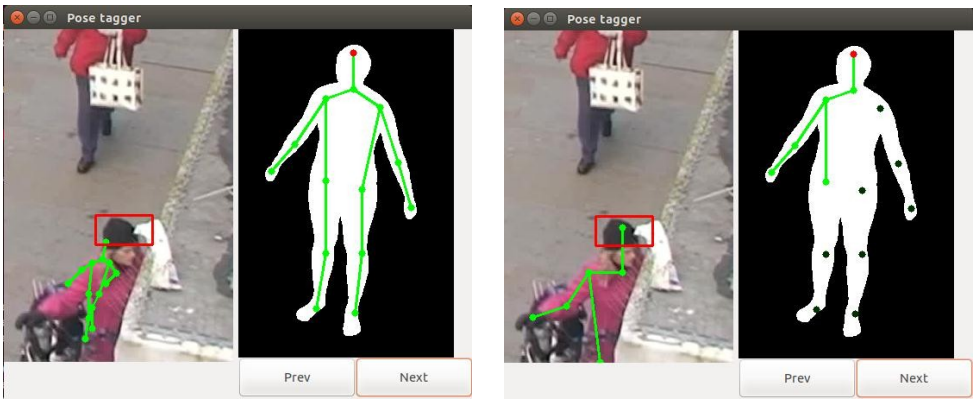
$$\delta(p_i^t, b_i^t) = \begin{cases} 0, & \text{如果} p^t \in b \\ +\infty, & \text{否则} \end{cases}$$

需要注意的是，拟议的额外因素$\delta(p^t, b^t)$

是单数的。因此，该模型在受到用户施加的额外约束后，等同于基础模型，其中不是单数的

联合定位因子$j_i(p^t, s^t)$用于$j'(p^t, s^t, m^t, b^t) = {}^t_i\ \phi m(p^t, s_t) + \delta(p^t, b^t)$。寻找能量最小值$E(P^t, \Theta)$的算法不取决于

录像不是一个简单的视频。因此，在考虑到引入的约束条件的情况下，不加修改地应用它们来寻找视频中的最佳人体姿态。



(a) (б)

图6.3 - 用户界面的例子和一个特定的人体姿势（a）之前和（b）在指定图像中可见的关节之后。

标记工具已经用python2实现。所开发的用于确定图像中人的姿势的算法被作为第三方模块使用boost::python库插入。wxWidgets GUI构建库通过wxPython模块用于用户交互。

图6.3显示了该应用程序的用户界面。屏幕的右侧显示了一个人的模型和该人身上的关节位置。在关节图像上使用鼠标左键，用户可以指定其可见性参数的值。使用鼠标右键，用户可以选择另一个关节（在模型上用红色标记），并突出显示图像中可以定位的区域。图6.3(a)显示了当一个关节被标出并且所有关节都被认为是可见的时候，姿势检测算法的结果。在图中

6.3 (b)显示了用户在图像中指定缺失的关节后的结果。用户界面上的按钮允许用户在指定人的轨迹内向前和向后移动。

在用户输入额外的信息后，该算法在视频序列的所有帧中更新人的姿势，使用之前的决定作为初始化。

与在每一帧中完全手动标记一个人的姿势相比，拟议的解决方案有几个优点：

– 来纠正算法错误，在许多情况下，用户只需要为框架中的少量人体骨骼关节标记允许的区域；

– 如果一个人的姿势在一段视频中被错误地检测到，在某些情况下，用户只需要在一帧中纠正结果，而不是标记整个片段；

– 在对一个序列的帧进行独立的手工打标时，会出现帧间关节随机 "抖动 " 的影响，这不是由于被打标者的运动，而是由于打标的不准确。建议的打标方法通过使用平滑系数E_T，减少这种影响。

这个软件工具已被实施，作为关于 "中国 "的工作的一部分。

RFBR项目16-29-09612 office_m "研究和开发视频监控数据中通过步态、手势和肤色进行生物识别的方法"。该项目用于建立一个视频监控数据的参考集合，其中每一帧都有一个人的已知姿态。 生成的集合用于解决通过步态识别人的问题。

总结

本论文研究中出现了以下主要发现：

1. 提出了一种原创的方法，根据人类检测的结果，确定静态摄像机在场景中的位置和方向。
2. 对于静态摄像机拍摄的视频序列，已经开发了一种人类追踪算法，利用摄像机的位置和方向来过滤掉检测器的假阳性。
3. 提出了一种在视频序列中估计人类姿态的算法，该算法同时考虑了视频序列帧中人体每个关节的位置和速度。
4. 在所提出的算法的基础上，开发了一个用于自动跟踪和确定视频序列中人类姿势的软件包和一个用于在每一帧中构建人类姿势的专家标记的自动化软件工具。

所提出的算法有可能按照以下思路进一步发展：

– 通过使用整个人的姿势检测的结果来估计摄像机的位置和方向以及焦距；
– 使用基于图像的人类再识别算法来可靠地匹配轨迹中的人；
– 在人体姿态变化的平滑度因素中加入对输入视频序列的依赖性。

参考资料清单

1. *Wang X.* 智能多机位视频监控：回顾 / X.Wang // Pattern recognition letters.- 2013.- T. 34, № 1.- C. 3-19.
2. *Caprile B.*使用消失点进行相机校准 / B．Caprile、V.Torre // International journal of computer vision.- 1990.- T. 4, № 2.- C. 127-139.
3. 单一图像的同时消失点检测和相机校准 / B. Li [等] / 国际视觉计算研讨会。Li [et al] // 视觉计算国际研讨会。- Springer.2010.- C. 151-160.
4. *Liu J.* Surveillance camera autocalibration based on pedestrian height distributions / J. Liu, R. T. Collins, Y. Y..Liu, R. T. Collins, Y.Liu // 英国机器视觉会议（BMVC）。- 2011.
5. 通过对地平面上移动的人的观察实现两台摄像机的精确自校准 / T．Chen [et al.] // Advanced Video and Signal Based Surveillance, 2007.AVSS 2007.IEEE会议。- IEEE.2007.- C. 129-134.
6. *Pflugfelder R*.跨越两个遥远的自校准相机的人员追踪/ R.Pflugfelder, H. Bischof // Advanced Video and Signal Based Surveillance, 2007.AVSS 2007。IEEE会议。- IEEE.2007.- C. 393-398.
7. 自动推断未经校准的不相干监控摄像机的几何摄像机参数和摄像机间的拓扑结构 / R. J. den Hollander [等] / SPIE Security+Defense.J. den Hollander [et al] // SPIE Security+ Defence.- 国际光学、光子学学会。2015.- pp. 96520d-96520d.

8. 体育转播中移动人群的Ptz摄像机网络标定/

J.Puwein [et al] // 计算机视觉的应用（WACV），2012年IEEE研讨会。-IEEE.2012.- C. 25-32.

9. *Dubsk´a M.* 用于交通理解的自动相机校准。 / M.Dubsk´a, A. Herout, J. Sochor // BMVC.- 2014.

10. *海岩D.*将物体纳入视野 / D.Hoiem, A. A. Efros, M. Hebert // International Journal of Computer Vision.- 2008.- T. 80, № 1.- C. 3-15.

11. *Viola P.*利用简单特征的提升级联进行快速物体检测 / 冯小刚 P.Viola, M. Jones // Computer Vision and Pattern Recognition, 2001.CVPR 2001.2001年IEEE计算机学会会议论文集。T.1.- IEEE.2001.- P. I-511.

12. *Bourdev L.* Robust object detection via soft cascade / L． Bourdev, J. Brandt // 2005年IEEE计算机学会计算机视觉和模式识别会议（CVPR'05）。T.2.- IEEE.2005.- C. 236-243.

13. *Doll´ar P.* The Fastest Pedestrian Detector in the West./ P. Doll´ar, S. Belongie 、 P.Perona // BMVC.T.2.- Citeseer.2010.- C. 7.

14. *Doll´ar P.* Crosstalk cascades for frame-rate pedestrian detection / P. Dollar R.Appel, W. Kienzle // Computer Vision-ECCV 2012.- Springer, 2012.- C. 645-659.

15. *Krizhevsky A.* Imagenet classification with deep convolutional neural networks / A． Krizhevsky, I.Sutskever, G. E. Hinton // 神经信息处理系统的进展。- 2012.- C. 1097-1105.

16. *Simonyan K.*用于大规模图像识别的超深度卷积网络 / K. Simonyan, A. Zisserman / arXiv preprint arXiv:1409.1556.Simonyan, A. Zisserman // arXiv preprint arXiv:1409.1556.- 2014.

17. 用于图像识别的深度残差学习 / K. He [et al] / arXiv preprint arXiv:1512.03385.He [et al] // arXiv preprint arXiv:1512.03385.- 2015.

18. 重新思考计算机视觉的初始架构 / C. Szegedy [等] / arXiv preprint arXiv:1512.00567.Szegedy [et al] // arXiv preprint arXiv:1512.00567.- 2015.

19. 用于精确物体检测和语义分割的丰富特征层次/R.Girshick [et al.] // IEEE 计算机视觉和模式识别会议论文集。- 2014.- C. 580-587.

20. *Girshick R.*快速r-cnn / R． Girshick // IEEE国际计算机视觉会议论文集。- 2015.- C. 1440-1448.

21. 更快的R-CNN：利用区域建议网络实现实时物体检测 / S. Ren [等] / 《神经信息处理系统研究进展》。Ren [et al] // 神经信息处理系统的进展。- 2015.- C. 91-99.

22. 用于交互式计算机图形的计算机视觉 / W. T.T.Freeman [et al] // IEEE Computer Graphics and Applications.- 1998.- T. 18, № 3.- C. 42-53.

23. *Isard M.*视觉跟踪的凝结-条件密度传播 / 凝结-条件密度传播。M.Isard, A. Blake // International journal of computer vision.- 1998.- T. 29, № 1.- C. 5-28.

24. 追踪的良好特征 / J． Shi [et al] // Computer Vision and Pattern Recognition, 1994.Proceedings CVPR'94., 1994 IEEE Computer Society Conference on.- IEEE.1994.- C. 593-600.

25. *Kolsch M.* Fast 2d hand tracking with flocks of features and multi-cue integration / M． Kolsch, M. Turk // Computer Vision and Pattern Recognition Workshop, 2004.CVPRW'04.Conference on.- IEEE.2004.- C. 158-158.

26. *Benfold B.*实时监控视频中稳定的多目标跟踪/ B.Benfold, I.Reid // 计算机视觉和模式识别（CVPR），2011年IEEE会议。- IEEE.2011.- C. 3457-3464.

27. (MP)2t: 多人多部位追踪器 / H. Izadinia [et al] // 欧洲计算机视觉会议。-

Springer.2012.- C. 100-114.

28. *Yoon J. H.* Visual tracking via adaptive tracker selection with multiple features / J. H. Yoon.H. Yoon, D.Y.Kim, K.-J.Yoon // European Conference on Computer Vision.- Springer.2012.- C. 28-41.

29. *Choi W.*多目标跟踪和集体活动识别的统一框架 / W. Choi, S. Savarese / 欧洲计算机视觉会议。Choi, S. Savarese // European Conference on Computer Vision.- Springer.2012.- C. 215-230.

30. *Leal-Taix´e L.* Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker / L． Leal Taix´e, G. Pons-Moll, B. Rosenhahn // Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on.- IEEE.2011.- C. 120-127.

31. *Butt A.A.*通过拉格朗日松弛对最小成本网络流的多目标跟踪 / A． A.Butt, R. T. Collins // IEEE计算机视觉和模式识别会议论文集.- 2013.- C. 1846-1853.

32. *Andriyenko A.*用于多目标跟踪的离散-连续优化/ A.Andriyenko, K. Schindler, S. Roth // Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. - IEEE. 2012. - C. 1926-1933.

33. *Milan A.* Detection-and trajectory-level exclusion in multiple object tracking/多目标跟踪中的检测和轨迹级排除。A.Milan, K. Schindler, S. Roth // IEEE计算机视觉和模式识别会议论文集。- 2013.- C. 3682-3689.

34. 视觉物体跟踪的多假设运动规划 / H. Gong [et al] // 2011年国际计算机视觉会议。- IEEE.2011.- C. 619-626.

35. 耦合检测和数据关联的多物体跟踪 / Z.Wu [et al.] // 计算机视觉与模式识别（CVPR），2012 IEEE会议。- IEEE.2012.- C. 1948-1955.

36. 追踪还是检测？ 一个优化选择的集合框架 / X. Yan [等] / 欧洲计算机视觉

会议。Yan [et al.] // 欧洲计算机视觉会议。- Springer.2012.- C. 594-607.

37. *Yang Y.*用灵活的部件混合物进行铰接式姿势估计 / Y. Yang.Yang、D.Ramanan // 计算机视觉和模式识别（CVPR），2011年IEEE会议。- IEEE.2011.- C. 1385-1392.

38. *Felzenszwalb P.*训练有素的、多尺度的、可变形的零件模型 / 辨别力强的、多尺度的、可变形的零件模型
P.Felzenszwalb, D.McAllester, D. Ramanan // Computer Vision and Pattern Recognition, 2008.CVPR 2008。IEEE会议。- IEEE.2008.- C. 1-8.

39. *Pirsiavash H.* Steerable part models / H. Pirsiavash, D. Ramanan / Computer Vision and Pattern Recognition (CVPR).Ramanan // 计算机视觉和模式识别（CVPR），2012年IEEE会议。- IEEE.2012.- C. 3226-3233.

40. Poselet conditioned pictorial structures / L． Pishchulin [et al] // IEEE计算机视觉和模式识别会议论文集。- 2013.- C. 588-595.

41. 解析被遮挡的人 / G． Ghiasi [et al] // IEEE计算机视觉和模式识别会议论文集。- 2014.- C. 2401-2408.

42. *Chen X.*通过灵活的组合来解析被遮挡的人 / X. Chen.Chen、A.L. Yuille // IEEE计算机视觉和模式识别会议论文集。- 2015.- C. 3945-3954.

43. 为识别建立实例外观模型--我们能比EM做得更好吗？ / A.Chou [et al.] // International Workshop on Structured Prediction: Tractability, Learning, and Inference.- 2013.

44. *Finley T.*当精确推断难以实现时，训练结构性SVMs / T.Finley, T. Joachims // 第25届国际机器学习会议论文集。- ACM.2008.- C.

304-311.

45. 可变形部件模型是卷积神经网络 / R． Girshick [et al] // IEEE计算机视觉和模式识别会议论文集。- 2015.- C. 437-446.

46. *Chen X.*通过与图像相关的成对关系的图形模型进行铰接姿势估计 / X. Chen, A. L. Yuille / 《神经信息处理系统进展》。Chen, A. L. Yuille // 神经信息处理系统研究进展.- 2014.- C. 1736-1744.

47. 卷积网络和图形模型的联合训练，用于人体姿势的估计 / J. J.J.J.Tompson [et al.] // 神经信息处理系统的进展。- 2014.- C. 1799-1807.

48. *Park D.* N-best maximal decoders for part models / D． Park, D.Ramanan // 2011年国际计算机视觉会议。- IEEE.2011.- C. 2627-2634.

49. *托舍夫A.*Deeppose：通过深度神经网络进行人体姿势估计 / A.Toshev, C. Szegedy // IEEE计算机视觉和模式识别会议论文集。- 2014.- C. 1653-1660.

50. 使用卷积网络进行有效的物体定位 / J. Tompson [et al.Tompson [et al] // IEEE计算机视觉和模式识别会议论文集。- 2015.- C. 648-656.

51. *Bulat A.*通过卷积部分热图回归进行人的姿势估计/ A.Bulat, G. Tzimiropoulos // European Conference on Computer Vision.- Springer.2016.- C. 717-732.

52. 为三维人体建模构建统计形状空间 / L. Pishchulin [et al.Pishchulin [et al.- 2015.- 三月。

53. *Prisacariu V*. fastHOG - a real-GPU implementation of HOG：技术报告 / V. Prisacariu, I.Prisacariu, I.Reid ; 牛津大学工程科学系.- 2009.- № 2310/09.

54. *Kingma D.*Adam: A method for stochastic optimization / D． Kingma, J. Ba

// arXiv preprint arXiv:1412.6980.- 2014.

55. *Thirde D.*PETS2006挑战概述/D.Thirde, L. Li、F.Ferryman // Proc.第九届IEEE跟踪和监视性能评估国际研讨会（PETS 2006）。- 2006.- C. 47-50.

56. *Shalnov E.*卷积神经网络用于从物体检测中估计摄像机的姿势。/ E.Shalnov, A. Konushin // International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.- 2017.- T. 42.

57. Caffe: 快速特征嵌入的卷积架构 / Y. Jia [等] // arXiv preprint arXiv:1408.5093.Jia [et al] // arXiv preprint arXiv:1408.5093.- 2014.

58. *Shalnov, E.*使用舞台几何学来提高检测器精度 / E. V. Shalnov, A. S. Konushin // 软件产品和系统。- 2017.- T. 30, № 1.- C. 106-111.

59. *Tomasi C.* Detection and tracking of point features / C． Tomasi, T. Kanade.- 1991.

60. *Fulkerson B.*用超级像素邻域进行类别分割和物体定位 / B. Fulkerson, A. Vedaldi, S. Soatto / Computer Vision, 2009 IEEE 12th International Conference.Fulkerson, A. Vedaldi, S. Soatto // Computer Vision, 2009 IEEE 12th International Conference on.- IEEE.2009.- C. 670-677.

61. *Bernardin K.* Evaluating multiple object tracking performance: the CLEAR MOT metrics / K． Bernardin, R. Stiefelhagen // EURASIP Journal on Image and Video Processing.- 2008.- T. 2008, № 1.- C. 1-10.

62. *Shalnov E.*对基于MCMC的视频追踪算法的改进 / 冯小刚 E.Shalnov, V. Konushin, A. Konushin // Pattern Recognition and Image Analysis.- 美国，2015年。- Vol.25.- P.532--540.

63. *Shalnov E.V.*基于MCMC的视频追踪算法的改进 / E.V.Shalnov, V. S. Konushin, A. S. Konushin // 第11届国际模式识别和图像分析会议：新信息技术（PRIA-11-2013）。萨马拉，2013年9月23-28日。会议论文集。Vol.2.- IPSI RAS Samara, 2013.- P.727--730.

64. *Ferrari V.*逐步减少人类姿势估计的搜索空间/

V.Ferrari, M. Marin-Jimenez, A. Zisserman // Computer Vision and Pattern Recognition, 2008.CVPR 2008。IEEE会议。- IEEE.2008.- C. 1-8.

65. *Shalnov E.*通过MCMC抽样对视频中的人的姿势进行估计 / E. Shalnov, A. Konushin / 第五届国际图像挖掘研讨会论文集。Shalnov, A. Konushin // 第五届国际图像挖掘研讨会论文集。理论与应用.- 2015.- P.71--79.

数字列表

表格列表