密级：_____

中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

基于 RGB-D 数据的多目标跟踪方法研究

作者姓名：高　山

指导教师：焦建彬　教授　　中国科学院大学

学位类别：工学博士

学科专业：计算机应用技术

研 究 所：中国科学院大学电子电气与通信工程学院

二零一六 年 四 月

Multiple Target Tracking based on RGB-D Data

By

Shan Gao

A Dissertation Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Doctor of Computer Application Technology

School of Electronics, Electrical and Communication Engineering

April, 2016

中国科学院大学直属院系

研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：高山

日　　期：2016.5.20

中国科学院大学直属院系

学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密的学位论文在解密后适用本声明。

作者签名：高山　　导师签名：[illegible]

日　　期：2016.5.20　　日　　期：2016.5.21

摘 要

多目标跟踪问题是计算机视觉领域的一个重要问题，涉及模式识别与智能系统、传感器、图像处理、统计与机器学习等多个领域的相关技术。多目标跟踪的主要任务是对视频序列中的多个目标进行关联，并使其身份标识在序列中保持不变。经过几十年的深入研究和发展，多目标跟踪技术已在智能交通系统、智能监控系统、机器人导航、人机交互、生物医学研究等多个领域广泛应用。近十年来，随着 RGB-D 传感器技术的成熟，基于 RGB-D 数据的多目标跟踪技术取得显著进展。但是，在复杂的实际应用场景中，多目标目标跟踪仍然面临频繁的目标遮挡、剧烈的光照变化、动态背景的切换等问题与难题。有效解决这些问题与难题才能进一步提升相关应用系统的性能。本文针对复杂背景中基于 RGB-D 数据的多目标跟踪问题进行了深入研究，内容与成果包括：

1）针对多目标跟踪的实时性要求，提出了一种深度结构关联模型（Depth Structure Association, DSA）。模型将场景中的多目标划分到不同的深度链状结构中进行三维分析。利用整数规划中的多维数据分配问题对多目标之间的数据关联进行建模。在面对多目标跟踪过程中的遮挡问题时，链状结构利用深度值对目标匹配代价进行重新加权，使得目标在场景的不同位置更具有区分性。RGB-D 数据集上的验证结果表明，DSA 模型可以在交通场景的多目标跟踪问题上实现实时处理。

2）针对多目标跟踪的遮挡问题，提出了一种分层图模型（Layered Graph Model，LGM）。模型将多目标跟踪与图论中的图模型相结合，将传统的基于离散-连续的轨迹级（tracklet-level）目标关联方式，提升到层级（layer-level）。LGM 利用深度数据构建目标在层内以及层间的图模型，利用目标之间的位置、运动和外形信息构造关联相似度。LGM 利用自身的分层关系，在层内利用加入虚拟点的策略解决交通场景中多目标之间的复杂遮挡问题。

3）针对多目标成组运动问题，提出了一种拓扑能量最小化（Topology Energy Minimization, TEM）模型。在多目标跟踪模型中引入行人的社会属性，利用行人组内组外的目标行人的运动相似度，进行能量形式的建模。目的是使组内的行人相似度尽可能的高，组间的行人相似度尽可能的低，并以“拓扑能量最小化”方式进行模型求解。在拓扑的变化过程中，TEM 通过加入虚拟点，将被遮挡的目标通过组内位置估计进行有效定位，减少了跟踪过程中的目标丢失次数。模型还对行人目标的 RGB-D 特征进行了创新，提出了更适合多目标跟踪的 RGB-D 特征。

4）针对行人组动态变化问题，提出了一种基于动态拓扑图模型（Graphical Social Topology，GST）。GST 模型将组内的行人看作是图中的节点元素，通过图中边的连接方式探索组内成员在行走过程中的运动的变化。模型通过离线与在线学习相结合的方式，对组的典型拓扑进行离线学习，并在在线跟踪中与行人组进行拓扑结构匹配。在组的运动过程中模型通过组的初始化、更新、合并和分裂的动态变化完成对行人组的跟踪。最后 GST 模型利用线性规划的方式完成组内行人的身份确认，得到目标的完整轨迹。将 GST 模型在 RGB-D 数据集和 RGB 数据集（MOT Benchmark）中进行测试，表明了其优越的性能。

本文还介绍了作者搭建的 RGB-D 数据采集平台以及在该平台上采集的多目标跟踪数据集。该数据集目前已公开，供多目标跟踪研究者使用。

关键词：多目标跟踪，RGB-D 数据，行人跟踪，数据关联，图模型，组模型

Abstract

Multiple Target Tracking (MTT) is one of the most important branches in the computer vision, which combines advanced technologies and research achievements in pattern recognition & intelligent system, sensor technology, image processing, statistics, machine learning and other relative fields. The goal of MTT is tracking each target in successive frames, keeping the same identity in sequences. After tens of years of research and development, MTT has been greatly advanced and widely applied in many practical systems including intelligent transportation, video surveillance, robotics perception, and human computer interface. With the development of RGB-D sensors, the MTT methods based on RGB-D data further improve the tracking performance. Despite such advances, however, the MTT problem is far from being solved, as some objects could be falsely tracked or missed. This is particular common in many complex wild scenes where frequent occlusions, strong illumination variation and dynamic backgrounds exist.

To solve the above mentioned challenging problems, this dissertation proposes four new MTT models and tracking methods. The contributions of this dissertation are summarized as follows.

1) Proposing a Depth Structure Association (DSA) model, which investigates the RGB-D data in data association framework, divides targets in different depth structures. DSA encodes the depth information in a chain structure, which is used together with appearance and motion information to address target occlusion issues in outdoor scenes. Additionally, the use of DSA has the advantages of regulating a much smaller solution space, reducing the computational complexity. DSA model can significantly reduce targets mismatch and tracking failure for long term occlusions.

2) Proposing a Layered Graph Model (LGM). The motivation is to investigate high level constraints in MTT and improve the optimization from the trajectory level to a layer level. To construct a layered graph, LGM defines pedestrian detection responses as the graph nodes, and integrates motion, appearance and depth features as graph edges. An online updating depth factor is defined to describe the depth relation among the observations in and out the layer, and the layer-level occlusion handling leads the serious occlusion issue to be solved in the neighbor layer region.

With a heuristic label switching algorithm, multiple pedestrian objects are optimally associated and tracked.

3) Proposing a Topology Energy Minimization (TEM) model. Inspired by sociological property of pedestrians, TEM adopts a learning approach to configure typical social-topology patterns, combines these spatial social-topology distributions with RGB-D cues to build a more reliable topology-energy variance model. TEM minimizes the variation of the topology energy function in the data association framework with a 3-step inference. This contributes stable group tracking and smooth transitions between groups and individuals.

4) Proposing a Graphical Social Topology (GST) model, which jointly estimates the group structure with the group & targets states using a topological representation based on the social affinity. With such a topology representation, targets are not assigned to groups, but connected to each other, which enables the cohesion of a group to be precisely modeled. We infer the birth/death and merging/splitting of groups by using online learned topology patterns and online topology updating modules. In-group individual identities are associated according to the group topology. Moreover, GST is able to naturally facilitate the self-occlusion problem by treating the occluded object and the other in-group members as a whole unit while leveraging overall state transition in the GST. Experiments on both RGB-D and RGB data sets confirm that GST model outperforms state-of-the-art trackers significantly.

Moreover, this dissertation introduces an RGB-D data acquisition system set up, as well as making the acquired RGB-D data set publically available.

Key Words：Multiple Objects Tracking, Pedestrian Tracking, RGB-D Data, Data Association, Graph Model, Group Model

目 录

图目录

表目录

第一章 绪论

1.1 研究背景和意义

随着人们对社会公共安全的关注和对智能生活方式的追求，计算机视觉的相关研究在信息科学领域与实际应用中的作用日益增强。建立在计算机视觉和机器学习上的视频分析技术已成为当前一个流行的研究领域。

视频分析的任务主要包括三个方面：目标检测、目标跟踪和目标行为分析。从三个任务的先后关系可以看出，目标跟踪技术处在目标检测与目标行为识别之间。它将目标检测的结果作为输入，同时又将跟踪结果向目标行为识别模块输出。目标跟踪一方面可以看做是对目标检测结果的修复完善，如弥补检测缺失和滤除检测误检；另一方面也可以看做是为目标行为识别提供有效的目标身份标识，使行为识别模块可以在连续帧中提取目标有效的行为特征。另外，目标跟踪的结果，即目标的运动轨迹，也可以是一种有效的行为分析特征。通过对某一场景内目标运动轨迹样本的聚类，不仅可以学习该场景中运动目标和运动群体的运动模板，还可以获取该场景的结构，如场景中的进入和离开区域等热点区域信息。所以，目标跟踪技术在视频分析中是不可或缺的重要一环，它在整个智能视频分析系统中起到承上启下的关键作用[1]。

近十几年来，随着智能视频分析技术的不断提高，目标跟踪技术得到了长足发展。从最开始的简单场景、单一种类单一数量的目标，到近几年的复杂场景、多种类多目标，其跟踪目标的种类、数量和难度都在不断上升，这也更加接近于实际生活中的应用场景。并且，随着新传感器技术和新机器学习理论方法的不断出现，多目标跟踪的正确率也在逐步提高。

多目标跟踪技术是对多个目标的状态进行测量并持续精确定位的过程。它通过综合利用统计估计、决策、控制以及智能优化算法等理论，将经过初级处理的观测数据依据一定规则进行判决和关联，从而判定被跟踪的目标数目并给出每个目标的状态参数。事实上，多目标跟踪过程是个递推迭代过程[2]。它首先在跟踪起始阶段创建目标特征库；接下来，通过数据关联规则实现目标观测和跟踪目标的配对，然后通过机动辨识、跟踪滤波与预测等方法估计各个目标的状态并对其进行更新；跟踪空间中不与任何已知目标关联的观测集合形成新的目标；当有目标消失时，由跟踪结束方法消除其信息；最后，利用目标预测状态确定下一时刻的目标状态。通过这样的递推循环，可以得到目标的连续状态。图 1-1 给出了多目标跟踪实现框架图。

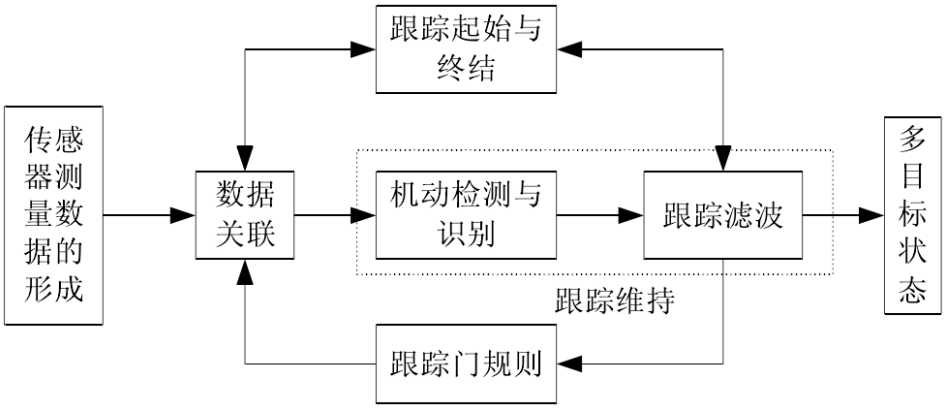


图 1-1 多目标跟踪实现框架图

由于目标跟踪技术在计算机视觉中所处的特殊地位，国内外有很多高校和研究机构陆续投入了大量的科研经费对其进行了深入研究，取得了丰富的成果。代表性的研究机构包括卡内基梅隆大学[3]、麻省理工学院[4]、美国南加州大学[5]、瑞士 ETH 实验室[6]、德国达姆施塔特工业大学[7]、德国 KITTI 研究所[8]、香港城市大学[9]、香港科技大学[10]、中科院自动化所[11]、清华大学[12]、北京大学[13]等等。同时，许多重要的权威期刊和国际学术会议都针对该领域的最新理论研究进展做出了专门讨论，如：International Conference on CVPR (Computer Vision and Pattern Recognition)、ICCV (International Conference on Computer Vision)、ECCV (European Conference on Computer Vision)、IJCV (International Journal of Computer Vision)、 PAMI (IEEE Transactions on Pattern Analysis and Machine Intelligence)、TIP (IEEE Transactions on Image Processing)等将多目标跟踪问题的研究作为主题内容之一，为该领域的研究人员提供了广泛的交流机会。

1.2 课题来源

本文研究工作受到国家重点基础研究发展计划（973 计划）、国家自然科学基金课题资助。

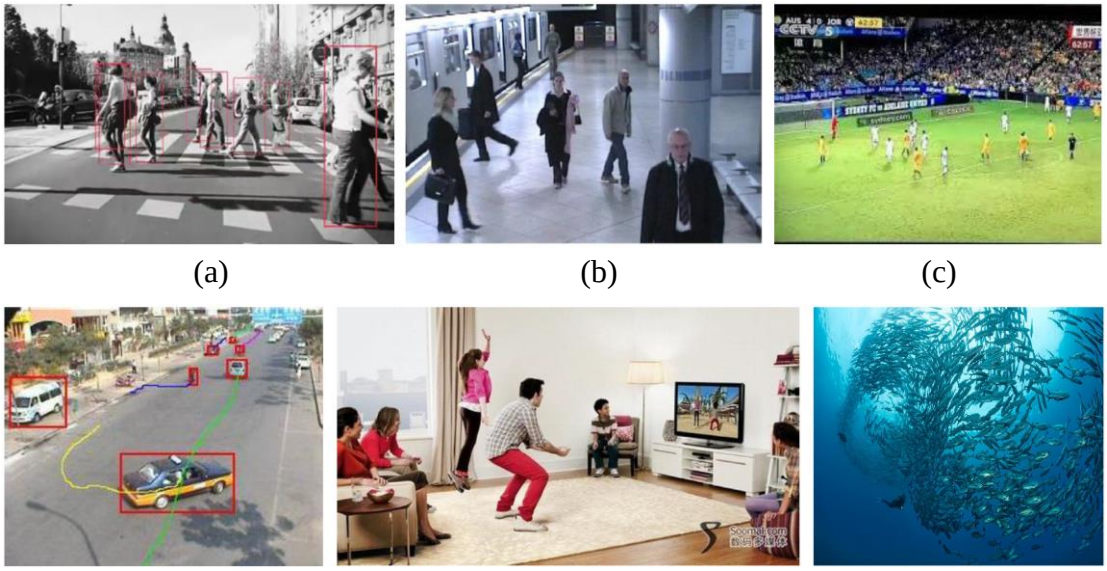
1、“基于多源数据的飞行器进近威胁目标检测跟踪及行为预测”，国家自然科学基金重点项目（课题编号：61039003），2011.01-2014.12，已结题。

2、“飞行器威胁目标识别与图像鲁棒匹配理论与方法”，国家 973 计划子课题（课题编号：2010CB731804-2），2010.01-2014.12，已结题。

3、“多视角多姿态人体目标检测”，国家自然科学基金面上项目，（课题编号：61271433），2013.01-2016.12，在研。

1.3 应用领域

多目标跟踪涉及到模式识别与智能系统、图像处理、统计学、机器学习等多个方面知识，是一门跨学科难度较高的技术。它的研究目的是以视频图像为基础，完成对场景中多个运动目标的跟踪及轨迹分析。其典型应用包括智能交通系统、智能监控系统、机器人感知系统、人机交互、生物医学研究等很多方面。



(a) (b) (c)

(d) (e) (f)

图 1-2 多目标跟踪应用场景

1.3.1 智能交通系统

交通事故和交通堵塞逐渐成为交通领域困扰人们的一大难题，交通问题浪费了国家财产，危及公民的生命安全以及社会的稳定。这些问题促使国家和科研机构开始着手智能交通的研究。如图 1-2 所示，（a）中展示驾驶环境下行驶道路上的行人车辆等障碍物进行检测跟踪，（d）中展示智能交通监控系统通过摄像头拍摄到的交通视频图像，实现对交通流量进行控制，对反常车辆或行人的行为进行检测、跟踪和分析。在自动驾驶方面，国内外众多汽车制造商、大学和科研院都将多种传感器应用于其开发的无人驾驶车辆上，并取得了积极的成果。谷歌公司在 2010 年开发的无人驾驶汽车采用摄像机、雷达传感器和激光测距仪，并通过详细的地图指引汽车在路上行驶[14]。谷歌公司于 2014 年 5 月 28 日推出自己的新一代的无人驾驶汽车。我国从 20 世纪 80 年代末开始智能车辆的研究，研究的领域主要集中于路径识别、自动转向等自主导航技术[15,16,17]。

国防科技大学以轻型越野汽车为平台，集成安装了激光雷达、视觉、组合导航系统等多种传感器，研制成功了我国第一辆自主驾驶汽车[17]。在对城市道路交通环境下的行人和车辆的监控中，行人检测和跟踪是其中的关键技术。在交通路口和关键路面上，通过检测和跟踪技术自动发现违章或事故停车、违章转弯及变线、行人横穿马路等常见不安全情况，并及时地报警。

1.3.2 智能监控系统

目前在我国智能监控系统占到了安防产业的 55%的规模，作为获取信息的最有效手段，视频监控系统越来越多的应用于各种场合。视频监控的数字化给视频监控系统带来了大量的视频资源，然而监控方式仍然主要为人工监控，这带来了很多问题，如监控人力有限、误检漏检多、监控检索困难、大量冗余数据等。智能监控系统将计算机视觉技术和机器学习等技术融入监控系统，对视频数据流进行目标检测、目标跟踪和目标行为分析等工作，判断目标的动作，并进行相关记录，同时对监控系统进行实时控制，发出报警，使智能分析算法代替人进行监控，给予监控系统智能性，变被动监控为主动监控。同时智能监控系统可以通过智能算法，完成全天候实时自动监控，有效降低误检和漏检，去掉冗余监控信息。甚至，在事发之前就能够有效的进行报警，提示监控人员关注监控画面，事后提高视频内容检索速度，使视频信息变得更加易用。

1.3.3 机器人感知系统

人类一直希望使用机器人代替人类进行生产劳动。基于视觉的传感器及视觉相关传感器（如红外，深度等）是智能机器人十分重要的信息源。为了能够自主进行运动，智能机器人在智能地进行生产劳动前，首先需要认识和跟踪环境中的物体[18,19]。例如机器人的“手眼”应用，目标检测和跟踪系统被安装在机器人的手臂上，在机器人对周围的事物进行拍摄，并对拍摄到影像进行分析完成对目标的检测识别后，来设计机器人的运行路径和轨迹，保证机器人能够以最佳的姿势来获取要抓的物体。为了使得机器人手臂能够成功的抓取物体，目标跟踪系统需要对目标进行准确并且实时的跟踪。无论是智能机器人像人一样进行生产劳动[20]，还是辅助人类进行生活服务，都需要通过传感器对周围事物进行检测、识别、跟踪和分析。

1.3.4 人机交互

自然人机交互的目的是让计算机与人类的交流可以像人与人交流那样通过语音，身体语言等来实现，从而代替传统的鼠标键盘等传统方式的人机交互触摸按键模式[1]。其本质是先对目标进行检测跟踪，然后对该目标进行行为识别[21]

将语义信息与计算机的对应指令联系起来。以微软，三星为代表的企业就这种交互模式进行了大量研究，并已经推出相应的产品。微软推出的 Kinect 通过跟踪目标的整个身体或某个身体局部，分析其运动特征，转化为设备的操作指令，使得用户可以在自己家的客厅进行游戏和体育运动。三星智能电视通过摄像头及其他传感器对用户的手势跟踪，并分析手势的运动方向达对电视进行操作目的。近期美国 Magic leap 公司即将推出虚拟现实 VR 产品，通过三维重建、检测、跟踪定位等技术将现实场景与虚拟的三维场景相结合对计算机视觉显示进行重新定位。从以上多个产品可以看出，现在基于计算机视觉的自然人机交互模式均以目标检测跟踪为基础。目标检测跟踪性能的高低直接影响了人机交互功能的准确性。

1.3.5 生物医学研究

多目标跟踪技术也可以服务于其他领域的科学研究。例如对大脑细胞中神经元内的多种蛋白质媒介物的运动进行跟踪分析，可以更好的了解大脑中神经传导的工作机制，帮助人类揭开人脑的工作原理[22]。对细菌等微生物的在药物作用下的运动进行跟踪分析，可以对药物的研究和人类疾病的发病原理进行有效分析[23,24]。对鱼群，蚁群，鸟类等进行多目标的群体性跟踪研究[25,26]，如图 1-2（f）所示，可以揭开其他生物物种之间的信息交流机制。针对群体性生物的相似性外形特征，如何在密集的遮挡中有效的区分出个体，准确的判断出种群生物的运动模式是待解决的难点问题。

1.4 基于 RGB 数据的多目标跟踪常见问题

多目标跟踪问题在上述的应用中得到了广泛的研究，也取得了很大的进步。但是，复杂场景中多目标跟踪技术距离实际应用还是有一定差距。在室外场景下，跟踪目标为行人的多目标跟踪系统中，频繁的目标之间互相遮挡、强烈变化的光照条件、复杂的动态背景、差异很大的图像分辨率等问题给基于视频分析多目标问题提出挑战。许多目标跟踪的实际应用，如室内环境下的智能监控，室内基于视觉的机器人导航，室内环境进行的 Kinect 体感游戏，手势识别等，限定在室内环境，他们无法处理复杂的光照环境带来的视觉不稳定性。目前被广泛研究的无人驾驶则是通过多种传感器，如激光传感器，GPS 等设备，以减少光照等不确定因素造成的影响。

目标检测问题不是本文重点讨论的对象，但是检测器的检测结果成为制约多目标跟踪的关键因素之一，尤其是基于“先检测再跟踪”策略的多目标跟踪模

型。输入的检测结果包含的目标漏检和误检越高，跟踪中数据关联的难度也越高。图 1-3 中展示了基于视觉的检测器在多目标跟踪数据集上的常见错误[27]。（a-b）中目标与背景有极低的对比度，因此被漏检；（c）中目标出现严重的姿态变化，（d）中的遮挡问题，以及（e-f）中与目标相似度高的伪目标。即使这些年的计算机视觉领域的研究中，DPM 检测器[101]和基于 CNN 特征的检测器[28,29]已经极大提高了目标检测的准确率，但是上述检测问题在复杂场景的人群中很大程度上还是难以避免。

(a)　(b)　(c)　(d)　(e)　(f)

图 1-3 基于 RGB 数据的检测器典型难题

具有上述问题的检测结果作为多目标跟踪技术的输入，对多目标跟踪技术本身就是一种挑战，但同时由于真实场景的复杂性，例如：多个目标在运动过程中的相互遮挡、监控场景光照变换、交通场景的复杂动态背景、摄像机角度不佳清晰度不足，再加上本身检测器输入的不精确的检测结果，使得正确获得多个运动目标的运动状态和完整的跟踪轨迹变得更加困难。图 1-4 中所展示了多目标跟踪过程中由于输入错误导致的多目标跟踪错误。其中，红色箭头代表误检测引起的误跟踪，黄色箭头代表漏检测引起的漏跟踪，白色箭头代表检测结果与真实值有较大偏差的检测带来的误跟踪，解决这些难题成为多目标跟踪领域所面临的严峻挑战。

基于传统视觉的检测和跟踪技术，我们可以称之为基于 RGB 数据的目标检测和跟踪技术。伴随着 RGB 摄像头的快速增长，基于视觉的其他类型传感器也得到了极大发展，例如 RGB-D 传感器，在 RGB 数据的基础上增加了深度维度的信息。RGB-D 传感器包括立体相机、双目摄像头、微软公司的 Kinect 传感器，以及包含激光距离信息的 RGB 传感器等。它们都能提供有效的深度信息。基于 RGB-D 数据的多目标跟踪方法能够利用 RGB 数据所没有的深度数据，可以在很大程度上能够规避上述的难题。例如，对于图 1-3（a-c）中的低对比度和高姿态变化错误，可以首先利用深度信息初步检测到目标，然后在视

觉检测中利用阈值较低的检测器找到目标；对于（e-f）中的相似度较高的伪目标，可以利用深度上的阈值对这些区域进行排除；对于(d)，在跟踪过程中，可以利用深度数据，有效关联处于部分遮挡和遮挡的的目标。随着 RGB-D 传感器的引入，以及针对 RGB-D 数据所设计的检测器和跟踪器，会使得多目标检测和跟踪的准确性有大幅提高。

图 1-4 基于 RGB 数据的多目标跟踪输入性错误

1.5 本文的研究内容

有效解决基于 RGB 传感器的方法普遍存在的有效探测范围小、数据可靠性低以及易受外部环境影响等问题，对于提高多目标跟踪的准确性、可靠性具有极其重要意义。本文利用 RGB-D 数据进行多目标跟踪，汲取了 RGB 图像信息和深度信息，可以弥补 RGB 图像缺失的像素深度值，解决依靠单一视觉技术处理图像的难题。本文提出的方法综合利用了 RGB 信息和深度信息，不仅可以弥补仅依靠视觉 RGB 传感器进行目标跟踪时易受天气状况和光照条件变化影响，无法得到跟踪对象的深度信息的不足，也可以克服深度数据无法判断障碍物类别、无法可视处理、冗余报警等缺点。在多目标跟踪模型中使用两种不同类型的数据，利用各自的优点完成在复杂环境下的多目标跟踪问题，可以实现优势互补，提高跟踪的准确性和实时性。本文的主要创新点包括以下几个方面：

1、提出了基于深度数据关联的多目标跟踪模型。利用多维分配模型推导出多行人跟踪模型，融合了 RGB 与 D（深度）数据特征，在深度域上进行深度链状结构的划分，在深度结构内利用目标的外形和运动特征，进行模型的在线更新。并利用图论中多部图匹配原理提出了基于标签交换的迭代求解方法。

2、提出了基于分层图模型求解多目标跟踪问题。采用图模型的方法，将多目标在深度域划分进场景中的多个深度层内，使得基于全图空间搜索的关联问题转化成为层内和层间的匹配。并在层内提出了基于深度权值的遮挡模型，有效解决目标遮挡状态下的关联错误问题。该方法做到了多帧内多目标跟踪问题

的实时求解。

3、提出了基于拓扑能量的多目标跟踪模型。基于人群行为研究，将自然场景中的行人目标划分为以组为单位的运动集合，每个组内的行人具有近似的运动模式。利用组内组外的行人运动相似度，进行了能量形式的建模，使得组内的行人相似度尽可能的高，组间的行人相似度尽可能的低。并以“拓扑能量最小化”的方式要求跟踪中进行模型求解，在拓扑的变化过程中，模型通过加入虚拟点的方式，将被遮挡的目标通过组内位置估计的方式进行有效定位，减少了多目标跟踪过程中目标丢失的次数。

4、提出了基于动态拓扑图的多目标跟踪模型。继承了以组为单位对行人进行研究的框架，将组内的行人看作是图中的节点元素，通过图中边的连接方式探索组内成员在行走过程中的运动的变化。模型通过离线与在线学习相结合的方式，对组的典型拓扑进行离线学习后，在线跟踪中与行人组进行拓扑结构匹配。在组的运动过程中，模型通过组的初始化、更新、合并和分裂的方式完成对行人组的跟踪，最后利用线性规划的方式完成组内行人的身份确认，得到每个目标的完整轨迹。

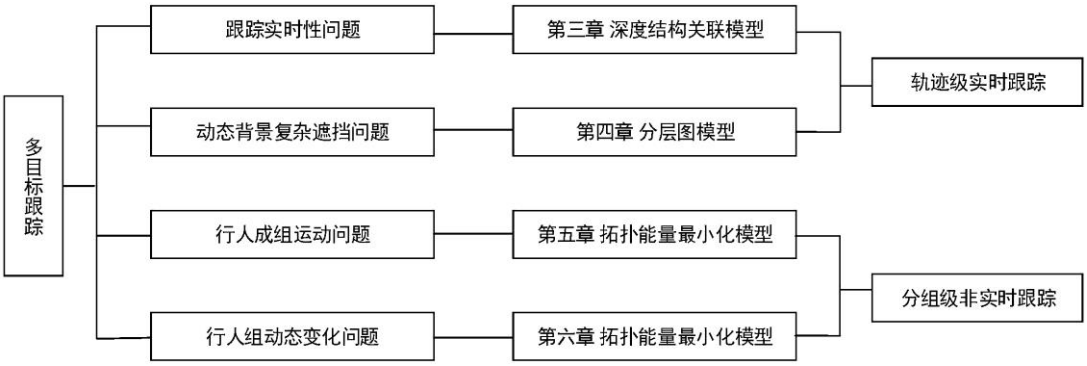


图 1-5 本文提出的多目标跟踪模型关系图

本文共提出上述四个基于 RGB-D 数据的多目标跟踪模型。针对每个模型，都进行了大量对比实验验证并与其他优秀的多目标跟踪方法做了对比分析，这些对比方法包括基于 RGB 数据的多目标跟踪模型，基于轨迹分析的多目标跟踪模型，基于行人组划分的多目标跟踪模型等。在公共的 RGB-D 数据集和 RGB 数据集中，实验结果表明本文提出的多目标跟踪模型的性能优于同类型的其他优秀模型。

此外，作者搭建了 RGB-D 数据采集原型平台，利用单目 RGB 视觉摄像头和激光传感器搭建了模拟移动平台，在真实交通场景中进行了 RGB-D 数据采集，将采集的数据集整理为 SDL-Garden、SDL-Crossing 和 SDL-Campus 数据集。数据集包括视频数据和对应的深度数据。这三个 RGB-D 数据集已公开[115]，供

研究交通场景内目标检测和跟踪的科研人员下载使用。

1.6 本文的组织结构

第一章，绪论。论述基于 RGB-D 数据的多目标跟踪的研究背景和意义，描述多目标跟踪在智能交通系统、智能监控系统、机器人导航、人机交互、生物医学研究等方面的应用。分析基于 RGB 数据的多目标跟踪研究方法的不足，以及基于 RGB-D 数据的多目标跟踪方法所带来的好处。明确本文的主要研究内容和贡献。

第二章，多目标跟踪方法概述。论述多目标跟踪方法的多种分类，针对基于实时和非实时、局部和全局、概率估计和图模型、个体轨迹分析和行人组分析等方面进行了全面详细分析。

第三章，数据集与实验评估方法。论述 RGB-D 数据采集平台的搭建以及数据获取方法，详细介绍采用的数据集，目标检测器和多目标跟踪方法性能评测指标等。

第四章，基于深度结构的多目标跟踪模型。论述如何利用 RGB-D 数据中的深度信息在场景中对行人的空间位置进行有效的划分，利用整数规划中的多维数据分配问题对多目标之间的数据关联进行建模。最后，在多个数据集进行与多种方法的对比实验。

第五章，基于分层图模型的多目标跟踪模型。描述多目标实时跟踪算法模型，详细介绍了模型的求解过程，在线更新，以及在复杂遮挡情况下，深度信息对遮挡问题的模型具体求解的作用。最后，在多个数据集分别对本章提出的跟踪模型进行实验，实验结果显示分层跟踪模型具有良好的稳定性与实时性。

第六章，基于拓扑能量的多目标跟踪模型。介绍如何利用目标之间的拓扑结构关系对行人组进行能量形式的建模，这种拓扑关系衡量组内和组间行人之间的运动行为相似性。模型利用拓扑能量变化最小的原理求解多帧之间组跟踪的解。本方法在 SDL-Campus 进行模型训练后，在多种 RGB-D 数据集进行测试。

第七章，基于动态拓扑图的多目标跟踪模型。提出基于拓扑图模型来构建行人组的运动特性。通过离线学习得到自然人群中最常见的拓扑结构，通过在线学习进行组的初始化、更新、合并和分裂，有效地将行人组的动态运动信息和拓扑图结合起来完成对行人组的跟踪。进而利用拓扑图内的拓扑关系对组内的成员进行身份识别。最后，将动态拓扑图模型在 RGB 和 RGB-D 数据集上进行实验验证。

最后，第八章总结本文的主要工作，提出对未来工作方向的展望。其中包括进一步提高行人、汽车等多种目标跟踪算法的鲁棒性和准确性，改善多目标跟踪框架，完善 RGB-D 多目标跟踪评测方法等热点问题。

第二章　多目标跟踪方法概述

传统的单目标跟踪方法主要解决的问题是：根据当前帧目标的位置与目标特征去确定下一帧中该目标所处的位置。当多个目标同时运动并发生相互干扰时，需要解决目标在被遮挡前后的匹配问题。这种匹配实际上就是判断遮挡前后的两个待匹配图像区域是否包含同一个目标。多目标跟踪是在单目标跟踪技术的基础上发展而来的，继承了单目标跟踪技术的思想，通过时序优化关联匹配一段视频中的多个目标。多目标跟踪方法涉及模式识别和智能系统、传感器技术、图像处理、机器学习以及计算机视觉等研究领域。

2.1 多目标跟踪方法的分类

文献中的多目标跟踪方法算法大致可以分为三类：一是面向目标的方法，假设跟踪过程中的目标数目是固定的，观测来自于已知的目标或杂波。典型算法包括概率数据关联算法[30]（PDA）和联合概率数据关联算法[31]（JPDA）；二是面向观测的方法，假设观测来自已知的目标、新目标或杂波，典型算法为多假设跟踪算法[32]（MHT）；三是面向轨迹的方法，该类方法假设轨迹未被检测、已经终结、与观测相关。

从应用的角度看，或者说站在跟踪的实时性角度，可以将多目标跟踪划分为两类：一类是实时性跟踪（Online tracking），一类时非实时性跟踪（Off-line tracking）。前者需要在当前帧做出决策判断，输出多目标跟踪结果。此类跟踪方法往往被用到对实时性要求较高的场景中，如自动导航领域的自动驾驶或者辅助驾驶、游戏应用、智能监控预警等。而非实时性跟踪一般对整个视频序列进行分析判断，通过全局的角度寻找到每一个目标在视频中的最优运动路径。非实时性跟踪主要用于监控场景中的行人轨迹分析、行为分类、以及对监控视频的分类检索。由于非实时性跟踪可以获得目标在整个视频中完整信息，所以其跟踪准确率显著高于实时性跟踪。从跟踪方法的角度讲，实时性跟踪主要采用基于贝叶斯滤波的概率分析方法，通过目标在前序帧的运动信息和外形信息估算其在后续帧中的相应位置，如图 2-2（a）所示。而非实时性跟踪主要采用基于轨迹分析的数据关联方法。对多目标的运动信息，外形信息以及场景信息，对目标建立精确和复杂的运动模型，但是此类跟踪模型往往具有较高的时间复杂度，如图 2-2（b）所示。介于实时性跟踪和非实时跟踪之间，部分多目标跟踪方法采用基于多帧或者一批次帧（batch）的目标关联方法，如图 2-2（c）所

示。此类方法融合了实时性和非实时性跟踪的部分优点，在多帧内进行目标匹配，既可以利用概率滤波方法，也可以利用目标局部时间内的信息进行数据关联，且此类方法也具有一定的实时性，能够满足实时性要求较高的应用。本文提出的四种多目标跟踪模型分别采用了如图 2-2（b）和（c）所示的跟踪方法。第四章基于深度结构的多目标跟踪模型和第五章基于分层图的多目标跟踪模型采用（b）中所示的基于多帧处理（batch）的多目标数据关联方式，主要应用于交通驾驶场景。而第六章和第七章基于能量最小化和动态图的拓扑图模型则采用如图（c）所示的全局数据关联的非实时跟踪方法，主要适用于非实时性要求的视频监控场景。

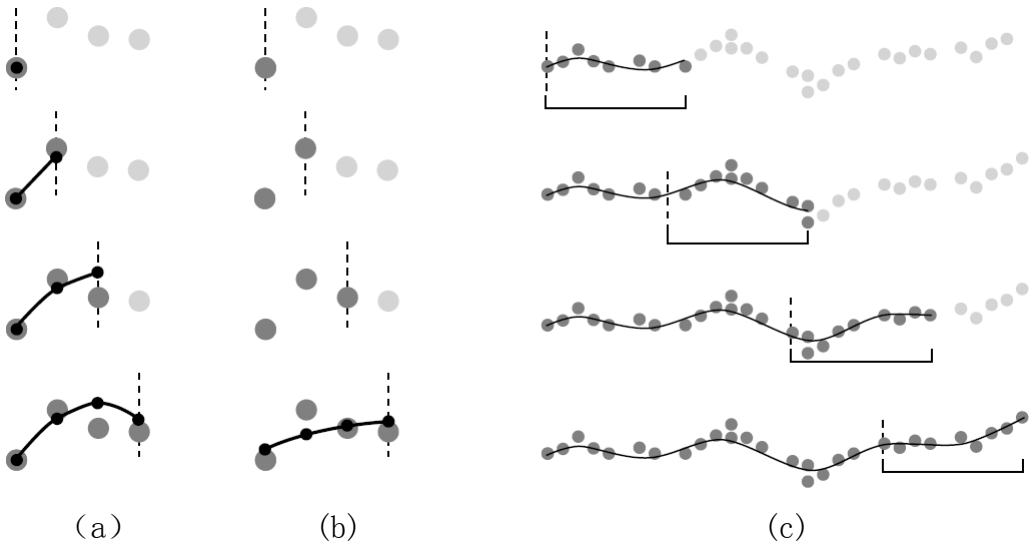


图 2-1 多目标跟踪方法的分类

以下内容综述了基于概率估计的实时性多目标跟踪方法，以及基于数据关联的非实时性多目标跟踪方法。由于本文的第六第七章涉及到基于分组模型的数据关联方法，我们对分组模型等多种方法也进行了详细讨论。

2.2 基于概率估计的多目标跟踪

在多目标跟踪问题中，目标的数目是未知的，且随时间无规律变化。例如，在驾驶过程中，车辆前方出现的行人数目和时间都是不确定的。另外，目前的行人检测技术仍然有很多虚警和漏检，传感器每时每刻都接收到随机数目的目标观测。在任何时刻，不能确定哪个观测应该用来更新哪个目标的状态。目标的不确定性与随机性激发人们采用基于概率估计得方法进行目标跟踪[33]。

- 全局最近邻

全局最近邻（Global Nearest Neighbor，GNN）滤波器[34,35,36]是一种简单的基于数据关联的跟踪模型，它将最近邻（Nearest Neighbor，NN）滤波器推广到了多目标跟踪情形中。在已知每个目标上一时刻状态估计的均值和协方差的情况下，GNN 滤波器首先运用卡尔曼滤波的预测部分得到每个目标观测的预测以及相对应的协方差。在数据关联部分，通过最小或最大化整体损失函数得到目标和观测唯一的联合关联。损失函数可以是距离也可是似然概率和，限制条件是每个观测最多与一个目标轨迹进行关联。在更新部分，该方法假设联合关联是正确的，然后直接运用卡尔曼滤波，用相关联的观测去更新目标状态。

很明显，这种朴素的跟踪方法很难做得到全局最优解。并且在跟踪过程中 GNN 滤波器会遇到和 NN 滤波器一样的问题和限制，每一次观测获取后，选择具有最高概率可能性的观测作为目标，而丢弃其他所有的目标观测。在复杂场景中，当漏检率和误检率很高时，GNN 滤波器的表现往往很差。

■ 联合概率数据关联

联合概率数据关联（Joint Probability Data Association, JPDA）滤波器是 PDA 滤波器在多目标跟踪中的推广，它能够处理多个目标数目已知的情况[37,38,39]。在该方法中，除了互联概率的计算以外，单个目标状态的迭代传播过程与 PDA 滤波器基本一致。JPDA 滤波器使用互联事件和互联概率来避免将多个观测值分配给多个目标时引起冲突。然而，互联概率的计算复杂度伴随目标数目和观测数目呈指数型增长，所以 JPDA 滤波器在实际场景的多目标跟踪中很少被直接采用。这种 JPDA 滤波器在计算上的不可行性导致了许多近似算法的产生，例如最优选择策略[40,41]以及基于马尔科夫链蒙特卡洛（MCMC）的策略[42,43]。此外，由于 JPDA 滤波器只能处理目标数目已知并且固定的情况，所以对于目标数目未知并且不断变化的场景，出现了许多新颖、实用的近似算法，例如说联合整体概率数据关联（JIPDA）滤波器[44]。此外，Vermaak[45]采用序贯蒙特卡洛（SMC）方法，将 JPDA 滤波器推广到了非线性、非高斯的数据模型中。

对跟踪问题的概率建模，JPDA 具有良好的理论模型，能够解决 GNN 滤波器模型中无法量化检测器误检率的问题，也能在模型中对跟踪目标的密度估计进行量化。但在复杂场景中，目标密度和检测器误检率的估计并不总是可靠，这导致了 JPDA 滤波器模型没有得到广泛应用。

■ 多假设跟踪

多假设跟踪（Multiple Hypothesis Tracking，MHT）滤波器[46,47]是一种基于非实时数据关联策略的多目标跟踪方法，MHT 滤波器通过寻找前序所有时刻的目标与观测的所有可能关联组合，构成目标所有可能的轨迹，来降低关联的

不确定性。所有的观测与真实目标或者虚假目标的所有关联集合称作假设，在任意时刻，MHT 滤波器会保留假设中后验概率较大的子集部分。当新的观测数据集到达后，会在原假设的基础上产生新的假设，使用贝叶斯准则更新相应的后验概率。值得注意的是，产生新假设时有三种选择：第一，一个观测可以与已存在的轨迹进行关联；第二，可以被当作噪声；第三，可以初始化一条新的轨迹。从这个方面来讲，MHT 滤波器自身具有初始化轨迹和终止轨迹的功能，所以它能够处理目标数目未知且随时间变化的情况。因此，MHT 滤波器的基本思想是向后传递后验概率较高的多个假设，在每个时刻寻找后验概率最高的假设作为当前的最优关联。得到最优的关联之后，运用卡尔曼滤波来更新各个目标的状态。R.Singer[48]提出传递关联假设的基本思想用于求解单目标跟踪模型，Reid[49]将这种思想系统地运用到了多目标跟踪中。

MHT 滤波器会遇到和 JPDA 相同的困境：假设的数目随时间呈指数型增长。在实际应用中，MHT 滤波器通常会设定关联阈值或者使用启发式规则来限制假设数目的过快增长，从而降低算法复杂度。

■ 粒子滤波

粒子滤波器（Particle Filter，PF）是一种基于贝叶斯原理的滤波器，使用粒子概率密度表示的序贯蒙特卡洛模拟方法。其基本思想是通过寻找一组在状态空间中传播的随机样本对后验概率分布进行近似，以样本均值代替积分运算，从而获得状态的最小方差估计的过程。在粒子滤波器被应用于多目标跟踪问题[50,51,52,53]时，通过对后验概率密度估计的一些统计要素（如混合概率密度估计的均值和协方差）的求解得到目标的状态估计。该方法不仅减少了估计方差，还减少了用于近似表示概率密度的标准粒子滤波器的粒子数目。在例子滤波器的使用过程中，粒子滤波器通常和其他的基于概率估计的跟踪模型相结合，文献[54,55]将粒子滤波和 JPDAF 结合，提出了一种蒙特卡洛联合概率数据关联方法（MC-JPDAF）；基于粒子滤波器的算法，所用的粒子数越多，跟踪表现的性能也越好，其计算复杂度也会随之越高。因此，在实验中选择粒子数时，需要在算法性能和计算复杂度之间进行折衷。目前，针对粒子数的选择往往根据跟踪场景来确定粒子数。文献[56]Karlsson 通过改善 MC-JPDAF 提出了序列采样粒子滤波算法（SSPF）和独立划分粒子滤波算法（IPPF），并与标准粒子滤波、MC-JPDAF 进行了对比。

粒子滤波器的优势在于强大的后验概率分布估计可以有效的描述多目标的运动状态，但是不能解决复杂跟踪场景中的数据关联问题，所以往往需要使用启发式算法来求解多目标之间的关联问题。Breitenstein[53]在多目标跟踪中，设计粒子滤波器在线收集多目标的外形特征，并且在数据关联中利用贪心策略决

定目标与目标观测之间的对应关系。

■ 马尔可夫链蒙特卡洛采样

马尔可夫链蒙特卡洛（Markov Chain Monte Carlo，MCMC）采样是一类依靠随机采样的概率统计方法。上面提到的粒子滤波器在序列采用中可以准确估计出复杂的概率分布。但是，它依赖于一阶马尔可夫假设，这在多目标跟踪中制约模型向高阶的扩展。所以在多目标跟踪过程中，很多方法采用更通用有效的采样方法。

Khan[57]提出使用马尔可夫随机场（Markov Random Field, MRF）作为运动先验对目标之间的运动行为进行建模，通过 MCMC 采样使模型复杂度达到指数级。预先设定的步数，可以使采样过程通过在变化的多维空间内完成跳转完成数目变化的多目标跟踪。在后续工作中 Khan[58]提出如何处理真实跟踪场景中一个目标产生多个观测，或者多个目标对应一个观测的方法。在线性运动模型的假设下，后验概率可以被混合高斯模型估计，并且可以估算出物体的连续运动位置，进而利用 MCMC 采样来完成数据关联。这两种方法在实验中都取得了较高的跟踪精度。但是，文献[54,55]实验所选用的场景来自于俯拍镜头，没有涉及或者处理遮挡问题。

在具有复杂遮挡的交通场景中，基于 MCMC 的数据关联框架近些年也得到了广泛研究[59]。Benfold 和 Reid[60]采用稳定的头肩检测器实现行人检测后，利用 MCMC 数据关联和以及多线程并行计算使交通场景多目标跟踪达到实时效果。Wojek[61]利用多帧之间的运动信息进一步提高了物体检测的准确度。同时利用交通场景的物体在同一水平面的假设，使用具有可逆跳转步的 MCMC（Reversible Jump MCMC，RJMCMC）采样方法在多帧内完成 3D 场景和相机位姿的联合建模，并对行人和车辆进行跟踪。Choi 和 Savarese[62]使用相似的基于 MCMC 采样的多目标跟踪框架，但是对目标之间的运动进行了重新建模，在新构建的图结构中加入目标之间的吸引力和排斥力，可以更精准地获得目标的轨迹。但是这种复杂的图结构极大的增加了模型的时间复杂度，使得每一帧的数据关联耗时高达几分钟。上述基于 MCMC 采样的多目标跟踪模型都在求解过程中使用到了“跳转”（Jump）和设定“步数”（Move）等策略，这样在求解过程中不断加入和删除目标，使得目标方程的求解在离散空间内完成。

2.3 基于多帧数据关联的多目标跟踪

基于多帧数据关联的多目标跟踪方法近年来取得了极大的进步。这类方法主要遵循“先检测再跟踪”（tracking-by-detection）策略，利用近年来检测技术

的进步，使得以检测结果作为跟踪的输入的框架变得鲁棒。利用多帧或者全局的目标观测信息，对目标的运动和外形信息、目标之间的相互作用、场景的先验等因素进行精确的建模，使得基于多帧的数据关联模型可以获得全局或者近似全局最优解。本文将这类多目标跟踪方法大致划分为以下几类[1]。

■ 基于整数规划的多目标跟踪

Morefield[63]早在 1977 年就提出利用 0-1 整数规划问题对多目标跟踪问题建模。在优化理论中，此类整数规划问题更接近于集合分配问题（Set Packing Problem）。该问题利用 0-1 二进制矩阵表示每一帧内的目标观测与前一帧内的目标观测的对应关系，并且该矩阵的维度与目标观测的个数相等。加入的线性约束保证轨迹数与目标的个数一致。枚举法常被用来做这种整数规划问题的解法，但是当目标增多时，这种解法无疑是低效的。后来，一种更有效的方式，拉格朗日松弛解法（Lagrange Relaxation, LR），被用来求解整数规划的目标方程。这一技术使得整数规划问题在多目标跟踪的求解中可以用线性规划的方式来完成，例如单纯型算法（Simplex Algorithm）[64]或者内在点解法（Interior Point Methods）[65]。这类方法使得数学规划更加有效的求解最短代价路径，进行多目标跟踪优化关联。实验结果表明，整数规划是一种有效求解多目标跟踪问题的方法，同时为设计性能更优的多目标跟踪器提供了一种新思路。

■ 基于启发式数据关联的多目标跟踪

近年来，比较流行的多目标跟踪算法是启发式的数据关联方法（Hierarchical Data Association，HDA），而且这些算法使用相似的策略——“先检测再跟踪”：首先，使用可靠的目标检测器对所有视频序列进行目标检测后输出检测结果；然后在相邻帧利用局部信息连接可靠的目标检测窗口组成短的轨迹短片段（Tracklets）；最后，利用全局信息连接这些短的轨迹片段成为长的轨迹（Trajectory）。在有些文献中，这种由短变长的轨迹连接方法称为轨迹关联法（Tracklet Association, TA）。

Kaucic[66]在 2005 年提出使用 Hungarian 算法[67]优化匹配矩阵，在航空交通场景中通过连接不同传感器间短的轨迹片段成为长的轨迹，达到多目标跟踪的效果。随后，这一框架被应用到视觉监控场景内完成具有遮挡情况下的多目标跟踪求解。Wu 和 Navatia[68]在使用具有轮廓的检测器完成视频内的行人检测后，在每一帧内最大化行人可见部分与遮挡部分的联合概率完成多目标识别。在求解过程中，相邻视频帧通过加入行人的运动信息和外形信息完成数据关联。随

[1] 值得注意的是，各类方法所综述的文献中可能包含多种跟踪策略，但这里只按照文献所描述的主体模型进行分类。

后，Huang[69]提出一种基于轨迹的具有三层结构框架的数据关联方法。在底层，通过判断相邻帧内检测结果区域的重叠面积连接短轨迹；在中层，基于目标的运动和外形特征，将短轨迹延长成为中等长度的轨迹；在高层，通过加入场景信息，例如场景内的遮挡物，场景入口和出口等信息，将中等轨迹连接成完整的行人轨迹。在继承这种轨迹关联的多目标框架后，Li[70]提出使用在线监督学习的方式，例如 Rankboost 或 Adaboost，通过判断轨迹之间相似度来区分不同目标的轨迹。与此方式类似，Yang[71]使用基于在线学习的条件随机场（Conditional Random Field，CRF）的能量方程区分具有较高相似度的轨迹。

■ 基于图模型的多目标跟踪

图模型在多目标跟踪领域得到广泛的应用。这主要得力于图模型和多目标跟踪问题之间具有天然的对应关系。在所构建的图模型中，目标观测可以被当作图中的一个节点，每一帧的图像上的所有目标观测可以看作是整个图中的一个部图。因此，多帧视频间（或者整个视频序列）的目标观测关联问题可以转化成多部图的节点匹配问题。最小团算法[72,77]，最大权重独立集算法[73]和网络流算法[74,75,76]等一些经典图模型算法相继被用来求解此类模型。Brendel[73]用最大独立集（Maximum Weight Independent Set）求解多目标的数据关联问题，将时域上的相邻成对目标观测看作是图中一个节点，图中边所连接的是具有相同高相似度的运动和外形信息的点。然后通过迭代的方式选择这些点的最佳独立集来延长目标观测的轨迹点，而最大独立集中的节点不被任何边所连接，这组独立集即为一帧内的有效目标观测点。Zamir[77]和 Dehghan[72] 对同一目标在一段时间内的所有观测构建成一张全连接图。每一帧内的点之间没有边的连接，不同帧的节点通过边相互连接，边的权值由目标的外形和运动特征相似度决定。为了获得同一目标在这段时间内的完整轨迹，使用广义最小团理论（Generalized Minimum Clique Problem，GMCP） 求解模型。此外，Wen[78]利用超图（hyper-graph）构建多目标在多帧之间的连接图，寻找多目标在超图之间的最短路径，找到多目标跟踪的最优解。

网络流模型：网络流模型（Network Flow）属于图模型的一种，并且近些年随着多种网络流模型的提出，极大降低了图模型中求解的时间复杂度，同时提高了跟踪的实时性。Jiang[79]提出使用网络流的方法来求解多标跟踪问题，利用整数线性规划求取近似全局最优解。这种引入计算机网络流原理的新颖思路同时也为基于视觉的为多目标跟踪建模开辟了新的方向。Zhang[74]在此基础上，构建出一个较理想的网络流模型，在网络结构中新加入两种节点——源点与汇点，并将目标检测得分做为边的权值，最后利用最小代价流算法（min cost flow algorithm）寻找全局最优解。Pirsiavash[76]重新定义了 Zhang[74] 的网络流模型，

使用连续最短路径算法（Successive Shortest Paths, SSP）寻找最短代价路径。这种局部贪婪算法在线性复杂度下获得了近似全局最优解，同时节省了多目标的匹配时间。在这种模型的基础上，Berclaz[75]将多目标跟踪问题描述成新的整数规划问题，并提出用 *K* 条最短路径算法在网络中快速搜索松弛线性规划后的全局最优解。但是，这种算法中所使用的路径由预先设置的网格决定，并非通过目标观测的信息获得。

2.4 基于分组模型的多目标跟踪

近年来，研究人员的关注点从对行人个体的检测跟踪逐步发展到对具有社会行为人群的跟踪，分析和行为识别[80,81,82,83,84]。群体行为学的研究和应用在多目标跟踪领域中引起了研究人员的重视。从社会心理学的角度来讲[85,86]，任何个体在任何环境中都想以最捷径的方式达到自己的目的地。例如无障碍环境中，个体总是愿意沿着直线运动。在集群的环境中，拥有相同目的的个体选择了一条最捷径的方式开始运动，运动发生的过程中与环境中其他个体相遇，为了避免碰撞，个体依据其他个体相互调整自己运动过程，最终达到一种稳定状态。换一种说法，随着时间的推移，个体的运动状态会渐渐弱化，取而代之的是一种“从众”的运动模式，这种运动模式具有自发性和无组织性。研究人员还发现，在自然行走的人群中，以组的形式运动的行人在总人群中占比高达 60-70%。组内的行人可能是朋友和家人，也有可能只是场景中具有相同行进目标的陌生人。基于这种分组的思想，在多目标跟踪中人们将对单一目标个体的运动行为的建模提升到对目标个体所在的组进行集体建模。

Pellegrini[87]基于目标的运动行为构建了高阶马尔可夫模型的复杂能量方程，同时在模型中加入了场景先验信息，目标避遮挡行为，以及行人间的社会属性。该模型还能完成了对行人的运动预测。Ge[88]基于行人轨迹的社会属性对行人进行了分组，在模型中定义轨迹的相似度后，利用自底向上的策略完成轨迹的聚类。Qin 和 Sheldon[89]用线性规划的模型描述将人群的组行为，并使用对偶优化的理论求解该模型完成多目标跟踪。Chen[90]采用相似的模型描述框架，但是利用在线学习的策略来完成组的划分，以及对组内的多目标的跟踪。Bazzani[91,92]提出了一种单个行人和组之间的联合跟踪框架，利用分散化的粒子滤波器对组和目标个体进行联合求解，证明了对组和目标个体的联合建模比单一的组跟踪或者个体跟踪有显著的性能提升。

2.5 本章小结

本章概述并分类了多目标跟踪方法。本文的第四章阐述的基于深度结构关联的多目标跟踪模型属于整数规划的方式完成跟踪建模。本文的第五章涉及的分层图模型求解多目标跟踪问题采用了基于图模型的数据关联方法。本文第六章第七章提出的基于能量和动态图的组跟踪模型采用了基于行人分组模型的多目标跟踪模型。

第三章 数据集与跟踪性能评估方法

本章首先介绍本文所使用的 RGB-D 数据集和 RGB 数据集，以及针对不同数据集所采用的目标检测方法；然后，阐述多目标跟踪器的性能评价方法，介绍多种评价多目标跟踪的性能指标。

3.1 RGB-D 数据采集平台

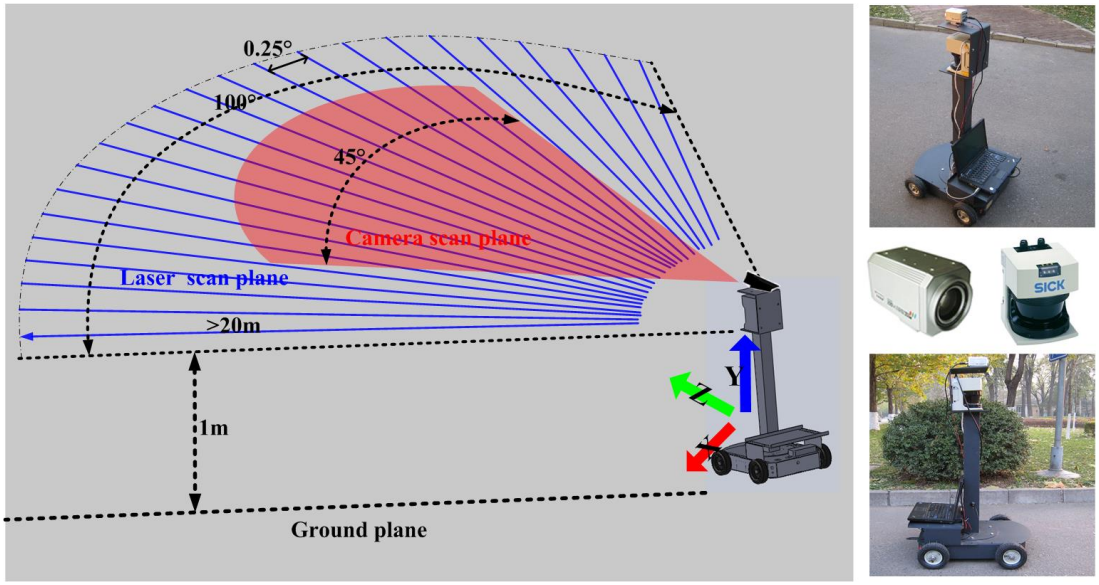


图 3-1 系统平台搭建

本文搭建了用于 RGB-D 数据采集的移动平台。系统如图 3-1 所示，由激光测距仪、视觉单目摄像头、数据存储设备、移动平台四部分组成。激光测距仪是德国 SICK 电子公司生产的脉冲式激光测距系统 LMS291。实验中选择参数是：视场角为 100°，精度为 0.25°，共 401 个激光数据点。系统选用维视公司 VS-870HC 工业数字相机，CCD 视野区应覆盖整个前方道路区域，水平视场角 45°，焦距为 16mm。该系统模拟驾驶员在驾驶场景下对交通场景的感知，所以本文激光测距仪安装高度为地面以上 1.2m 处，摄像头的高度在激光测距仪正上方 20cm 处，此高度与驾驶员的观测高度相同。在系统搭建完成后，采用作者在硕士期间实现的激光测距仪与单目摄像头联合配准的方法[93]完成对整个系统深度和视觉的联合参数配准工作。随后，该移动平台在中国科学院大学玉泉路校区的多个道路场景进行 RGB-D 数据采集，其中包括行人、机动车、自行车等多种障碍物。本文选用了 SDL-Campus，SDL-Crossing 和 SDL-Garden

三组视频和深度序列进行多目标跟踪算法的训练和测试。

3.2 数据集

为了验证所提出的多目标跟踪模型的准确性，本文采用真实场景中的多种类型的数据集作为测试数据集。而且场景的多样化可以避免模型对单一数据集的过拟合。但是，过多的测试数据集又会带来新的问题，每一个数据集都需要进行人工标定以获取真实行人轨迹，复杂且耗时。所以，本文采用的数据集来自于被广泛使用的经典多目标跟踪数据集，包括 RGB-D 数据集和 RGB 数据集。

3.2.1 RGB-D 数据集

RGB-D 数据集包括两部分：深度数据和图像数据。深度数据来自于激光扫描仪的场景真实数据。图像数据为传统的单目摄像头所得到二维图像信息。两种数据经过了相机和激光扫描仪联合标定，通过坐标转换矩阵从激光数据转换到图像的二维像素坐标，完成深度数据和图像数据对齐。使用的 RGB-D 数据集包括利用上文描述的 RGB-D 数据集采集系统所采集的 SDL 数据集，还有两个公用数据集：LIPD 数据集和 Sync 数据集。数据集的具体属性参考表 3-1。

表 3-1 RGB-D 数据集属性

数据集	序列	帧数	帧率	分辨率	行人密度	相机运动	深度扫描精度	深度距离精度
SDL	Campus	1000	10	中	低	是	0.25	1.0CM
	Crossing	356	10	中	中	否	0.25	1.0CM
	Garden	145	10	中	中	否	0.25	1.0CM
LIPD	Campus	4823	15	高	高	是	0.25	1.0CM
Sync	Sync	2147	15	高	低	是	0.25	1.0CM

LIPD 数据集[94]的深度数据来源于 ibeo 激光扫描仪，视频数据来源于单目鱼眼摄像头。该系统安放在雅马哈电动车上，并在真实街道上进行了数据采集。由于该数据录制于黄昏时分，所以在摄像头采集的视频中强烈的光照变化和低对比度给目标的检测和跟踪带来了极大挑战。

Sync 数据集[95]的数据构成与 LIPD 数据集相似，每一帧的图像数据对应深度距离数据。数据中出现的行人穿着了相似颜色的衣服，使得多目标跟踪时的目标外形特征区分不大，在目标匹配关联中容易出现错误的轨迹关联。

3.2.2 RGB 数据集

本文所采用的RGB数据集为著名的MOT Benchmark[96]多目标跟踪数据集。它是阿德莱德大学、苏黎世联邦理工学院及达姆施塔特工业大学联合发布的一个算法测评平台，旨在评测多个行人对象跟踪技术在视频监控环境下的算法性能。其中部分数据集来自于监控场景，部分来自于类似本文所采用的移动平台，还有来自于真实驾驶环境所拍摄的视频。有些测试场景跟踪对象多，人群密度大，相互之间遮挡严重，造成对象检测与跟踪难度非常大。该 Benchmark 提供公共的行人检测结果，供测试者作为多目标输入使用，还提供了统一标定的真实行人轨迹数据，供测试者检验跟踪模型使用。MOT Benchmark 包含 ETH 数据集、PETS 数据集、TUD 数据集、KITTI 数据集等。

ETH 数据集[97]是用水平前视摄像机在繁忙的街道上拍摄而成。视频分辨率为 640*480，帧率为每秒 15 帧。由于摄像机采用水平前视角度，加之摄像机的位置较低，目标被完全遮挡的情况经常发生。

PETS 数据集[98]源于 2000 年跟踪监控领域第一届国际研讨会（The first international workshop on Performance Evaluation of Tracking and Surveillance, 简称 PETS）。随后，逐年完善扩充。其中，使用最广泛的是 2009 年加入的 PETS09 数据集，它是一个利用多摄像机录制的具有重叠视域的跟踪评价视频。在基于单视角的多目标跟踪评价中，一般视角 1 经常被使用，其更接近于真实的视频监控视角。该视频一共 795 帧，分辨率 768*576。视频中行人的数目最多高达每帧 42 人，且具有长时间且复杂的运动轨迹。众多的行人也引起严重的的遮挡问题，这给多目标跟踪带来很大挑战。

TUD 数据集[7]录制于德国达姆施塔特市繁忙的街道以及大学校园。区别于其他数据集中很多行人是参加录制的志愿者，行走轨迹轨迹经过预先设定，该数据集中的三段视频虽然很短，但其中的行人完全来源于真实场景，并且视频的拍摄角度很低。这会带来两个挑战：1、视频的拍摄角度很低，使得行人处在被遮挡的时间加长；2、低的拍摄角度使 3D 位置估计很不准确，给准确重建轨迹带来很大困难。

KITTI 数据集[8]是德国卡尔斯鲁厄理工学院和芝加哥丰田技术研究所联合创办的一个算法评测平台，其视频由深度数据和视觉数据组成，旨在评测目标（机动车、非机动车、行人等）检测、目标跟踪等计算机视觉技术在车载环境下的性能，为机动车辅助驾驶应用做技术评估与技术储备。整个数据集中的多段视频常被用作多目标跟踪使用。

表 3-2 MOT 基准数据集包含的训练集和测试集属性

数据集	视频序列	帧数	帧率	分辨率	行人密度	相机运动	相机标定
PETS09	S2L1	794	7	中	高	否	是
	S2L2	436	7	中	高	否	是
TUD	Campus	71	25	中	中	否	否
	Crossing	201	25	中	中	否	否
	Stadtmitte	179	25	中	中	否	是
AVG	TownCentre	4500	25	高	高	否	是
ETH	Bahnhof	1000	14	中	高	是	否
	Sunnyday	354	14	中	高	是	否
	Pedcross2	840	14	中	高	是	否
	Linthescher	1194	14	中	高	是	否
	Crossing	219	14	中	高	是	否
	Jelmoli	440	14	中	高	是	否
KITTI	KITTI-13	340	10	高	低	是	是
	KITTI-17	145	10	高	中	是	否
	KITTI-16	209	10	高	中	是	否
	KITTI-19	1059	10	高	中	是	否
Venice	Venice-1	450	30	高	中	否	否
	Venice-2	600	30	高	中	否	否

3.3 多目标检测器

本文提出的基于 RGB-D 数据的多目标的跟踪方法和其他一流的多目标跟踪方法一样，采用“先检测再跟踪”的框架。这也意味着在进行多目标跟踪前需要先进行多目标的检测。本节则简要介绍系统所采用检测方法[93,99]。

HOG 特征提取：将图像（人体样本为 64×128 像素的训练样本）按 8×8 个像素分割为若干个单元（cell）；将相邻的 4 个 cell（如田字结构）划分为一个 block（利用每个 cell 进行窗口滑动而生成）。将$[-\pi/2,\pi/2]$的梯度方向平均划分，其中规定每 20 度分成一个 bin，即 9 个方向 bins。对每个 cell，将所有像素的所有梯度方向投影以建立各自的梯度方向直方图。将每个 block 中含有的 4 个 cell 的梯度方向直方图连接起来，形成 36 维的向量。再将所有的 block 内的 36 维向量归一化后在向量中串联起来，由此得到每个训练样本的 HOG 特征向量。图 3-2 为 HOG 特征示意图，反映了行人的轮廓。HOG 特征提取之后，

训练一个 SVM 分类模型用于目标检测。对于候选行人区域，通过逐像素扫描的图像块来判断该图像块内是否存在行人目标，如图 3-2 所示。

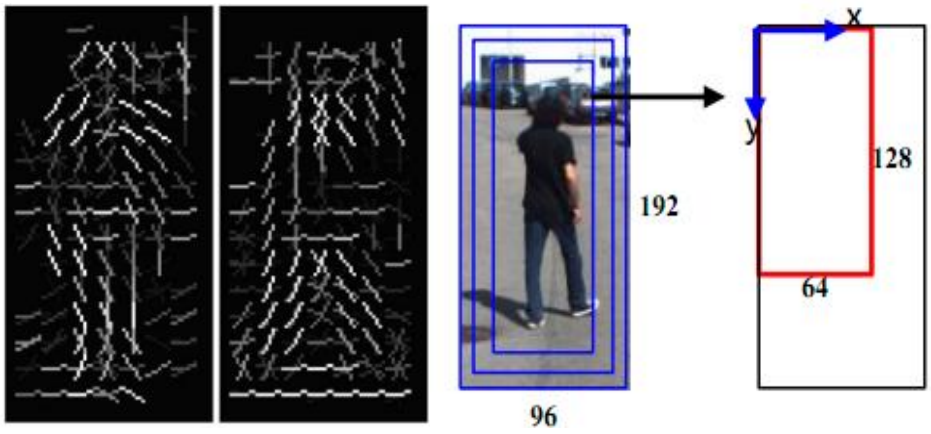
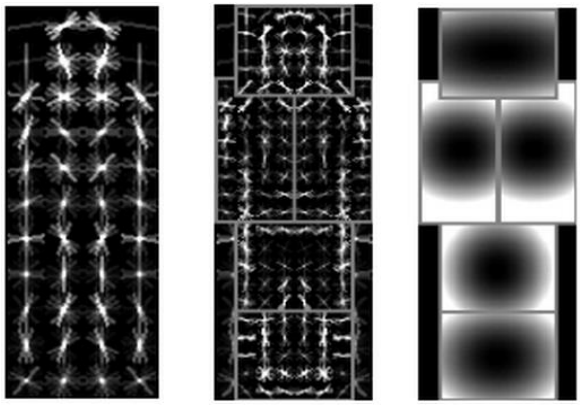


图 3-2 HOG 特征和人体目标检测示意图

SVM 分类器：SVM 是一种基于结构风险最小化准则的学习方法，其核心思想是在样本特征空间建立一个超平面作为决策曲面，构造最优平面，使得正例样本和反例样本之间的分隔边界最大化，从而达到最大的泛化能力。SVM 的主要优点是对目标模式变化的鲁棒性，已经被证明是一种更系统的学习线性和非线性的分类边界的方法[100]，并且在目标检测等应用中表现出了良好的性能。

DPM 检测器：DPM 算法由 Felzenszwalb[101]于 2008 年提出，是一种基于部件的检测方法，对目标的形变具有很强的鲁棒性。目前 DPM 已成为众多分类、分割、姿态估计等算法的核心部分。DPM 算法采用了改进后的 HOG 特征，SVM 分类器和滑动窗口检测思想，针对目标的多视角问题，采用了多组件的策略，针对目标本身的形变问题，采用了基于图结构的部件模型策略。图 3-3（a）和（b）是其中一个组件的根模型和部件模型的可视化的效果。图 3-3（c）为部件模型的偏离损失，越亮的区域表示偏离损失代价越大。

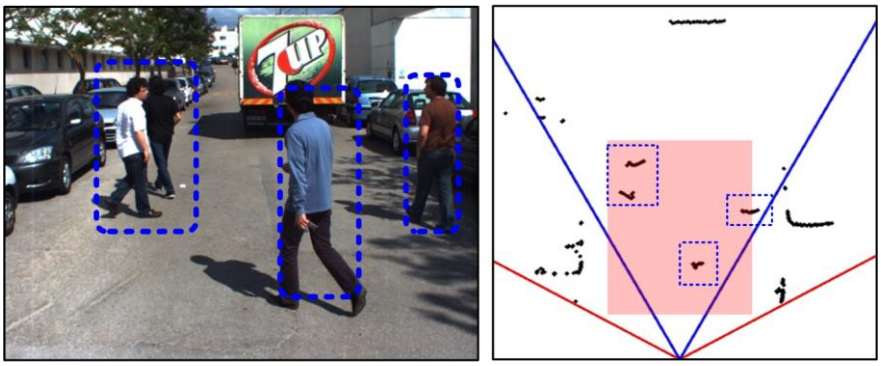


(a) (b) (c)

图 3-3 DPM 行人模型

在 RGB-D 数据集中，本文通过视觉图像信息和深度信息串行检测目标。本文采用的基于行人目标检测算法共分为三个步骤：深度滤波，图像投影和视觉检测。

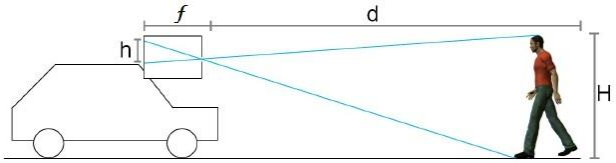
步骤一：深度滤波。在深度数据中，障碍物表现为相互靠近且长度在一定阈值范围内的一组深度点，因此首先需对离散深度信息进行聚类分析。由于不能事先确定深度数据中包含的类别数目，本文使用层次聚类方法，自下而上划分。类与类之间的距离采用固定阈值进行划分。



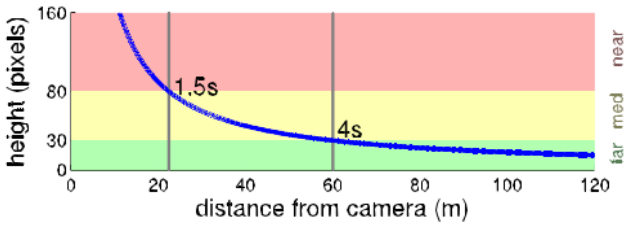
(a) (b)

图 3-4 从激光点中产生候选目标

步骤二：图像投影。深度信息只能提供目标的宽度，但是目标在图像上的高度不能获得。行人目标在图像上的特征与真实的行人有所不同，行人在图像中的大小不仅受实际大小的影响，且反比于行人到镜头的距离。所以利用聚类后的深度信息根据坐标转换可以求得目标在二维图像上的宽度，再经过小孔成像的原理，可以求得在特定深度上的行人在图像上的高度，如图 3-5（a）所示。假设行人的深度为 d，摄像机的焦距为 f，假设行人的高度 170cm，则在图像上的高度理论上应该是 $h \approx Hf/d \approx 170f/d$。但是，图像上存在像素的畸变，矫正后的行人在图像上的投影高度应该满足如图 3-5（b）所示的曲线[32,102]。



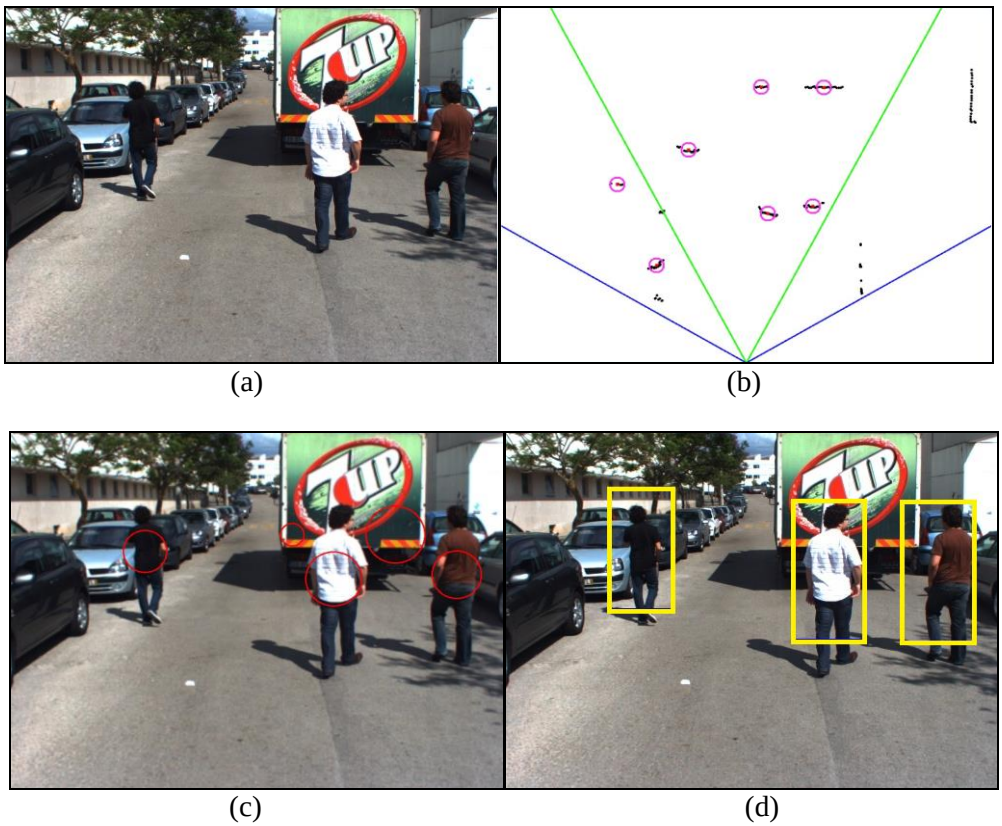
(a)



(b)

图 3-5 人体在图像上的显示比例

步骤三：视觉检测。在经过图像投影后，可以获得图像上的行人候选区，这一过程和单纯的不同尺度下滑动窗口扫描策略（sliding window）相比，可使目标检测达到实时处理。进而利用 DPM 检测算法对行人候选区域内，在小尺度下可以完成检测。检测结果如图 3-6 所示，其中(a)为原始图像，(b)为经过聚类分析后的激光片段，(c)为利用激光特征选取的候选目标在图像层上的投影，(d)为利用图像检测方法进行确认行人检测结果，可见图中在激光层聚类结果中与行人类似的车体片段在图像检测确认中被排除。



(a) (b)

(c) (d)

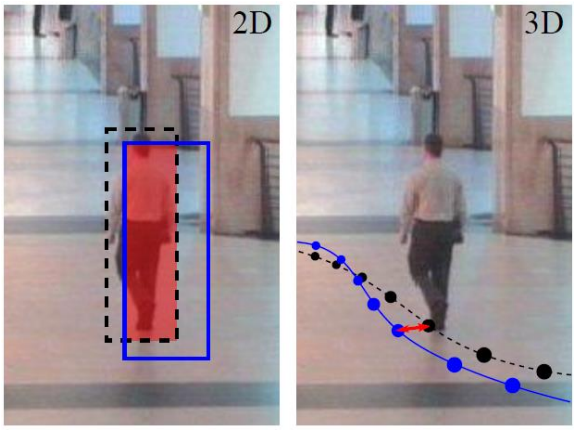
图 3-6 检测结果示意图

在 RGB-D 数据集中，本文采用上述方法对多目标进行检测，并将检测结果

作为多目标跟踪器的输入。在 RGB 数据集中，利用数据集提供的检测结果作为输入，这样为各种多目标跟踪方法提供公平的对比标准。

3.4 评测指标

要设定一种评价指标，首先要分析算法的输出。一个理想的多目标跟踪算法应该具有如下的输出：能精确检测到每个目标的位置；能保持对目标的连续跟踪；每个目标对应一个唯一的 ID，能对遮挡具有一定的鲁棒性。目前在目标跟踪的评价标准中，CLEAR-MOT 评价标准[103]是一套能够满足上面三点的评价指标体系。它起源于 2006 年 Classification of Event，Activities and Relationships（CLEAR）研讨会，现在被多目标跟踪领域广泛采用的评测工具。其中两项综合指标 MOTA（Multiple Object Tracking Accuracy）和 MOTP（Multiple Object Tracking Precision）不仅记录了跟踪中出现错误的次数，还反映了目标估计位置与其真实位置之间的距离关系。



（a）　　（b）

图 3-7 RGB 数据集和 RGB-D 数据集中跟踪正确的评判方法

在跟踪过程中一个目标是否被跟踪器跟踪和成功，决定于跟踪器的输出位置 H 和真实位置 GT 之间的关系。所以会计算这两者之间的距离。在 RGB 数据集中利用目标的输出框与 GT 给定框的重叠面积来计算，

$$\bar{d}\left(H,GT\right)=\frac{bbox\left(H\right)\bigcap bbox\left(GT\right)}{bbox\left(H\right)\bigcup bbox\left(GT\right)},\qquad(3\text{-}1)$$

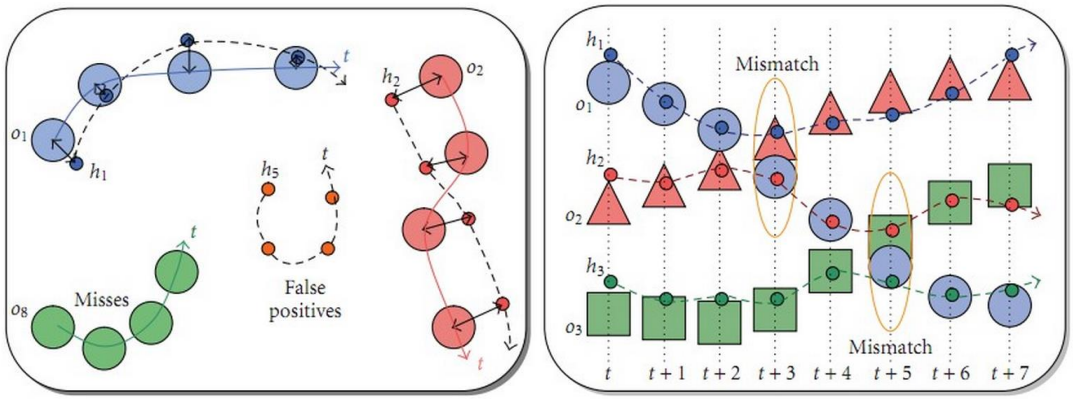
其中，*bbox* 表示目标的边界框（Bounding Box），如图 3-7（a）中所示。$\bar{d}=0$ 表示没有覆盖，$\bar{d}=1$ 表示完全覆盖。通常设置 0.5 为阈值，当 $\bar{d}>0.5$ 表示轨迹点被成功跟踪到。在 RGB-D 数据集中，更合理的方式是采用真实的世界坐标来表示物体位置，如图 3-7（b）所示。采用 bbox 底边的中心点与实际位置之

间的欧式距离作为两者的距离度量值，通常设定 $\bar{d}<1$ meter 为跟踪成功的阈值。

MOTA：多目标跟踪中有两个指标的定义与目标检测中类似，漏检（False Positive, FP）和误检（False Negative，FN）。前者表示跟踪器输出的轨迹点未能和任何真实轨迹（Ground Truth，GT）匹配成功。后者表示真实轨迹中存在的轨迹点，但是没有跟踪到。在多目标跟踪中，不仅需要跟踪对所有目标（降低误检率）和抑制虚警错误（降低漏检率），还需要使得整条轨迹尽量保持完整，换句话说，一个目标从进入场景到走出场景的轨迹完整性。当跟踪器的输出目标和真实轨迹点之间不能匹配，就记做一个误匹配（Identity Switch，IDS）错误。如图 3-8（b）中，三条轨迹都是完整的，并没有出现漏检，也没有出现误检。但是，跟踪中两次出现了误匹配，使得轨迹与目标并没有做到正确匹配。因此，MOTA 得分是建立在上面三项指标：漏检，误检，误匹配的一个综合得分：

$$\text{MOTA}=1-\frac{\sum_t\left(FP(\text{t})+FN\left(t\right)+IDS(\text{t})\right)}{\sum_t N_{GT}(\text{t})}. \qquad (3\text{-}2)$$

此时，如果出现的错误跟踪次数为零，式中的分子求和则为零，所以此时的准确度为 100%。出现的跟踪轨迹错误越多，得分则会越低。



（a） （b）

图 3-8 连续帧的错误匹配情况

MOTP：MOTA 指标是对多目标跟踪过程中的轨迹点错误次数做了精确的统计，而 MOTP 反映了多目标跟踪中轨迹定位精确度。它统计了跟踪器输出的轨迹点与 GT 轨迹点之间的平均距离。

$$\text{MOTP}=\frac{\sum_{t,i}\bar{d}\left(GT_i\left(t\right),H_i\left(t\right)\right)}{\sum_t m_t}, \qquad (3\text{-}3)$$

式中，$\bar{d}\left(GT_i(t),H_i(t)\right)$表示每个轨迹点到真实轨迹点的距离，$m_t$ 表示在 t 时刻轨迹点个数。在 RGB 数据集中，如图 3-8（a）中，MOTP 表示轨迹点与目标区域的覆盖区域的平均值，在 RGB-D 数据集中，则表示 3D 区域在轨迹投影区的平均距离值。从这个定义可以看出，MOTP 指标比较依赖 GT 中的轨迹位置，而真实的轨迹位置是往往是人在数据集中手工标出来的，所以 GT 中的位置信息可能存在误差甚至错误。

上面描述的几项指标是 CLEAR-MOT 准则基于整个视频计算所有轨迹中所有目标的平均准确率，而不是针对跟踪轨迹的完整程度来衡量。如图 3-8（b）中，仅出现了两次 IDS 错误，但是三条轨迹的输入和输出都不是同一个目标。所以从识别的角度讲，这种输出虽然 MOTA 达到将近 100%，MOTP 得分也很高，但是没有一个目标的跟踪输出是与开始是的目标对应正确的。对于轨迹较长的目标，其在整个跟踪样本中所占比重就会比较大，一旦跟踪出现错误，就会对整个系统的评价产生较大的影响，而轨迹较短的目标由于帧数较少，即使完全在跟踪中丢失，对于整个数据集来讲影响也是很小的[1]。但是这样的评判标准不符合客观规律。对于每一条轨迹，应该公平对待，即在结果中所占比重相同。基于此原因，CLEAR-MOT 准则又加入了多个基于轨迹完整度的指标。成功跟踪 MT（Mostly Tracked）、部分跟踪 PT（Partially Tracked）、跟踪丢失 ML （Mostly Lost）和 Frag.（Fragments）等。表 3-3 列出了基于 CLEAR-MOT 准则的多项评价指标，其中的箭头↑表示得分越高越好，↓表示得分越低越好。

表 3-3 多目标跟踪评价指标

指标	含义
MOTA（↑）	目标跟踪的准确度
MOTP（↑）	目标跟踪的精确度
GT	轨迹的真实位置
FP（↓）	输出的轨迹点未能和任何真实轨迹匹配成功
FN（↓）	真实轨迹中存在，但是没有跟踪到的轨迹点
MT（↑）	轨迹被成功跟踪 80%以上
PT（↓）	轨迹被成功跟踪 20%到 80%之间
ML（↓）	轨迹被成功跟踪少于 20%
IDS（↓）	目标之间 ID 交换的次数
Frag.（↓）	轨迹断裂的次数

3.5 本章小结

本章介绍了 RGB-D 数据采集平台的搭建以及数据获取，简要描述了如何在 RGB-D 数据集进行目标检测，详细介绍了多目标跟踪所采用的 RGB-D、RGB 数据集与实验评估方法及性能评价指标。后四章中所用到的数据集及评测标准均可参考本章内容。

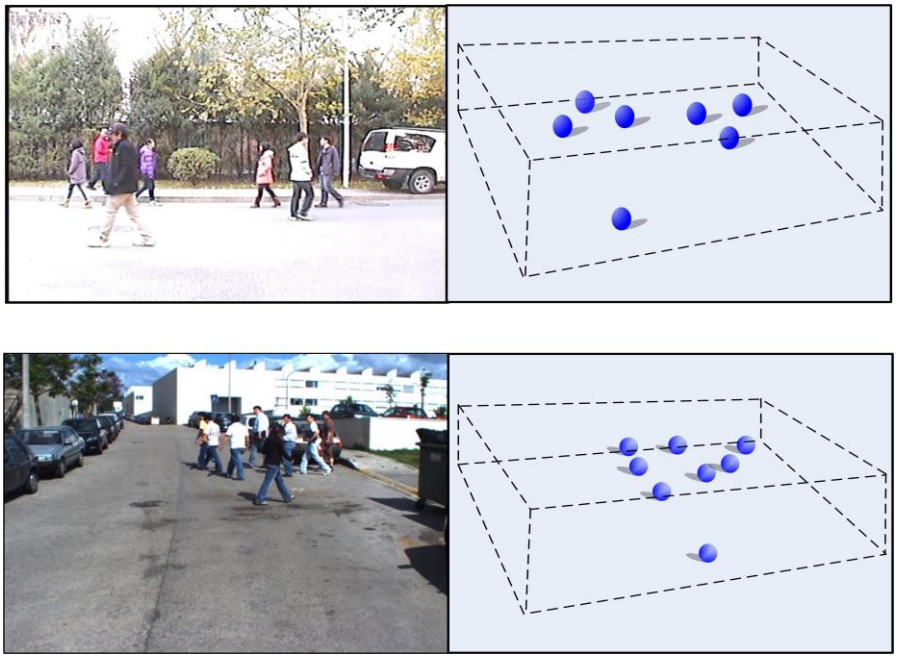
第四章　基于深度结构关联的多目标跟踪方法

4.1 模型概述与创新点

由于 RGB-D 数据加入了深度信息，目标观测可以被放到世界坐标中进行建模，如图 4-1 所示。本章提出了基于深度结构关联的多目标跟踪模型，将场景中的多目标划分到不同的深度链状结构中进行三维分析。利用整数规划中的多维数据分配问题对多目标之间的数据关联进行建模。在面对多目标跟踪过程中的遮挡问题时，链状结构利用深度值对目标匹配代价进行了重新加权，使得目标在场景的不同位置更具有区分性。经过 RGB-D 数据集验证，本章提出的模型可以在交通场景的多目标跟踪问题上做到实时处理。方法的创新点在于：

1）提出了深度数据关联模型对基于 RGB-D 的目标观测进行多目标跟踪；

2）提出了基于深度结构的遮挡问题处理方法。



（a）　　（b）

图 4-1 目标的 RGB-D 信息

表 4-1 本章使用的符号及含义

符号	含义
A	目标观测之间的相似度矩阵
X	目标连接关系的二进制矩阵
o_i	目标 i 的观测，特征表示为 $o_i=\{\pi_i=(u_i,v_i,z_i),\varphi_i,\theta_i\}$
d_i^t	目标观测 o_i 所属的深度结构标号
D^t	记录 t 时刻，目标观测所属深度结构的标号矩阵
s_i^t	目标观测 o_i 的深度得分
S^t	记录 t 时刻，目标观测深度得分的矩阵
ε_i	目标观测 o_i 相对于所属深度结构中心的偏移值
β_i	目标观测 o_i 在所属深度结构中的深度因子
$\bar{\pi}_m$	第 m 条链状结构的中心位置
$\Phi(\pi_i)$	从图像坐标到世界坐标的坐标转移函数

4.2 多维数据分配问题

在第二章多目标跟踪方法综述中提到，多目标跟踪问题可以使用基于多帧的数据关联方法求解。本章介绍的深度结构关联模型将多目标跟踪问题转化成基于改进的多维分配（Multi-Dimensional Assignment，MDA）问题进行建模。MDA 问题属于基于整数规划数据关联的求解方法，但是当面对三帧及三帧以上的数据关联时属于 NP 难问题。本方法结合图像数据和深度数据改进传统的多维数据分配问题，探索了多目标在持续多帧图像之间的时空关系，同时利用目标的外形和运动特征，来解决复杂场景内的多目标跟踪问题。在多目标跟踪问题中，假设有 k 帧图像，每帧含有 N 个观测目标 $\{o_1,o_2,\cdots,o_n\}$，则对于这 N 个目标的最优关联匹配可以划分到 k 部结构中求解 MDA 问题。

$$\begin{aligned}&\max\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\cdots\sum_{i_k=1}^{N}A_{i_1i_2\cdots i_k}X_{i_1i_2\cdots i_k},\\&s.t.\sum\sum_{I\setminus i_t}\cdots\sum X_{abcd}\le 1;X_{i_1i_2\cdots i_k}\in\{0,1\}\end{aligned}\qquad(4\text{-}1)$$

其中，$\mathrm{A}_{i_1i_2\cdots i_k}$ 表示相似度矩阵，矩阵中的元素 A_{ij} 表示相邻帧目标观测 o_i 和 o_j 之间相似度。$X_{i_1i_2\cdots i_k}$ 表示二进制指示矩阵，矩阵中的元素 X_{ij} 表示相邻帧的观测 o_i

和o_j是否连接。如果连接，则$X_{ij}=1$，表示o_i和o_j是同一个目标在相邻帧的不同观测。$X_{i_1 i_2 \cdots i_k}$中的连接关系由目标观测之间的相似度矩阵$\mathrm{A}_{i_1 i_2 \cdots i_k}$决定。约束中的$I \setminus i_t$表示约束作用于$k$帧的每一个目标观测，并且约束中的符号是小于等于号，这意味着每个目标观测最多只属于一个目标。如果$\sum_i X_{ij}=0$，则意味着目标o_j未能与前一帧中的任何目标观测匹配，它是新一段轨迹的开始。所以$X_{i_1 i_2 \cdots i_k}$矩阵记录了目标观测所连接成的轨迹的起止时间。

4.3 深度结构关联模型

在观测场景中，一个行人目标的观测向量为：三维特征$\pi_i=(u_i,v_i,z_i)$，外形特征φ_i和运动特征θ_i。其中π_i是检测器输出的目标的空间的中心坐标，(u_i,v_i)是目标观测的图像域的坐标，z_i是目标观测的深度值。

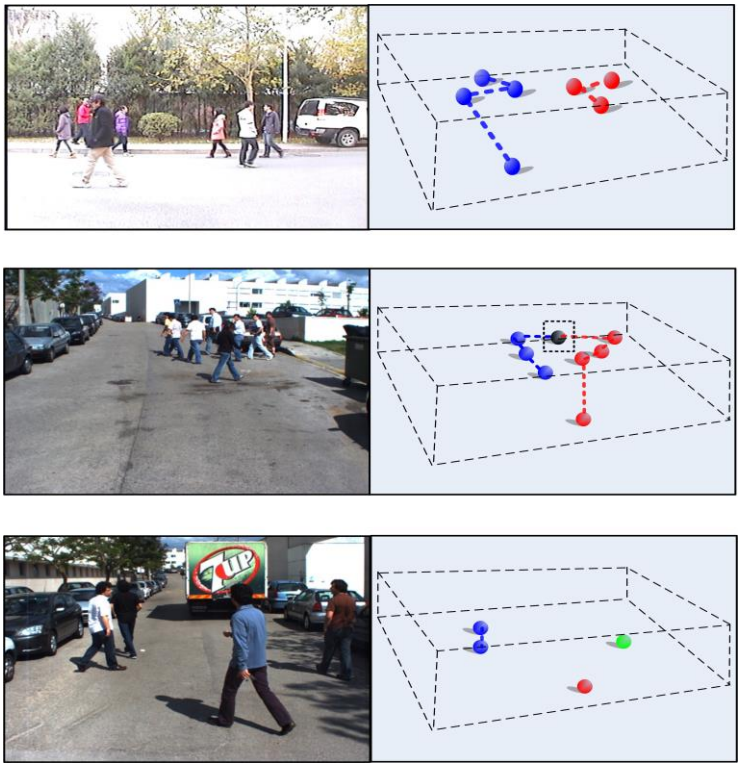


图 4-2 多行人目标间的深度关系

4.3.1 深度结构标号矩阵

对于从$t-k$帧到t帧中共$k+1$帧图像中的目标观测，定义矩阵D^t描述各个目标观测之间的深度结构关系。假设每一帧有N个目标观测，用如图 4-2 所示的链状结构来表示目标观测之间的深度关系，矩阵D^t记录了每一个观测所属的

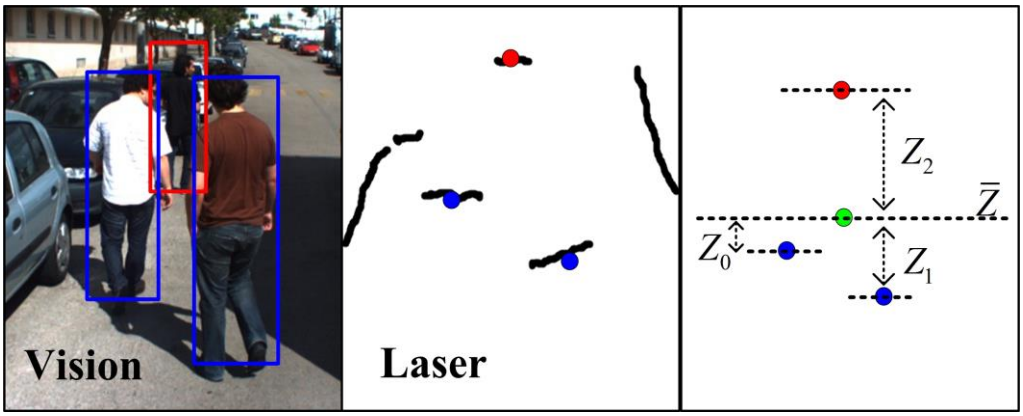
链状结构的标号。当 N=5，k=3 时：

$$D=\begin{bmatrix}1 & 1 & 2 & 2 & 2 & 3\\ 1 & 1 & 1/2 & 2 & 2 & 3\\ 1 & 1 & 2 & 2 & 3 & 3\\ 1 & 2 & 2 & 2 & 2/3 & 3\end{bmatrix}, \tag{4-2}$$

其中的“1”，“2”，“3”分别代表目标观测所属的链状结构标号。同一个目标观测也可能同时属于两个不同的链状结构，如公式（4-2）中的“1/2”和“2/3”均表示该目标观测属于两个链状深度结构。图 4-2 第一行中目标属于两个不同的深度链状结构，第二行中出现了同一个目标属于不同的深度链状结构的情况，第三行中出现了一个链状结构仅含有一个目标的情况。为了把同一场景中的目标划分到不同的链状结构中，模型定义了每一个目标观测相对于链状结构的偏移值 ε_i，

$$\varepsilon_i=\sqrt{\frac{1}{|C_m|}\sum_{\pi_i\subset C_m}|\Phi(\pi_i)-\bar{\pi}_m|^2}, \tag{4-3}$$

其中，$\Phi(\pi_i)$ 表示从图像坐标到世界坐标的坐标转移函数，$\bar{\pi}_m$ 表示第 m 条链状结构的中心位置，C_m 表示该中心位置在图像上的投影坐标。通过这种定义，当目标观测比较密集，具有很高的概率相互遮挡时，他们会被划分进同一条链状结构。此时，偏移值 ε_i 起到控制链状结构区域大小，以及区域之间的重叠区域大小的作用。



（a）　　（b）　　（c）

图 4-3 遮挡问题求解示意图

4.3.2 深度权值矩阵

在同一链状结构体内，为了区分目标观测之间的深度关系，模型利用目标

观测深度值 z_i 对每一个观测进行深度加权， s_i 表示每个目标观测在链状深度结构加权后的深度得分，

$$s_i = \log\frac{\beta_i}{1-\beta_i}, \tag{4-4}$$

$$\beta_i = \left[1+\exp(\frac{1}{\mathrm{m}}\sum_{j=1}^{m} z_j^{t-1} - \hat{z}_i^t)\right]^{-1}, \tag{4-5}$$

其中， $\hat{z}_i^t$ 表示 t 时刻目标观测的深度 z_i^t 与链状结构的平均深度 $\bar{z}^t$ 之间的相对深度，$\hat{z}_i^t = z_i^t - \bar{z}^t$。$\frac{1}{\mathrm{m}}\sum_{j=1}^{m} z_j^{t-1}$ 表示该链状深度结构在 t-1 时刻的平均深度值。因此，β_i 反映了目标观测在同一链状深度结构内相邻时刻的深度关系，给处在位置距离接近的目标观测一个具有区分性的深度权重值。如图 4-3 所示的深度结构由三个处于不同深度的目标观测组成，绿色的点表示该结构的平均深度值 $\bar{z}^t$。根据公式（4-5）可求的三个目标观测的 β_i，i=1,2,3。进一步分析可知，当目标观测具有较小的深度值时（如图 4-3（c）中的 Z_0 和 Z_1），$0.5<\beta_i<1$；当具有较大的深度值时（如图 4-3（c）中的 Z_2）,$0<\beta_i<0.5$。将 β_i 带入到公式（4-4）时发现，深度权值的值域为（-1,1），同时具有较小深度值的目标观测的深度得分 s_i 为正值，而具有较大深度值的则为负值。利用深度得分 s_i 和链状结构标号矩阵 D^t，我们定义了深度得分矩阵 S^t：

$$S^{(t)}=\begin{bmatrix} 0.5 & -0.5 & 0.8 & -0.4 & -0.5 & 0 \\ 0.7 & -0.2 & -0.4/0.7 & -0.3 & -0.5 & 0 \\ 0.5 & -0.5 & 0.5 & -0.5 & 0.5 & -0.5 \\ 0 & -0.2 & 0.4 & 0.2 & -0.4/0.5 & -0.5 \end{bmatrix}. \tag{4-6}$$

深度权值矩阵 S^t 与记录链状结构标号的矩阵 D^t 的大小一致，S^t 记录了每个目标观测在所在链状深度结构中的深度得分。其中，“-0.4/0.7”和“-0.4/0.5”表示目标观测在不同链状深度结构的深度得分。在深度权值矩阵 S^t 中，有的目标观测的深度权值为“0”，通常属于这种情况是单一目标观测构成一个深度结构，如图 4-2 第三行所示， $\beta_i \approx 0.5$，进而求得 $s_i \approx 0$。

4.4 模型求解

本方法利用深度结构矩阵 D^t 和深度权值矩阵 S^t 结合迭代估计的方法来求解 MDA 问题。用四帧之间的多目标关联来阐述求解过程。假设在 t_a~t_d 四帧中各有 n_a~n_d 个目标观测，则 MDA 问题的目标方程（4-1）可以写成如下形式：

$$\begin{aligned}&\max \sum_{a=1}^{n_a}\sum_{b=1}^{n_b}\sum_{c=1}^{n_c}\sum_{d=1}^{n_d} A_{abcd}X_{abcd},\\ &s.t.\ X_{abcd}\in\{0,1\};\\ &\sum_{b=1}^{n_b}\sum_{c=1}^{n_c}\sum_{d=1}^{n_d} X_{abcd}\le 1;\quad a=1,2,\cdots n_a.\\ &\qquad\vdots\qquad\qquad\qquad\vdots\\ &\sum_{a=1}^{n_a}\sum_{b=1}^{n_b}\sum_{c=1}^{n_c} X_{abcd}\le 1;\quad d=1,2,\cdots n_d.\end{aligned} \tag{4-7}$$

其中，A 和 X 矩阵的定义与公式（4-1）相同。A 代表相似度矩阵，A_{ab} 表示 t_a 和 t_b 之间目标观测的相似度。X 是一个二进制关联矩阵，表示两帧之间的一组目标观测是否为同一个目标。$X_{ab}=1$ 表示目标观测 o_b 的轨迹与目标观测 o_a 是在同一目标的相邻两段。这种四部图结构使得轨迹和相似度可以在多帧嵌套计算。同时，这种图结构使得一段由（a,b,c,d）组成的长轨迹在多帧之内分解成每一帧内的小段轨迹组合（a,b），（b,c），（c,d）。此时对关联矩阵 X_{abcd} 进行相似的分解操作 $y_{ab}\times g_{bc}\times h_{cd}$。则公式（4-7）可分解为

$$\begin{aligned}&\sum_a\sum_b\sum_c\sum_d A_{abcd}X_{abcd}\\ &=\sum_a\sum_b\sum_c\sum_d A_{abcd}y_{ab}g_{bc}h_{cd}\ .\\ &=\sum_a\sum_b y_{ab}\sum_c g_{bc}\sum_d h_{cd}A_{abcd}\end{aligned} \tag{4-8}$$

这种分解方式，使得原本需要在所有帧之间两两进行的边连接变成了相邻帧之间的连接。这种矩阵分解策略使得计算复杂度从 $O\left(n^k\right)$ 降为 $O\left((k-1)\times n^2\right)$。

本方法所用的目标观测相似度 A 矩阵由 RGB 和 D 信息结合计算。其中的相似度矩阵 A 定义为

$$A_{ij}=\begin{cases}A_a\left(\varphi_i^{t-1},\varphi_j^t\right)A_p\left(\theta_i^{t-1},\theta_j^t\right)A_m\left(X_i^{t-1},X_j^t\right), & if\ \ d_i^{t-1}=d_j^t\\ 0, & otherwise\end{cases} \tag{4-9}$$

其中，$A_a(\cdot)$、$A_p(\cdot)$、$A_m(\cdot)$ 分别表示观测之间的外形、位置、运动特征相似度，并归一化三种特征满足高斯分布。其中 $d_i^{t-1}=d_j^t$ 表示目标观测 o_i 和 o_j 在相同的

深度链状结构内。这种相同深度链的约束将进一步减小计算复杂度。

本方法将深度得分矩阵加入到目标方程（4-8）中，通过固定其他帧的关联矩阵参数，更新相邻帧相同深度链状结构内“成对”的关联参数。因此，关联矩阵 X 在加入深度得分 S 后可以进行如下分解，

$$\begin{aligned}&\sum_{a}\sum_{b}\sum_{c}\sum_{d}(A_{abcd}+S_{abcd})X_{abcd}\\&=\sum_{a}\sum_{b}y_{ab}(A_{ab}+S_{ab})\sum_{c}g_{bc}(A_{bc}+S_{bc})\sum_{d}h_{cd}(A_{cd}+S_{cd})\end{aligned}. \tag{4-10}$$

在求解目标方程（4-10）时，采用迭代策略来搜索多帧内的全局最小值，得到关联矩阵 X。首先，将搜索匹配缩小在多帧内的相同深度链状结构内完成，这样可以排除匹配在深度跨度大的目标观测之间进行。在同一深度链状结构内，搜索沿着深度下降的方向进行，我们先匹配深度小的目标观测，再沿着深度链向深度大的目标观测进行匹配。这样可以使得深度大的目标不会错误地匹配到深度小的目标。同时，深度大的目标观测往往处在被遮挡的位置，因此，我们可以避免被遮挡的目标观测先被搜索。对于一个深度链状结构只有一个目标观测的情况，直接对其进行运动位置信息匹配，省去计算外形特征相似度。这样可以进一步减小搜索时间复杂度。

4.5 实验验证

在实验中采用了多种类型的方法与提出的 DSA 模型进行实验对比：1）基于 RGB 的多目标跟踪方法：Berclaz[75]和 Zhang[74]使用基于网络流的多目标跟踪模型；Andriyenko[104]使用基于连续的轨迹分析法进行多目标跟踪建模；Milan[105]使用基于能量的轨迹建模方式; Yang[71]使用基于在线学习的 CRF 模型。 2）基于深度的多目标跟踪方法：“NN”使用第二章中基于全局最紧邻的方法；“MDA”使用基于 K 部图求解多目标数据关联，在此方法中不使用本章使用的深度链状结构划分，直接在全图内进行匹配搜索。3）使用本章提出的 DSA 模型进行建模求解。前两种类型的跟踪方法在第二章多目标跟踪方法综述中有详细的描述。

本章使用的多目标跟踪评测方法在第三章中已经详细介绍。本章采用四个 RGB-D 数据集进行实验验证分别是：Sync 数据集，SDL-Garden 数据集，SDL-Crossing 数据集和 SDL-Campus 数据集。此外，为了达到实验公平的原则，对所有的方法实验采用相同的目标检测结果作为输入。详细的对比实验结果如表 4-2 所示。表中所有用粗体标明的结果为该组实验中最好的结果。

4.5.1 实验结果分析

从表 4-2 中可以看出，在四个数据集上，利用本章提出的 DSA 模型在所有指标中都做到了最好。在速度方面，虽然没有达到前面三种方法相同的速度，但是仍然能够满足交通场景中实时跟踪的要求。图 4-4，4-5 是实验方法在三个数据集上的跟踪结果，三个图中第一行均为世界坐标成像。其中两条蓝线之间代表摄像机的视场范围，两条红线之间是深度扫描范围。

表 4-2 对比实验结果

数据集	**方法**	**Recall**	**Prec.**	**GT**	**MT**	**PL**	**ML**	**Frag.**	**IDS**
Sync	Berclaz[75]	69.6	74.8	66	64.5	22.7	12.8	45	23
	Andriyenko[104]	73.4	78.3	66	69.7	19.7	10.6	39	18
	Milan[105]	75.6	80.2	66	71.2	18.2	10.6	37	16
	NN	54.5	64.3	66	45.5	30.3	24.2	52	31
	MDA	73.1	78.6	66	68.2	15.2	16.6	39	17
	DSA	**85.0**	**89.7**	66	**80.3**	**10.6**	**9.1**	**21**	**7**
SDL-Garden	Zhang[74]	67.8	72.5	10	60.0	20.0	20.0	7	5
	Andriyenko[104]	76.6	81.0	10	70.0	20.0	10.0	5	4
	NN	56.4	64.9	10	40.0	30.0	30.0	8	6
	MDA	68.7	73.5	10	60.0	20.0	20.0	5	4
	DSA	**94.5**	**98.3**	10	**90.0**	**10.0**	**0.0**	**2**	**0**
SDL-Crossing	Berclaz[75]	68.9	74.5	92	60.9	17.4	21.7	58	31
	Andriyenko[104]	70.4	76.4	92	63.0	20.7	16.0	51	29
	Yang[71]	72.3	77.8	92	64.1	21.7	14.2	47	26
	NN	59.4	65.6	92	43.5	23.9	32.6	69	38
	MDA	67.0	73.5	92	59.8	22.8	17.4	49	25
	DSA	**82.4**	**87.3**	92	**76.1**	**15.2**	**8.7**	**28**	**14**
SDL-Campus	Zhang[74]	76.4	79.8	74	71.6	18.9	9.5	30	16
	Milan[105]	80.0	84.5	74	75.7	16.2	8.1	26	14
	NN	67.7	74.6	74	60.8	21.6	17.6	37	19
	MDA	78.3	82.9	74	73.0	17.6	9.4	26	15
	DSA	**85.6**	**89.3**	74	**81.1**	**12.2**	**6.7**	**14**	**8**

SDL-Garden 数据集：整个序列中的四个行人出现了反复的位置遮挡关系，

如图 4-4，如果只利用单纯的视觉信息进行跟踪，穿红色衣服的行人与紫色衣服的行人由于在很长一段时间内是处于遮挡和被遮挡关系，容易被混淆。但 DSA 模型中，加入深度动态权重因子后，紫色衣服行人的转移代价为负值，在进行模型求解中优先会被选择。而处在远处的红色行人和白色行人则具有比较大的转移代价，处在末位被匹配选择的位置。所以 DSA 模型在处理遮挡问题时，在具有深度权值后能够更加有效的解决复杂遮挡问题，使得跟踪的轨迹的 Frag. 和 IDS 错误更少。因此，我们观测到表 4-2 中的 MT 比其他对比试验的结果高，ML 则比对比实验结果低。

SDL-Crossing 数据集：该数据集记录了长时间内，位于丁字路口出现在车辆前方的行人运动状况，实验对每个行人进行了标号，DSA 模型，面对处于反复遮挡位置的行人，完成了正确的标号保持。尤其处在车辆正前方的区域内，DSA 模型没有失误，只是在较远处的区域内出现了行人的标号前后不一致的 IDS 错误。

Sync 数据集：在这个数据集中，目标数目不确定的行人一直处在背景比较复杂的道路上进行交错行走，图 4-5 中（b-e）分别展示了四种对比方法：Berclaz 等[75]，Andriyenko[104]，Milan[105]和 MDA 方法，从图中（b-d）的实验结果可以看出仅利用 RGB 信息作为目标关联时能量和代价的参考，结果要明显差于 RGB-D 数据进行目标关联的跟踪结果。而加入到动态深度方向的权重因子后，遮挡情况出现后， DSA 模型仍然可以做到不丢失目标，而在加入了多种颜色，运动，方向等信息后能够找到全局最优解。DSA 模型的结果如图 4-5（e）所示，在多目标的集群点内，模型对于不同深度的行人基于区分性的深度动态权重因子后，即使出现多次反复的近距离遮挡，红色和绿色框内的行人仍然可以做到正确跟踪。

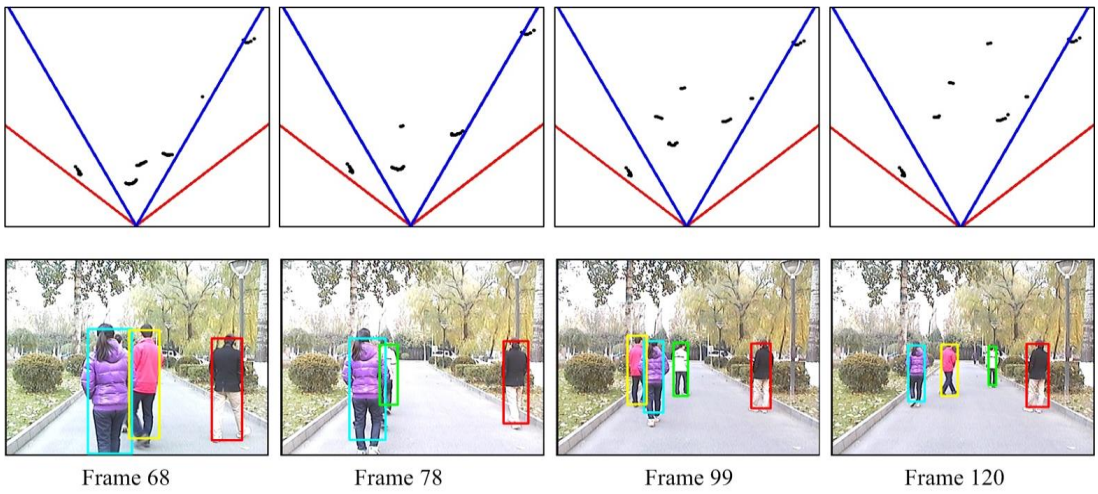


图 4-4 DSA 模型在 SDL-Garden 数据集的跟踪结果

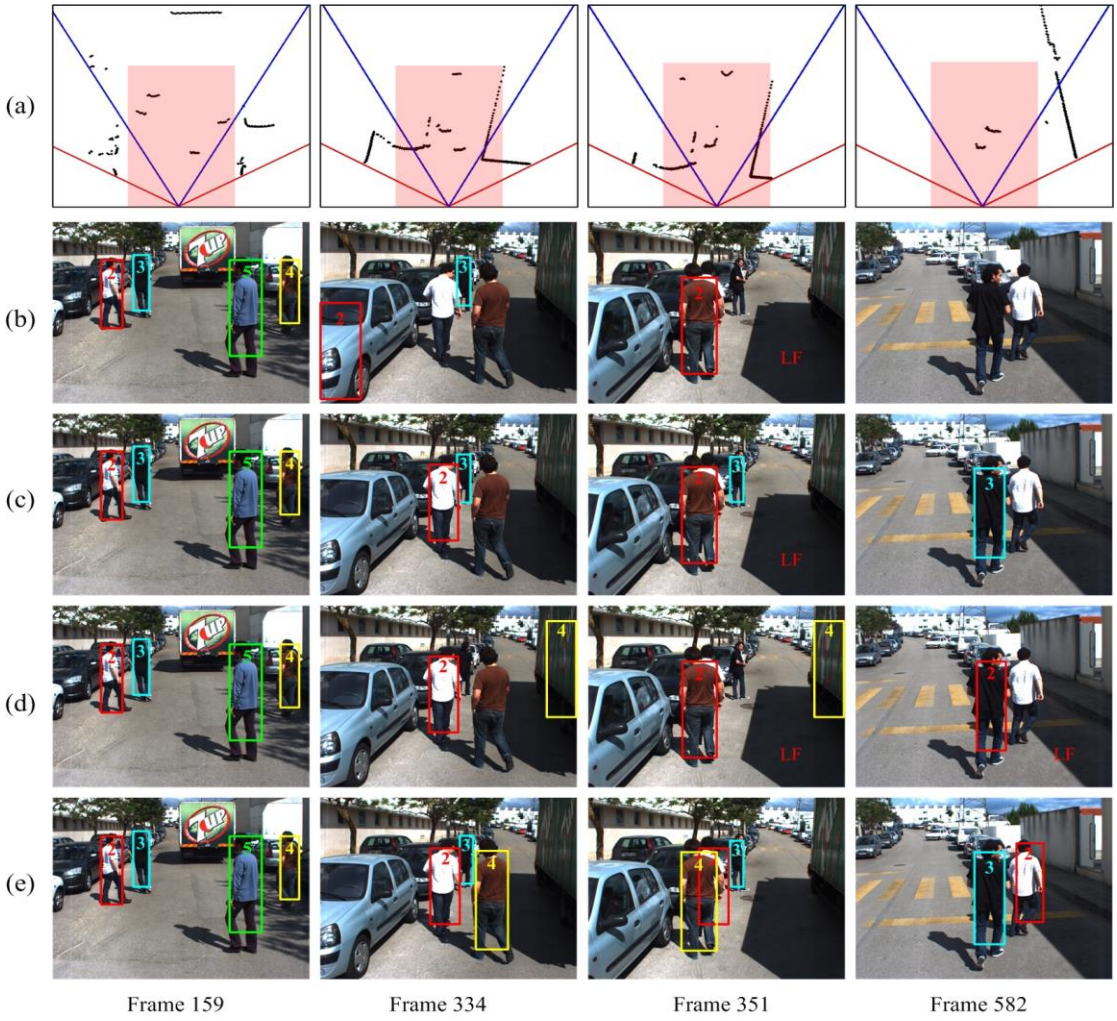


图 4-5 DSA 模型在 Sync 数据集的跟踪结果

4.6 本章小结

本章提出了基于深度结构关联的多目标跟踪模型，利用交换标签算法对该模型进行了多项式复杂度的求解，并实现了模型如何利用目标观测的外形，运动和深度的信息进行在线更新。此外，本章介绍了该模型在复杂遮挡情况下，在深度方向利用深度信息进行深度权重的调节，进行了遮挡问题的求解。最后，利用四个数据集对提出的算法进行了验证，结果显示所提出的多目标跟踪模型是鲁棒高效的。

第五章　基于分层图模型的多目标跟踪方法

5.1　模型概述与创新点

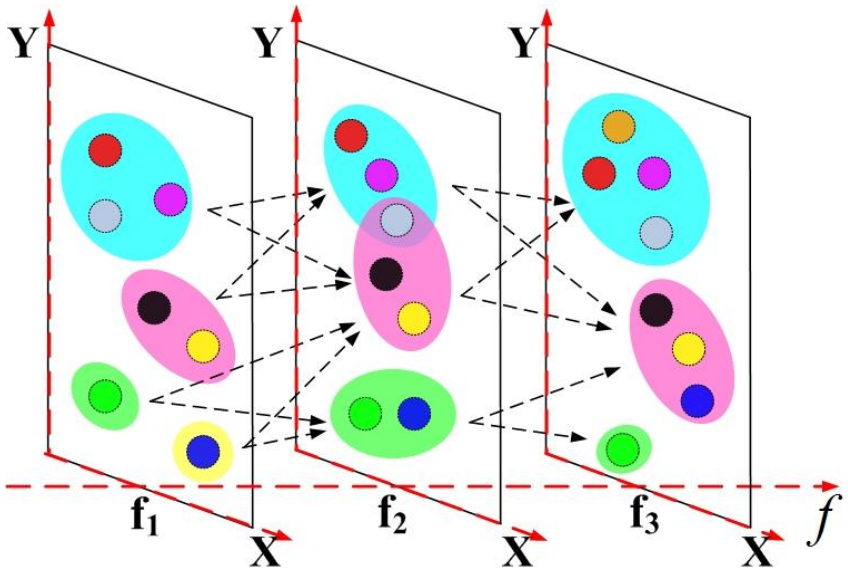


图 5-1 分层图模型示意图

如第二章综述所提到，基于数据关联的多目标跟踪问题与图模型之间具有天然的对应关系。例如，一个目标观测可以被当作图模型中的一个点，而目标观测之间的相似度可以用图中边的权值表示。如果图中两点之间有边进行连接，说明这两个目标观测属于一个目标的轨迹中相邻的两个目标观测。本章用分层图模型求解基于 RGB-D 数据的多目标跟踪问题。将传统的基于离散-连续的轨迹级（tracklet level）关联方式，提升到基于层级（layer level）。利用深度数据构建目标在层内以及层间的图模型，利用目标之间的位置、运动、外形信息构成他们之间的关联相似度。此外，所提出的分层图模型利用自身的层关系，在层内利用加入虚拟点的策略解决目标之间的遮挡问题。本章的创新点如下：

1） 利用 RGB-D 数据在交通场景内构建分层图模型解决多目标跟踪问题；

2） 将传统的数据关联方式从轨迹级升到层级；

3） 根据模型的层关系提出了有效的解决遮挡问题的策略。

表 5-1 本章使用的符号及含义

符号	含义
N_i^k	第 k 帧中第 i 个目标观测，$N_i^k=\{X_i^k=(\mathrm{u}_i^k,\mathrm{v}_i^k,\mathrm{z}_i^k),\varphi_i^k,\theta_i^k\}$
X_i^k	第 k 帧中第 i 个目标观测位置坐标，$X_i^k=(\mathrm{u}_i^k,\mathrm{v}_i^k,\mathrm{z}_i^k)$
φ_i^k	第 k 帧中第 i 个目标观测的外形特征
θ_i^k	第 k 帧中第 i 个目标观测的运动特征
S	目标组成的轨迹片段集合，s_k 表示 S 中的一个轨迹片段
A	目标观测之间的相似度函数
e_i	二进制指示向量，在图中表示边的连接关系
G	$G=(N, L, E)$; 图模型包括 点，层，边三种元素
C	节点之间的匹配代价，Cj 层代价，Cji 边代价
ε_i	目标观测 o_i 相对于所属深度结构中心的偏移值
p_i	目标观测 o_i 的深度因子
X_m	第 m 条链状结构的中心位置

5.2 分层图模型

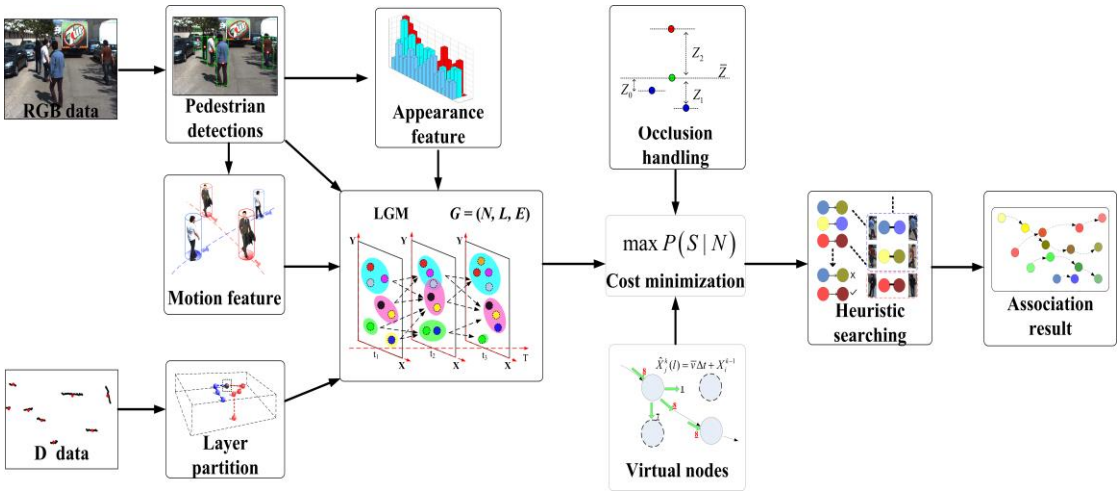
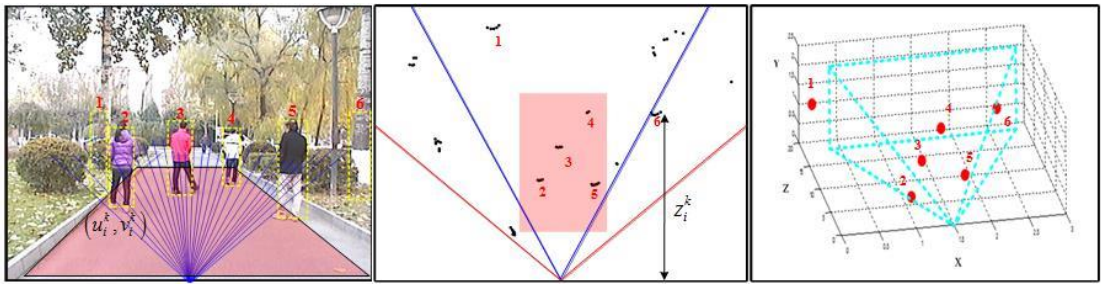


图 5-2 分层图模型求解多目标跟踪问题流程图

分层图模型表示为$G=(N, L, E)$。其中 N，L 和 E 分别代表图模型中的点，层和边。

点（N）：G 由 K 部图组成，并且每一部图代表一帧图像。分层图中的每一个点代表一个目标观测。n_i^k 表示第 k 帧中的第 i 个目标观测，其中 $i\in\mathbb{Z}^+$，$1\le k\le K$。n_i^k 由 3D 特征 $X_i^k=(\mathrm{u}_i^k,\mathrm{v}_i^k,\mathrm{z}_i^k)$，外形特征 φ_i^k 和运动特征 θ_i^k 表示。X_i^k 表示 n_i^k 的三维空间坐标的中心，$(\mathrm{u}_i^k,\mathrm{v}_i^k)$ 表示 n_i^k 在图像上的像素坐标位置如图 5-3 中（a）所示，z_i^k 表示 n_i^k 的实际深度值，如图 5-3 中（b）所示。外形特征 φ_i^k 是由梯度直方图（HOG）和颜色特征组成的特征向量。运动特征 θ_i^k 由 n_i^k 的速度和方向组成的特征向量。



(a) (b) (c)

图 5-3 行人目标在不同观测空间的表示

层（L）：同一帧内的目标观测被划分到 l 个层中，本文用 l_i^k 表示 n_i^k 的层号，并且 $1\le l_i^k\le l$。在图 5-4（b）中，红色的方框代表一个层，每层至少包含一个目标观测。在 l 层中的点可以表示为 $N^{(1)},N^{(2)},\cdots,N^{(m)}$。

边（E）：图中的边定义为 $\mathrm{E}=\left\{\left(n_i^{k-1},n_j^k\right)\mid |l_i^{k-1}-l_j^k|\le 1\right\}$，表示只有在相同或者相邻层内的目标之间具有边的连线。从分层图模型中，在图 5-4（a）中可以看出，并不是处在相邻帧的所有目标之间都有边进行相互连接。边的权值 $\omega\left(n_i^{k-1},n_j^k\right)$ 表示两个目标关联的相似度，其中包含 X_i^k，φ_i^k 和 θ_i^k 的特征信息。

分层图的输出是对应多个目标的多条长轨迹片段，一条长轨迹片段对应图中的一个可行解。一个可行解表示为整个图的一个子图（sub-graph），$G_s=\left(N_s,L_s,E_s\right)$，其中 $N_s=\left\{n_a^1,n_b^2,n_c^3,\cdots\right\}$ 表示第一帧中的第 a 点，第二帧中的第 b 点，第三帧中的第 c 点等其他点构成了该子图。他们的层标号则对应为 $L_s=\left\{l_a^1,l_b^2,l_c^3,\cdots\right\}$，根据上面分层图模型对点和层的定义，边连接关系表示为 $\mathrm{E}_s=\left\{\mathrm{E}\left(n_a,n_b\right)\mid n_a,n_b\in N_s\right\}$。整个分层图包含多条长轨迹片段，则对应着多个子图，且子图之间没有重叠覆盖。

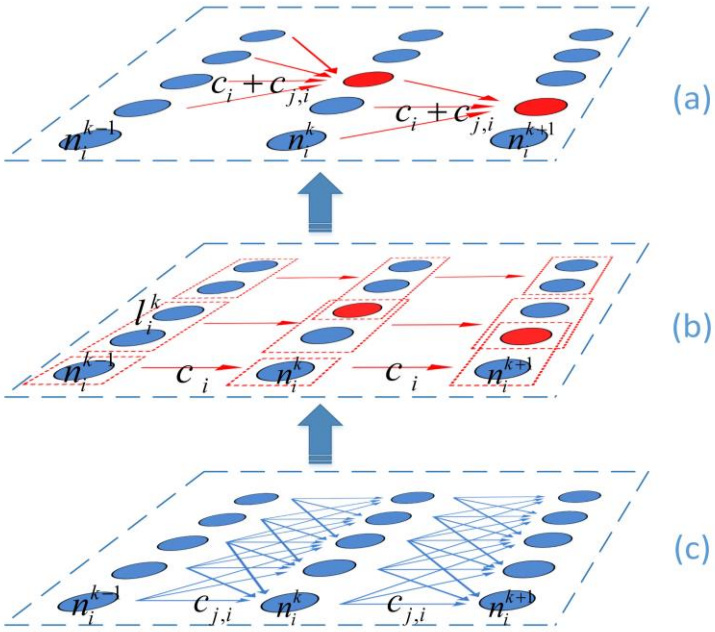


图 5-4 分层图示意图

5.2.1 最大后验概率

跟据分层图中点的定义，目标的观测向量为：$N_i^k=\{X_i^k=(\mathrm{u}_i^k,\mathrm{v}_i^k,\mathrm{z}_i^k),\varphi_i^k,\theta_i^k\}$。$N$ 为目标观测的集合。对于同一个目标的跟踪轨迹估计可以使用一串的轨迹片段集合来表示，例如 $S=\{s_k\}$，其中，s_k 表示轨迹片段。对于给定的观测向量 N，数据关联的目标就是最大化后验概率 S，即：

$$S^*=\arg\max_S P(S\mid N)=\arg\max_S P(N\mid S)P(S). \tag{5-1}$$

假设场景中行人目标的运动是相互独立的，则可以将式（5-1）分解成

$$S^*=\arg\max_S \prod_i P(T_i\mid S)\prod_{s_k\in s} P(S_k). \tag{5-2}$$

然后进行 $S=-\log S^*$ 操作，则式（5-2）可以重新写成

$$S=\arg\min_S \sum_i\left(-\log P_{sim}\left(n_i\mid S\right)\right)-\sum_{s^k\in S}\log P\left(s^k\right). \tag{5-3}$$

式（5-3）中的条件概率 $P(n_i\mid S)$ 用一个二项分布来描述：

$$P(n_i\mid S)=\begin{cases}\beta_i & \exists s_k\in S,T_i\in s_k\\ 1-\beta_i & otherwise\end{cases}. \tag{5-4}$$

其中，β_i 与上一章中基于深度结构关联模型中定义的相同。式（5-3）中，概率 $P\left(s^k\right)$ 可用两个目标观测之间的相似度函数来表示，即 $P\left(s^k\right)=\omega\left(n_i^{k-1},n_j^k\right)$，$k=1,2,...,K$。它表示两个相邻帧之间目标观测在 3D 空间特征，外形特征和运动特征上的相似度：

$$\begin{aligned}\omega\left(n_i^{k-1},n_j^k\right)&=P(n_i^{k-1}|n_j^k)\\&=\begin{cases}A_{app}\left(\varphi_i^{k-1},\varphi_j^k\right)A_{ori}\left(\theta_i^{k-1},\theta_j^k\right)A_{mot}\left(X_i^{k-1},X_j^k\right), & if\ \ |l_i^{k-1}-l_j^k|\le 1\\ 0, & otherwise\end{cases}\end{aligned} \tag{5-5}$$

两个目标观测之间的外形相似值 A_{app} 定义为

$$A_{app}(\varphi_i^{k-1}\,|\,\varphi_j^k)=G\left(sim\left(\varphi_i^{k-1},\varphi_j^k\right);0,\delta_c\right), \tag{5-6}$$

其中，$sim\left(\varphi_i^{k-1},\varphi_j^k\right)$ 表示两个外形特征向量的距离，使其满足高斯分布。本文将方向空间分到 0-8 九个区域内，“1-8”表示 8 个不同的运动方向，而“0”表示目标观测在两帧之间保持静止，如图 5-5（a）所示。在交通场景中，目标运动过程在相邻帧具有连续性，所以具有相同或者相邻方向往往具有较高的相似度。基于此点，方向特征之间的相似度被定义为

$$A_{ori}(\theta_i^{k-1}\,|\,\theta_j^k)=G\left(\left|\theta_i^{k-1}-\theta_j^k\right|;0,\delta_o\right), \tag{5-7}$$

其中，$\left|\theta_i^{k-1}-\theta_j^k\right|$ 表示两个方向之间的相对距离。运动特征被定义成

$$A_{mot}(X_i^{k-1}\,|\,X_j^k)=G(X_j^{k-1}+\bar{v};X_i^k)G(X_i^k-\bar{v};X_j^{k-1}), \tag{5-8}$$

其中，$\bar{v}$ 表示最近 K 帧内该目标观测的平均速度值，并且目标沿着 K 帧内的平均速度运动的概率最高，如图 5-5（b）所示。目标观测的位置与预测位置之间的距离差值满足高斯分布。

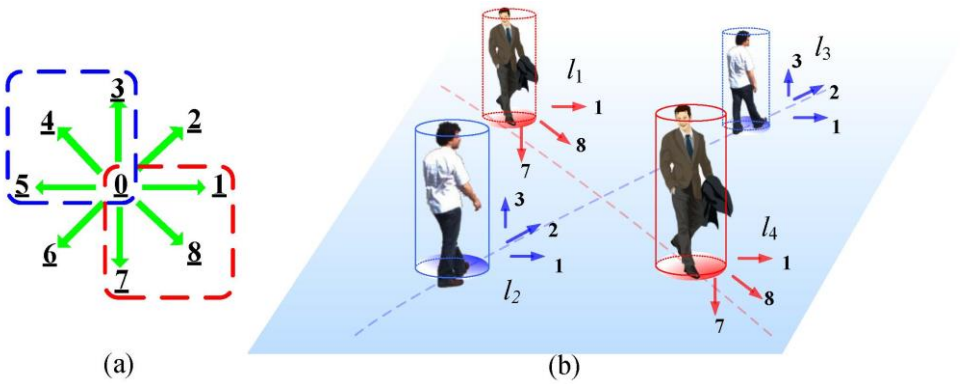


图 5-5 方向特征和运动特征示意图

为了满足一个轨迹片段只属于一个行人的约束，引入指示变量 e 表示轨迹片段属于某一个行人的轨迹点。

$$e_{j,i}=\begin{cases}1 & if\ \ T_j\ \ is\ \ right\ \ after\ \ T_i\\ 0 & otherwise\end{cases},$$
$$e_i=\begin{cases}1 & if\ \ T_i\in S_k\\ 0 & otherwise\end{cases}. \qquad (5\text{-}9)$$

其中，$e_{j,i}$ 表示目标观测 o_j 是否为 o_i 的下一帧观测，e_i 表示 o_i 是否属于目标完整轨迹。对于行人轨迹的划分，需要保证轨迹是非重叠的，即一个目标观测只能属于一条轨迹，则轨迹的指示变量满足 $e_i=\sum_j e_{j,i}\le 1$。此时，将式（5-9）带入式（5-3）中，则目标方程可以重写为：

$$S=\arg\min_S \sum_{i,j} C_{j,i}e_{j,i}+\sum_i C_i e_i$$
$$s.t.\ \ e_i=\sum_j e_{j,i}\le 1. \qquad (5\text{-}10)$$

对照公式（5-3），新引入的参数 $C_{j,i}$ 和 C_i 为边的权值，可以表示为：

$$C_{j,i}=-\log P\left(S^k\right), \qquad (5\text{-}11)$$

$$C_i=-\log P(n_i\mid S)=\log\frac{1-\beta_i}{\beta_i}. \qquad (5\text{-}12)$$

其中，$C_{j,i}$ 与 $e_{j,i}$ 对应，表示目标观测 n_i^{k-1} 和 n_j^k 之间的相似度代价。C_i 与 e_i 对应，表示层的匹配代价。

5.2.2 分层约束

为了避免错误的轨迹关联，减小轨迹的匹配搜索的时间复杂度，利用 RGB-D 数据的深度信息对所要跟踪的目标进行分层。使用目标的深度 z_i^k 值，以及目标的三维空间信息 X_i 值作为分层依据。定义一个均方偏移值，$\varepsilon_i^{(m)}$，表示某个目标观测的 3D 位置相对于所在层中心的 3D 位置的偏移值，

$$\varepsilon_i^{(m)}=\left(\frac{1}{\left|N^{(m)}\right|}\sum_i\left\|X_i^{(m)}-\bar{X}^{(m)}\right\|^2\right)^{1/2}, \qquad (5\text{-}13)$$

其中，$m=1,2,\ldots,M$ 是层的标号，$\bar{X}^{(m)}$ 表示第 m 层中心的三维坐标值，$\left|N^{(m)}\right|$ 表示第 m 层所包含的观测目标的个数。依据以上定义，当目标观测处在同一个区域被划分进同一层，并且处在该层的目标观测相互之间比较容易遮挡。在进行

目标匹配时，处在同一层的目标观测在相邻帧优先进行匹配。通过以上的层划分方式可以看出，同一个目标观测可以被划分到不同的层中。这是由于该目标观测处在两层之间的位置，模型并不强行将其划分到某一层中，而是将该观测划分进两个不同的层。在搜索匹配时，对于该目标将在两层内进行搜索。

从分层的角度，本方法将目标观测集合 N 划分为 $N^{(1)},N^{(2)},\ldots,N^{(m)}$，层与层之间的关系可以总结为 $N^{(1)}\bigcup N^{(2)}\bigcup\ldots\bigcup N^{(m)}=N$ 和 $N^{(1)}\bigcap N^{(2)}\bigcap\ldots\bigcap N^{(m)}\geq\varnothing$。根据分层理论，目标方程（5-10）可以重新定义为

$$S=\arg\min\sum_{m=1}^{M}\sum_{i,j}c_{j,i}^{(m)}e_{j,i}^{(m)}+\sum_{m,l}^{M}\sum_{i,j}c_{j,i}^{(m,l)}e_{j,i}^{(m,l)}+\sum_{i}c_{i}e_{i}\,. \qquad (5\text{-}14)$$

根据此公式，图中的边可以分为层内和层间两种形式，分别对应这该目标方程中第一和第二项。$e_{j,i}^{(m)}$ 表示第 m 层内的边，$e_{j,i}^{(m,l)}$ 表示第 m 层和第 l 层之间的边。并且 m 和 l 满足 $|m-l|\leq 1$。$c_{j,i}^{(m)}$ 和 $c_{j,i}^{(m,l)}$ 分别表示两种边对应的权值。第三项表示层之间的连接。这种分层结构降低了目标观测搜索匹配的空间。因为 $\sum_{m}\sum_{i,j}e_{j,i}^{(m)}+\sum_{m,l}\sum_{i,j}e_{j,i}^{(m,l)}\ll\sum_{i,j}e_{j,i}$。所以利用分层图求解图模型极大缩小了求解搜索空间，提高了跟踪的效率。

虚拟点：由于遮挡或者漏检测的原因，可以发现在某些层中目标观测的数量少于实际的目标数。为了解决这个问题，在这些层中加入虚拟点。当层中待匹配点数目少于目标观测数目时，给该层加入虚拟点。虚拟点的放置位置由目标观测前后帧的该目标观测的运动模型估计完成 $\hat{X}_j^k(l)=\bar{v}\Delta t+X_i^{k-1}$，其中 $\hat{X}_j^k(l)$ 表示该虚拟点在 l 层的空间位置，$\bar{v}$ 为该目标观测的在 K 帧内的平均速度。该虚拟点的运动方向与前帧的该目标观测的运动方向相同。对正常目标观测点和虚拟点之间设置一个较小的相似值，使得虚拟点在层内匹配时只能在正常目标观测搜索匹配完成后进行。

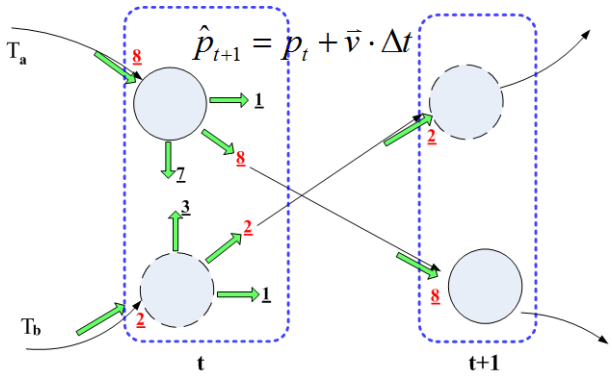


图 5-6 目标位置更新示意图

由于虚拟点是被遮挡的目标观测的假想观测，优先匹配具有较小深度值的正常目标观测点。即优先求解目标方程（5-14），在每一次迭代求解过程中，优先匹配场景中存在的目标观测，当层内的真实目标观测被匹配完成后，加入虚拟点进行搜索匹配，直到算法收敛，完成G_s输出。

5.3 模型求解

求解公式（5-14）等同于在分层图中寻找匹配代价最小的路径。图中的一个点代表一个目标观测，目标观测之间的连线的权值代表两个目标观测之间的相似性。其中图 5-4 中的红色方框代表不同的深度层，红色的圆点代表在两个不同深度层的同一个目标观测。在匹配的过程中模型对蓝色的圆点在相同层内进行匹配。而红色的圆点处在两个不同的深度层，对于这样的目标观测，在两个层内进行匹配搜索。

在求解分层图模型中，用启发式算法在分层后的空间内寻找相似度最大的匹配搜索。这里采用标签交换来求解：对于每层中选中的点，尝试在相同或者相邻层内不断交换标签。如果交换标签后的使得公式（5-14）中的整体代价更低，保留新的标签；如果交换标签后使得总体相似度降低，放弃此次标签交换。具体的交换标签法见表 5-2。

表 5-2 交换标签算法

算法 1 交换标签算法

Input: 分层图 G; 代价 $\{C_i\},\{C_{j,i}\}$;

Output: G中边连接 E_s;

Initialization: 使用贪心算法法寻找公式（5-14）中具有最小代价的边连接 E_s;

1: **for** $i < N$ **do**

2:设置最小代价和 $C_{\min} = +\infty$;

3: **for** $j = i, \cdots, k$ **do**

4: -根据公式（5-13），交换层中边连接 n_i^{k-1} 和 n_i^k，并计算新的代价 C_{temp};

5: - **If** $C_{temp} < C_{\min}$，$C_{over} = C_{temp}$;

6: **end for**

7: $C_{over} < C_{\min}$，$C_{\min} = C_{over}$，在 E_s 中更新这组边;

8: **end for**

5.3.1 时间复杂度

在每一次迭代过程中，模型都需要寻找目标方程（5-14）的最小代价。在图中搜索最短论经的解法中，Dijsktra 算法可以做到在每次迭代中的时间复杂度为$O(N\log N)$，则在整个循环中的复杂度为$O(MN\log N)$，其中 M 为目标的个数。但是，在本章提出的分层图模型中，边的代价有可能为负值，所以 Dijsktra 算法就会失效。而本章提出的基于启发式的交换标签算法可以做到时间复杂度为$O(MN^2)$，是多项式时间复杂度，这在实时跟踪过程中是可以接受的。另外，多目标跟踪的时间复杂度还与目标的数目有很大的关系。本模型使用分层求解，所以目标会被划分进不同层内，在寻找最短路径时，不需要对不同层内的目标进行匹配，这样会有效减少搜索路径。因此在实验中可以发现，模型的时间复杂度和目标的数目成线性相关性。

5.4 实验验证

本章在实验中采用了多种类型的方法与本章所提出的分层图模型进行实验对比：1）基于 RGB 的多目标跟踪方法：Berclaz[75]和 Zhang[74]使用基于网络流的多目标跟踪模型；Andriyenko[104]使用基于连续的轨迹分析法进行多目标跟踪建模；而 Milan[105]使用基于能量的轨迹关联建模；Yang[71]使用的基于在线学习的 CRF 模型。2）基于深度的多目标跟踪方法：“NN”使用第二章中基于全局最紧邻的方法；“K-partite”使用基于 K 部图求解多目标数据关联，在此方法中未使用分层图进行深度层的划分，直接将多个目标视为在全图需要关联的节点，然后在整个图空间内进行目标匹配搜索。3）使用本章提出的完整分层图模型（LGM）进行建模求解。前两种类型所使用的多目标跟踪方法在第二章多目标跟踪方法综述中有详细分类描述。

本章使用的多目标跟踪性能评测方法在第三章中已经详细介绍，用于评测的每一个指标的具体含义请参考表 3-3。本章采用五个公开的 RGB-D 数据集进行实验验证。他们分别是：Sync 数据集、SDL-Garden 数据集、SDL-Crossing 数据集、SDL-Campus 和 LIPD 数据集，数据集的属性介绍请参考 3.2 节以及表格 3-1。此外，为了达到公平实验的原则，在实验中对于所有的方法采用相同的目标检测结果输入，用第二章所介绍的基于 RGB-D 数据的 DPM 检测器[101]对所有数据集进行目标检测，以此作为本章实验中多目标跟踪的输入。详细的对比实验结果如表 5-3 所示。表中所有用粗体标明的结果为该组实验中最好的结果。

表 5-3 对比实验结果

数据集	方法	Recall	Prec.	GT	MT	PL	ML	Frag.	IDS
Sync	Berclaz[75]	69.6	74.8	66	64.5	22.7	12.8	45	23
	Andriyenko[104]	73.4	78.3	66	69.7	19.7	10.6	39	18
	Milan[105]	75.6	80.2	66	71.2	18.2	10.6	37	16
	NN	54.5	64.3	66	45.5	30.3	24.2	52	31
	K-partite	73.1	78.6	66	68.2	15.2	16.6	39	17
	LGM	**85.0**	**89.7**	66	**80.3**	**10.6**	**9.1**	**21**	**7**
SDL-Garden	Zhang[74]	67.8	72.5	10	60.0	20.0	20.0	7	5
	Andriyenko[104]	76.6	81.0	10	70.0	20.0	10.0	5	4
	NN	56.4	64.9	10	40.0	30.0	30.0	8	6
	K-partite	68.7	73.5	10	60.0	20.0	20.0	5	4
	LGM	**94.5**	**98.3**	10	**90.0**	**10.0**	**0.0**	**2**	**0**
SDL-Crossing	Berclaz[75]	68.9	74.5	92	60.9	17.4	21.7	58	31
	Andriyenko[104]	70.4	76.4	92	63.0	20.7	16.0	51	29
	Yang[71]	72.3	77.8	92	64.1	21.7	14.2	47	26
	NN	59.4	65.6	92	43.5	23.9	32.6	69	38
	K-Partite	67.0	73.5	92	59.8	22.8	17.4	49	25
	LGM	**82.4**	**87.3**	92	**76.1**	**15.2**	**8.7**	**28**	**14**
SDL-Campus	Zhang[74]	76.4	79.8	74	71.6	18.9	9.5	30	16
	Milan[105]	80.0	84.5	74	75.7	16.2	8.1	26	14
	NN	67.7	74.6	74	60.8	21.6	17.6	37	19
	K-partite	78.3	82.9	74	73.0	17.6	9.4	26	15
	LGM	**85.6**	**89.3**	74	**81.1**	**12.2**	**6.7**	**14**	**8**
LIPD	Berclaz[75]	72.8	76.4	77	68.9	18.1	13.0	24	19
	Yang[71]	77.6	80.2	77	72.7	13.0	14.3	19	14
	NN	64.4	70.1	77	64.9	15.6	19.5	27	19
	K-partite	71.9	75.4	77	70.1	18.2	11.7	25	16
	LGM	**84.9**	**88.3**	77	**77.9**	**11.7**	**10.4**	**16**	**10**

5.4.1 实验分析

从表 5-3 中可以看出，利用本文提出的方法在五个数据集上所有指标中都做到了最好。在速度方面，虽然没有达到前面几种方法相同的速度，但是仍然

能够满足交通场景中实时跟踪的要求。图 5-7 和 5-8 是实验方法在不同数据集上的跟踪结果。

SDL-Garden 数据集：整个序列中的四个行人出现了反复的相互遮挡。如图 5-7 所示，如果只利用单纯的视觉信息进行跟踪，穿红色衣服的行人与紫色衣服的行人由于在很长一段时间内是处于遮挡和被遮挡关系，容易被混淆，在本文提出的跟踪模型中，加入深度动态权重因子后，紫色衣服行人的转移代价为负值，在进行模型求解中优先会被选择。而处在远处的红色行人和白色行人则具有比较大的转移代价，处在末位被匹配选择的位置。所以本文提出的 LGM 模型在处理遮挡问题时,在具有深度权值后能够更加有效的解决复杂遮挡问题，使得跟踪的轨迹的 Frag.和 IDS 错误更少。因此，可以看到表 5-3 中的 MT 比其他对比试验的结果高，ML 则比对比实验低。

SDL-Crossing 数据集：该数据集记录了长时间内，位于丁字路口出现在车辆前方的行人运动状况，实验对每个行人进行了标号。本文提出的跟踪模型，面对十个左右数目的且具有不同行进方向行人，进行了正确的标号保持。尤其处在车辆正前方的危险区域内，本章的检测跟踪方法没有失误，只是在危险区域外围的潜在危险区域内出现了行人的标号前后不一致的标号错误。

Sync 数据集：此数据集具有目标数目不确定的行人，而且他们一直处在背景比较复杂的道路上进行交错行走，可以看出只利用 2D 运动信息和外形信息的 RGB 跟踪方法，在面对复杂情况时会出现多次的跟踪丢失和 IDS 错误。而使用分层原理并将目标划分进不同的深度层后的 LGM 模型的结果，如图 5-7 第三行，这两种错误的情况明显减少。并且 LGM 模型对于不同深度的行人基于区分性的深度动态权重因子后，即使出现多次反复的近距离遮挡，算法仍然可以做到不丢失目标。

LIPD 数据集：有别于第四章所使用的实验数据集，LIPD 数据集是本章中新加入的 RGB-D 数据集。从图 5-8 可以看出，该数据集也具有更高的挑战：1）该数据集录制于黄昏时分，低对比度使得基于传统方式的行人检测技术几乎于失效，如图 5-8 第二行的结果。2）更多的行人，低对比图像使得具有相似外形的行人从外形特征上更加难以区分，我们在模型中对外形之间的相似度设置了比较低的权重。在 LGM 模型中，通过分层的方式将目标划分到不同层后，使得目标匹配不会在深度差异比较大的目标观测间进行，这样既减少了匹配错误的可能性，又极大地节省了用来匹配求解的时间。而对于基于 RGB 的跟踪方法，在面对图 5-8 所示的环境时，只能通过检测窗口的位置和运动方向进行匹配，所以在跟踪过程中出现了大量的轨迹碎片，造成跟踪错误。

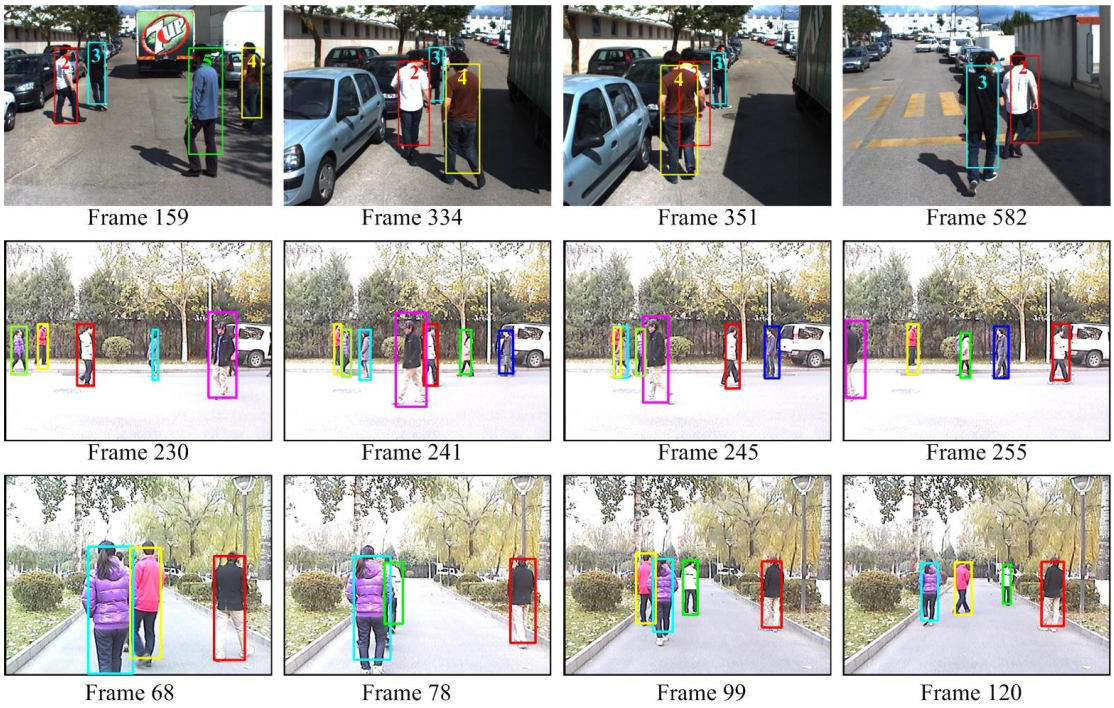


图 5-7 LGM 模型在 Sync, SDL-Crossing 和 SDL-Garden 数据集的跟踪结果

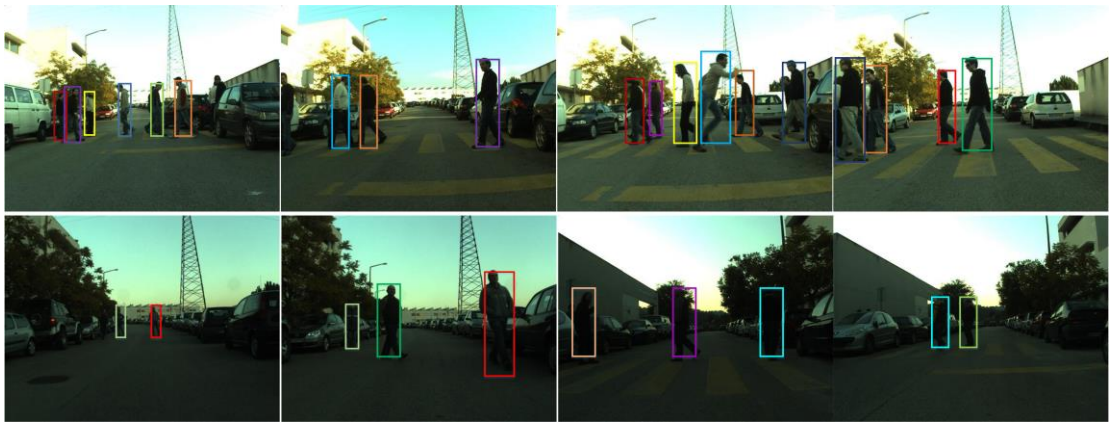


图 5-8 LGM 模型在 LIPD 数据集的实验结果

5.5 本章小结

本章提出了基于分层图的多目标跟踪模型，将目标跟踪与图模型理论进行了有效结合。模型将场景中处在不同深度的行人进行了基于深度值的划分。使得目标之间首先具有层约束。因此，也将传统的轨迹级的关联方提升到层级的关联。进而利用最大后验概率公式将目标在图中的关联匹配划分为层代价和节点代价之和。最后，利用交换标签算法对该模型进行了多项式复杂度的求解。

在求解过程中，模型会对被遮挡的目标加入虚拟点进行替代，表示在遮挡过程中目标的位置信息，有效的解决多目标跟踪过程中的遮挡问题。本章利用五个公用数据集对提出的算法进行了充分的实验验证，验证结果显示本章提出的分层图模型是实时高效的。

第六章　基于拓扑能量最小化的多目标跟踪方法

6.1 模型概述和创新点

社会行为学研究发现，自然场景中的人群有高达 70%的行人以组的形式运动[85,86]。例如：以家人，朋友为单位运动的行人之间位置更近，运动模式相同，这更利于他们之间进行交流。这些以组为单位的人群有相似的运动模式，且他们的行走方式与自身所处的环境、周围行人有很大关系。

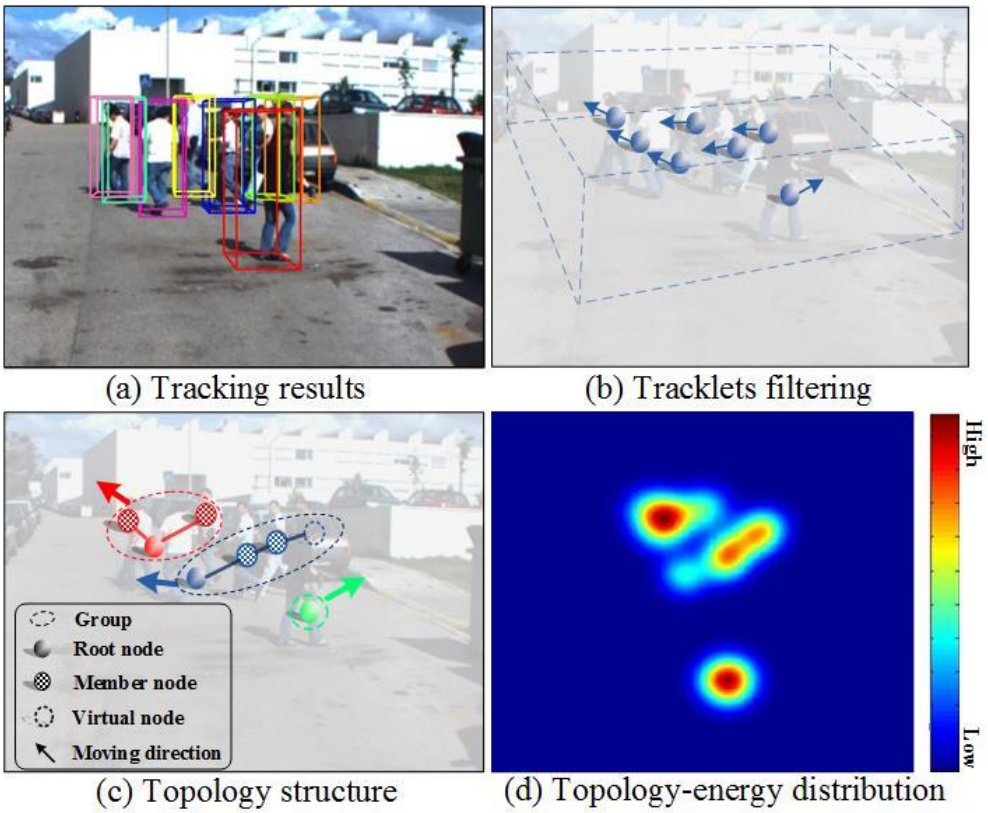


图 6-1 拓扑能量模型示意图

为了合理描述行人组的运动变化和空间拓扑结构，我们提出在图模型中加入基于拓扑的约束。这里，“拓扑”用来展现组内成员和组间行人的空间关系。本章提出了基于拓扑能量的多目标跟踪模型。模型基于人群行为研究，将自然场景中的行人目标划分为以组为单位的运动集合，每个组内的行人具有近似的运动模式。模型利用组内组外的行人运动相似度，进行能量形式的建模，使得

组内的行人相似度尽可能的高，组间的行人相似度尽可能的低，并以“拓扑能量最小化”的方式求解模型。在拓扑的变化过程中，模型通过加入虚拟点的方式，将被遮挡的目标通过组内位置估计的方式进行有效定位，减少了多目标跟踪过程中目标丢失问题。本方法在 SDL-Campus 进行了模型训练后，在多种 RGB-D 数据集进行了测试。本章提出的基于拓扑能量的模型和现有的分组模型的最大区别在于：

1）使用拓扑级的约束在拓扑内和拓扑间描述行人组的运动；

2）使用拓扑能量变化最小的方式求解行人组跟踪问题。

6.2 拓扑能量模型

表 6-1 本章使用的符号及含义

符号	含义
X_{ij}	二进制指示向量，表示目标观测之间的连接关系
T_{ik}	二进制指示向量，表示单个轨迹与组之间所属关系
A_{ij}	轨迹片段l_i与l_j之间的特征相似度
l_i	第 i 个轨迹片段，$l_i^t=\left[x_i^t,y_i^t,z_i^t,1\right]^T$
G_k	第 k 个行人组
Ψ_{ik}	轨迹片段l_i和组G_k的拓扑相似度
V_{ik}	轨迹片段l_i和组G_k的速度相似度
ϕ_{ik}	轨迹片段l_i和组G_k的运动方向相似度
φ_n	轨迹片段的运动方向
Γ	行人之间的拓扑关系，$\Gamma=\{G,R,T\}$
π_k	组 k 内的马尔可夫链状态参数向量，$\pi_k=\{C_k,Q_k,\mu_k,\delta_k\}$
τ_k	组内运动稳定度偏差
E_{topo}	拓扑能量值
ΔE_{topo}	拓扑能量的变化量值

如第四章的深度结构关联模型求解多目标跟踪问题一样，本章提出的模型也从基于整数规划的 MDA 问题入手。不过本章提出的拓扑能量模型不是在 K 帧之间进行轨迹的数据关联，而是使用旨在达到全局最优的轨迹求解，采用基

于先检测再跟踪的策略，在对整个视频序列中的目标进行检测后，在利用检测器的输出结果，完成场景内的多目标跟踪。本章使用 $\mathcal{L}=\{l_1,l_2,\cdots,l_n\}$ 表示视频序列中的所有轨迹片段。每一个轨迹片段 l_i 是同一目标的一小段短轨迹，这些轨迹片段由于检测器漏检和误检不能形成长的完整轨迹。所以基于全局的多目标关联的任务是将这些同一目标的短轨迹连接成完整的长轨迹。基于全局的轨迹关联问题可以表示为：

$$\begin{aligned}&\arg\max_{X}\underbrace{\sum_{i,j}A_{ij}X_{ij}}_{tracklet},\\&s.t.\ X_{ij}=\begin{cases}1, & if\ l_j\ associated\ after\ l_i,\\0, & otherwise,\end{cases}\\&\quad\sum\nolimits_i X_{ij}\le 1,\ \ \sum\nolimits_j X_{ij}\le 1.\end{aligned}\tag{6-1}$$

其中，A_{ij} 表示轨迹片段 l_i 和 l_j 之间相似度。X_{ij} 是二进制指示参数，表示轨迹片段 l_i 和 l_j 是否连接。如果连接则 $X_{ij}=1$，表示 l_i 和 l_j 是同一个目标的轨迹片段。为了满足全局唯一性要求，每一个轨迹片段只能属于一个目标，所以加入约束条件：$\sum_i X_{ij}\le 1$ 和 $\sum_j X_{ij}\le 1$。

定义拓扑关系为 $\Gamma=\{G,R,T\}$，其中 $G=\{G_k\}$ 表示拓扑中的所有行人组，$R=\{r_k\}$ 表示每一个组内的中心节点，$T=\{T_{ik}\}$ 表示轨迹片段相对于组的所属关系。如果轨迹 l_i 属于组 G_k，则 $T_{ik}=1$，反之则为零。将这种拓扑关系加入到公式（6-1）中，得到

$$\begin{aligned}&\arg\max_{X,T}\underbrace{\sum_{i,j}A_{ij}X_{ij}}_{tracklet}+\underbrace{\alpha\sum_{i,k}\Psi_{ik}T_{ik}}_{topology},\\&s.t.\ T_{ik}=\begin{cases}1, & if\ l_i\ belongs\ to\ group\ G_k,\\0, & otherwise,\end{cases}\\&\quad X_{ij}=\begin{cases}1, & if\ l_j\ associated\ after\ l_i,\\0, & otherwise,\end{cases}\\&\quad\sum\nolimits_i X_{ij}\le 1,\ \ \sum\nolimits_j X_{ij}\le 1,\\&\quad\sum\nolimits_k T_{ik}=1,\ \ \sum\nolimits_i T_{ik}\ge 1.\end{aligned}\tag{6-2}$$

其中，Ψ_{ik} 表示轨迹片段 l_i 和组 G_k 的拓扑相似度。这里加入了关于组的约束 $\sum_k T_{ik}=1$ 表示每个轨迹片段只能属于一个组，$\sum_i T_{ik}\ge 1$ 表示每个组至少含有一个轨迹片段。α 为归一化参数，用来平衡轨迹项和拓扑项的相似度。对于行人之间的行为描述[85,86]，在自然人群场景中控制单个行人运动的社会学因素包括以下三点：

1）去往相同目的地的行人通常具有相同的运动轨迹；

2）每个行人根据周围场景选择最合适的运动速度；

3）行走过程中，行人为了便于组内成员口头交流，避免与其他行人发生碰撞而不断调整自己的位置。

在拓扑模型中我们将前两个因素通过用运动方向 ϕ_{ik} 和运动速度 V_{ik} 对组内成员进行运动建模，第三个因素通过拓扑距离 D_{ik} 对组间进行空间建模。则公式（6-2）中的第二项拓扑项可以展开为

$$\underbrace{\sum_{i,k}\Psi_{ik}T_{ik}}_{topology}=\underbrace{\sum_{i,k}V_{ik}\phi_{ik}T_{ik}}_{intra-T}-\underbrace{\sum_{i,k}D_{ik}T_{ik}}_{inter-T}. \tag{6-3}$$

该式共包含两项：组内相似度和和组间相似度。模型构建中，我们希望其具有较高的组内相似度和较低的组间相似度。

组内相似度：模型从运动速度和运动方向来来衡量轨迹片段 l_i 和组 G_k 的拓扑相似度。运动速度的相似性表示为

$$V_{ik}=e^{\frac{-\|v_i-\bar{v}_k\|^2}{2\sigma_{\bar{v}_k}^2}}, \tag{6-4}$$

其中，v_i 表示 l_i 的运动速度，$\bar{v}_k$ 表示 G_k 内成员的平均运动速度。因此，V_{ik} 衡量 l_i 和组 G_k 的运动相似度。在运动方向相似度中，本方法采用 Potts 模型[106]来衡量运动方向之间的相似度：

$$\begin{aligned}\phi_{ik} &= \cos(\varphi_i-\bar{\varphi}_k),\\ \varphi_n &= \frac{2\pi n}{q},\end{aligned} \tag{6-5}$$

其中，ϕ_{ik} 衡量 l_i 和组 G_k 的运动相似度。这里，模型将轨迹片段 l_i 的运动方向 φ_n 划分到 8 个以 45 度为单位的运动区间。n 代表 l_i 的运动区间编号。和第五章中方向定义相同，仍然用“0”表示静止的轨迹。此时 q=9。

组间相似度：距离特征被用来描述运动行人的组间相似度，所以模型期望行人的组间距离越大越好。公式（6-3）距离相似度可以表示为

$$D_{ik}=\frac{D(l_i,G_k)}{\overline{T}^{(i,k)}}, \tag{6-6}$$

其中，$D(l_i,G_k)$表示轨迹片段 l_i 到组 G_k 中心的距离。只有当轨迹片段 l_i 满足 $D(l_i,G_k)<$ d时，才被考虑划分进组 G_k。参数 d 在第 6.4 节详细描述，由训练集中得到。$\overline{T}^{(i,k)}$ 表示当 l_i 划分进组 G_k 时，所有组的平均组半径。

6.2.1 拓扑稳定性

在行人组的运动过程中，模型要求组内成员空间运动尽可能的稳定一致。假设组 G_k 内有 m 条马尔可夫链，并且每一条链时域上的模型满足 $l_i^t = C_k l_i^{t-1} + \xi^t$，其中，$t$ 时刻组内的轨迹片段 l_i^t 由马尔可夫链中状态转移矩阵 $C_k \in \mathbb{R}^{4\times4}$ 更新。$\xi^t \sim \mathbb{N}(0,Q)$ 表示满足高斯分布的噪声。$l_i^t = \left[x_i^t, y_i^t, z_i^t, 1\right]^T$ 表示目标观测的 3D 坐标，其中 z_i^t 表示深度信息。初始的位置观测满足高斯分布 $\mathbb{N}(\mu,\delta)$。此时参数 $\pi_k = \{C_k, Q_k, \mu_k, \delta_k\}$ 代表了整个马尔可夫链的参数集合，其中 C_k 表示 G_k 内的所有成员的运动状态，$\{\mu_k, \delta_k\}$ 保证组内成员在空间上的尽可能靠近。模型中，定义组内运动稳定性偏差为

$$\tau_k = \frac{1}{|G_k|}\sum_{l_i \in G_k} \left\| l_i^t - C_k l_i^{t-1} \right\|^2, \tag{6-7}$$

其中，$|G_k|$ 表示组 G_k 内的轨迹片段数量，也就是组内的成员数。τ_k 值越小，表示组内的所有的成员运动越具有一致性。

6.2.2 拓扑能量变化

在拓扑模型的构建过程中，我们放弃了很多轨迹级的约束，希望使得模型具有更多的应用场合。而能量方程的引入可以使目标方程（6-2）的求解更加简单有效。根据方程（6-3），我们将这种拓扑能量定义为

$$E_{topo} = \sum_{i,k,\ l_i \in G_k} \sum \left\| V_{ik}\phi_{ik} - D_{ik} \right\|^2, \tag{6-8}$$

其中，E_{topo} 表示整个拓扑结构的能量，它包含了组内和组间的综合能量。根据前面提到的自然场景中性人的社会属性，行人的运动具有连续性，则拓扑能量的变化也具有连续性。这使得大多数行人在行进过程中跟随自己所在组进行运动，而只有少数行人会在组之间进行跳变。因此，在目标方程的求解过程中，我们并不计算拓扑的总能量，而是计算在连续帧拓扑能量的变化量

$$\Delta E_{topo} = \left| E_{topo}^t - E_{topo}^{t-1} \right|. \tag{6-9}$$

此时的 ΔE_{topo} 为连续帧的能量变化值，在下面的模型求解中，我们利用求解 ΔE_{topo} 的最小值，寻找的轨迹片段的分组关系。

6.3 拓扑能量模型求解

本节将介绍如何利用拓扑能量最小化求解多目标跟踪问题。传统的多目标跟踪问题使用解析求解的方法完成对目标方程的求解。但是本章提出模型所含的目标方程式（6-2），很明显是非凸函数。模型可以采用 Hungarian 算法[67]和

对偶优化求解[89]。但是这样会使得求解的时间复杂度变得很高，所以本章我们采用启发式的“拓扑能量变化”的方式对目标方程进行优化求解。图 6-2 展示了拓扑能量变化的过程，（a）为正确的跟踪示意图，分组和组内的成员均做到正确匹配，此时第二行中的组能量变化很平稳。（c）为错误的跟踪示意图，当分组和目标跟踪错误时，第二行中整个拓扑能量的变化很大。（b）和（d）对拓扑能量的变化和组的大小进行了量化描述，可以发现红色曲线代表的正确跟踪的能量和大小变化很平稳，而蓝色代表的错误跟踪曲线的抖动比较大。根据这种“正确的跟踪拓扑能量变化小”的观点，模型采用三步对方程（6-2）求解。

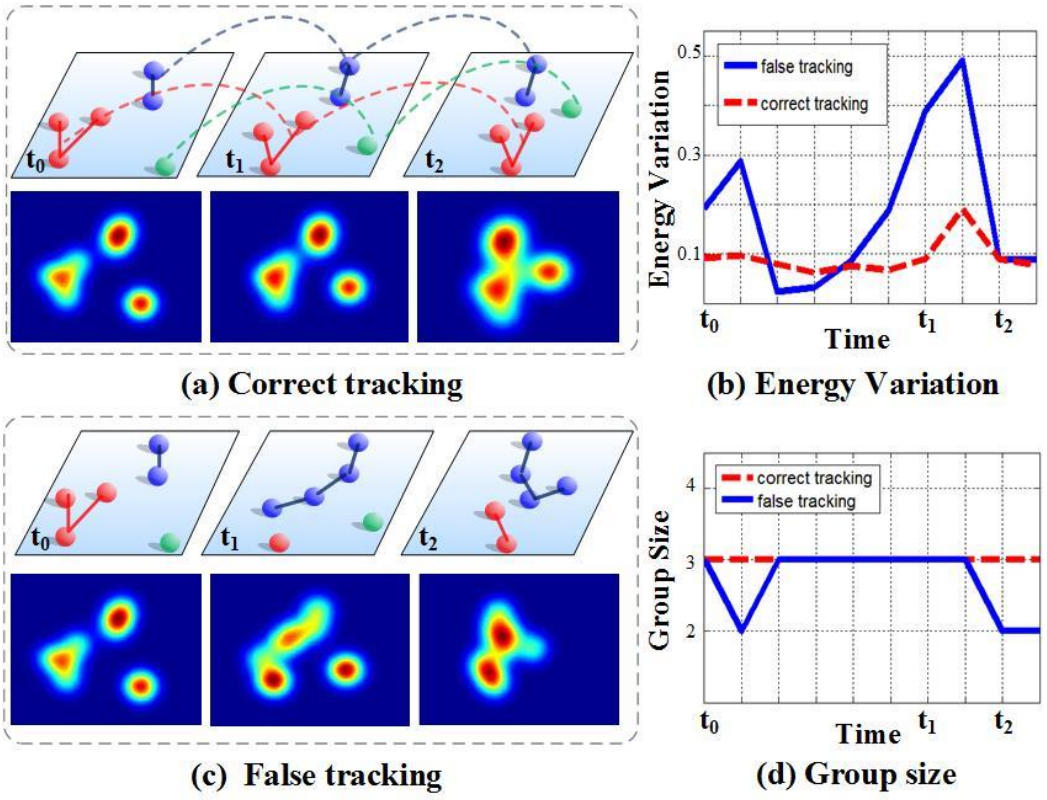


图 6-2 拓扑能量变化示意图

步骤一：分组初始化。在“先检测再跟踪”的框架下，每一个轨迹片段都是检测结果串联而成。但是，用于进行组分析的轨迹片段必须保证准确有效，所以在本方法中模型使用具有高置信度的轨迹片段。错误的轨迹片段往往很短，我们使用持续帧数大于 5 的轨迹片段作为高置信度轨迹片段。在产生初始分组时，我们用 $\Psi_{ik} > \tau$ 作为阈值进行组的划分依据，τ 在下一节中在训练数据集中训练得到。

步骤二：组跟踪。使用最小化能量变化 ΔE_{topo} 完成组的跟踪，

$$\arg\min_{T}\sum_{t}^{\Delta t}\Delta E_{topo} \tag{6-10}$$

其中，ΔE_{topo} 如公式（6-9）介绍，表示连续帧的能量变化。Δt 表示所在组的存在时长，一般我们用组内最长轨迹的时长表示该时长 Δt。使用和第四章相似的标签交换法求解公式（6-10），获得分组关系的解 T_{ik}。

步骤三：组内跟踪。在获得了组跟踪的结果 T_{ik} 后，需要对组内的轨迹片段进行匹配求解。和全局的组跟踪不同，此时的组内目标跟踪只需在组 G_k 内完成子空间的搜索求解，

$$\arg\max_{X^{(k)}}\sum_{i,j}A_{ij}^{(k)}X_{ij}^{(k)},k=1,...,K, \tag{6-11}$$

其中 A_{ij} 和 X_{ij} 的定义和公式（6-1）相同，但只需要在 K 个组遍历求解即可。这里模型采用 Hungarian[67]算法完成内部匹配。

表 6-2 能量拓扑模型求解算法

算法 1 能量拓扑模型求解
Input: 视频序列在检测器的输出结果；
Output: 组关联结果 T，轨迹关联结果 X；
1: 通过检测框覆盖区域计算拥有高置信度的轨迹片段 L；
2: **for** $L\neq\varnothing$ **do**
3: **Step-1:** 根据公式（6-2~6-6）计算 T_{ij}，与离线拓扑匹配后产生粗分组G；
4: **Step-2:** 根据公式（6-8和6-10）计算最小拓扑能量变化，得到 $\{G\}_{k=1}^{K}$；
5 组分裂与合并;
6: **Step-3: for** k = 1…K
7: 确认拓扑中的中心节点 r_k；
8: 添加虚拟点 $\hat{l}_i$；
9: 根据公式6-11计算拓扑内轨迹关联；
10: **end for**
11: **end for**

6.4 模型训练

本节研究在 RGB-D 训练数据集下，自然场景中行人以组为运动的常见的拓扑结构和相对应的能量分布，以及随着行人密度的增加，这些拓扑结构和能量

分布如何变化。此外，我们还在训练中学习到行为相似度的阈值δ和社会距离阈值d和稳定度阈值τ。

我们用 RGB-D 数据集作为训练数据集，其中包括 13732 帧视频数据和对应的深度数据。根据训练数据集提供的真实 GT 数据，给每个检测框标上其正确的轨迹标号，进而选取轨迹长度大于 10 帧的轨迹作为训练使用。训练中我们记录拓扑的以下信息：1）大小；2）速度；3）运动方向和 4）行人组运动稳定度。部分基本拓扑结构和能量分布如图 6-3 所示。

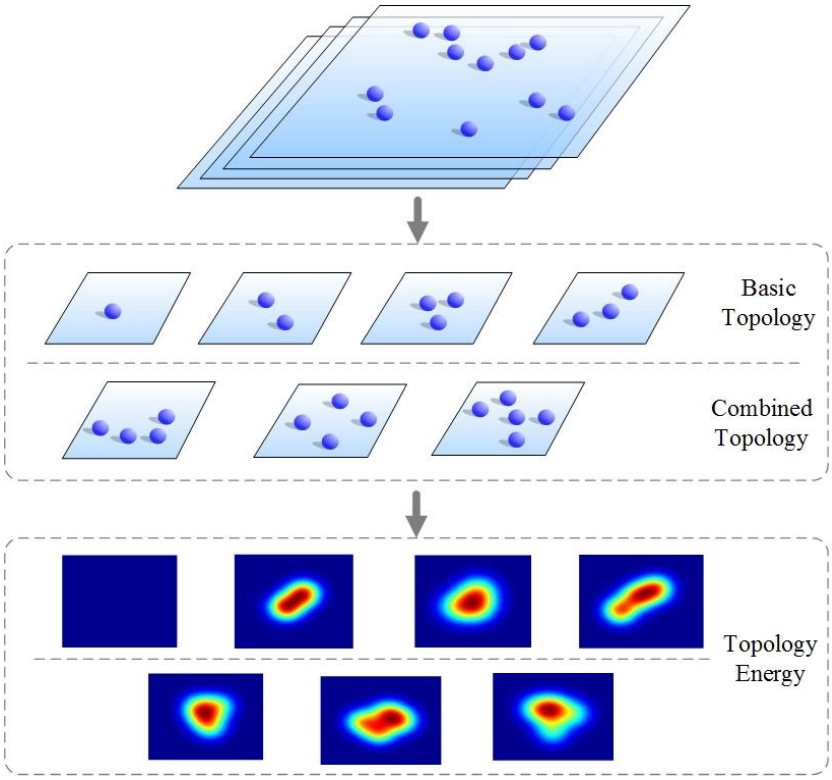


图 6-3 典型拓扑结构及其能量分布

根据训练集场景中行人的密度和拓扑的能量分布，可以观测到拓扑的组成中，行人的密度可以分为三个级别。低密度：当组内的仅有两位行人，他们一般呈现肩并肩的分布，并且垂直于他们的运动方向，因此在街道上会占据较大的位置。中等密度：当行人的密度增加时，组内成员之间的距离在减小。当组的大小为 3 时，可以发现站在中间的行人往往会退后半步，而左右两人在前面，呈 V 性分布。当组的大小为 4 时，会延续这一现象，结构呈现出 U 形分布。根据研究表明，这样的结构分布比起横线或者纵线的结构更有利于组内的所有成员进行信息交流。高密度：当组内的密度继续提高时，物理约束会比社会性约束表现的更强，组内的行人呈队列式前进，这种队列式分布也更符合流体运动的特征。此时的场景往往比较拥挤，组内行人会调整自己的位置，避免和组内

成员以及组外的行人发生碰撞。

表 6-3 训练集所学的模型参数

行为相似度阈值	距离阈值	稳定性偏差阈值
τ =0.3（公式 6-1）	d=2 米（公式 6-5）	δ =1.0（公式 6-6）

参数训练：表 6-3 显示了在训练过程中获得模型的重要参数。根据获得的基本拓扑结构，设定拓扑组内的成员的距离为 d=2 米，只有当某个轨迹片段和组内其他任意的某个目标的轨迹片段之间的距离满足此阈值，才会考虑将其划分进该组。而组的初始化中，只有当$\Psi_{ij} > \tau$时，将其划分进同一组。在组的运动过程中，当稳定性偏差值δ>1.0 时，进行组的分裂。

确定根结点：在每一个组内，并不是每一成员都对整个组的能量贡献值相同。所以模型考虑使用根节点作为拓扑组中的重要节点。模型并不是使用靠近中心位置的目标作为根节点$\{r_k\}$，而是从以下三方面特征进行考虑 1）持续时间长，2）低的稳定性偏差值，3）具有较小的深度值（不容易被其他成员遮挡）。

加入虚拟点：目标在运动过程中，往往由于位置的前后关系产生部分遮挡或者全部遮挡问题。本方法通过在组内加入虚拟节点的策略，对被遮挡的目标进行位置估计。当组内的成员数小于该组在前后帧的大小时，在该组G_k内加入虚拟点$\hat{l}_i^t$，虚拟点的位置通过该组的状态转移矩阵C_k求得，$\hat{l}_i^t = C_k l_i^{t-1}$。

组的分裂与合并：当不同组内的目标之间满足$\Psi_{ij} > \tau$时，将会考虑合并该组，但是合并后的新组仍然需要满足运动的稳定性δ<1.0。当组内成员的运动模式不一致时，如运动的方向和速度不相同，此时δ>1.0 考虑将组进行分裂。我们将从具有最小行为相似度Ψ_{ij}的目标之间分裂该组。

6.5 RGB-D 特征提取

本节在多目标跟踪框架下探索了基于 RGB-D 数据的行人目标特征。由于 RGB-D 数据对行人目标的呈现主要包括两个部分：来自图像域的目标边界框（bounding box）以及来自激光的深度数据。模型需要通过这两种性质的数据完成目标的 3D 位置、外形和运动信息的计算。

如图 6-4 所示，假设 3D 空间内的每一个目标都有一个独立的 3D 边界框，则目标的特征主要是提取基于 RGB-D 的特征向量，同时在外形和 3D 空间能够区分不同的目标。传统的基于 2D 目标边界框的特征提取会面对一个框中含有两个或更多目标的情况，并且这些目标之间往往出现遮挡，尤其当框中的目标

具有较高颜色相似度时，这种特征提取的方式基本无法区分目标。当加入深度信息之后，模型将 RGB-D 数据进行两个平面的投影如图 6-4 所示，X-Y 平面和 Y-Z 平面，来消除这种遮挡造成的特征提取错误的情况。模型对通用的 RGB-D 特征提取方式[107,108,109]进行了修改。在 X-Y 平面内不是简单的利用 2D 特征提取方式，如 HOG 和 Haar-like 特征完成特征提取，而是对目标的深度数据和 RGB 数据进行对齐归一。根据深度数据，将边界框中的图像分为两部分，目标区域和背景区域。通过计算框内深度平均值，可以将大于平均深度的像素区域去掉，从而得到目标区域的像素值。在目标区域内计算 HOGC 特征[110]得到目标的外形特征。模型将 Y-Z 平面作为一个辅助平面，在该平面内通过聚类来计算目标的平均深度，以此作为目标的深度值 z_i^t。

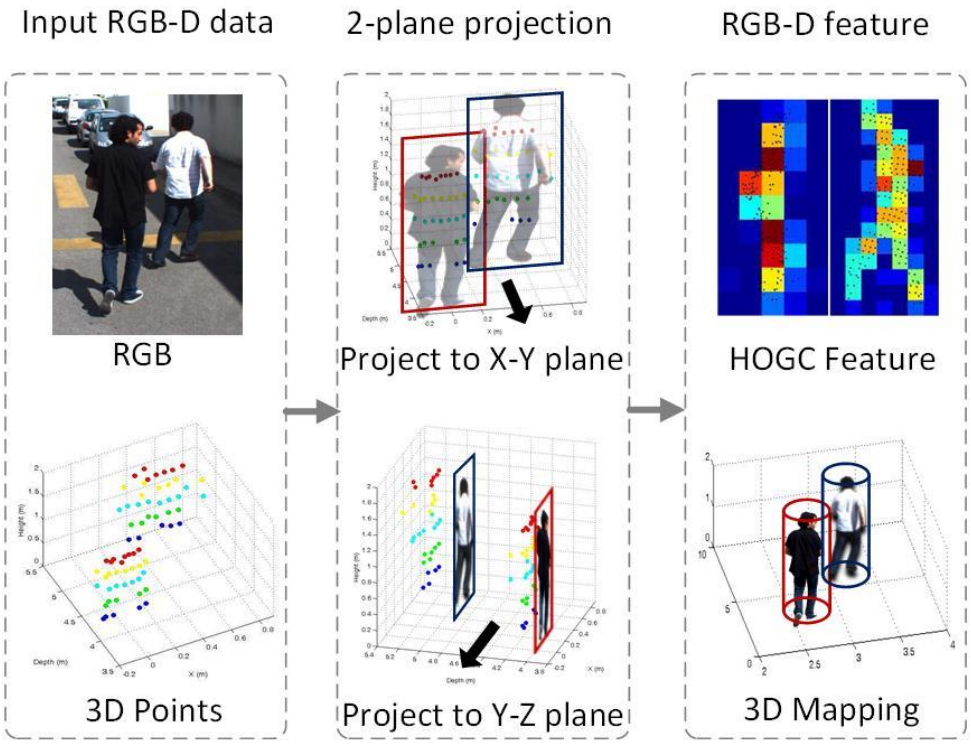


图 6-4 基于 RGB-D 数据集的目标特征

外形相似度：两个轨迹片段之间的外形相似度 A_{ij} 通过高斯方程计算

$$A_{ij} = \mathbb{N}\left(S_{cos}\left(w_i, w_j\right); 0, \Sigma\right), \tag{6-12}$$

其中，$S_{cos}(\cdot)$ 计算两组 RGB-D 特征 w_i 和 w_j 之间的余弦距离。

6.6 实验验证

通过在 RGB-D 数据集与其他多目标跟踪模型的对比实验，验证了所提出的基于拓扑能量模型的精确性。所用的四个公开的 RGB-D 数据集包括：Sync 数据集、SDL-Crossing 数据集、SDL-Campus 数据集和 LIPD 数据集。这些数据集在第三章中有详细介绍。为了使对比实验在公平性原则下进行，实验采用统一的检测输入，这里使用 DPM 行人检测器[101]，在实验评测时采用相同的真实数据和评测标准。

表 6-4 对比实验结果

数据集	类型	方法	Recall	Prec.	GT	MT	ML	Frag.	IDS
Sync	RGB	Berclaz[75]	69.6	72.8	66	9.0	24.2	345	323
		Andriyenko[104]	73.4	78.3	66	19.6	19.6	89	125
		Milan[105]	76.2	79.2	66	25.8	16.7	102	147
	RGB+G	Chen[90]	76.6	80.3	66	25.8	**15.1**	121	133
	RGB+T	Ours*	80.8	83.4	66	27.2	**15.1**	97	124
	RGBD+T	Ours	**87.5**	**92.2**	66	**37.8**	16.7	**77**	**109**
SDL-Crossing	RGB	Berclaz[75]	68.9	70.5	92	9.8	25.0	168	189
		Andriyenko[104]	70.4	76.4	92	21.7	19.6	89	**69**
		Yang[71]	73.7	75.1	92	17.3	25.0	106	79
	RGB+G	Chen[90]	74.6	77.5	92	18.4	19.6	103	98
	RGB+T	Ours*	79.5	82.1	92	25.0	17.3	84	77
	RGBD+T	Ours	**87.4**	**89.0**	92	**30.4**	**10.9**	**55**	83
SDL-Campus	RGB	Berclaz[75]	66.4	69.8	74	12.2	33.8	130	146
		Milan[105]	74.0	76.5	74	20.2	21.6	78	97
	RGB+G	Chen[90]	75.1	77.3	74	18.9	20.2	80	110
	RGB+T	Ours*	81.9	83.1	74	21.6	18.9	77	80
	RGBD+T	Ours	**89.2**	**91.2**	74	**33.8**	**13.5**	**53**	**58**
LIPD	RGB	Berclaz[75]	72.8	72.4	77	10.4	33.8	324	219
		Yang[71]	76.3	81.2	77	19.5	22.1	141	164
	RGB+G	Chen[90]	77.3	81.5	77	13.0	23.4	150	172
	RGB+T	Ours*	82.3	85.6	77	24.7	19.5	102	147
	RGBD+T	Ours	**88.7**	**90.0**	77	**37.7**	**13.0**	**89**	**92**

选用多种类型的多目标跟踪方法作为对比方法：

1）基于 RGB 的方法（RGB）：网络流模型[75]，在线 CRF 模型[71]，轨迹分析模型[104]和连续能量优化模型[105]；

2）基于 RGB 数据的分组方法（RGB+G）：分组跟踪模型[90]；

3）基于 RGB 数据的拓扑能量方法（RGB+T）：采用本章提出的模型但不使用第 6.5 节提出的 RGB-D 特征，在表 6-4 中用“Ours*”表示；

4）基于 RGBD 数据的拓扑能量方法（RGBD+T）：本章提出的完整拓扑能量模型，在表 6-4 中用“Ours”表示。

表 6-4 中展示了本章提出的基于拓扑能量的模型与其他所有类型的跟踪方法在各种评测指标下的详细结果。在表格中可以看到“RGBD+T”方法的性能优于其他对比方法。在准确率和覆盖率指标 Prec.，Recall 分别提高将近 12%。即使不使用数据中的深度数据（RGBD 特征），只采用本章提出的“RGB+T”的拓扑能量方法，仍然在 Recall 和 Prec.指标上平均提高 5%和 4%。这充分证明基于拓扑能量的模型对于求解多目标跟踪问题是准确有效的。因为新加入的拓扑级约束比传统的轨迹级约束的跟踪方法能够更准确地刻画以组为单位运动的行人目标。分别对组内成员和组外行人的拓扑建模方式能够将不同运动模式下的行人区分性对待。通过组内成员的拓扑构建，当目标之间存在遮挡问题时，能够通过组内统一的运动方程估算出被遮挡目标的位置。

从实验结果中可以发现，拓扑能量模型在实验中有更少的 IDS 和 Frag.错误，这样可以关联出更完整的行人轨迹。拓扑能量模型的实验结果如图 6-5 和 6-6 所示，图中的实验结果包括三个部分，真实世界坐标下的视图，图像上的跟踪结果以及拓扑能量分布图。从图 6-5 可以很明显地看出，由于行人都是成组运动的，拓扑能量在多帧内的变化平稳缓慢。在图 6-6 中的世界坐标系视图下，可以看到行人组的合并以及分裂现象的发生。

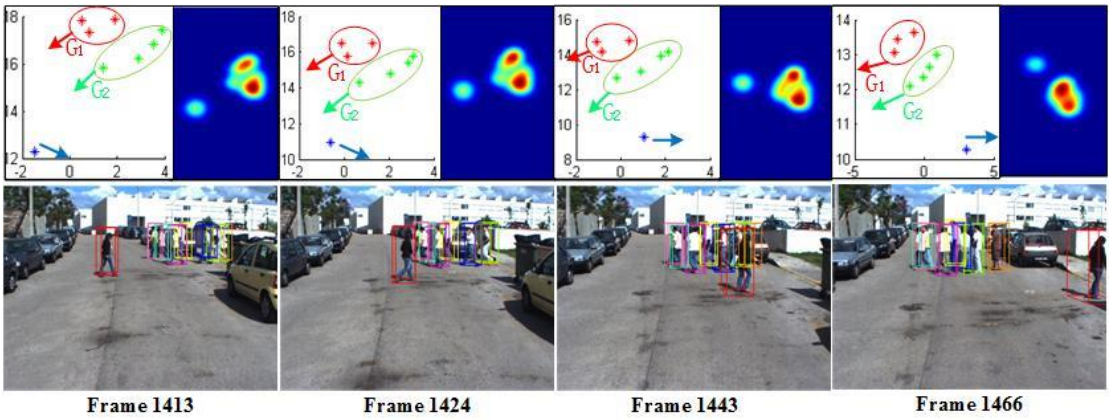


图 6-5 拓扑能量模型在 Sync 数据集实验结果

6.6.1 实验分析

从实验结果可以看出，本章提出的基于拓扑能量最小化求解多目标的跟踪模型性能优于基于轨迹分析的模型，分组模型。模型的性能提升主要来源于全局拓扑结构和 RGB-D 特征的引入。本节从这两方面分析模型的优点。

拓扑结构与组结构：即使去掉 RGB-D 特征，在基于 RGB 数据的模型上作比较，本章的拓扑能量模型（表 6-4 中标记为“ours*”）仍然比基于分组进行多目标建模的方法[90]性能更好。同样是对自然场景的行人进行分组，分组模型[90]只是考虑将具有相同运动模式且距离较近的行人进行聚类，并没有考虑组与组之间的运动关系。而本章提出的基于拓扑的方法则要考虑组间的拓扑能量关系，将组内与组间的运动关系用能量变化形式进行表征，并且以全局的眼光在持续帧之间保持总的拓扑能量近似相等。这使得每帧之间的目标本组状况差异不会很大，同时不会在两帧之间形成差异较大多目标关联结果。这能保证自然场景中多目标跟踪与实际情况的行人运动是一致的。

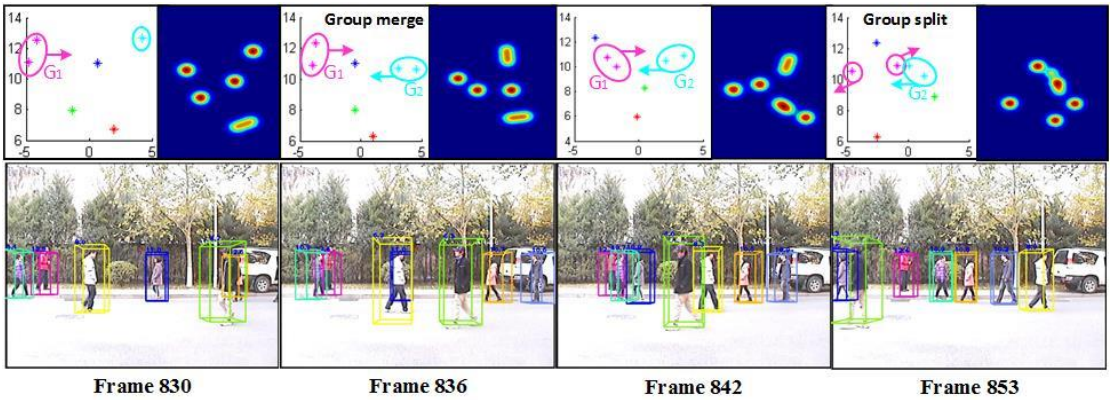


图 6-6 拓扑能量模型在 SDL-Crossing 数据集实验结果

RGB 特征与 RGB-D 特征：本章还对 RGB-D 特征进行了探索，改进了传统的 RGB-D 特征，使其在多目标跟踪框架中更加有效。区别与传统的方形检测框中直接提取目标外形特征，本章通过深度信息只提取检测框内的前景区域目标自身的外形特征。这对具有动态背景的交通场景应用更加有效，使得目标的外形特征不会被复杂多变的背景特征所干扰。而且模型利用深度信息进行距离和位置的判断比使用图像域的像素距离更可靠。往往在平视视角的场景中使用像素距离作为距离度量会使得目标之间不具有区分性，因为平视视角下的行人在同一水平面，几乎所有的检测框的中心处在同一高度，在检测结果不准确时，使用这样的像素距离会直接导致跟踪错误。因此，RGB-D 特征中的深度特征是

对 RGB 特征的有效补充。

6.7 本章小结

本章提出了基于拓扑能量的多目标跟踪模型，用能量形式衡量拓扑结构中组内组外目标的运动相似性。在分组过程中达到组内目标的运动尽可能相似，组间的目标运动尽可能相异。在求解模型时，模型使用拓扑能量变化最小的方式寻找帧之间拓扑能量变化最小的解作为组跟踪的解。进而在组内进行组内目标的匹配求解。通过这种二步求解方式找到所有目标的对应轨迹。在四个 RGB-D 数据集中，与多种类型的多目标跟踪方法进行对比实验后发现，本章提出的多目标跟踪模型提高了多目标跟踪算法的性能。

第七章　基于拓扑图模型的多目标跟踪方法

7.1 模型概述与创新点

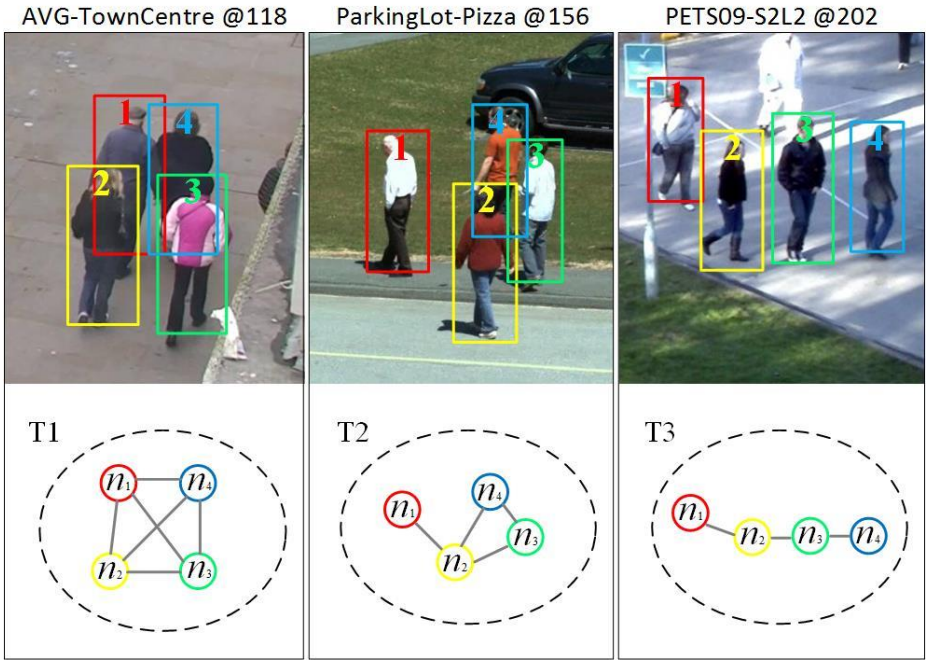


图 7-1 目标的拓扑图表示

本章继承了第六章中以组为单位对行人目标进行跟踪的研究框架，提出了基于动态拓扑图的多目标跟踪模型。将组内的行人看作是图中的节点元素（如图 7-1 所示），通过图中边的连接方式探索组内成员在行走过程中的运动变化。模型通过离线与在线学习相结合的方式，对组的典型拓扑进行离线学习后，在线跟踪中与行人组进行拓扑结构匹配。在组的运动过程能够中，模型通过组的初始化、更新、合并和分裂的动态变化完成对行人组的跟踪。最后，利用线性规划的方式完成组内行人的身份确认，得到每个目标的完整轨迹。实验中将动态拓扑图模型在 RGB 和 RGB-D 数据集上进行实验验证。本章的创新点包括以下几个方面：

1）提出了在线学习的动态拓扑图模型；

2）使用离线学习方式在训练集中获取典型的拓扑结构；

3）通过离线学习和在线学习相结合的方式进行行人组和组内行人跟踪。

7.2 拓扑图模型

本章首先介绍建立在组动态和空间拓扑结构上的行为相似度矩阵$T=\{T_{ij}\}$，该矩阵计算了场景中所有目标的社会行为相似度（social behavior affinity）。接着分别从空间和时间角度介绍了拓扑模型的两种特性，空间一致性和时间一致性。这两种拓扑特性与行为相似度矩阵被用来在线更新拓扑结构，以及离线获得组的典型拓扑结构。

表 7-1 本章使用的符号及含义

符号	含义
n_i^f	第 f 帧中第 i 个轨迹片段，$n_i^f=(p_i^f,v_i^f,o_i^f,A_i^f)$
X_{ij}	二进制指示向量，表示轨迹片段之间的连接关系
T_{ij}	两个轨迹片段的行为相似度包括T_d，T_o，T_v和T_t
L_{ij}	两个轨迹片段共同出现的帧数
o_i	轨迹片段n_i的运动方向
w_i	轨迹片段n_i的在图像上的宽度
v_i	轨迹片段n_i的速度
G	拓扑图，包含边和节点
D_{n_i}	拓扑图中节点n_i的度
A_{ij}	两个轨迹片段的外形相似度
π_k	组G_k的马尔可夫链状态参数向量，$\pi_k=\{C_k,Q_k,\mu_k,\delta_k\}$
l	轨迹共同出现的时间阈值
d	距离阈值
τ	相似度阈值
P_{ij}	组内每个目标相对于中心位置的位置相似度

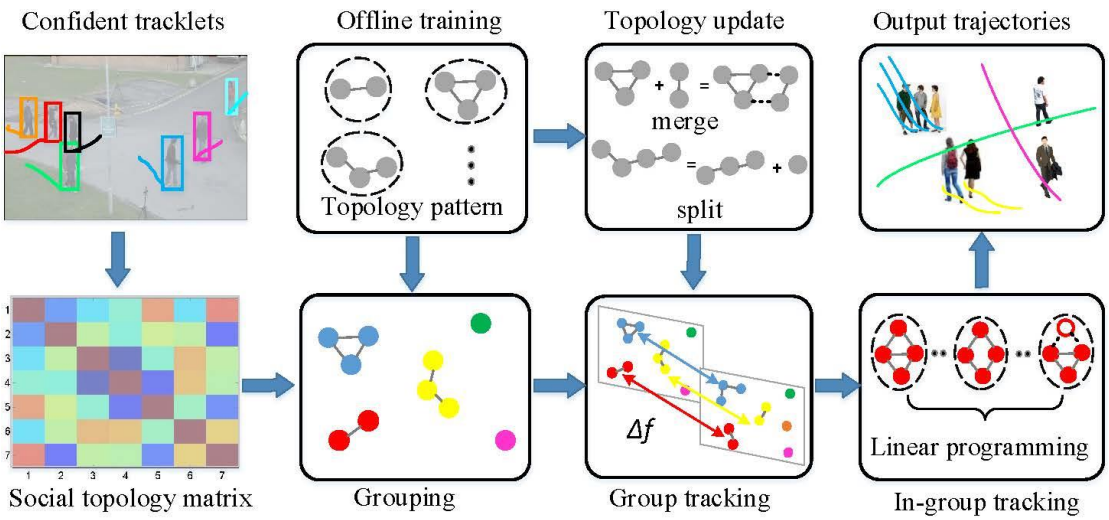


图 7-2 拓扑图模型求解多目标跟踪的流程图

7.2.1 行为相似度

在视频帧 f 中，目标 i 所对应的轨迹片段 n_i^f 可以表示为 $n_i^f=(p_i^f,v_i^f,o_i^f,A_i^f)$，其中 p_i^f 、 v_i^f 、 o_i^f 和 A_i^f 分别表示位置、速度、方向和外形特征。社会相似度矩阵 $T=\{T_{ij}\}$ 衡量每对轨迹片段之间的相似度，表示为：

$$T_{ij}=T_d\left(n_i,n_j\right)\cdot T_t\left(n_i,n_j\right)\cdot T_v\left(n_i,n_j\right)\cdot T_o\left(n_i,n_j\right), \tag{7-1}$$

其中， $T_d(\cdot)$ 、 $T_t(\cdot)$ 、 $T_v(\cdot)$ 和 $T_o(\cdot)$ 表示距离、时间、速度和方向的相似度。这里为了使公式的简洁，省略了上标 f。

距离相似度：模型根据不同性质的数据集，对于在距离度量采取不同的策略。在 RGB-D 数据集，模型采用真实的距离值， d_{ij} 表示两个目标观测之间以米为单位的真实距离值。在 RGB 数据集中，模型采用像素距离 $d_{ij}=\left(w_i+w_j\right)/2$，其中 w_i 和 w_j 表示目标观测在图像上的宽度。 δ 为距离阈值，将在本章训练模型中求得，距离相似度可以表示为

$$T_d\left(n_i,n_j\right)=\frac{w_i+w_j}{2\delta}. \tag{7-2}$$

时间相似度：时间相似度衡量两个目标的轨迹片段在多长时间内具有相似的运动特征和相近的距离。本方法用 L_{ij} 表示两个轨迹片段共同出现的帧数，他们的距离 d_{ij} 满足： $d_{ij}<2\delta$ ，并且要求这个共同出现的帧数至少持续 l 帧。他们之间的时间相似度可以表示为

$$T_d\left(n_i,n_j\right)=\frac{L_{ij}}{L_{ij}+l}. \tag{7-3}$$

速度相似度：同一组内的目标往往具有相似的运动速度。这里用 v_i 和 v_j 表示两个轨迹片段的速度，则他们之间的速度相似度表示为

$$T_v\left(n_i,n_j\right)=\mathbb{N}\left(\left\|v_i-v_j\right\|\right), \tag{7-4}$$

其中，$\mathbb{N}(\cdot)$表示归一化操作。这里将两段轨迹的速度线性归一在[0, 1]范围内。

方向相似度：模型采用改进的 Potts 模型[106]来定义两段轨迹方向之间的相似度，

$$T_o\left(n_i,n_j\right)=\cos\left(o_i-o_j\right), \tag{7-5}$$

其中，$o_b=\frac{2\pi q_b}{q},b=i,j$。目标的运动方向被划分为 q 个单位中，和前几章运动方向定义相似，模型中使用 $q=9$。

上面的速度和方向特征的相似度与参考文献[88,90]不同，本方法将速度和方向分成两项单独定义。我们发现将目标的运动特性分成速度和方向分别进行描述定义，在多种数据集上，如 RGB 和 RGB-D 数据集，具有鲁棒性。同时，在 RGB 数据集上，我们发现在面对输入的检测结果不好时，分开定义的速度和方向相似度项，使得目标观测之间的相似性更具有区分性。例如我们在方向相似度中，使用“0”区间代表静止速度，使得目标在检测结果在一定距离阈值出现抖动时，仍能将目标看作静止的物体。此外，社会行为相似度矩阵 T 可以作为一种研究社会行为相似度的工具用于不同应用场合。

拓扑图模型：当获得了场景内的目标之间的行为相似度矩阵 T，我们定义动态拓扑图为 $G=\left(\{n_1,\cdots,n_N\},E\left(T_{ij}\right)\right)$，其中 N 个目标观测由点集 $\{n_1,\cdots,n_N\}$ 构成，n_i 作为图中的点代表目标在一帧内的轨迹点。图中的边的集合 E 代表观测之间的相似度，它的权值为 T_{ij}。G 中的边连接关系代表了每个组的拓扑结构。在后面章节中，我们利用相似度矩阵 T 对图 G 进行在线的学习更新。

7.2.2 拓扑图性质

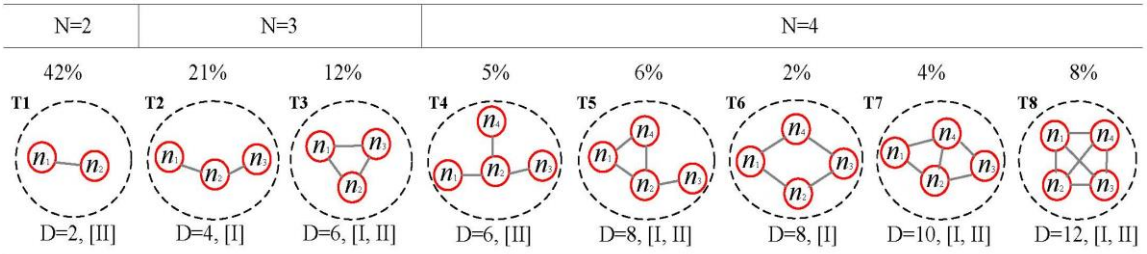


图 7-3 典型拓扑结构

在本节中，我们定义拓扑图在时间和空间上的特性：松紧度（tightness）和

稳定性（consistency）。松紧度代表每个组内的成员在“行为距离”要足够的近；稳定度要求组内每个成员在一定时间内的运动方向和速度要尽可能一致。

松紧度：我们用图内节点之间边的密度来衡量组内的成员“行为距离”上的远近。在图理论中，一个节点的度（degree）表示有多少条边与该节点相连。这里我们使用符号 D_{n_i} 表示节点 n_i 的，则对拓扑图的松紧度约束表示为：

$$s.t.\begin{cases} I:\sum_{i}^{N} D_{n_i} > 2(N-1), \\ II: \ \max D_{n_i} = N-1, \end{cases} \tag{7-6}$$

其中，N 代表组内成员的个数。约束 I 保证拓扑内有较高的边密度，满足该条件的拓扑结构的节点之间有足够多的边进行连接。这样能保证如图 7-1 中 T3 的“线形结构”能被排除。约束 II 保证拓扑呈现出一个团结构，往往组内的成员分布在中心节点的四周，呈现出一个“星形结构”。本模型要求组的拓扑结构至少满足上述两种约束中的一条。如图 7-3 中所示，T2 和 T6 满足约束 I，T1 和 T4 满足约束 II，T3，T5，T7 和 T8 同时满足两种约束。在实验中发现，同时满足两种约束的拓扑结构表现出更强的稳定性。

稳定性：我们用稳定性要求组内成员空间运动尽可能的一致，假设组 G_k 内有 m 条马尔可夫链，并且每一条链满足时域上的模型满足

$$p_i^f = C_k p_i^{f-1} + \varphi^f \ , \tag{7-7}$$

其中，组内成员 n_i^f 的空间位置 p_i^f 由马尔可夫链中状态转移矩阵 $C_k \in \mathbb{R}^{4\times4}$ 更新。$\varphi^f \sim \xi(0,Q)$ 表示满足高斯分布的噪声。$p_i^f = \left[x_i^f, y_i^f, z_i^f, 1\right]^T$ 表示目标观测的 3D 坐标，其中 z_i^f 表示深度信息（在 RGB 数据中，z_i^f 被设置为 0）。初始的位置观测满足高斯分布 $\xi(\mu,\delta)$。此时参数 $\pi_k = \{C_k, Q_k, \mu_k, \delta_k\}$ 代表了整个马尔可夫链的参数集合，其中 C_k 表示 G_k 内的所有成员的运动状态转移矩阵，$\{\mu_k, \delta_k\}$ 保证组内成员在空间上尽可能靠近。

7.3 拓扑图在线学习

动态拓扑图模型反应场景中目标的动态变化，同时需要根据跟踪过程中目标之间的行为关系不断更新。为了达到这一目的，在组内加入新的成员、删除某些成员、组之间的合并和分裂等活动都需要在动态拓扑图内的边和节点进行调整变化。本方法将跟踪中在线进行组的初始化（Birth），更新（Update），合并（Merge）和分裂（Split）放到四个模块中。其中，上一节介绍的两种性质也将在模块中用于对组在时间和空间中的更新。松紧度约束用于找到空间结构上

符合要求的组，稳定性约束会在线对组的运动方程进行更新。与其他方法相比，本方法提出的动态拓扑图模型能够自动地对组进行划分，并且通过行为相似度矩阵将目标之间的行为相似度与组的更新进行自然结合。每个模块的详细算法参考表 7-2。

初始化模块：在组的初始状态中，它的边集合为 $E=\{\varnothing\}$。目标之间的行为相似度矩阵 T 被用来作为构建边的依据。当目标观测 n_i 和 n_j 的轨迹片段之间的相似度满足 $T_{ij}<\tau$，则设定此对目标之间存在一条边。得到整个帧内的图中的 E 后，我们将目标之间的连接关系与图 7-3 中学到的八种典型拓扑结构进行对比，如果有目标之间的结构符合这些典型拓扑结构吻合，则把这些目标划分为一组。在初始化组的过程中，不用复杂的结构描述一个组，这样可以规避由于初始化中的组过大导致后面不断出现组分裂的风险。如果这些基本组具有很高的行为相似度，将在合并模块中对小的组进行合并。

更新模块：在跟踪过程中，组的结构是随时间进行更新的，因此拓扑图中节点之间的边的连接也随着目标的运动进行在线更新。此模块仍然使用相似度矩阵 T 来更新边的权值。当不同组内的目标之间满足 $T_{ij}<\tau$ 时，将会跳转到合并模块；当组内原本存在边的两个目标之间的相似度 $T_{ij}>\tau$ 时，将会跳转到分裂模块。

合并模块：两个组具有较高的行为相似度时，考虑将这两个组进行合并。但这并不意味着所有具有较高行为相似度的组会被成功合并。首先，模块需要计算新的组是否满足公式（7-6）中的松紧度约束，如果不满足该约束，则两个组不能被合并。进一步，模块需要计算新组的松紧度 $D=D_{G1}+D_{G2}+D_{new}$，其中 D_{new} 是组内因为合并产生的新边所产生的度。此时定义这个新的度满足

$$D_{new}>\min(N_{G1},N_{G2}), \tag{7-8}$$

其中，N_{G1} 和 N_{G2} 表示组 G1 和 G2 的大小。D_{new} 则要求大于较小的组的大小，$\min(N_{G1},N_{G2})$。图 7-4 展示了两个组合并的一个实例，G1 和 G2 是两个具有较高行为相似度的组，可以合并成如 G3 和 G4 两种不同结构的组，但是只有 G4 是正确合并的组。因为此时 $\min(N_{G1},N_{G2})=2$，而只有 G4 中 $D_{new}>2$。

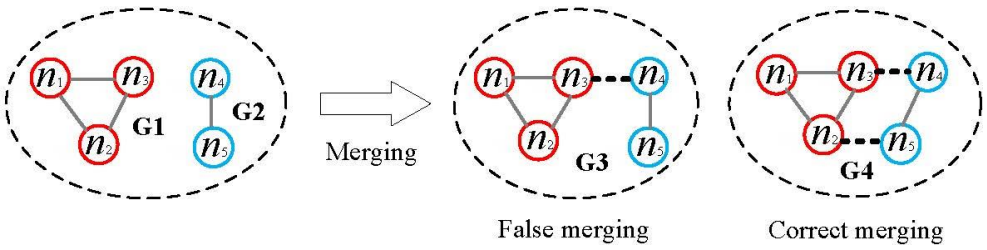


图 7-4 组合并实例

分裂模块：当组内成员的运动模式不一致时，如运动的方向和速度不相同，当他们之间的行为相似度$T_{ij}>\tau$，考虑将组进行分裂。我们将切断具有最小行为相似度T_{ij}的边。此时，仍然要求分裂后的新组满足结构要求公式（7-6）。

表 7-2 在线学习的四种模块

算法 1：拓扑图模型的在线学习

初始化模块（Birth-Module）

Input: $E=\{\varnothing\}$

Output: $G=\{G_k\}$

Step-1：

根据公式（7-1）计算$T=\{T_{ij}\}$；

For 每一对连接 T_{ij}；

IF $T_{ij}<\tau$，$E=E\cup\{(n_i,n_j)\}$；

End for

Step-2：

聚类产生 E；

For 每一个分组G_k

IF D_k 不满足松紧度约束 7-6，**GOTO** 分裂组块;

根据图（7-3）产生标准的初始分组；

根据公式（7-7）计算每一个组的动态转移矩阵C_k；

End for

更新模块（Update-Module）

Input: $G^{f-1}=\{G_k^{f-1}\}$，$T^f=\{T_{ij}^f\}$

Output: $G^f=\{G_k{}^f\}$

For 每一个组 G_k^{f-1}

Step-1:

更新边的权值T_{ij}，$i,j\in G_k$；

根据公式（7-7）计算组的动态转移矩阵C_k；

Step-2:

IF 组的大小 $N_k^f<N_k^{f-1}$，加入虚拟点；

IF 边与组外目标有链接, **GOTO** 合并模块；

IF 松紧度D_k不满足公式（7-6），**GOTO** 分裂模块；

End for

合并模块（Merge-Module）

Input: G_{k1}，G_{k2}

计算 $G_{k1} \cup G_{k2}$ 的松紧度 D_{new}；

IF $D_{new} > \min(N_{k1}, N_{k2})$

根据公式（7-7）计算组的动态转移矩阵 C_{new}；

Output: $G_{new} = G_{k1} \cup G_{k2}$；

ELSE

Output: G_{k1} 和 G_{k2}；

分裂模块（Split-Module）

Input: G_k ,T

Output: $\{G_k'\}$

While D_k 松紧度 D_k 不满足公式（7-6）；

切断具有最低相似度 T_{ij} 的边；

$G_k' = G_{k1} + G_{k2}$，$D_k = D_{k1} + D_{k2} + D_{new}$；

End While

计算状态转移矩阵 $\{C_k'\}$ 和 $\{G_k'\}$；

7.4 拓扑图模型训练

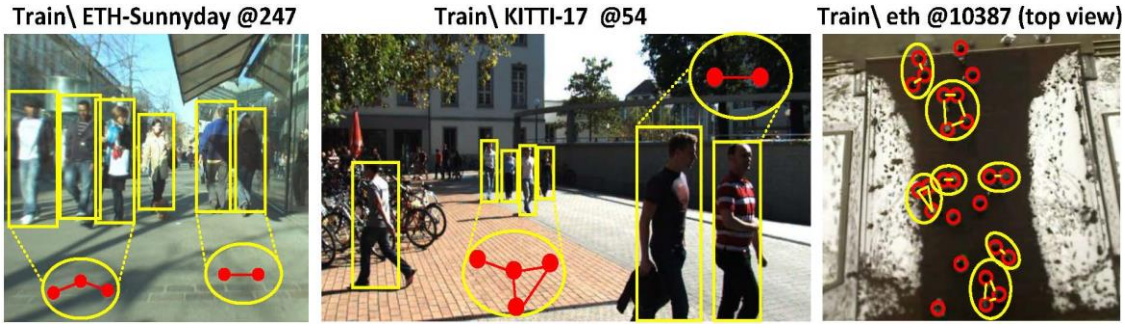


图 7-5 典型拓扑结构的离线学习

本节研究多种场景内以组为单位运动的行人目标在行走过程中的组内结构变化，并探索典型的组拓扑结构。这些典型的结构将被用于上一节中组的初始化。此外，模型中两个参数需要在训练集中通过离线训练求得：公式（7-2）所用到的距离阈值 δ，以及行为相似度阈值 τ。

在训练数据集中，目标的正确检测结果以及每一个行人的完整轨迹在真实数据（Groundtruth）中已经给出。首先，将真实数据中的正确标号赋值给每一

个行人目标。假设场景内有 N 个行人，并且他们同时存在时长超过 l 帧，可以通过公式（7-1 至 7-5）计算他们的行为相似度矩阵 T，则得到一个关于场景内所有目标连接情况的图，其中用来连接这些目标的边的权值仍然是T_{ij}。然后，我们设定一个逐渐增大的相似度阈值τ，来决定这张图中节点目标的是否被连接，这时能够得到对应于每个τ的分组情况。最后，通过与真实数据比较计算分组的总数错误ε_g，以及目标在组之间错误跳变的次数ε_s，并求出

$$arg\min_{\tau}\sum \varepsilon_g + \varepsilon_s \tag{7-9}$$

此时的τ为训练数据集下最优的相似度阈值。对应于此阈值，可以求得目标之间的距离值δ。

表 7-3 记录了不同训练数据集下，对应的相似度阈值和距离阈值。在 RGB 数据集中，如图 7-5（a）和（b）所示，使用目标的区域窗口中心位置之间的像素距离，$\left(w_i + w_j\right)/2$，作为目标之间的距离值，在 RGB-D 数据集中，使用目标的真实世界坐标下的距离值。另外，所用到的 RGB 数据集 eth 和 hotel 为顶视角下的拍摄视频，目标的检测窗口为圆形区域，如图 7-5 所示。所以我们采用该区域的半径 r 作为度量值。在获得典型的拓扑结构时，当$d_{ij} < 2\delta$，我们计算两个目标之间的行为相似度T_{ij}，而当$T_{ij} < \tau$时，在相对应的两个目标之间存在一条边进行连接。按照如上所述的规则，在训练数据集中我们得到了八种最常见的拓扑结构，他们的大小在 2-4 之间。

表 7-3 训练集所学的模型参数

数据集	性质	距离阈值	相似度阈值	时间阈值	数据集密度
MOT test	RGB	$\delta = \left(w_i + w_j\right)/2$	τ = 0.27	l=10	高达每帧 30 个行人目标
eth 和 hotel	RGB(顶视角)	$\delta = r_i + r_j$	τ = 0.30		
SDL-Campus	RGB-D	$\delta = 1$ meter	τ = 0.34		

7.5 基于拓扑图的多目标跟踪求解

本节将会介绍利用上文所述的拓扑图模型，自底向上的完成多目标跟踪问题求解。假设在场景中存在轨迹片段集合$L = (n_1, \cdots, n_n)$，其中每一个轨迹片段n_i是通过检测结果在空间上的覆盖率计算出的目标轨迹片段。与前面章节所描述的多目标跟踪方法目标一致——找到每个目标的完整轨迹。下面模型采用四个

步骤先完成对组的跟踪，再完成对组内每个成员的跟踪。

步骤一：寻找高置信度的轨迹片段。在“先检测再跟踪”的框架下，每一个轨迹片段都是检测结果串联而成，但是用于进行组分析的轨迹片段，在本方法中只是用具有高置信度的轨迹片段。因为错误的轨迹片段往往很短，所以模型使用持续帧数大于 5 的轨迹片段作为高置信度轨迹片段。

步骤二：组跟踪。如拓扑图的在线学习一样，通过在置信度高的轨迹片段之间计算行为相似度矩阵 T 并和训练求得的八种典型拓扑结构匹配进行组的初始化。然后通过组的在线更新模块完成组的更新，合并和分裂，并完成对组的跟踪。

步骤三：解决自遮挡问题。多目标在运动过程中，往往由于位置的前后关系对观测位置产生部分遮挡或者全部遮挡问题。在这种情况下，即使最好的检测器也无法完成目标的检测，所以轨迹片段关联到这里会被切断。然而，本方法通过在组的拓扑图内加入虚拟节点的策略，对被遮挡的目标进行位置估计。当组内的成员数小于该组在前帧的大小时，在该组内加入虚拟点 $\hat{n}_i$，虚拟点的位置通过组的状态转移矩阵 C_k 求得，$\hat{p}_i^f = C_k p_i^{f-1}$。

步骤四：组内目标跟踪。在获得了组跟踪的结果后，模型利用线性规划方法求解组内目标跟踪问题。和全局的组跟踪不同，组内跟踪完成组的“生存周期”内子图（sub-graph）搜索求解，

$$\begin{aligned} &\arg\max_X \sum_{i,j} A_{ij} X_{ij} + P_{ij} X_{ij} \\ &s.t. \sum\nolimits_i X_{ij} \le 1, \sum\nolimits_j X_{ij} \le 1 \end{aligned} . \qquad (7\text{-}10)$$

其中 X_{ij} 表示 G_k 内的二进制矩阵，决定 p_i^{f-1} 和 p_i^{f-1} 是否属于同一个目标。A_{ij} 表示两个目标观测之间的外形特征相似度。这里，我们使用 HSV 颜色直方图和 HOG 特征联合串联特征。P_{ij} 表示组内每个目标相对于中心位置的位置相似度，这里使用他们相对于中心位置的角度偏移，并且将这个由角度偏移计算的位置相似度归一化到满足正态分布的[0,1]区间。最后，采用 Hungarian 算法[67]和迭代估计算法[111]求解方程（7-10）。

7.6 实验验证

在 RGB 和 RGB-D 两种性质的数据集上分别验证本章提出的动态拓扑图（GST）模型的准确性和有效性。我们将所提出的 GST 模型与近年来的性能最好的多目标跟踪器在相同数据集，相同初始条件下进行比较。这些跟踪器包括：DP[76]，SSP[75]，CEM[105]，SegTrack[112]，MotiCon[113]和 MDP[114]。

7.6.1 RGB 数据集评估

选用 MOT Benchmark[96]平台作为实验所需的 RGB 数据集。其中 MOT Benchmark 包含 11 段训练集和 11 段测试集，共有 11286 帧视频数据（大约 16.5 分钟），并且视频的帧频各有不同。部分视频序列由搭建在移动平台上摄像头录制完成，另一部分视频序列来自于监控视频。在实验中使用该平台提供的 11 段训练集进行参数的训练，通过训练得到$\tau = 0.27$ 作为图中边的阈值。为了使所有方法在相同条件下进行比较，我们使用 MOT Benchmark 提供的公共检测结果作为多目标跟踪的输入。此外，实验采用相同的跟踪评测标准 CLEAR-MOT 评价体系做多目标跟踪的评判标准。

表 7-4 RGB 数据集对比实验结果

数据集	方法	MOTA	MOTP	MT	ML	FP	FN	IDS	Frag.
MOT Benchmark	DP[76]	14.5_13.9	70.8	6.0	40.8	13,171	34,814	4,537	3,090
	SegTrack[112]	22.5_15.2	**71.7**	5.8	63.9	**7,890**	39,020	**697**	**737**
	MotiCon[113]	23.1_16.4	70.9	4.7	52.0	10,404	35,844	1,018	1,061
	MDP[114]	30.3_14.6	71.3	**13.0**	38.4	9,717	32,422	680	1,500
	GST	**33.8_13.6**	71.1	12.1	**34.8**	9,232	**31,743**	722	1,257
AVG-TownCentre	DP[76]	6.6	69.4	4.4	35.8	876	4,482	1,317	562
	SegTrack[112]	3.3	69.3	0.9	86.3	**235**	6,528	151	108
	MotiCon[113]	11.9	**70.3**	0.9	69.9	353	5,872	**74**	**75**
	MDP[114]	25.4	69.7	17.7	33.6	1,517	**3,691**	122	264
	GST	**33.7**	70.2	**22.1**	**30.1**	942	3,756	113	163
PETS09-S2L2	DP[76]	33.8	69.4	7.1	**9.5**	948	4,410	1,029	705
	CEM[105]	44.9	70.2	11.9	14.3	657	4,506	**150**	165
	MotiCon[113]	46.6	67.6	9.5	14.3	**560**	4,354	238	264
	SegTrack[112]	46.1	**70.6**	**26.2**	16.7	1,213	**3,773**	211	211
	GST	**51.8**	70.4	16.7	11.9	715	3,812	172	**161**

表 7-4 总结了本章所提出的 GST 模型和其他性能最好的多目标跟踪器在 MOT 测试集上的结果对比。从表格中可以发现，本章所提出的 GST 模型很多性能优于对比跟踪模型。尤其在最能反映多目标跟踪的整体性能指标——跟踪准确率（MOTA），超过所有对比模型。此外，表 7-4 展示了 MOT 测试集上难度最高的两个数据集 AVG-TownCentre 和 PETS09-S2L2 视频序列的对比结果：

在 AVG-TownCentre 视频序列上，MOTA 提高大约百分之八；在 PETS09-S2L2 视频序列上，MOTA 提高大约百分之五。这两段视频序列包含很多以组为单位进行运动的行人目标。这证明 GST 模型先进行以组为单位的跟踪后再进行每一个行人的跟踪的策略是行之有效的。

相似度分析：在图 7-6 展示了在 MOT Benchmark 测试集中行为相似度矩阵中不同项对跟踪结果的贡献。公式（1）中的行为相似度测量包括：距离（公式（7-2）），时间（公式（7-3）），速度（公式（7-4））和方向（公式（7-5））四种相似度。实验中，每次通过每次去掉一种特征相似度，并计算其他相似度对模型的贡献。以此来衡量被去掉的特征相似度对总体相似度的贡献。此外，我们将速度项和方向项合成一项，$T_{v+o}=\alpha\mathbb{N}\left(\left\|v_i-v_j\right\|\right)+(1-\alpha)\mathbb{N}\left(\left\|v_i-v_j\right\|\right)$，其中$\mathbb{N}(\cdot)$表示在[0,1]区间归一化的分布。另外两项，距离和时间项，仍然保持与公式（7-2）和（7-3）相同。从图 7-6 中可以看出，新的相似度矩阵会降低整体的跟踪准确率 MOTA，这也证明了将速度和方向特征分成两项进行描述比合成一项在跟踪中有更好的效果。

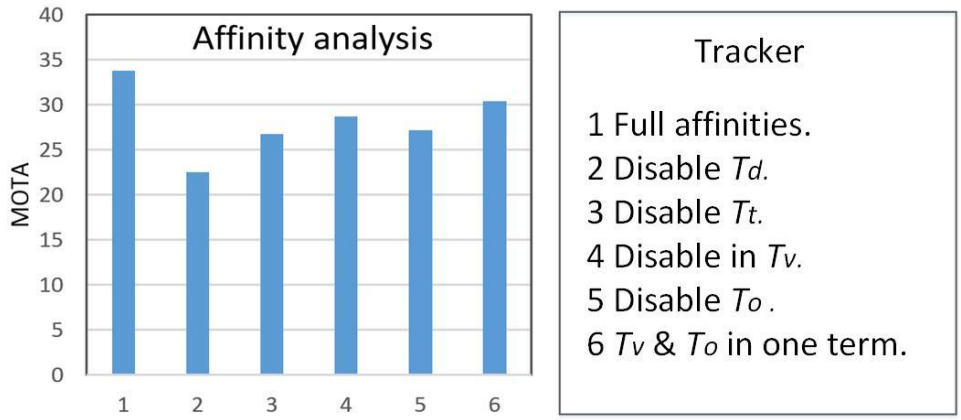


图 7-6 不同相似度组合的贡献比较

动态组跟踪分析：从实验结果可以发现，本章所提出的 GST 模型比其他对比实验在 MOTA 上有显著提高。这主要来源于动态拓扑图模型能很好的反应行人组的动态变化，能通过在线和离线不同的手段以合理的方式进行组的初始化、更新、合并和分裂。图 7-7 展示了在 PETS09-S2L2 测试视频序列中，对组的分裂和组内部的遮挡处理。在（b）和（c）中，GST 模型通过在组内加入虚拟点估计被遮挡的目标。并且在（d）中，当目标走出遮挡时，模型可以给目标相同的 ID。通过这种方式，GST 模型可以使得整个跟踪过程中有更少的目标丢失，所以跟踪到的完整轨迹也比其他对比实验多。

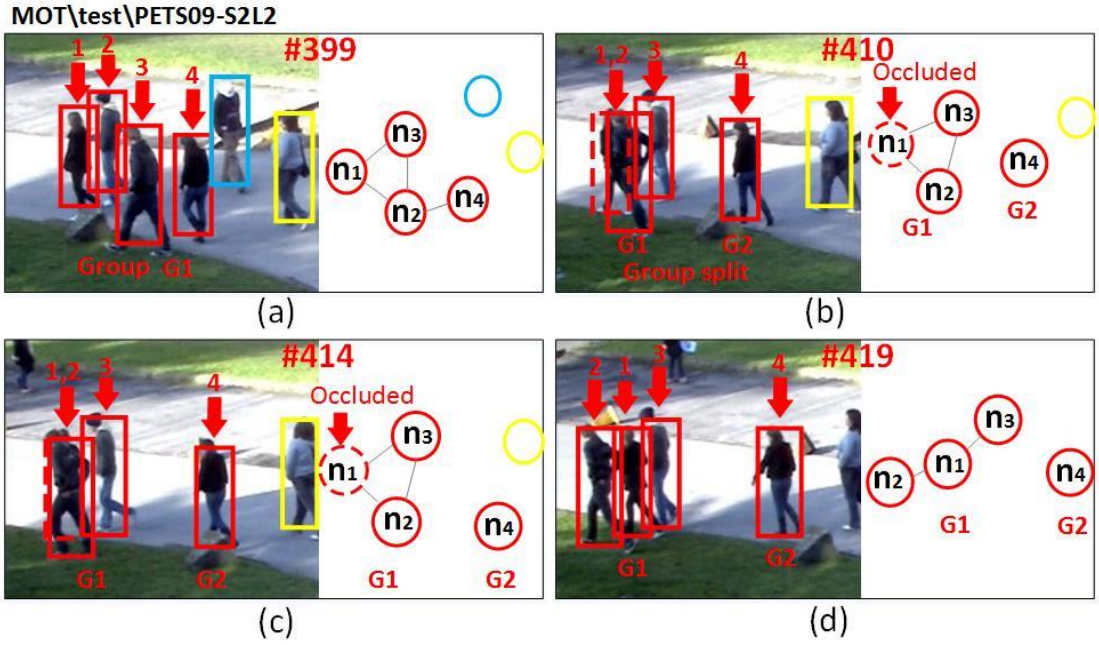


图 7-7 组的在线更新实例

7.6.2 RGB-D 数据集评估

为了验证本章提出的模型的普适性，实验在 RGB-D 数据集中同样完成和其他性能出色的多目标跟踪器的对比实验。在实验中使用 SDL-Campus 数据集进行模型训练，使用 $\tau = 0.34$ 作为边的阈值。同时使用另外三种 RGB-D 数据集作为测试集：Sync 数据集、SDL 数据集（将 SDL-Crossing 和 SDL-Garden 合成一个数据集）和 LIPD 数据集。如第三章所述，这三个数据集都录制于真实道路交通驾驶场景，每个数据集包括视频数据以及深度数据。每个数据集都包含多种目标：行人、自行车以及机动车。对于跟踪的挑战性在于 1）复杂的场景（包含遮挡和动态背景）；2）非静止的摄像头；和 3）目标以不同尺度出现在场景的任意位置。因为上述挑战，所以很多适用于监控场景的跟踪技术并不适用于真实交通场景，例如对场景进出口分析、背景建模等。

表 7-5 展示了本章提出的 GST 模型与其他对比方法的实验结果。其中基于 RGB 数据的多目标跟踪方法包括 SSP[75]、CEM[105]和 SegTrack[112]。基于 RGB-D 的多目标跟踪方法包括 DSA[115]；GST 是本章提出的拓扑图模型，但是并没有用到数据集中所提供的深度数据，而是将它作为 RGB 跟踪方法；GST+D 为完整的使用深度数据的拓扑图模型，实验中我们加入了真实深度数据作为目标观测的特征。从表 7-5 中，可以看出 GST+D 比其他基于 RGB 数据的多目标跟踪方法，在 Recall 上提高大约 12%，在 Prec.提高大约 11%。并且在 MT、PL、

ML、IDS 和 Frag.指标上，GST+D 方法也基本做到了性能最优。即使和基于深度的跟踪方法 DSA[115]相比，也在 Recall 和 Prec.上提高了大约 2%。图 7-8 中展示了跟踪过程中组的初始化、合并和分裂。同时，我们在实验中也发现，本章所提出的 GST+D 方法面对同一目标的多个检测窗口时，可以根据深度数据排除错误的短轨迹片段。

即使只用 RGB 方法进行横向比较，也可以发现本文提出的 GST 方法优于其他的基于 RGB 数据的多目标跟踪方法。分别在 Recall 和 Prec.指标上提高了大约 4%。这些对比结果可以证明本章所提出的基于动态拓扑图的多目标跟踪方法无论在 RGB 数据集或者 RGB-D 数据集，监控场景还是交通驾驶场景都是稳定可靠的。

表 7-5 RGB-D 数据集对比实验结果

数据集	**方法**	**Recall**	**Prec.**	**GT**	**MT**	**PL**	**ML**	**IDS**	**Frag.**
Sync	SSP[75]	69.6%	72.8%	66	9.0%	66.8%	24.2%	345	323
	CEM[105]	73.4%	78.3%	66	19.6%	60.8%	19.6%	**89**	125
	SegTrack[112]	76.2%	79.2%	66	25.8%	57.5%	16.7%	102	147
	DSA[115]	85.0%	89.7%	66	28.8%	57.5%	**13.7%**	90	108
	GST	83.9%	85.4%	66	19.6%	65.3%	15.1%	92	**103**
	GST+D	**87.5%**	**92.3%**	66	**31.8%**	**47.0%**	21.2%	118	134
SDL	SSP[75]	62.9%	70.5%	92	9.8%	59.8%	30.4%	168	189
	CEM[105]	70.4%	76.4%	92	19.6%	**55.4%**	25.0%	65	74
	SegTrack[112]	72.3%	77.8%	92	18.4%	70.8%	10.8%	55	71
	DSA[115]	82.4%	87.3%	92	25.0%	59.8%	15.2%	60	68
	GST	79.5%	85.1%	92	19.6%	69.6%	10.8%	61	**60**
	GST+D	**84.4%**	**89.0%**	92	**30.4%**	59.8%	**9.8%**	**58**	71
LIPD	SSP[75]	72.8%	76.4%	77	10.4%	**55.8%**	33.8%	324	219
	CEM[105]	78.4%	78.6%	77	19.5%	58.4%	22.1%	92	123
	SegTrack[112]	77.6%	80.2%	77	13.0%	67.5%	19.5%	75	118
	GST	82.3%	86.6%	77	20.2%	60.3%	19.5%	86	**62**
	GST+D	**86.7%**	**90.0%**	77	**33.8%**	55.8%	**10.4%**	**71**	65

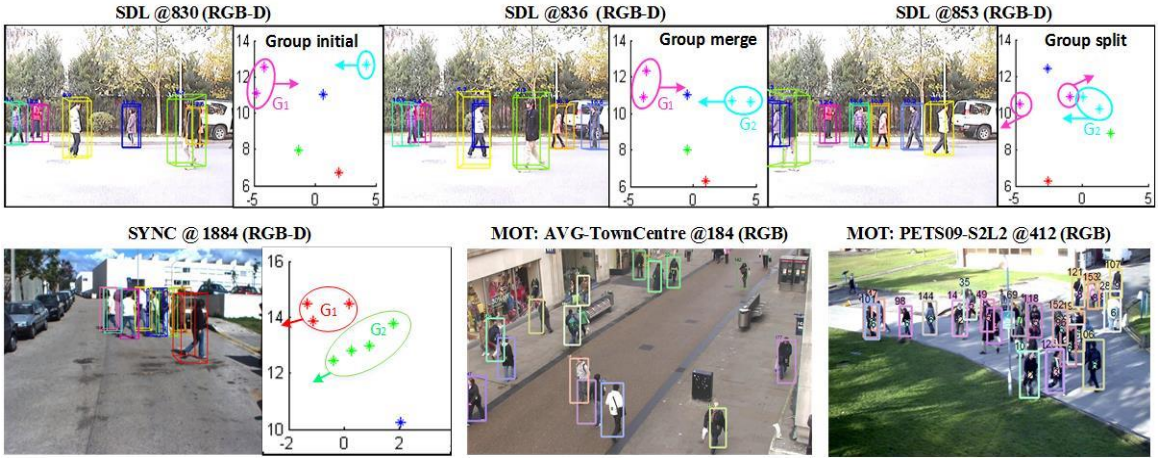


图 7-8 动态拓扑图模型在 RGB 和 RGB-D 数据集上的跟踪结果

7.7 本章小结

本章继承了以组划分的方式求解多目标跟踪问题的思路，并在此基础上进行扩展延伸，将动态图模型与多目标跟踪问题相结合，提出了基于拓扑图模型来构建行人组的运动特性。通过离线学习得到自然人群中最常见的拓扑结构，并通过在线学习进行组的初始化、更新、合并和分裂。最后，通过在线和离线相结合的方式进行组跟踪和组内行人身份识别。这种离线学习拓扑结构和在线进行组更新的方式有效的将行人组的动态运动信息和拓扑图结合起来完成对行人组的跟踪。实验中将动态拓扑图模型在 RGB 和 RGB-D 数据集上与多种类型的多目标跟踪方法进行比较。实验结果证明动态拓扑图模型不仅适用于 RGB-D 数据集，在 RGB 数据集上也同样表现优异。

第八章　总结与展望

多目标跟踪是计算机视觉领域的一个非常重要的问题，涉及到模式识别与智能系统、传感器技术、图像处理、统计学、机器学习等多个方面知识。经过近二三十年的深入研究和发展，多目标跟踪技术已在智能交通系统、智能监控系统、机器人导航、人机交互、生物医学研究等多方面得到广泛应用。近年来，随着 RGB-D 传感器技术的兴起，基于 RGB-D 数据的多目标跟踪技术有了突飞猛进的发展。本文针对基于 RGB-D 数据的多目标跟踪问题进行了深入研究，提出了基于 K 帧深度结构关联 DSA 的多目标实时跟踪方法、适用于交通驾驶领域的基于分层图模型 LGM 的全局多目标实时跟踪方法、基于拓扑能量模型 TEM 对行人进行组划分的全局多目标跟踪方法、以及基于动态拓扑图 GST 的全局多目标跟踪方法。在本章中，我们将对上述章节所讨论的内容进行总结，并对将来的基于 RGB-D 数据的多目标跟踪技术的发展方向进行展望。

8.1 全文总结

多目标跟踪的主要任务是对视频序列中的多个目标进行连续追踪，使其身份标识在序列中始终保持不变。虽然多目标跟踪问题得到了深入广泛的研究，也取得了很大的进步，但是在真实场景中，复杂的多目标跟踪技术距离实际应用还有一定差距。例如，很多目标跟踪应用场景要求在室内环境下进行来规避复杂的光照环境带来的视觉不稳定性。在室外场景下跟踪目标为行人的多目标跟踪系统中，频繁的目标之间互相遮挡、强烈变化的光照条件，复杂的动态背景，差异很大的图像分辨率等问题不光给基于视频分析多目标问题提出挑战，同时也为目标检测技术带来很大困难。本文利用日益兴起的 RGB-D 数据解决上述 RGB 多目标跟踪中出现的问题。本文搭建了 RGB-D 数据采集平台完成数据获取，提出了四种基于 RGB-D 数据进行多目标跟踪的方法。创新点总结如下：

1）在交通场景中，目标的运动背景和光照环境剧烈变化导致基于传统 RGB 数据的目标检测器与跟踪器出现严重漏检和误检。再加上因为行人的非刚性特征，导致其在场景内的运动变化复杂多样，而多行人之间又需要处理复杂的遮挡问题，使得跟踪也更加复杂。本文在第三章提出了基于深度结构关联的多目标跟踪模型（DSA），将场景中的多目标划分到不同的深度链状结构中进行三维分析。利用整数规划中的多维数据分配问题对多目标之间的数据关联进行建模。

在面对多目标跟踪过程中的遮挡问题时，链状结构利用深度值对目标匹配代价进行了重新加权，使得目标在场景的不同位置更具有区分性。经过 RGB-D 数据集验证，DSA 模型可以在交通场景的多目标跟踪问题上做到实时处理。

2）将多目标跟踪与图论中的图模型相结合的求解模式由来已久。本文的第五章提出利用基于 RGB-D 数据的分层图模型（LGM）求解多目标跟踪问题。本方法将传统的基于离散-连续的轨迹级（tracklet level）目标关联方式提升到深度层级（layer level）。模型利用深度数据构建目标在层内以及层间的图模型，利用目标之间的位置、运动、外形信息构成他们之间的关联相似度。此外，LGM 利用自身的分层关系，在层内利用加入虚拟点的策略解决目标之间的遮挡问题。

3）与前两种方法不同，本文的后两种方法在多目标跟踪模型中加入行人的社会属性。其基本出发点是：在自然场景中，人群中有高达 70%的行人以组的形式行走。本文第六章提出了基于拓扑能量的多目标跟踪模型。模型利用组内组外的行人运动相似度，进行能量域的建模，使得组内的行人相似度尽可能的高，组间的行人相似度尽可能的低，并以“拓扑能量最小化”的方式要求跟踪中进行模型求解。在拓扑的变化过程中，模型通过加入虚拟点的方式，将被遮挡的目标通过组内位置估计的方式进行有效定位，减少了多目标跟踪过程中的目标丢失问题。本方法还在目标的 RGB-D 特征上进行了创新性探索，提出了更加适合于多目标跟踪的行人 RGB-D 跟踪特征。模型在 SDL-Campus 进行模型训练后，在多种 RGB-D 数据集进行了测试。

4）在第七章中，本文提出了动态拓扑图模型（GST）及基于 GST 的多目标跟踪方法。本方法继承了第六章中以组为单位对行人进行研究的框架，但是提升了对行人空间拓扑关系的描述。将行人组模型与动态图相结合，将组内的行人看作是图中的节点元素，通过图中边的连接方式探索组内成员在行走过程中的运动的变化。模型通过离线与在线学习相结合的方式，对组的典型拓扑进行离线学习后，在线跟踪中与行人组进行拓扑结构匹配。在组的运动过程中，模型通过组的初始化、更新、合并和分裂的动态变化完成对行人组的跟踪。最后，利用线性规划的方式完成组内行人的身份确认，得到每个目标的完整轨迹。GST 模型在本章中不仅在 RGB-D 数据集上进行了实验验证，还在著名的 RGB 数据集 MOT Benchmark 中进行了测试，模型性能优于基于 RGB 的方法。

本文还介绍了作者搭建的 RGB-D 数据采集平台，描述了多种公开使用的 RGB-D 数据集。并将采集的数据集进行整理和公开，供多目标跟踪研究者使用。

8.2 未来工作展望

随着 RGB-D 传感器的日益普及，基于 RGB-D 数据的目标检测和跟踪方法在不断提高改善中。从著名的多目标跟踪方法评测平台 MOT Benchmark 上的跟踪结果中我们可以看出，即使现在最好的多目标跟踪检测方法的性能在 MOTA 指标项上也只能到达 50%左右。相对较低的跟踪性能制约了多目标跟踪技术在实际场景中的应用。未来可以从以下几个方面进行后续研究，完善基于 RGB-D 数据的多目标跟踪系统：

1）更有效地利用目标的外形和运动特征：现有的多目标跟踪方法中，无论是基于 RGB 数据还是基于 RGB-D 数据的方法在对目标特征的探索上都远远不够，研究思路主要是利用轨迹的各种关联方式寻找轨迹的相似度。即使是对外形特征进行提取或者在线学习，也是基于比较传统的表观特征。但是，在检测框内和多帧之内，前景与背景、时域与空域上的综合上下文信息并没有用到，光流信息等基于图像本身的信息也没有有效利用。在行人目标的密度比较大的拥挤场景中，RGB-D 数据也很难对所有目标的深度给出准确的数值。这时候上下文、时空、前景背景等特征信息就有很大作用。

2）实现基于 RGB-D 的目标检测器：本文第三章给出了基于 RGB-D 数据进行目标检测的简要框架。但是该框架是利用深度数据与 RGB 数据串联进行目标检测。也就是说，在深度空间内缩小检测的区域，图像上的场景内进行一定的过滤，只在有意义的深度值上进行基于图像的行人检测，而且可以使用一个比较低的阈值进行目标检测。这样做的优点是缩短了目标检测时间，也可以在一定意义上提高目标检测的准确率。但是，本文并没有提出一个有效的框架将 RGB-D 数据放在一个特征向量中，再进行目标检测。RGB 目标检测器在文中所使用的数据集上的检测性能并不是很高，这也导致了多目标跟踪的初始化并不准确。当使用这些数据集的人工标注目标位置作为多目标跟踪的输入时，多目标跟踪的准确率提升到 90%以上。因此，在以后工作中，如何将 RGB-D 数据进行统一特征提取，在一个分类器下进行检测，提高检测性能具有极高的研究价值。

3）完善 RGB-D 多目标跟踪评价方法：本文在第三章中介绍了基于 CLEAR-MOT 的多目标跟踪评价方法。但是该评价体系是建立在 RGB 数据集上提出的，图像上的评价指标是根据 2D 数据，计算跟踪框和 Groundtruth 框之间的重叠区域，在 3D 数据上则只计算投影区的像素距离。这样的评价指标很明显是建立在 RGB 图像数据的，当我们用 RGB-D 数据进行多目标跟踪时，它不能反映目标之间在 3D 空间中的真实位置信息。需要在以后的工作中考虑改

进这一评价体系，完成对基于 RGB-D 数据多目标跟踪更加准确的性能评测。

4）完善“Tracking-by-Detection”跟踪框架：无论是本文中的四种多目标跟踪方法，还是现有的绝大多数优秀多目标跟踪方法，都是采用“tracking-by-detection”（先检测再跟踪）这一经典的多目标跟踪框架。通常会先使用检测器进行目标检测，然后使用关联的方法进行目标关联。这样的框架造成的结果是跟踪的准确性严重依赖于检测器是否准确。一旦检测器的误检率漏检率高时，多目标跟踪性能就会大幅度下降。如果从应用场景来考虑，利用“先检测再跟踪”的方法会制约多目标跟踪的应用场景。往往这个框架只被视频目标检索这一应用场景所采纳，因为这个应用场景不用考虑跟踪的实时性。我们在计算算法的时间效率时，通常只考虑了多目标跟踪算法的时间。而对于一个系统来说，应该考虑的是整个系统处理算法的总时长，一旦加上目标检测的时间，系统难以做到实时处理。在后续的研究中，尤其是考虑不同场景的应用情况时，可以完善经典的“先检测再跟踪”的框架，适用于不同的多目标跟踪场景。

参考文献

[1] 申远. 基于小轨迹关联的多人跟踪方法研究[D]. 北京交通大学, 2014.

[2] 瑚成祥. 基于随机集理论的多目标跟踪方法[D]. 北京交通大学, 2014.

[3] Yu S. and Hauptmann A.. Harry Potter's Marauder's Map: Localizing and Tracking Multiple Persons-of-Interest by Nonnegative Discretization[C]. In Proceedings of the IEEE International Conference on Computer vision and Pattern Recognition, 2013:3714-3720.

[4] Chang J. and Fisher J. W.. Topology-Constrained Layered Tracking with Latent Flow[C]. In Proceedings of the IEEE International Conference on Computer Vision, 2013:1-8.

[5] Zhao X., Gong D. and Medioni G.. Tracking using Motion Patterns for Very Crowded Scenes[C]. In Proceedings of the European Conference on Computer Vision, 2012:315-328.

[6] Milan A., Schindler K. and Roth S.. Detection- and Trajectory-Level Exclusion in Multiple Object Tracking[C]. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2013:3682-3689.

[7] Andriluka M., Roth S., and Schiele B.. Monocular 3D Pose Estimation and Tracking by Detection[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010:623-630.

[8] Geiger A., Lenz P., and Urtasun R.. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:3354-3361.

[9] He S., Yang Q., Lau R. W. H., Wang J. and Yang M.. Visual Tracking via Locality Sensitive Histograms[C]. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2013:2427-2434.

[10] Wang N., Wang J. and Yeing D.. Online Robust Non-negative Dictionary Learning for Visual Tracking[C]. In Proceedings of the IEEE International Conference on Computer Vision, 2013: 1-8.

[11] Shi X., Ling H., Xing J. and Hu W.. Multi-target Tracking by Rank-1 Tensor Approximation[C]. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2013:2387-2394.

[12] Duan G., Ai H., Cao S. and Lao S.. Group Tracking: Exploring Mutual Relations for Multiple Object Tracking[C]. In Proceedings of the European Conference on Computer vision, 2012: 129-143.

[13] Cui J., Zha H., Zhao H. and Shibasaki R.. Multi-modal Tracking of People using Laser Scanners and Video Camera[J]. Image and Vision Computing, 2008, 26(2):240-252.

[14] 谷歌测试无人驾驶汽,行程 14 万英里[J/OL]. 新华网(2010-10-12). http://news.xinhuanet.com/world/2010-10/12/c_12648406.htm.

[15] 徐有春, 王荣本, 李兵. 世界智能车辆近况综述[J]. 汽车工程, 2001, 23(5):289-295.

[16] 高德芝, 段建民, 郑榜贵. 智能车辆环境感知传感器的应用现状[J]. 现代电子技术, 2008(19).

[17] 刘大学. 用于越野自主导航车的激光雷达与视觉融合方法研究[D]. 国防科技大学, 2009.

[18] Davison A. J., Murray D. W.. Mobile Robot Localisation Using Active Vision[C]. In Proceedings of the 5th European Conference on Computer Vision, 1998:385-392.

[19] Espiau B., Chaumette F., Rives P.. A New Approach to Visual Servering in Robotics[J]. Robotics and Automation, 1992, 8(3): 313-326.

[20] Dla F., Jute F., Ferri F., Ricens M.. Color Segmentation Based on a Light Reflection Model to Locate Citrus Fruits for Robust Harvesting[J]. Computer and Electronics in Agriculture. 1993, 9(1):53-70.

[21] 郭萍. 基于视频的人体行为分析[D]. 北京交通大学, 2012.

[22] Sarmad A., Min X., Thomas J. W., and Arnold D. B.. Differential trafficking of transport vesicles contributes to the localization of dendritic proteins[J]. Cell Reports, 2012, 2(1):89–100.

[23] Lou X. and Hamprecht F. A.. Structured learning for cell tracking[J]. In Advances in Neural Information Processing Systems, 2011, 24:1296–1304.

[24] Kausler B. X., Schiegg M., Andres B., Lindner M., Leitte H., Hufnagel L., Koethe U., and Hamprecht F. A.. A discrete chain graph model for 3d+t cell tracking with high misdetection robustness[C]. In Proceedings of the 12th European Conference on Computer Vision, 2012:144–157.

[25] Zhou B., Wang X., Tang X.. Understanding collective crowd behaviors: Learning a Mixture Model of Dynamic Pedestrian-agents[C]. In Proceedings of IEEE conference on computer vision and pattern recognition, 2012: 2871-2878.

[26] Moussaid M., Garnier S., Theraulaz G., et al. Collective Information Processing and Pattern Formation in Swarms, Flocks, and Crowds[J]. Topics in Cognitive Science, 2009, 1(3): 469-497.

[27] Milan A.. Energy Minimization for Multiple Object Tracking[D], Technology University of Darmstadt, 2014.

[28] Ross B. G., Jeff D., Trevor D., Jitendra M.. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation[J]. IEEE Transactions on Pattern Analysis

Maching Intelligence, 2016, 38(1): 142-158.

[29] Ross B. G.. Fast R-CNN[C]. In Proceedings of the IEEE International Conference on Computer Vision, 2015:1440-1448.

[30] Bar-Shalom Y., Tse E.. Tracking in a Cluttered Environment with Probabilistic Data Association[J]. IEEE Transactions on Automatic control, 1975, 11(9): 451-460.

[31] Bar-Shalom Y., Fortmann T. E.. Tracking and Data Association[M]. Academic Press, 1988.

[32] Reid D. B.. An Algorithm for Tracking Multiple Targets[J]. IEEE Transactions on Automatical Control, 1979, AC-24:843-854.

[33] 乔向东. 信息融合系统中目标跟踪技术研究[D]. 西安电子科技大学, 2003.

[34] Samuel S. B. and Robert P.. Design and Analysis of Modern Tracking Systems. Artech House, 1999. ISBN 9781580530064.

[35] Rachid D. and Faugeras O. D.. Tracking line segments[C]. In Proceedings of the First European Conference on Computer Vision, 1990: 259–268.

[36] Leung H., Hu Z. and Blanchette M.. Evaluation of Multiple Radar Target Trackers in Stressful Environments[J]. IEEE Transactions on Aerospace & Electronic Systems, 1999, 35(2): 663-673.

[37] Yakov B. and Jaffer A. G.. Adaptive Nonlinear Filtering for Tracking with Measurements of Uncertain Origin[C]. In Proceedings of the IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes, 1972:243 –247.

[38] Thomas E., Yaakov B., and Molly S.. Sonar Tracking of Multiple Targets Using Joint Probabilistic Data Association[J]. IEEE Journal of Oceanic Engineering, 1983, 8(3):173–184.

[39] Yaakov B., Kuo-Chu Chu Chang, and Henk Blom A. P.. Tracking of Splitting Targets in Clutter Using an Interacting Multiple Model Joint Probabilistic Data Association Filter[C]. In Proceedings of the 30th IEEE Conference on Decision and Control, 1991:2043–2048.

[40] Collins R. T., Liu Y., and Leordeanu M.. Online Selection of Discriminative Tracking Features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10):1631–1643.

[41] Stauffer C. and Grimson E.. Learning Patterns of Activity Using Real-Time Tracking[J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 747-757.

[42] Oh S., Russell S., and Sastry S.. Markov Chain Monte Carlo Data Association for General Multiple Target Tracking Problems[C]. In Proceedings of the IEEE Conference on Decision & Control, 2004:735-742.

[43] Oh S. and Sastry S.. A Polynomial-time Approximation Algorithm for Joint Probabilistic Data Association[C]. In Proceedings of American Control Conference, 2005: 1283-1288.

[44] Musicki D.. Joint Integrated Probabilistic Data Association[J]. IEEE Transactions on Automatic Control, 2004, AC-40(3): 1093–1099.

[45] Vermaak J., Godsill S., and Perez P.. Monte Carlo Filtering for Multi-target Tracking and Data Association[J]. In IEEE Transactions on Aerospace & Electronic Systems, 2005, 41(1):390–332.

[46] Blackman S.. Multiple Target Tracking with Radar Applications[M]. Norwood: Artech House, 1986.

[47] Blackman S.. Design and Analysis of Modern Tracking Systems[M]. Norwood: Artech House, 1999.

[48] Singer R. and Stein J.. An optimal Tracking Filter for Processing Sensor Data of Imprecisely Determined Origin in Surveillance Systems[C]. In Proceedings of the IEEE Conference on Decision & Control, 1971: 171–175.

[49] Reid D.. An Algorithm for Tracking Multiple Targets[J]. IEEE Transactions on Automatic Control, 1979, AC-24(6): 843–854.

[50] Isard M. and Blake A.. Condensation-Conditional Density Propagation for Visual Tracking[J]. International Journal of Computer Vision, 1998, 29(1):5–28.

[51] Vermaak J., Doucet A., and Pérez P.. Maintaining multimodality through mixture tracking[C]. In Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003:1110–1116.

[52] Okuma K., Taleghani A., Freitas O. D., Little J. J., and Lowe D. G.. A Boosted Particle Filter: Multitarget Detection and Tracking[C]. In Proceedings of the 8th European Conference on Computer Vision, 2004:28–39.

[53] Breitenstein M. D., Reichlin F., Leibe B., Esther K., and Gool L. V.. Robust Tracking-by-detection using a Detector Confidence Particle Filter[C]. In Proceedings of the 12fth IEEE International Conference on Computer Vision, 2009:1515-1522.

[54] Karlsson R., Gustafsson F.. Monte Carlo Data Association for Multiple Target Tracking[M]. Target Tracking: Algorithms and Applications (Ref. No. 2001/174), IEE, 2001, 1：13/1-13/5.

[55] Schulz D., Burgard W., Fox D. et al. Tracking Multiple Moving Targets with a Mobile Robot using Particle Filters and Statistical Data Association[C], In Proceedings of IEEE International Conference on Robotics and Automation, 2001:1665-1670.

[56] Vermaak J., Simon J. Godsill, Perez P.. Monte Carlo Filtering for Multi-target Tracking and Data Association[J], IEEE Transactions on Aerospace and Electronic Systems, Jan 2005, 41(1):309-332.

[57] Khan Z, Balch T., and Dellaert F.. MCMC-based Particle Filtering for Tracking a Variable

Number of Interacting Targets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (11):1805–1918.

[58] Khan Z., Balch T., and Dellaert F.. MCMC Data Association and Sparse Factorization Updating for Real time Multitarget Tracking with Merged and Multiple Measurements[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12):1960–1972.

[59] Yu Q., Medioni G. G., and Cohen I.. Multiple Target Tracking using Spatio-temporal Markov Chain Monte Carlo Data association[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007:1-8.

[60] Benfold B. and Reid I.. Stable Multi-Target Tracking in Real-time Surveillance Video[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011:3457-3464.

[61] Wojek C., Roth S., Schindler K., and Schiele B.. Monocular 3D scene modeling and inference: Understanding Multi-object Traffic Scenes[C]. In Proceedings of the 11th European Conference on Computer Vision, 2010:467–481.

[62] Choi W. and Savarese S.. Multiple Target Tracking in World Coordinate with Single, Minimally Calibrated Camera[C]. In Proceedings of the 11th European Conference on Computer Vision, 2010:553–567.

[63] Morefield C.. Application of 0-1 Integer Programming to Multitarget Tracking Problems[C]. IEEE Transactions on Automatic Control, 1977, 22(3):302–312.

[64] Dantzig G.. Linear Programming and Extensions[M]. Princeton University Press, August 1998.

[65] Karmarkar N.. A new polynomial-time algorithm for linear programming[C]. In Proceedings of the sixteenth annual ACM symposium on Theory of computing, STOC’84, 1984:302–311.

[66] Kaucic R., Perera A. G. A., Brooksby G., Kaufhold J. P., and Hoogs A.. A unified framework for tracking through occlusions and across sensor gaps[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 990–997.

[67] Kuhn, H.W.: The Hungarian method for the assignment problem[J]. Naval research logistics, 1955, 2(1-2):83-97.

[68] Wu B. and Nevatia R.. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors[J]. International Journal of Computer Vision, 2007, 75(2):247–266.

[69] Huang C., Wu B., and Nevatia R.. Robust object tracking by hierarchical association of detection responses[C]. In Proceedings of the Tenth European Conference on Computer Vision,

2008:788–801.

[70] Li Y., Huang C., and Nevatia R.. Learning to associate: Hybridboosted multi-target tracker for crowded scene[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009:2953-2960.

[71] Yang B. and Nevatia R.. An online learned CRF model for multi-target tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012: 2034–2041.

[72] Dehghan A., Assari S. M., Shah M.. GMMCP tracker: Globally optimal Generalized Maximum Multi Clique problem for multiple object tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015: 4091-4099.

[73] Brendel W., Amer M., and Todorovic S.. Multiobject tracking as maximum weight independent set[C], In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011:1273–1280.

[74] Zhang L., Li Y., and Nevatia R.. Global data association for multi-object tracking using network flows[C], In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008:1–8.

[75] Berclaz J., Fleuret F., Turetken E., and Fua P.. Multiple object tracking using k-shortest paths optimization[J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9):1806–1819.

[76] Pirsiavash H., Ramanan D., and Fowlkes C. C.. Globally-optimal greedy algorithms for tracking a variable number of objects[C], In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011:1201–1208.

[77] Zamir A. R., Dehghan A., and Shah M.. GMCP-Tracker: Global multiobject tracking using generalized minimum clique graphs[C], In Proceedings of the European Conference on Computer Vision, 2012:343–356.

[78] Wen L., Li W., Yan J., Lei Z., Yi D. and Li S. Z.. Multiple target tracking based on undirected hierarchical relation hypergraph[C], In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014:1282–1289.

[79] Jiang H., Fels S., and Little J. J.. A linear programming approach for multiple object tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007:1-8.

[80] Moussaid M., Perozo N., Garnier S., Helbing D., Theraulaz G.. The walking behaviour of pedestrian social groups and its impact on crowd dynamics[J]. PloS one 5(4) (2010) e10047.

[81] Singh, H., Arter, R., Dodd, L., Langston, P., Lester, E., Drury, J.. Modelling subgroup

behaviour in crowd dynamics dem simulation[J]. Applied Mathematical Modelling, 2009, 33(12): 4408-4423.

[82] Rodriguez M., Laptev I., Sivic J., and Audibert J. Y.. Density-aware person detection and tracking in crowds[C]. In Proceedings of the Thirteenth IEEE International Conference on Computer Vision, 2011:2423-2430.

[83] Lempitsky V. and Zisserman A.. Learning to count objects in images[J]. In Advances in Neural Information Processing Systems, 2010:1324–1332.

[84] Alahi A., Ramanathan V., Fei-Fei L.. Socially-aware large-scale crowd forecasting[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014:2211-2218.

[85] Helbing D., Molnar P.. Social force model for pedestrian dynamics[J]. Physical review E 1995, 51(5): 4282-24.

[86] Treuille A., Cooper S., and Popovic Z.. Continuum crowds[J]. In ACM Transactions on Graphics, 2006, 25:1160–1168.

[87] Pellegrini S., Ess A., Schindler K., and Gool L. V.. You'll never walk alone: Modeling social behavior for multi-target tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009:261–268.

[88] Ge W., Collins R.T., and Ruback B.. Vision-based analysis of small groups in pedestrian crowds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(5): 1003–1016, 2012.

[89] Qin Z. and Shelton C. R.. Improving multi-target tracking via social grouping[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1972-1978.

[90] Chen X., Qin Z., An L., Bhanu B.. An online learned elementary grouping model for multi-target tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014:1242-1249.

[91] Bazzani, L., Zanotto, M., Cristani, M., Murino, V.. Joint individual-group modeling for tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(4): 746-759.

[92] Bazzani L., Cristani M., Murino V.. Decentralized particle filter for joint individual-group tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012: 1886-1893.

[93] 高山，基于激光和视觉信息融合的行人跟踪研究[M]. 中国科学院大学, 2013.

[94] Lipd dataset in urban environment[OL], http://www2.isr.uc.pt/cpremebida/dataset.

[95] Oliveira L., Nunes U., Peixoto P., Silva M., Moita F.. Semantic fusion of laser and vision in pedestrian detection[J]. Pattern Recognition, 2010, 43(10): 3648-3659.

[96] MOT Benchmark: Multiple object tracking benchmark[OL], http://motchallenge.net.

[97] Ess A., Leibe B., Schindler K., and Gool L. V.. A mobile vision system for robust multi-person tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008:1-8.

[98] Ferryman J. and Shahrokni A.. PETS2009: Dataset and challenge[C]. In 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009.

[99] 武博. 融合激光深度和图像特征的快速行人检测研究[D]. 中国科学院研究生院, 2012.

[100] Vapnik V.. The nature of statistical learning theory[J]. Springer Verlag, 1995.

[101] http://www.cs.berkeley.edu/rgb/latent.

[102] Dollar P., Wojek C., Schiele B., and Perona P.. Pedestrian Detection: An Evaluation of the State of the Art[J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4):743-761.

[103] Bernardin K. and Stiefelhagen R.. Evaluating multiple object tracking performance: The CLEAR MOT metrics[J]. Image and Video Processing, 2008(1):1–10.

[104] Andriyenko A., Roth S., and Schindler K.. An analytical formulation of global occlusion reasoning for multi-target tracking[C], In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2011:1839–1846.

[105] Milan A., Roth S., Schindler K.. Continuous energy minimization for multitarget tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(1):58-72.

[106] Wu F.Y.. The potts model [J]. Reviews of modern physics, 1982, 54(1):235.

[107] Chen L-C., Fidler S., Yuille A. L., and Urtasun R.. Beat the mturkers: Automatic image labeling from weak 3d supervision[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014:3198-3205.

[108] Luber M., Spinello L., and Arras. K. O.. People tracking in rgb-d data with on-line boosted target models[C]. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011:3844–3849.

[109] Navarro-Serment L. E., Mertz C., and Hebert M.. Pedestrian detection and tracking using three-dimensional ladar data[J]. The International Journal of Robotics Research, 2010, 29(12):1516-1528.

[110] Han Z., Jiao J., Zhang B., Ye Q., and Liu J.. Visual object tracking via sample-based adaptive sparse representation (adasr)[J]. Pattern Recognition, 2011, 44(9):2170–2183.

[111] Collins R.T.. Multitarget data association with higher-order motion models[C]. In

Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1744-1751.

[112] Leal-Taixe L., Fenzi M., Kuznetsova A., Rosenhahn B., Savarese, S.. Learning an image-based motion context for multiple people tracking[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014:3542-3549.

[113] Milan A., Leal-Taix L., Schindler K., Reid I.. Joint tracking and segmentation of multiple targets[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2015:5397-5460.

[114] Xiang Y., Alahi A., Savarese S.. Learning to track: Online multi-object tracking by decision making[C]. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015:4705-4713.

[115] Gao S., Han Z., Doermann D., Jiao J.. Depth structure association for rgb-d multi-target tracking[C]. In Proceedings of the IEEE International Conference on Pattern Recognition, 2014:4152-4157.

个人简介及发表文章目录

- 2013 年 9 月至 2016 年 7 月　　中国科学院大学　　工学博士
- 2014 年 11 月至 2015 年 11 月　　德国弗劳恩霍夫协会　　访问博士
- 2010 年 9 月至 2013 年 7 月　　中国科学院大学　　专业硕士
- 2006 年 9 月至 2010 年 7 月　　南开大学　　工学学士

已发表论文：

- **Shan Gao,** Zhenjun Han, Ce Li, Qixiang Ye, and Jianbin Jiao, “Real-time Multi-pedestrian Tracking in Traffic Scenes via an RGB-D based Layered Graph Model”, IEEE Transactions on Intelligent transportation system, 16(5): 2814-2825 (2015). (SCI)
- **Shan Gao,** Zhenjun Han, D. Doermann, Jianbin Jiao, “Depth Structure Association for RGB-D Multi-Target Tracking”, IEEE 22th International Conference on Pattern Recognition, 2014:4152-4157. (EI)
- **Shan Gao,** Zhenjun Han, Qixiang Ye, and Jianbin Jiao, “Real-time Multi-pedestrian Tracking Based on Vision and Depth Information Fusion”, Proc. Pacific-Rim Conf. Multimedia, 2013:708-719. (EI)
- **Shan Gao**, Zhenjun Han, Yang Xu, Qixiang Ye, and Jianbin Jiao, “Real-time Pedestrian Learning-Tracking with Information Fusion”, Proc. International Conference Internet Multimedia Computing and Service, 2012:88-91. (EI)
- Ce Li, Zhenjun Han, Qixiang Ye, **Shan Gao**, Lijin Pang, Jianbin Jiao “Locality-constrained Sparse Reconstruction for Trajectory Classification”, In proceedings of the IEEE 22th International Conference on Pattern Recognition. 2014:2602-2606. (EI)

已投稿论文：

- **Shan Gao,** Q. Ye, A. Kuijper, Z. Han, and J. Jiao, “Multi-group tracking for multi-object tracking”, submitted to European Conference on Computer Vision, 2016.
- **Shan Gao,** Q. Ye, A. Kuijper, Z. Han, and J. Jiao, “Beyond Group: RGB-D Multi-Target Tracking using Minimal Topology-Energy-Variation”, submitted to IEEE Transactions on Image Processing.

专利与著作权：

- 基于信息融合的行人快速检测跟踪方法， 中华人民共和国专利权，专利受理号：201510071310.7.
- 无线上网身份认证系统，中华人民共和国软件著作权，登记号：2013SR010247.

致　谢

读博士前，自己考虑过很久读博士学位究竟值不值的问题。当完成博士研究工作写完毕业论文，再来回答这个问题时，我想说很值。不读博士永远不会明白认真做科研的价值和意义，不明白多年苦寒路的默默坚守，不明白面对问题深入思考的力量，不明白获得实验进展的那份喜悦，同时也不会明白自己真正想要什么生活。也许真的是在自己读了博士之后，看待问题才会多一份从容淡定与深邃。对我来说，这些比起博士学位本身更加受益终身。

本论文的研究工作是在焦建彬教授、叶齐祥教授、韩振军副教授的悉心指导下完成的。感谢导师焦建彬教授在我攻读博士学位期间从理论和实验方面给予的大量的、极其有益的指导，对我的每一个实验和每一篇论文都给予极大的支持。感谢叶齐祥教授对我实验中的每一步，论文中的每一句都精益求精。感谢韩振军副教授，对我学习生活中的无论是小灵感、小思路，还是小困难、小抱怨都关怀备至。三位恩师在科研上精益求精、在学术上认真严谨、生活中的认真负责都令我敬佩。

本论文的部分研究工作在德国弗劳恩霍夫协会计算机图形学研究中心（Fraunhofer Institute IGD，Darmstadt，Germany）完成。感谢 Arjan Kuijper 教授以及 GRIS 组上的各位老师和博士，让我在德国交流访问的一年时光里，明白了德国人严谨认真的科研工作态度和劳逸结合的生活方式。

感谢中国科学院大学模式识别与智能系统实验室秦飞副教授，以及师兄师姐师弟师妹。我们一起科研、一起奋斗、一起成长、一起锻炼、一起欢笑、一起流泪。没有与你们的并肩作战，我很难想象怎样在实验室里一周度过 60 个小时。

感谢我的父母、亲戚挚友，你们给我无私的爱和无条件的支持，永远是我前进中最坚强的后盾和依靠，在别人关心我飞得高不高的时候，是你们关心我飞得累不累，愿你们永远幸福安康。

感谢参加开题、中期和毕业答辩的各位指导老师专家，你们丰富的经验和细致的指导对论文方向和研究进度的指点给整个研究工作带来了巨大的帮助。

最后，感谢自己。十年后的我，会感谢今天没有放弃的自己。

高　山

2016 年 4 月