工学博士学位论文

基于序列蒙特卡洛滤波算法的视觉目标跟踪

王建宇

哈尔滨工业大学

2006 年 2 月

国内图书分类号：TP391.41
国际图书分类号：681.39

工学博士学位论文

基于序列蒙特卡洛滤波算法的视觉目标跟踪

博 士 研究生：王建宇
导　　师：高文教授
申 请 学 位：工学博士
学 科、专 业：计算机应用技术
所 在 单 位：计算机科学与技术学院
答 辩 日 期：2006 年 2 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.41
U.D.C: 681.39

Dissertation for the Doctoral Degree in Engineering

SEQUENTIAL MONTE CARLO FILTERING BASED VISUAL TRACKING

Candidate: **Wang Jianyu**
Supervisor: **Prof. Gao Wen**
Academic Degree Applied for: **Doctor of Engineering**
Speciality: **Computer Application**
Affiliation: **School of Computer Science and Technology**
Date of Defence: **February, 2006**
Degree-Conferring-Institution: **Harbin Institute of Technology**

摘要

基于图像序列的目标跟踪作为计算机视觉领域的一个核心问题，得到了广泛而深入的研究。视觉跟踪研究的主要目的是模仿生理视觉系统的运动感知功能，赋予机器辨识图像序列中物体运动及其相互关系的能力，为图像序列理解提供重要途径。视觉跟踪技术具有广阔的应用前景，如视频监控、视频分析、视频检索、基于视频的运动分析和合成、基于运动信息的身份识别等。经过四十多年尤其是近十年的不懈研究，视觉跟踪技术取得了长足的进步，但实践表明一般意义上的视觉跟踪技术还远未成熟，要开发出真正鲁棒、实用的视觉跟踪应用系统还需要更为鲁棒的核心算法并需要解决大量的算法实现问题。

本文在序列蒙特卡洛滤波算法的框架下，以人脸和人体跟踪为研究对象，针对其中涉及的关键问题进行了探讨，研究了开发鲁棒实用的视觉跟踪系统所需要的核心技术和关键问题解决方案，重点探讨了目标表观建模，复杂运动的建模和推断，融合低端模型和高端模型的运动描述方法等几个关键问题。具体的研究内容如下：

1) 提出了可区分性目标表观模型的自适应建模和更新算法。表观建模是视觉跟踪算法性能的决定性因素之一。实践表明：图像特征选择和基于图像特征的目标表观描述模型从根本上决定了算法的鲁棒性和计算复杂性。虽然这一问题得到了领域内学者的极大重视和不懈努力，其仍是阻碍视觉跟踪技术进入实际应用的最困难问题之一。本论文中提出了一种自适应目标表观建模和更新算法。该算法在动态建模过程中不仅考虑目标表观信息，同时对目标所处环境中的背景信息进行考察，从而可对目标/背景的差异信息进行有效建模，在根本上保证了模型具有从变化的背景中区分前景的能力。实验结果表明，相比于目前最具代表性的跟踪算法之一 Mean Shift，提出的算法在公开的测试序列上取得了更好的跟踪结果。
2) 提出了集成多运动模型的复杂运动建模和推断算法。由于计算复杂性的限制，视觉跟踪算法通常基于局部搜索的策略确定目标的运动状态。所以，根据目标运动规律确定其以较高概率出现的局部区域成为算法效率的关键因素之一。如何针对复杂运动描述目标运动规律，是很多现实跟踪问题的效率瓶颈所在。本论文提出了采用多运动模型对目标复杂运动

进行建模和估计的基本框架。在此基础上，针对具有多种运动模式和具有高维运动状态的两类常见的复杂运动模式，将多模型的估计框架融入序列蒙特卡洛滤波算法中，从而针对两类复杂运动问题提出了标准序列蒙特卡洛滤波算法的两个改进：基于多模型切换和基于多模型协同的序列蒙特卡洛滤波算法。在人脸跟踪和面部表情估计问题上分别验证了改进的算法。实验结果表明，相对比于标准序列蒙特卡洛滤波算法，在计算复杂度降低的同时，改进的算法得到了更高的跟踪精度。

3) 提出了融合光流和特定模型的面部特征点跟踪算法。面部特征点跟踪是基于特征点的运动感知研究的典型应用，也是基于特征点运动感知任务中的困难问题。现有的面部特征点跟踪方法主要可以分为基于特定描述模型(以下简称模型)和基于光流的方法。本论文在序列蒙特卡洛滤波算法的框架下融合了基于光流和基于特定模型的方法来解决面部特征点跟踪问题，以克服单独采用一类方法的不足，从而达到鲁棒跟踪面部特征点的目的。在基于尺度空间理论改进 KLT 光流算法的基础上，以光流估计结果约束基于模型的形变特征点估计的起始搜索位置，大大加速了序列蒙特卡洛滤波算法的搜索过程。对于估计结果中存在的跟踪误差，进一步采用特征点运动轨迹的子空间约束来迭代求精跟踪结果。相比于广泛使用的 KLT 特征点跟踪算法，实验结果证实了提出算法的有效性。

本论文的三个主要创新点，分别对应视觉跟踪中的两个关键问题：目标表观的建模和目标运动的描述。其中创新点一提出了动态建模前景/背景差异的理念，使其不同于已有的大部分视觉跟踪算法。创新点二和三则分别从显式地采用特定模型描述目标运动和隐式地采用离散特征点描述目标运动方面进行了创新尝试。三种方法互为补充，并适合不同的应用情境。

关键词 视觉目标跟踪；序列蒙特卡洛滤波算法；在线特征选择；运动建模；基于光流的跟踪

Abstract

Image sequence based object tracking is a fundamental problem for computer vision research and has been widely studied. The main goal of visual tracking is to imitate the motion sensibility of physical visual system, empower the machine with the ability of perceiving the object motion and their relations in the scene and provide an important way for image sequence understanding. Visual tracking technique has many applications, such as video surveillance, video analysis, video indexing, video based motion analysis and synthesis, motion-based human identification. After more than 40 years' development, visual tracking technique has made great progress especially in the past ten years. However, practical experience has shown that visual tracking technologies are currently far from mature. A great number of challenges need to be solved before one can implement a robust visual tracking system for commercial applications.

Under the framework of sequential Monte Carlo filtering algorithm, this thesis try to get insights on some key issues in visual tracking with application scenarios on face and human tracking. Some important technologies and solutions are studied which are necessary for robust and practical tracking systems, especially concentrate on how to model the object appearance variation, how to model and estimate the object complex dynamics, how to combine low level and high level motion estimation methods to enhance tracker robust and efficiency. The main contributions of this thesis can be concluded as follows.

1) An algorithm for online modeling and adapting discriminative object appearance model is proposed. How to model the appearance of the object is one of the key factors determining the performance of a visual tracking system. Practical experience shows that feature selection strategy and how to model features fundamentally determine the robustness and computational complexity of a tracker. This problem has been extensively studied by many researchers. However, it is still one of the biggest difficulties to prevent the tracking technique into practical applications. An adaptive appearance model with online updating process is proposed. The algorithm considers both object appearance and its relevant background when constructing object

model. The constructed model encodes the difference between the object and background dynamically. Therefore, during the tracking process, the discriminability of the updated model is guaranteed basically. Compared with one of the state-of-the-art tracking algorithms, Mean Shift, experimental results show that our algorithm performs better on publicly available test sequences.

2) An algorithm for modeling and estimating complex object dynamics by integrating multiple models is proposed. Due to the limitation of computational resource, tracking algorithms are almost based on local search methods to find object motion state. Therefore, predicting the object future positions according to its motion trajectory is one of the key factors to determine the algorithm efficacy. How to model these complex dynamics is the bottleneck of these kinds of tracking tasks. This thesis proposes a framework to model and estimate complex motion by incorporating multiple motion models. Based on the proposed framework and aim to solving two kinds of complex motions, two new variations of sequential Monte Carlo filter are proposed, termed as multi-model switching sequential Monte Carlo filter and multi-model cooperation sequential Monte Carlo filter respectively, by combining the proposed framework with the standard sequential Monte Carlo filter. Experimental results show that the proposed algorithms perform better than standard sequential Monte Carlo filter and simultaneously lower the computational burden.
3) A facial feature tracking algorithm by combining optical flow and specific description model is proposed. Facial feature tracking is one of the classical applications of local feature based motion perception. Tracking facial feature is also a challenging problem. Existing facial feature tracking algorithms can be categorized into two kinds: specific description model based and optical flow based methods. The thesis proposes to combine the optical flow based and specific model based methods under the sequential Monte Carlo filter to solve the facial feature tracking problem. The classical KLT feature tracker is improved with the scale space theory. Based on the fine initial conditions constrained by improved KLT tracker, mouth description model is employ for those deformable features and the searching process of sequential Monte

Carlo filtering algorithm can be accelerated significantly. Considering remained tracking errors, subspace constraint on motion trajectories of all features is furtherly adopted to iteratively refine tracking results. Compared with original KLT tracker, experimental results confirm the effectiveness of the proposed method on facial feature tracking task.

Three proposed novel ideas in the thesis are try to solve two basic probelems in visul tracking research: object appearance modeling and object dynamics description. The first idea proposes to dynamically modeling the difference between foreground/background, which make it different from most exsiting visual tracking algorithms. The second and third ideas try to describe object dynamic explicitly from the motion model or implicilty from discrite feature points' motions respectively. Therefore, the three proposed algorithms are somewhat complementary and can be choosen for different application scenarios.

Keywords Visual object tracking, sequential Monte Carlo filtering algorithm, online feature selection, motion modeling，optical flow based tracking

目录

Contents

图表目录

List of Figures and Tables

第1章 绪论

1.1 研究背景

作为视觉的一项基本功能，人类视觉系统可以敏锐地感知所处场景中的物体运动。除非在很困难的情况下，如辨识乒乓球的运动轨迹，对生理视觉系统而言，运动感知已经成为一种自然而然的无意识视觉行为。然而对机器来说，迄今为止，赋予其感知运动的能力仍然是一项极大的挑战和尚未解决的问题，并作为视觉跟踪领域的主要研究目标而具有广泛的商业应用前景。

随着硬件成本的降低和制造工艺的改进，摄像机等成像设备的性价比快速提高，并迅速普及到生产生活中的方方面面，在安全保障等方面发挥着越来越重要的作用。但是，当前的视频监控系统只能简单地记录事情发生的映像。自“911”事件以来，恐怖主义的猖獗使国家安全危机凸显，赋予视觉监控系统能够理解所拍摄场景的能力成为日益迫切的需求。在 2000 年，美国 DARPA 启动了 HumanID 项目，其目标是试图通过视觉方法远距离地辨认人的身份信息，而步态等运动信息在其中扮演了重要角色。除视觉监控之外，视觉跟踪技术还可以在基于视频的运动分析等领域发挥重要作用，如可以采用跟踪技术分析运动员的技术动作与理想动作之间的差别。另外运动分析技术还可以用来指导真实感运动的合成技术，从而为虚拟现实技术的广泛应用提供技术支持。

视觉跟踪技术还经常作为很多系统的重要组成部分。在视频分析领域，系统输入是一组在时间轴上具有关联的图像集合。如果计算机能够将图像集合分解为事件的集合，从而得到视频的高层描述和语义特征，将对视频检索等领域产生重要影响。跟踪技术可以通过寻找在时间轴上相同物体的对应，不仅将图像集合分解为以物体为中心的对象集合，同时能够得到物体的运动信息，从而为场景语义的提取提供了重要支持。近些年提出的视频压缩领域的国际标准 MPEG-4 和 MPEG-7 将基于对象的编码和描述列为视频处理的研究方向。而采用跟踪技术的运动分层技术可以对这一策略提供有效支持。另外，在机器人自动导航、智能人机交互、视频分割、军事目标定位等领域，都迫切需要成熟稳定的视觉跟踪技术。

除了上述应用方面的需求之外，研究视觉跟踪技术对理解人类视觉的机

制和探讨人工智能的实现手段也具有重要的指导意义。

1.2 跟踪研究中面临的主要问题

基于视觉的目标跟踪可以定义为根据时间轴连续的图像集合推断目标运动状态的问题[1]。目标运动状态是问题相关的，可以是目标的 2D 位置、3D 位置、大小、运动方向以及更为复杂的运动姿态等，而目标表观则指包含了目标图像投影的可见光图像。

对目标跟踪来说，首先需要面对的问题是能够在图像中将目标的投影完整而清晰地分割出来。但是，从目标的图像表观推断目标的状态，首先就是一个具有病态解的问题。当现实世界中的三维物体投影到二维图像平面时，深度信息丢失，而深度信息是将目标从纷繁芜杂的背景图像中分割出来的重要依据之一。另外，相对于理论研究中的成像模型，现实世界中的投影矩阵异常复杂，不仅受着诸多随机因素的影响，同时在求解的过程中必须考虑噪声和误差的作用。例如，对图像中的一个像素，其成像过程要受到物体本身的材质、外界的光源、摄像机的视角、透明度和遮挡关系等诸多因素的影响，并且像素值与诸多影响因素的关联关系并非都可采用线性函数表达。虽然外界的变化因素是繁杂的，但反映到图像上，却仅仅是像素值的简单变化，从而掩盖了造成变化的众多原因。对视觉跟踪算法来说，其输入往往是以万为数量级的像素在时间轴上的变化，算法需要从如此高维的输入中选择信息进行推理，同时需要抵抗噪声和不确定性因素的干扰。

视觉跟踪问题的研究经过了几十年的积累和诸多学者的不懈努力之后，取得了不俗的成果并出现了一些进入某些特定应用场合的商业系统，但总体来说，对大规模应用和达到工业标准，仍面临重重困难：

1) **应用场合中具有复杂的背景变化：**复杂背景是引起跟踪失效的最重要因素之一。背景本身是不断变化的因素。另外，出于实时计算的考虑，要求刻画目标表观所采用的图像特征及其模型必须简单，从而有可能造成背景在所选择的图像特征上与目标相似，在根本上使算法无法区分目标和背景，造成跟踪的失效。例如，选择颜色直方图作为目标表观刻画方法，如果背景中有区域和目标在颜色分布上相近时，则存在算法被背景“吸引”从而丢失目标的危险。需要指出，所谓“复杂背景”是与所选择的图像特征和应用场合紧密相关的。
2) **目标具有复杂的运动模式：**基于计算性和鲁棒性的考虑，跟踪算法主要

采用局部搜索的策略，即只关心目标以较大概率出现的区域。但当目标具有复杂的运动模式时，如目标的运动速度或者方向突然发生改变，未对特殊情况进行建模的简单预测机制容易失效，造成算法在目标并不存在的区域中进行目标的搜索，从而造成目标丢失。

3) **场景中物体对目标的遮挡和目标的自遮挡(如图 1-1(c)、(e)、(f)所示)：**目标遮挡情况更多出现在多目标跟踪的问题中。遮挡是造成目标的图像表观突然变化的重要原因之一，并且使图像表观变化具有突然性和不连续性等非线性特征，从而容易引起既有目标表观模型的失效。

4) **光照变化：**如图 1-2 所示，相同的目标状态在不同的光照条件下具有差异明显的目标表观。尤其当目标处于室外环境中，复杂的光照变化如阴影等会引起目标图像表观的剧烈变化。光照影响是很多实际应用中阻碍系统性能提升的瓶颈因素问题。

5) **视点变化和目标的非刚性形变(如图 1-1(b)、(d)所示)：**该两因素会造成目标图像表观的非线性变化，而对非线性变化的建模本身就是困难的问题。

图 1-1 视角，尺度、遮挡等因素变化而引起的目标表观变化示例

Fig.1-1 Examples of observation variations due to viewpoint, scale, occlusion changed

上面的诸多困难因素中，在根本上可以归结为两类问题，首先是如何建立和维护正确的目标表观模型的问题。前面描述的困难 1、3、4 和 5 中，都需要鲁棒准确的目标表观建模和更新过程。而现实世界中的跟踪问题，由于

目标(比如人耳，头发等)形状的不规则和遮挡、光照等外界不确定性因素的客观存在，能够获取的目标表观通常是不完整或者含有噪声信息的，采用这样的数据更新目标模型，很快将使目标模型与实际的目标表观发生偏离，累积的误差最终造成模型不能很好描述不断变化的目标表观，从而引起著名的“漂移”问题，造成跟踪的失效。所以，如何正确有效地提取目标的图像表观，建立和维护目标模型以反映目标表观的变化成为解决困难 1、3、4 和 5 的关键；其次是如何描述目标运动以作出有效预测的问题。这通常需要解决如何有效地融合目标运动规律的先验知识并且同时尽量降低运动模型的复杂度的两难问题。

图 1-2 光照对目标表观的影响示例，上行图片在室内环境下采集，下行图片在室外环境下采集(NIST-NSWC-USF 数据库)

Fig.1-2 Illustration the impact of illumination variations on target appearance, top row and bottom row are recorded in indoor environment and outdoor environment respectively (from database NIST-NSWC-USF)

上述两类问题，体现到序列蒙特卡洛滤波算法的实现中，就是如何定义算法中状态观测关联模型和状态转移模型的问题。状态观测关联模型不仅定义了目标物体的状态模型和表观模型，还定义了两模型之间的关系，并将状态和观测之间的关系以条件概率分布的形式进行了约定，从而能够使算法系

统地处理不确定性问题和多信息融合问题。状态转移模型则编码了目标物体的运动状态转移规律。因此，序列蒙特卡洛滤波算法为解决视觉跟踪中现存的困难问题提供了系统有效的理论框架。但在理论框架的基础上，根据具体问题来有效定义和实例化模型才是算法成败的关键。所以，本论文将人脸和人体跟踪问题纳入到基于序列蒙特卡洛滤波的理论框架中来系统地研究，并针对该具体问题进行了序列蒙特卡洛滤波算法的改进。

1.3 视觉目标跟踪研究概述

在视觉跟踪研究领域，众多学者发表了大量研究成果并产出了一些著名的视觉跟踪系统[2-7]。根据构成视觉跟踪系统的关键技术进行分类，本节分别对视觉跟踪研究中的代表性工作进行简单总结和评论，而与论文密切相关的前人工作将在相关章节中进行集中评述。

视觉跟踪系统主要由四项关键技术支撑：目标表观的建模，目标表观的提取，数据关联方法和滤波方法[8]。需要指出的是，跟踪问题在雷达，声纳等军事领域具有更长远的研究历史和更丰富的研究成果[9]，视觉跟踪领域中滤波框架和数据关联等理论方法大多借鉴自上述领域中。相对于数据关联和滤波框架更多地根植于自动控制和信号处理领域，目标表观的建模和提取则与计算机视觉研究中的图像表示、图像匹配等技术有更密切的关联，但由于目标运动的时间关联特性，又使其与静态图像匹配等领域的问题不同，形成单独的研究子领域。

总体上说，目标表观的建模和提取更多地是视觉跟踪研究中的独特问题，也是主要的算法性能决定因素。数据关联问题更多地用来解决多目标跟踪中目标/图像表观之间的对应问题，而滤波算法则提供了解决跟踪问题的理论框架。所以本节的前人工作综述更多偏重于视觉跟踪独特问题的总结，即目标表观的建模和提取。同时也对数据关联方法和滤波框架进行了简单总结。

1.3.1 目标表观的建模和提取

视觉跟踪中，目标表观表现为某图像区域中像素的集合。目标表观受自身运动(如旋转，形变、平移等)，成像参数变化(如观察角度和距离等)和外部自然条件变化(如光照等)等因素的影响，并随着上述因素的变化而改变。并且随着目标的运动，目标所处的背景也不断变化。所以，视觉跟踪问题主

要处理的是时变信号的分类问题：即将时变的前景(目标)从时变的背景中正确地，连续地分离出来。

如何排除种种不确定和变化因素的影响，在“未来”的场景映像中推断出目标在时间轴上的对应关系，最主要的倚重因素就是正确地建立目标表观的刻画模型，从而对目标和背景的分类提供根本的保证。表观的提取过程则主要指在目标表观模型的基础上找到相邻帧间目标表观的对应，提取方法与所采用的目标模型是密切相关的。

本小节按照目标表观的建模及其提取策略的不同对前人的工作进行分类总结。

1.3.1.1 基于颜色分布的方法

图像区域颜色分布具有对旋转、小幅度仿射变换等变化和图像噪声较鲁棒的优点，并且计算简单，符合实时处理的计算要求。但是，颜色分布在对目标表观的表面颜色进行统计的同时，丢失了其颜色分布的几何特性，无法根据图像区域的颜色直方图重建图像区域，从而使完全不同的图像区域可能具有相同的颜色分布，所以，其对图像区域的刻画能力是偏弱的，只能编码目标相关的部分图像信息。

但颜色分布仍然是视觉跟踪中应用最广泛和最成功的目标刻画方法之一[10-12]。通过将人的头部用椭圆近似，Birchfield 提出了采用直方图结合椭圆圆周梯度进行头部跟踪的方法[13]。作者认为，颜色直方图刻画了椭圆内部的纹理信息，椭圆圆周梯度则表示了椭圆轮廓的信息，根据集合论的观点，二者在目标刻画上具有互补性，即当一种表示方法失效时，另一种方法往往能起到矫正的作用。该方法计算简单，满足实时跟踪的要求。所采用的颜色空间为 RGB 空间的变形，三个颜色分量分别为色度分量 G-R 和 B-G，光照强度分量(R+G+B)/3。

相对于 RGB 颜色空间，Bradski 认为 HSV 色彩空间中的 H 分量(色调分量)能够更好地区分人的肤色和其它自然色[14]。算法将在 H 分量上肤色隶属度的分布用直方图概率密度估计的形式求取，然后对整幅图像中每个像素的肤色隶属度进行计算，最后采用 Camshift 算法找到一个与人脸形状相似的具有最大概率可能的图像区域作为人脸跟踪结果。

为了弥补直方图的刻画能力不足的弱点，Comaniciu 采用了空间加权直方图。Enpanolkov 核函数使靠近图像区域中心的像素具有进行较大的权值从而部分编码了像素颜色分布的空间信息，使加权直方图相对于标准直方图具有了更好的目标定位精度[15]。该算法根据直方图是密度分布估计的一

种，采用 Bhttacharyya 距离来度量模型和图像区域直方图之间的相似度，并引入 mean-shift 梯度上升算法来迭代匹配模型到正确的图像区域。不同于局部穷举搜索的方式，基于梯度上升的 mean-shift 方法降低了搜索的计算复杂性。为了满足 mean-shift 算法的应用条件，算法要求相邻帧间的目标表观具有重叠的区域，该缺陷可以采用基于滤波的预测框架加以弥补。

与上述方法采用单一色彩空间计算直方图的方法不同，自适应地选择合适的色彩空间的方法首先在[16]的工作中得到了体现。在该工作中，作者注意到没有任何色彩空间在所有人脸跟踪任务中总是优于其它色彩空间的。在算法实现中，预定义了多个候选色彩空间，并以对被跟踪目标和背景的区分度作为选择色彩空间的标准，从而通过对背景的评估，自适应地从多个色彩空间中选择一个构建面部直方图来保证跟踪算法的鲁棒性。

自适应地直方图特征选择思想在[17]的工作中被用来进行一般图像区域的跟踪。在算法中，作者通过对 RGB 颜色空间中的三个色彩分量进行加权 $w_1R+w_2G+w_3B$ 得到不同包含直方图的特征空间，其中，$\{w_i \mid i=1,2,3\} \in [1,5]$，共有 49 种不同的特征空间可供选择。与[16]中的自适应选择标准类似，算法定义了似然比来度量特定直方图对前景和背景的区分能力，作为特征空间选择的标准。

国内方面，刘明宝博士在[18]中提出了一种在复杂背景中实时跟踪人脸自由运动的方法。在人脸直方图的表达上，将图像的 RGB 空间变换到色度空间,然后利用最大能量坐标和矩表征色度空间的直方图分布。该方法中的运动模型则结合了运动检测与运动预测，以减少搜索区域，提高检测速度。

在[19]中，姚鸿勋教授分析得出了关于色度与彩色坐标系的关联关系,建立了肤色和唇色色系的坐标变换映射，从而得到了对象姿态、背景鲁棒的人脸面部定位与跟踪的方法。通过进一步地在线学习,还可以去除光照条件变化的影响及摄像设备参数变化的影响。

1.3.1.2 基于模板的方法

模板可以有效编码目标表观的信息，而且通常仅依靠较少的初始图像数量就可以进行构造[20, 21]。但是由于模板通常带有全局性的信息，并以像素为基本单位，难以有效地得到更新来反映目标表观的实时变化，所以基于模板的跟踪方法更适合于短时跟踪任务，而在长时序的跟踪任务中，基于模板的方法容易产生目标“漂移”现象[22]。

Frey 在[23]中对传统的模板方法做了改进。对目标区域的每一像素，算法根据一段时间内的目标表观求取其变化范围。该模板对噪声具有更好的鲁棒性，但是增加了在训练集合上对图像像素的灰度变化进行预学习的过程。在随后的工作中，Jojic 针对场景中存在多物体的情况，建立了层次性的基于模板的目标模型。由于该生成式模型中包含了叠加性高斯噪声和阿尔法通道，具有处理多目标深度信息和透明性遮挡等情况的能力[24]。模型的学习采用了效率优于最大后验概率方法的迭代算法。在迭代过程中，在不同运动层间传递关于灰度的概率分布信息，在层内水平地传递多目标的标识信息。但在算法中，运动层的个数和目标物体的数目都是假设预知，从而限制了算法的实用性。

Rucklidge 在[25]中提出了一种有效率地定位灰度模式的方法。算法基于块匹配策略，采用平方差之和(Sum of Squared Differences, SSD)作为图像匹配度准则。算法将特定模式匹配到图像区域的函数变换空间划分为若干子空间。在每个子空间中，考虑所有可能的变换形式，对变换后的特定模式和图像区域之间的差值根据经验设定阈值，从而忽略那些最佳匹配值亦大于阈值的子空间来降低计算量。对最佳匹配值低于阈值的子空间则进行进一步划分，并采用深度优先的搜索策略进行求解。

Olson 等人认为广泛使用的模板匹配准则 SSD 对所有像素等同看待，所以对噪声等外部影响敏感，并且该准则不随目标表观和表观的偏离情况呈线性关系[26]。算法将模板匹配过程纳入到基于极大似然估计的概率求解框架中。为达到亚像素的匹配精度和处理遮挡等诸多非确定因素，算法中采用高斯分布来拟合模型表观间相似性分布的峰值区域，并将高斯分布的峰值作为匹配的位置，高斯分布的方差则给出了估计结果的信度。在相同条件下的实验表明，该算法优于基于 SSD 匹配准则的方法。

Morris 分析了运动范围约束对跟踪具有关节运动特性的运动物体的作用，认为当运动与摄像机的视角同方向时，会造成约束矩阵的奇异[27]。如果矩阵奇异，算法的收敛性变差甚至会丢失被跟踪目标。作者针对该问题提出了二维尺度棱镜模型(2D Scaled Prismatic Model, SPM)。相对于三维运动模型，该模型更少发生矩阵奇异问题，并且不需要预先学习特定的三维运动规律。算法还可以对发生矩阵奇异的情况进行判断从而在发生奇异时通过借助三维模型进行跟踪。

在面部图像分析领域，高文教授在国内研究领域做出了早期的工作[28]。该方法建立了基于部件分解组合的人脸图像模型。通过对部件的分

析，采用分类树建立表情模型的向量表示，从而可以根据能量优化原理，利用模板匹配方法提取目标特征，得到人脸表情的表征向量从而实现表情的识别。

苗军博士提出了一种基于重心模板的实时人脸检测方法[28]。并将该方法应用于人脸检测和跟踪任务中。该方法首先将图像马赛克化，通过计算马赛克图像中显著边缘的重心，利用规则的方法确定人脸的位置。该方法能够检测水平和深度旋转的人脸。

1.3.1.3 基于轮廓的方法

轮廓是目标刻画中广泛使用的图像特征，轮廓点通常包含目标的重要信息。基于轮廓的方法一般求解策略是通过迭代方法趋近于问题相关的能量函数的局部最优解。基于轮廓的方法通常对搜索的起始位置比较敏感，所以初始位置的选择是基于轮廓方法性能的重要倚赖。在视觉目标跟踪中，算法通常将前帧的搜索结果作为后帧的起始搜索位置。

Snake 算法是视觉跟踪中最著名的轮廓模型之一，其采用样条曲线连成的封闭轮廓来逼近目标物体的形状[29]。算法中的能量函数包含三个子项：控制轮廓光滑程度的内力，控制初始位移的外力和控制边缘、特征点吸引程度的图像力。通过梯度下降来迭代最小化上面三个能量项的加权和，最终收敛到能量一维分布中的某局部极小作为算法的输出，其中三个能量项的权重往往需要根据具体应用问题进行调整。

Leymarie采用Snake算法进行细胞轮廓的跟踪[30]。细胞轮廓的初始位置在参考帧中由人工确定。并采用前帧中的估计结果作为后帧中的初始搜索位置。当相邻帧尖的细胞位移和形变较小，能观察到算法具有较好的跟踪性能。与原始Snake 算法中的能量项不同，作者采用轮廓上的分段能量平均最小值作为优化标准，并认为相对于求取[29]中定义的目标函数最小，新的能量函数在避免搜索中的振荡和定位到最优解的频率上都有提高。

Paragios 根据水平集(Level Set)理论提出了测地主动轮廓模型(Geodesic active contour) [31]。在该算法中，根据水平集理论递归地改变预定义轮廓的位置和形状来拟合目标的轮廓，达到同时跟踪和检测目标的目的。

上述的基于轮廓的目标跟踪方法主要采用经验定义的能量函数，并试图利用通用的能量函数解决不同的跟踪问题。在[32, 33]的工作中，学习的机制被引入到主动轮廓模型中，虽然得到模型的过程需要繁琐的学习训练，并且学到的模型是应用相关的，但这种针对特定问题的模型表现出了比通用模型优秀的定位和跟踪特性。

主动形状模型(Active Shape Model, ASM)和主动表面模型(Active Appearance Model, AAM)都是基于点分布模型(Point Distribution Model, PDM)的算法[32-35]。在 PDM 中，特定类别物体(比如人脸、人手)的轮廓形状通过若干关键的特征点进行定义，这些特征点的坐标串接构成描述目标的原始形状向量。对训练集中的所有形状向量进行对齐操作后，对他们进行 PCA 分析建模，保留的主成分形成最终的形状模型，形状模型的参数个数反映了形状的主要可变化模式[36]。ASM 搜索首先通过局部纹理模型匹配得到各个特征点的更佳的位置，经过相似变换对齐后，通过统计形状模型对其进行约束，而后再进行局部纹理模型的匹配，形成一个迭代过程，以期形状模型最终匹配到输入的形状模式上去。在 ASM 中，仅使用了特征点局部纹理特征作为启发式信息，没有使用全局的纹理约束，实践中发现 ASM 很容易陷入局部极小。

而在 AAM 中，则采用了形状和纹理二者融合的统计约束，即所谓的统计表观模型。AAM 搜索借鉴了基于合成的分析技术(Analysis-By-Synthesis, ABS)的思想，通过模型参数的优化调整使得模型能够不断逼近实际输入模式。模型参数的更新则放弃了 ASM 中的局部纹理搜索过程，仅使用一个线性预测模型根据当前模型和输入模式之间的差别来预测和更新模型参数。AAM 尽管利用了全局纹理，但却抛弃了局部纹理匹配过程，因此会在一定程度上降低关键特征点配准的精度，而且其线性预测模型也有较大的局限性，在初始位置偏离目标位置过大时，则很难收敛到正确位置。

除了在轮廓模型本身的改进，Isard 将主动轮廓模型融入到 CONDENSATION(Conditional Density Propagation)算法中[3]。在算法实现中，轮廓形状的后验分布采用一组离散的粒子及其相应的权重表示。这种非参数的离散表示方法可以求解任何形式的后验分布。而前述算法中主要采用基于梯度下降的策略，相当于力图求解目标轮廓形状的后验概率分布的期望。当后验分布呈尖峰状分布的时候，期望具有较好的解性质，否则，很容易造成被跟踪目标的丢失。而在[3]中，采用粒子模拟后验概率分布的方法不仅解决了由于暂时的误差所造成的误差传递问题，同时也在一定程度上解决了主动轮廓模型对其实搜索位置要求较苛刻的问题。

1.3.1.4 基于子空间的方法

从本质上说，跟踪问题处理的是非静态信号，前景目标和背景都在随着时间的变化和目标的运动而改变。虽然现在有众多的方法能在短时间隔内和可控环境中很好地跟踪目标，但是当目标表观本身发生剧烈变化或者环境

(比如光照)发生不可预知改变时，往往会造成跟踪中的“漂移现象”，造成跟踪任务的失败。在基于子空间建模的方法中，由于其能够充分地利用历史数据或者训练集得到目标物体表观变化的子空间，从而提供了完整描述前景目标变化的能力，减少因为目标的表观变化所产生“漂移现象”的可能。

Black 提出了基于视角的子空间表示方法并将两不同视角的子空间匹配过程归结为求解优化问题[37]。该工作的主要贡献有两点：首先，对匹配中具有较大误差的特征项进行降权，提高了误差范数的鲁棒性，从而避免了等权使用特征系数时对噪声和奇异点敏感的缺点，比如发生部分遮挡等情况。其次，算法中泛化了光流计算中的灰度恒定假设，提出了子空间一致性假设。该假设认为，目标的视角图像，如果找到正确的参数变换来调整图像，可以用相同的特征基组进行重建，并且可以保证重建图像和原图像具有相同的图像灰度分布。在该假设的基础上，通过求取目标视角图像和子空间重构图像之间的形变参数，做到视角无关的目标跟踪。在其后续工作中[38]，作者进一步采用混合模型构造子空间来建模目标表观的变化。

Torre 在[37]的基础上，将基于子空间的表示方法应用到人体跟踪任务中[39]。根据当前需处理的帧数据和训练集中图像的相似性，选择训练集合中的相似度高的一个子集来构造所跟踪人体的特定子空间。算法中采用肤色进行面部区域的分割，并采用了卡尔曼滤波算法来估计仿射运动模型的参数。

Hager 和 Belhumuer 提出了基于参数化运动模型的模板匹配方法。通过梯度下降策略来求解运动模型的参数从而将固定的模板匹配到正确的图像位置[7]。由于是应用在视觉跟踪中，为了避免每帧都要计算雅可比矩阵，该工作将矩阵分为由模板的亮度梯度乘以其变换的空间微分的常量项和针对运动模型的参数进行微分的可变项，从而大大简化了计算复杂度。该方法在模板的基础上利用子空间方法处理光照问题，通过对涵盖各种光照情况的训练图像的学习得到光照描述的子空间。算法中队部分遮挡问题也采用 M 估计算子(M-estimator)进行了简单地处理。

Jepson 等人提出了一种融合稳定部分，噪声部分和帧间变化部分的目标模型[40]。该模型采用小波基响应作为图像特征，采用期望最大化(Expectation-Maximization)算法估计混合模型中三部分的权重并且采用稳定部分进行仿射运动的估计。在该模型能够处理目标表观变化和光照变化的同时，作者指出在背景变化平缓的情况下，该模型也可能同时估计了背景的稳定部分，从而有可能造成被跟踪中常见的“漂移”问题。

虽然基于子空间的方法能够更好地刻画被跟踪目标的表观变化，但是由于需要事先的训练，从而阻止了在一些条件不满足场合的应用。在[41]中，Ross 等人提出了渐进获取子空间基的方法，并且子空间基随着表观数据的不断增加而得到完善。在相邻帧间，算法采用类似粒子滤波的方法来对仿射运动的参数进行采样并求取最大后验估计。该算法可以处理较大的光照、姿态和尺度变化。但在当前作者给出的实验中，都是在比较简单的背景下进行的，如何能够正确完整地分割目标物体从而正确地更新子空间，在复杂背景中，仍是一个比较困难的问题。

1.3.1.5 基于三维模型的方法

除基于二维图像的方法之外，基于三维模型的目标刻画方法也得到了广泛的研究。这类方法主要采用“基于合成的分析”的策略，通过在三维模型和二维图像之间进行匹配来推算目标物体在三维世界中的状态。

在[42]中，Koller 等人提出了可以描述不同形状汽车的参数化三维汽车模型表示方法，用在交通场景中车辆跟踪的任务中。算法中采用匀速和匀角速度两种运动模型，并对速度的波动用噪声项加以补偿。在跟踪过程中，首先以检测出的运动区域作为种子点展开推断，将三维模型投影到二维图像中，对模型和图像中的边缘片段特征进行比较来计算模型和图像区域的相似度，通过最大后验概率方法不断修正得到的目标表观和状态。算法中采用远光源假设来处理阴影问题。

在[43]中，作者采用了三维模型进行人体自由运动的跟踪。该方法较早地摒弃了在被跟踪人物身上粘贴标志物的方法，直接从视频序列中估计人每时刻的姿态达到运动推断的目的。在每帧视频的处理中，通过对比人体模型所生成图像和实际图像之间的相似度来进行优化搜索。由于人体姿态的复杂性，作者采用了贪婪搜索算法并采用 Chamfer 距离度量图像之间的相似度。该方法同时采用了多个摄像头并假设所有摄像头都是良好定标的来简化问题的难度。

Sidenbladh 等人在贝叶斯推断的框架下研究人体跟踪问题[44]。人体模型由关节相连的圆柱体和关节处的圆球构成，并在肢体上映射可以得到的最新纹理。通过对训练集合的学习得到人体表观和运动的约束模型并假设形状和速度符合一阶马尔科夫模型的假设，利用粒子滤波来近似人体姿态的后验分布。在工作中，即采用了通用的运动模型，如匀速运动模型，也采用了特定运动的模型，如采用 PCA 从训练集合中得到的人行走的特定描述模型。

1.3.1.6 基于高斯或者高斯混合模型的方法

和基于前景的建模方法相对应，背景建模的方法也在跟踪问题中得到了广泛的应用，尤其在拍摄视角固定或者视频序列的整体运动可以得到有效补偿的场合。

在著名的 MIT 的 PFinder 系统中[4]，为了对背景的渐变和各种噪声进行建模，Wren 等人对每一个背景像素在 YUV 空间的取值范围采用高斯函数进行建模，其中高斯均值和方差从视频的初始若干帧中估计并且随着新表观的到来进行平滑的更新。在背景模型的基础上，对当前帧，可以从高斯分布计算其隶属于背景的概率，从而将那些明显偏离背景模型的像素分类为前景，达到目标跟踪的目的。

虽然高斯模型在表示，更新和处理缓慢变化背景方面能够满足要求，然而，在很多室外应用场合中，比如树枝随风摇摆，云层遮住太阳等情况下，高斯模型会产生大量的虚假前景 [45]。在[4]的基础上，Stauffer 和 Grimson 提出了基于混合高斯的背景像素建模策略。由于多高斯模型的特性，该方法允许表示具有多态性质的背景像素，例如树枝随风摆动等情况可以用两态的混合高斯模型对其建模。在分类阶段，对当前帧中的每一个像素，分别采用混合高斯中的每一个分量对其进行解释，如果能够解释则认为其为背景，并采用该像素值更新该高斯分量，否则对该像素值建立一个新的高斯分量。

1.3.1.7 基于非参数概率密度估计模型的方法

不同于基于高斯等分布假设的参数化模型，在 W4 系统中[46]，采用帧内和帧间统计量(中值、最大值、最小值、最大帧间差等)来描述背景像素的变化范围，该模型计算简单，并且作者在在模型的初始阶段采用了中值滤波算法，从而能够滤除背景建模阶段场景中的运动。

Elgammal 同样采用了非参数的概率密度估计方法[47]，其利用 Parzen 窗方法进行概率密度的估计。相对于使用参数的高斯混合模型来逼近真实概率密度，Parzen 窗方法由于没有假定像素取值的分布形式而适用范围更广泛，但也于非参数的计算问题，在模型更新上带来计算量的增加。

1.3.2 数据关联技术

在所有的已提取表观中，如何选择与目标相关的子集来推断目标的运动状态是数据关联技术所要解决的主要问题。尤其对于多目标跟踪问题，正确地将多个有效表观映射到多个目标物体成为完成跟踪任务的主要困难。

在跟踪研究领域，直到[48]的工作出现之前，数据关联都是采用简单的近邻法，门限法等。

在视觉跟踪领域，Rasmussen 等将根植于雷达和声纳跟踪领域的联合统计数据关联滤波器(Joint Probabilistic Data Assiciation Filter，JPDAF)引入到视觉跟踪领域中来[6]，并根据视觉跟踪问题不同于雷达中点跟踪的特性，提出了联合可能性滤波器(Joint Likelihood Filter, JFL)。在 JFL 中，通过对目标之间相对深度的推测，从而可以对目标的表观进行其是否可见或者被遮挡的预测，并在不同的目标表观之间进行信息共享从而解决多目标的数据关联问题。

MacCormick 提出了基于 CONDENSATION (CONditional DENsity propagaTION)算法的数据关联方法[49]。在多目标跟踪的过程中，算法将每个目标的深度信息嵌入到联合状态中，并且约定当目标重叠的时候其相对深度位置是不可改变的。

1.3.3 滤波框架

滤波框架在随机信号处理领域得到广泛的应用和研究。该领域中具有完善的根据系统表观估计系统状态的理论框架和算法。虽然视觉跟踪问题是计算机视觉中的研究子领域，但从理论上，视频仍是随机信号的一种。所以，众多学者将随机信号处理中的成熟理论引入到视觉跟踪研究中来，并且结合本领域的研究问题提出了改进。实践证明，在滤波算法为视觉跟踪研究提供了完整、有效地问题解决框架。

1.3.3.1 卡尔曼及扩展卡尔曼滤波

卡尔曼滤波主要针对线形系统和高斯分布的状态估计问题[50]。应用卡尔曼滤波主要分为两个步骤：采用过程模型对下一时刻的状态预测过程和采用表观模型关联状态和观测，从而修正预测状态的过程。当过程模型或者表观模型中具有非线性因素的时候(需要状态仍满足高斯分布)，通过在局部点上采用泰勒展开式可以得到扩展卡尔曼滤波算法[51]。卡尔曼滤波算法在视觉跟踪领域得到了广泛的应用。Terzopoulos 将卡尔曼滤波和 Snake 算法相结合[52]。Broida 和 Chellappa 采用卡尔曼滤波算法来跟踪噪声图像中的点[53]。在基于立体视摄像机的目标跟踪中， Beymer 和 Konolige 利用卡尔曼滤波来预测目标的位置和速度[54]。Rosales 和 Sclaroff 采用扩展卡尔曼滤波算法从二维运动中估计目标的三维运动轨迹[55]。

1.3.3.2 序列蒙特卡洛滤波算法

当跟踪问题具有强非线性或者状态向量不符合高斯分布时，或者图像中具有较严重的噪声或者目标遮挡时，卡尔曼滤波算法和扩展卡尔曼滤波往往无法有效工作。

由于计算资源的逐渐丰富和在其它领域中的成功应用，序列蒙特卡洛滤波算法在视觉跟踪问题中得到了越来越多的重视和研究。自 Isard 将其引入到视觉目标跟踪领域以来(并重新命名为 CONDENSATION)[3]，Deutscher 和 Sullivan 引入层次性采样法，使其适用于跟踪具有关节型运动的目标类型[56, 57]。Li 利用序列重要性采样框架来达到同时进行目标跟踪和验证的目的[58, 59]。Sidenbladh 则应用粒子滤波解决步行人的跟踪问题[44]。

粒子滤波解决跟踪问题的主要思想是采用数量为 N 的加权采样集合 $\{(x_t^{(i)}, w_t^{(i)}) \mid i = 1,...,N\}$ 来逼近目标状态的后验概率，其中 $x_t^{(i)}$ 表示采样(称为粒子)，$w_t^{(i)}$ 是相应的粒子权重。根据在时间轴上到来的图像表观，对粒子及其权重进行不断地更新以获取时间轴上目标状态后验分布的变化，于是，目标状态的进化可以通过粒子的模拟传递来进行近似。由粒子模拟后验分布的主要好处是无需假设后验分布的形式，从而可以用来求解任何非线性的跟踪问题。

1.4 本工作的主要技术路线和目标

作为视觉研究的一个重要子领域，视觉跟踪技术最早的需求来自于军事应用领域。早期的视觉跟踪研究主要基于图像匹配的策略，即在给定时间轴采样率的情况下，通过传统的图像匹配方法找到目标物体在相邻帧间的对应，从而通过微分等算子根据目标在采样点时刻的位置信息推断目标的运动状态。

由于视觉跟踪的应用环境通常要求算法能够实时处理图像序列，80 年代早期的研究集中于采用梯度下降或者能量最小化等方法来降低视觉跟踪算法匹配过程中的搜索复杂性和提高算法鲁棒性。代表性研究如 LK 光流算法[60]，Snake 轮廓模型[29]等，都通过定义目标能量函数，从而采用基于梯度下降的策略来最小化能量函数进行图像的匹配。相对于局部穷举搜索策略，基于梯度下降的搜索方法明显降低了算法的复杂性和计算强度。

进入 90 年代，人们越来越认识到视觉跟踪问题的复杂性和不确定性，采用简单的搜索匹配策略已经无法处理诸多复杂的视觉跟踪问题。诸多学者将视觉跟踪问题纳入到贝叶斯时序滤波的理论框架下，不仅使视觉跟踪中的不确定性因素得到了系统完整的处理，也为视觉跟踪问题的研究提供了完整的理论基础。著名的卡尔曼滤波算法就是贝叶斯时序滤波框架的一种具体实现。卡尔曼滤波器由 Rudolph E. Kalman 于 1960 年提出并广泛应用于雷达跟踪等领域[50]。90 年代初，其在视觉跟踪领域中也得到了广泛的应用并在很多问题上取得了良好的效果。然而，卡尔曼滤波算法要求所处理的系统是线性的，并且状态和观测向量必须满足高斯分布，这限制了其应用范围。虽然扩展卡尔曼滤波算法能够处理部分非线性和非高斯分布的问题，但由于其基于局部线性化的策略，在具有本质非线性化的复杂视觉跟踪问题上很难取得令人满意的效果。

蒙特卡洛方法最早应用于 60 年代的统计物理学研究中，虽然在该领域中被奉为经典方法，但由于其计算复杂性很高，在早期计算机视觉研究中始终无法得到应有的重视。进入 90 年代后期，随着台式机算设备计算能力的提升，以往受限于此的蒙特卡洛方法，逐渐得到了计算机视觉跟踪领域研究者的重视。

序列蒙特卡洛滤波算法是将贝叶斯框架和蒙特卡洛方法有机融合的时序信号分析理论框架，是贝叶斯时序滤波算法的一种实现。由于其本身能够处理非线性问题和具有坚实完整的理论框架，使该算法可以广泛地应用于复杂视觉跟踪问题的求解中。目前，该方法在视觉跟踪领域得到越来越多的重视和应用。

本论文以序列蒙特卡洛滤波算法为框架，将人脸和人体跟踪问题纳入到基于贝叶斯理论的求解框架之中，以人脸和人体跟踪为主要研究问题，深入研究了序列蒙特卡洛滤波算法在特定问题上的应用并提出了序列蒙特卡洛滤波算法的改进。

1.5 本论文的主要贡献

针对当前视觉跟踪中的困难问题，本文在序列蒙特卡洛滤波算法的理论框架下，以人脸和人体跟踪为应用背景，提出了下列算法：

1. 可区分性目标模型的构建和更新算法

目标表观建模是视觉跟踪算法性能的决定性因素之一。构建目标表观模型主要需要解决两方面的问题：如何选择图像特征和如何在图像特征取值分布的基础上描述被跟踪目标的表观。视觉跟踪研究的实践表明：图像特征选择和基于图像特征的目标描述从根本上决定了跟踪算法的鲁棒性和计算复杂性。虽然这一问题得到了领域内学者的极大重视和不懈努力，其仍是阻碍视觉目标跟踪技术进入实际应用的最困难问题之一。本论文提出了一种新的自适应目标表观建模和更新方法，该方法在建模的过程中不仅仅考虑目标表观信息，同时对目标所处环境中的背景信息进行考察，可以对目标/背景的差异信息进行有效建模，从根本上保证了模型的可区分性。在图像特征选择上，将 Haar 小波特征首次引入到视觉跟踪领域。基于 Haar 小波特征，采用分类器组合的方式构建目标表观的可区分性模型。另外，注意到在采用序列蒙特卡洛滤波算法进行目标跟踪的过程中，由于序列蒙特卡洛滤波算法本身的性质，从而有大量的“背景”粒子的存在。在以往工作中，这些粒子被认为对最终结果贡献很小而被忽视。本算法则利用了“背景”粒子中蕴含的背景分布信息，从而为目标表观模型随背景的变化而实时更新提供了有效的方法。在人体跟踪问题上验证提出的算法，实验结果表明，相比于目前最有代表性的跟踪算法之一 Mean Shift[15]，提出的算法在公开测试序列上取得了更优的跟踪效果。

2．集成多运动模型的运动建模和推断方法

由于计算复杂性的限制，视觉跟踪算法通常基于局部搜索的策略确定目标的运动状态。所以，根据目标运动规律确定目标在当前时刻以较高概率出现的区域成为跟踪算法成败的关键问题。现实世界中的很多物体都具有复杂的运动模式。如何针对复杂运动建立模型，始终是解决这些目标跟踪问题的瓶颈所在。本论文提出了采用多运动模型对目标复杂运动进行建模和估计的框架。在此基础上，针对具有多种运动模式和具有高维运动状态的两类常见的复杂目标运动，将多模型的估计框架融入到序列蒙特卡洛滤波算法中，从而针对两类复杂运动问题提出了标准序列蒙特卡洛滤波算法的两个改进：基于多模型切换和基于多模型协同的序列蒙特卡洛滤波算法。在汽车跟踪和面部表情估计问题上分别验证了提出的基于多模型切换和多模型协同的序列蒙特卡洛滤波算法。实验表明，在降低计算复杂度的同时，算法相对比于标准序列蒙特卡洛滤波算法得到了较高的性能提升。

3. 融合光流和模型的面部特征点跟踪算法

面部特征点跟踪是基于特征点的运动感知研究的典型应用。但是由于面部具有典型的非刚性形变等复杂运动模式，并且具有多以肤色为主调的平滑纹理，所以面部特征点跟踪是基于特征点运动感知任务中的困难问题。现有的面部特征点跟踪方法主要可以分为基于特定描述模型(以下简称模型)和基于光流的方法。基于模型的方法通常对特征点的变化进行描述性建模，具有鲁棒准确的优点，但通常计算复杂度高，对起始位置比较敏感。与基于模型的方法不同，基于光流的方法计算复杂性低，不需要特定的运动模型，但跟踪结果中往往含有较多的噪声。本论文在序列蒙特卡洛滤波算法的框架下融合了基于光流和基于特定模型的方法来解决面部特征点跟踪问题，来克服单独采用一类方法不足，从而达到鲁棒跟踪面部特征点的目的。基于尺度空间理论改进的 KLT 光流算法能够准确估计面部具有刚性或者近似刚性运动特征点的位移运动。光流估计结果为基于模型的形变特征点估计提供了更好的起始搜索位置，同时加速了基于序列蒙特卡洛滤波算法的搜索过程。由于基于光流的估计算法中存在跟踪误差，进一步采用特征点运动的子空间约束来求精跟踪结果。相比于广泛使用的 KLT 特征点跟踪算法，实验结果证实了提出的算法的有效性。

1.6 本文组织及各章间关系

第 1 章为绪论，介绍了研究背景，研究意义，前人工作概述，面临的主要问题，主要技术路线和主要贡献。第 2 章对本文理论框架进行了阐述和分析。第 3 章描述了基于目标/背景信息差的自适应目标建模和更新方法。第 4 章给出了基于多运动模型的复杂运动建模和推断方法。第 5 章探讨了融合高端和低端运动信息进行面部特征点跟踪的问题。最后在对全文进行总结的基础上，讨论了可能的后续扩展工作。

表 1-1 本论文主要章节之间的关系

Table 1-1 Relations among main chapters in this thesis

核心问题	主要困难	关联章节	应用背景
目标建模	复杂背景对跟踪系统性能的影响	第三章	人体区域跟踪
复杂运动的建模和运动推断	目标复杂运动模式对跟踪系统性能的影响	第四章	人脸区域跟踪和面部复杂运动估计
	目标的非刚性形变对跟踪系统性能的影响	第五章	面部特征点跟踪

理论层面：目标建模和运动建模是决定跟踪系统性能的核心因素。

问题层面：复杂背景、复杂运动模式和目标非刚性形变是当前研究中的主要困难问题；

应用层面：不仅研究了人脸和人体的位移运动推断问题，同时研究了面部复杂形变和复杂表情等运动的推断问题。所有算法的应用背景构成了对面部所有运动的完整描述。

第2章 序列蒙特卡洛滤波算法

2.1 引言

在视觉跟踪问题中，由于目标自身的运动，背景的变化以及成像条件的复杂多样性等诸多不确定性因素的存在，采用简单的匹配搜索策略已经无法处理日益复杂的视觉跟踪问题。从自动控制领域引入的卡尔曼滤波算法首先为视觉跟踪问题提供了完整的求解理论框架。由于卡尔曼滤波算法基于贝叶斯理论，其能够处理各种非确定性因素和融合多通道信息。卡尔曼滤波算法目前仍是视觉跟踪研究中应用最广泛的算法，但由于算法基于线性系统和高斯分布假设，随着视觉跟踪问题研究的深入，卡尔曼滤波算法在诸多复杂视觉跟踪问题上越发力不从心。

近年来引入视觉跟踪研究领域的序列蒙特卡洛滤波算法与卡尔曼滤波算法一样根植于贝叶斯时序滤波框架。在某种意义上，卡尔曼滤波算法可以看作序列蒙特卡洛滤波算法的特例。序列蒙特卡洛滤波算法由于下述优点而得到了越来越多研究人员的重视：

1) 相对于卡尔曼滤波要求后验分布满足高斯形式，算法可以求解后验分布为任意函数形式的视觉跟踪问题；
2) 算法在理论上能够解决带有任意非线性特性的复杂视觉跟踪问题；
3) 算法能够在时间轴上传递条件分布的变化，从而系统地处理不确定性问题。
4) 当粒子数目足够多时，算法总能够得到收敛的跟踪结果；
5) 由于在贝叶斯的框架下进行推断，算法能够有效融合多通道信息。

由于本论文以序列蒙特卡洛滤波算法为理论框架来系统地研究以人脸和人体跟踪为例的视觉跟踪问题。本章依据[61]简要介绍序列蒙特卡洛滤波算法，从而为阐述提出的相关算法奠定基础。

2.2 跟踪问题的定义和贝叶斯时序滤波框架

视觉跟踪的主要研究目标是如何通过输入的图像时间序列产生对目标物体运动信息的感知。考虑视觉跟踪问题中需要处理各种不确定性因素的发

生，本节从基于概率统计和贝叶斯框架的角度来定义，阐述和分析视觉跟踪问题。

从贝叶斯统计推断的角度，算法的理想输出是关于物体运动状态的后验概率分布。如何在存在噪声和诸多不确定因素的目标表观基础上，准确而有效地表示和求取目标运动状态的后验概率分布是跟踪算法面临的主要任务。用 $\mathbf{x}_t$ 表示目标物体在时刻 t 的运动状态，用 $\mathbf{z}_t$ 表示时刻 t 的目标表观。用 $\mathbf{x}_{1:t}$ 表示状态序列 $\{\mathbf{x}_1,...,\mathbf{x}_t\}$，类似有 $\mathbf{z}_{1:t}$。

在视觉跟踪研究中，主要需要分析和解决如下三个基本问题：

1) **目标运动状态的预测**：假设已经知道 $p(\mathbf{x}_{1:t-1}|\mathbf{z}_{1:t-1})$，如何预测 $\mathbf{x}_t$？要解决这一问题，需要定义 $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$。

2) **数据关联问题**：在时刻 t 提取的所有表观，只有部分与目标运动状态有关联，需要通过定义 $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$ 来分辨与目标相关的表观，从而依据正确的表观求取目标的运动状态。

3) **状态校验问题**：在得到 t 时刻表观 $\mathbf{z}_t$ 后，需要根据最新表观调整后验概率 $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ 得到解的更新估计。

由于上面的求解后验概率问题的复杂性，首先进行如下的条件独立性假设：

1) 目标的运动状态只与其相邻状态条件相关，即

$$p(\mathbf{x}_t|\mathbf{x}_{0:t-1}) = P(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2-1}$$

2) 目标的当前表观唯一条件依赖于目标的当前运动状态，即当 $\mathbf{x}_t$ 给定，$\mathbf{z}_t$ 与其他时刻的表观条件无关，表示为

$$p(\mathbf{z}_{t:t-m}|\mathbf{x}_t) = p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{z}_{t-1:t-m}|\mathbf{x}_t) \tag{2-2}$$

在解决实际问题中积累的经验表明，上述问题简化的假设是合理的，而且极大地简化了视觉跟踪问题的求解过程。根据上述假设，对状态预测问题，可

以导出，

$$\begin{aligned} & p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1}) \\ &= \sum_{\mathbf{x}_{t-1}} p(\mathbf{x}_t, \mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1}) \\ &= \sum_{\mathbf{x}_{t-1}} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_{t-1}) p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1}) \\ &= \sum_{\mathbf{x}_{t-1}} p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1}) \end{aligned} \tag{2-3}$$

对状态校验问题，可以导出：

$$\begin{aligned} p(\mathbf{x}_t \mid \mathbf{z}_{1:t}) &= \frac{p(\mathbf{x}_t, \mathbf{z}_{1:t})}{p(\mathbf{z}_{1:t})} \\ &= \frac{p(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{z}_{1:t-1}) p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1}) p(\mathbf{z}_{1:t-1})}{p(\mathbf{z}_{1:t})} \\ &= p(\mathbf{z}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1}) \frac{p(\mathbf{z}_{1:t-1})}{p(\mathbf{z}_{1:t})} \\ &= \frac{p(\mathbf{z}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1})}{\sum_{\mathbf{x}_t} p(\mathbf{z}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1})} \end{aligned} \tag{2-4}$$

假设在初始时刻，有 $p(\mathbf{x}_0 \mid \mathbf{z}_0) = p(\mathbf{x}_0)$，则 $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$ 可以通过预测和状态更新的两步迭代运算得到。设 t-1 时刻的目标状态后验概率分布为 $p(\mathbf{x}_{t-1} \mid \mathbf{z}_{1:t-1})$。在预测阶段，采用公式(2-3)到目标在 t 时刻的先验概率分布 $p(\mathbf{x}_t \mid \mathbf{z}_{1:t-1})$。在 t 时刻，提取目标表观 $\mathbf{z}_t$，根据公式(2-4)得到目标状态的后验概率 $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$。

贝叶斯时序滤波框架依据预测校验步骤，在贝叶斯框架下进行迭代后验概率估计的一类滤波器的总称，公式(2-3)和(2-4)构成了贝叶斯滤波器的核心步骤。通过在时间轴上递归地传递后验概率分布，可以得到贝叶斯规则下的最优序列解。贝叶斯滤波器包括广泛使用的卡尔曼滤波器，扩展卡尔曼滤波器，序列蒙特卡洛滤波器等。

2.3 序列蒙特卡洛滤波算法

2.3.1 核心思想

作为蒙特卡洛方法(Monte Carlo)的一种，序列蒙特卡洛滤波器在很多研究领域中都有应用，如 bootstrap filtering，CONDENSATION，particle filtering，interacting particle approximation 和 survival of fittest 都是序列蒙特卡洛滤波算法在不同领域中的名称。序列蒙特卡洛滤波器是通过蒙特卡洛模拟对贝叶斯滤时序滤波器的实现，其核心思想是采用对特定概率分布的一组随机采样和其相应权重的集合来表示待求后验分布，并根据这些采样计算后验分布的某种测度。根据蒙特卡洛特性，当采样数据趋于无穷多，算法可以无穷逼近后验概率的真实分布，从而可以得到无限接近最优解的估测结果。

设 $\{\mathbf{x}^i_{0:t}, w^i_t \mid i=1,...,N_s\}$ 是对后验分布 $p(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t})$ 的随机采样集合，其中 $\{\mathbf{x}^i_{0:t} \mid i=1,...,N_s\}$ 是采样点集合，这里称为粒子集合，$\{w^i_t \mid i=1,...,N_s\}$ 是与采样点相关联的权重集合，并满足 $\sum_i w^i_t = 1$。于是，后验分布 $p(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t})$ 可以用该粒子集合离散地近似为

$$p(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} w^i_t \delta(\mathbf{x}_{0:t} - \mathbf{x}^i_{0:t}) \tag{2-5}$$

公式(2-5)中的权重集合 $\{w^i_t \mid i=1,...,N_s\}$ 依据重要性采样原理确定[62, 63]：假设 $p(x) \propto \pi(x)$ 是一个难以进行采样的分布例如无法得到 $p(x)$ 的解析解或者只了解 $p(x)$ 的部分信息等情况。用 $\mathbf{x}^i \sim q(\mathbf{x}), i=1,...,N_s$ 表示从分布 $q(.)$得到的采样，称 $q(.)$为提议分布(proposal distribution)或者重要性密度(importance density)。于是，$p(.)$可表示为

$$p(\mathbf{x}) \approx \sum_{i=1}^{N_s} w^i \delta(\mathbf{x} - \mathbf{x}^i) \tag{2-6}$$

其中，

$$w^i \propto \frac{\pi(\mathbf{x}^i)}{q(\mathbf{x}^i)} \tag{2-7}$$

是第 i 个采样的权重。

于是，如果粒子 $\mathbf{x}^i_{0:k}$ 是从提议分布 $q(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k-1})$ 的采样，则在公式(2-7)中的权重可以写做：

$$w^i_t \propto \frac{p(\mathbf{x}^i_{0:t} \mid \mathbf{z}_{1:t})}{q(\mathbf{x}^i_{0:t} \mid \mathbf{z}_{1:t})} \tag{2-8}$$

考虑视觉目标跟踪问题，在时刻 t，已经有后验分布 $p(\mathbf{x}_{0:t-1} \mid \mathbf{z}_{1:t-1})$ 的粒子集合逼近，现在需要更新粒子集合来逼近 $p(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t})$。如果提议分布可以分解为如下的形式：

$$q(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t}) = q(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}, \mathbf{z}_{1:t}) q(\mathbf{x}_{0:t-1} \mid \mathbf{z}_{1:t-1}) \tag{2-9}$$

则可以得到 $\mathbf{x}^i_t \sim q(\mathbf{x}_t \mid \mathbf{x}_{0:t}, \mathbf{z}_{1:t})$，在此基础上融合 $\{\mathbf{x}^i_t, x^i_{0:t-1} \sim q(\mathbf{x}_{0:t-1} \mid \mathbf{z}_{1:t-1})\}$ 从而得到 t 时刻的采样 $\mathbf{x}_{0:t} \sim q(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t})$。

注意到有

$$\begin{aligned} p(\mathbf{x}_{o:t} \mid \mathbf{z}_{1:t}) &= \frac{p(\mathbf{z}_t \mid \mathbf{x}_{0:t\text{-}1}, \mathbf{z}_{1:t\text{-}1}) p(\mathbf{x}_{0:t} \mid \mathbf{z}_{1:t\text{-}1})}{p(\mathbf{z}_t \mid \mathbf{z}_{1:t\text{-}1})} \\ &= \frac{p(\mathbf{z}_t \mid \mathbf{x}_{0:t}, \mathbf{z}_{1:t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}, \mathbf{z}_{1:t-1})}{p(\mathbf{z}_t \mid \mathbf{z}_{1:t-1})} \times p(\mathbf{x}_{0:t-1} \mid \mathbf{z}_{1:t-1}) \end{aligned} \tag{2-10}$$

$$\begin{aligned} &\frac{p(\mathbf{z}_t \mid \mathbf{x}_{0:t}, \mathbf{z}_{1:t-1}) p(\mathbf{x}_t \mid \mathbf{x}_{0:t-1}, \mathbf{z}_{1:t-1})}{p(\mathbf{z}_t \mid \mathbf{z}_{1:t-1})} \times p(\mathbf{x}_{0:t-1} \mid \mathbf{z}_{1:t-1}) \\ &\propto p(\mathbf{z}_t \mid \mathbf{x}_t) p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \times p(\mathbf{x}_{0:t-1} \mid \mathbf{z}_{1:t-1}) \end{aligned} \tag{2-11}$$

将(2-10)和(2-8)代入(2-7)，可以得到权重的更新方程：

$$\begin{aligned} w^i_t &\propto \frac{p(\mathbf{z}_t \mid \mathbf{x}^i_t) p(\mathbf{x}^i_t \mid \mathbf{x}^i_{t-1}) p(\mathbf{x}^i_{0:t-1} \mid \mathbf{z}_{1:t-1})}{q(\mathbf{x}^i_t \mid \mathbf{x}^i_{0:t-1}, \mathbf{z}_{1:t}) q(\mathbf{x}^i_{0:t-1} \mid \mathbf{z}_{1:t-1})} \\ &= w^i_{t-1} \frac{p(\mathbf{z}_t \mid \mathbf{x}^i_t) p(\mathbf{x}^i_t \mid \mathbf{x}^i_{t-1})}{q(\mathbf{x}^i_t \mid \mathbf{x}^i_{0:t-1}, \mathbf{z}_{1:t})} \end{aligned} \tag{2-12}$$

更进一步，如果有 $q(\mathbf{x}^i_t \mid \mathbf{x}^i_{0:t-1}, \mathbf{z}_{1:t}) = q(\mathbf{x}^i_t \mid \mathbf{x}^i_{t-1}, \mathbf{z}_t)$，则提议分布仅仅依赖于

$\mathbf{x}_{t-1}$和$\mathbf{z}_t$。如果只关心$p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$形式的后验分布，该假设极大简化了滤波过程，在后面的叙述中，如果没有特别说明，都以此为假设。于是，权重更新方程简化为

$$w^i{}_t \propto w^i{}_{t-1} \frac{p(\mathbf{z}_t \mid \mathbf{x}^i{}_t)p(\mathbf{x}^i{}_t \mid \mathbf{x}^i{}_{t-1})}{q(\mathbf{x}^i{}_t \mid \mathbf{x}^i{}_{t-1}, \mathbf{z}_t)} \tag{2-13}$$

可以用

$$p(\mathbf{x}_t \mid \mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} w^i{}_t \delta(\mathbf{x}_t - \mathbf{x}^i{}_t) \tag{2-14}$$

模拟关心的后验分布。通过不断接受表观数据来递归地传递粒子和更新其相应的权重，序列蒙特卡洛滤波不断更新后验分布。算法伪代码如图 2-1。

标准序列蒙特卡洛滤波算法的主要流程

输入：视频序列$I(u,v,t)$，目标物体的初始状态$\mathbf{x}_0$

输出：目标物体的序列状态$\mathbf{x}_{1:T}$

For 时刻 t=1:T,

预测：$\mathbf{x}_{t-}{}^i = f(\mathbf{x}_{(t-1)+}{}^i) + \xi^i$，其中$f(\cdot)$是运动模型，$\xi^i \sim N(0, \Sigma_{d_x})$。

$w_{t-}{}^i = w_{(t-1)+}{}^i$。用集合$\{(x_{t-}{}^i, w_{t-}{}^i) \mid i = 1,...,N\}$表示先验分布$p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$。

校验：$\mathbf{x}_{t+}{}^i = \mathbf{x}_{t-}{}^i$，$w_{t+}{}^i = p(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{x}_{t-}{}^i) \cdot w_{t-}{}^i$。用粒子集合$\{(x_{t+}{}^i, w_{t+}{}^i) \mid i = 1,...,N\}$表示后验分布$p(\mathbf{x}_t \mid \mathbf{z}_t)$。

End

图 2-1 标准序列蒙特卡洛滤波算法的流程

Fig.2-1 Flowchart of standard sequential Monte Carlo filtering algorithm

2.3.2 算法的退化问题

在视觉跟踪过程中，通常经过一段时间的循环运行之后，会出现某个粒子有接近于 1 的权重，而其余所有粒子的权重都非常小，从而意味着绝大部分计算量都对最终结果只有很小的贡献，即粒子群发生退化问题。有工作证明[63]，粒子集合的方差只能随着时间增加，所以粒子退化问题几乎是不可避免的。

为解决粒子退化问题，首先定义粒子集合退化程度的测量[64]，

$$N_{eff} = \frac{N_s}{1+Var(w^{*i}{}_t)} \tag{2-15}$$

公式(2-15)中，$w^{*i}{}_t = p(\mathbf{x}^i{}_t \mid \mathbf{z}_{1:t}) / q(\mathbf{x}^i{}_t \mid \mathbf{x}^i{}_{t-1}, \mathbf{z}_{1:t})$ 指粒子的真实“权重”。由于其无法得到，将度量近似为

$$\hat{N}_{eff} = \frac{1}{\sum_{i=1}^{N_s} (w^i{}_t)^2} \tag{2-16}$$

其中，$w^i{}_t$ 是由公式(2-13)求得的归一化权值。由公式(2-15)，有 $N_{eff} \le N_s$。越小的 N_{eff} 说明粒子集合退化越严重。根据公式(2-15)，可以通过增大 N_s 大的方法避免退化问题，但同时所带来的计算量增长影响了该策略的实用性。

在实际的视觉跟踪问题中，主要采用以下两种策略：

a) 选择合适的提议分布；
b) 对粒子集合进行重新采样。

2.3.3 选择合适的提议分布

在 $\mathbf{x}^i{}_{t-1}$ 和 $\mathbf{z}_t$ 确定的情况下，能够控制公式(2-16)中权重方差增加的最优提议分布应满足如下的形式[63]：

$$q(\mathbf{x}_t \mid \mathbf{x}^i{}_{t-1}, \mathbf{z}_t)_{opt} = p(\mathbf{x}_t \mid \mathbf{x}^i{}_{t-1}, \mathbf{z}_t) \tag{2-17}$$

$$w^i{}_t \propto w^i{}_{t-1} \frac{p(\mathbf{z}_t \mid \mathbf{x}^i{}_t) p(\mathbf{x}^i{}_t \mid \mathbf{x}^i{}_{t-1})}{q(\mathbf{x}^i{}_t \mid \mathbf{x}^i{}_{t-1}, \mathbf{z}_t)} \tag{2-18}$$

将(2-18)带入(2-13)有，

$$w^i_t \propto w^i_{t-1} p(\mathbf{z}_t \mid \mathbf{x}^i_{t-1})$$

$$= w^i_{t-1} \int p(\mathbf{z}_t \mid \mathbf{x}'_t) p(\mathbf{x}'_t \mid \mathbf{x}^i_{t-1}) d\mathbf{x}'_t \tag{2-19}$$

于是可以得到，无论如何从 $q(\mathbf{x}_t \mid \mathbf{x}^i_{t-1}, \mathbf{z}_t)_{opt}$ 中采样，只要 $\mathbf{x}^i_{t-1}$ 确定，则 $w^i_t \propto w^i_{t-1}$，即有 $Var(w^i_t) = Var(w^i_{t-1})$。也就是说，最优的提议分布也仅仅能使粒子集合的方差不增大。

但是，找到最优提议分布存在两个困难，需要从未知的分布 $p(\mathbf{x}_t \mid \mathbf{x}^i_{t-1}, \mathbf{z}_t)$ 中进行采样和求取积分 $\int p(\mathbf{z}_t \mid \mathbf{x}'_t) p(\mathbf{x}'_t \mid \mathbf{x}^i_{t-1}) d\mathbf{x}'_t$。最优提议分布一般只有在两种特殊情况下是可得的。一是当 $\mathbf{x}^i_t$ 的取值属于一个有限集合，则积分问题变为求和问题。另一种情况是当状态表观关联模型 $p(\mathbf{z} \mid \mathbf{x})$ 是线性变换的时候，则 $p(\mathbf{x}_t \mid \mathbf{x}^i_{t-1}, \mathbf{z}_t)$ 满足高斯分布，即

$$\mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_{t-1}) + \mathbf{v}_{t-1} \tag{2-20}$$

$$\mathbf{z}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{n}_t \tag{2-21}$$

其中，$\mathbf{f}_t : R^{n_x} \to R^{n_x}$ 是非线性映射，$\mathbf{H}_t \in R^{n_x \times n_x}$ 是观测矩阵。

$$\mathbf{v}_{t-1} \sim N(\mathbf{v}_{t-1}, \mathbf{o}_{n_v \times 1}, \mathbf{Q}_{t-1}) \tag{2-22}$$

$$\mathbf{n}_t \sim N(\mathbf{n}_t, \mathbf{o}_{n_n \times 1,}, \mathbf{R}_t) \tag{2-23}$$

定义

$$\mathbf{\Sigma}_t^{-1} = \mathbf{Q}^{-1}{}_{t-1} + \mathbf{H}^T{}_t \mathbf{R}_t^{-1} \mathbf{H}_t \tag{2-24}$$

$$\mathbf{m}_t = \mathbf{\Sigma}_t (\mathbf{Q}_{t-1}{}^{-1} \mathbf{f}_t(\mathbf{x}_{t-1}) + \mathbf{H}^T{}_t \mathbf{R}_t^{-1} \mathbf{z}_t) \tag{2-25}$$

可以推导得到，

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t) = N(\mathbf{x}_t; \mathbf{m}_t, \mathbf{\Sigma}_t) \tag{2-26}$$

$$p(\mathbf{z}_t \mid \mathbf{x}_{t-1}) = N(\mathbf{z}_t; \mathbf{H}_t \mathbf{f}_t(\mathbf{x}_{t-1}), \mathbf{Q}_{t-1} + \mathbf{H}_t \mathbf{R}_t \mathbf{H}_t^T) \tag{2-27}$$

对其他的各种情况，通常无法得到解析解。然而，通过局部线性化获得次优解是可能的[63]。比如用高斯模型逼近分布 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t)$ 或者采用 unscented transform 来估测 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t)$ 对高斯模型的近似程度[65]。通常，采用这样的提议分布，可以在取得相同性能的情况下大大减少所需粒子的数量。

需要指出的是，由于直观和容易实现，一种常用的提议分布为状态转移模型，即

$$p(\mathbf{x}_t \mid \mathbf{x}^i_{t-1}, \mathbf{z}_t) = p(\mathbf{x}_t \mid \mathbf{x}^i_{t-1}) \tag{2-28}$$

将(2-28)代入(2-18)，可以得到

$$w^i_t \propto w^i_{t-1} p(\mathbf{z}_t \mid \mathbf{x}^i_t) \tag{2-29}$$

2.3.4 粒子重采样技术

另一种避免粒子集合退化的方法是在退化现象发生的情况下采用重采样技术更新粒子集合(比如，当 $\hat{N}_{eff}$ 小于某一指定阈值)。重采样的基本思想是去除权值较小的粒子从而使权值较大的粒子获得进化或者生存的机会。其基本方法是通过对 $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$ 在离散点的近似表示形式 $p(\mathbf{x}_t \mid \mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} w^i_t \delta(\mathbf{x}_t - \mathbf{x}^i_t)$ 进行 Ns 次有放回的独立采样生成新的粒子集合 $\{\mathbf{x}_t^{i^*} \mid i = 1, ..., N_s\}$，并且满足 $\Pr(x_t^{i^*} = x_t^j) = w_t^j$。既然新粒子集合是从 $p(\mathbf{x}_t \mid \mathbf{z}_{1:t})$ 的独立采样，可以重置所有粒子的权值为 $w_t^{i^*} = 1/N_s$。于是，标准序列蒙特卡洛滤波的伪代码算法可以改写为图 2-2 中的算法。

基于重采样的标准序列蒙特卡洛滤波算法的主要流程

输入：视频序列 $I(u,v,t)$，目标物体的初始状态 $\mathbf{x}_0$;

输出：目标物体的序列状态 $\mathbf{x}_{1:T}$。

For 时刻 t=1:T,

预测：$\mathbf{x}_{t-}^{\ i} = f(\mathbf{x}_{(t-1)+}^{\ i}) + \xi^i$，其中 $f(\cdot)$ 是目标运动模型，$\xi^i \sim N(0,\Sigma_{d_x})$。$w_{t-}^{\ i} = w_{(t-1)+}^{\ i}$。用粒子集合 $\{(x_{t-}^{\ i}, w_{t-}^{\ i}) \mid i = 1,...,N\}$ 表示先验分布 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$。

校验：$\mathbf{x}_{t+}^{\ i} = \mathbf{x}_{t-}^{\ i}$，$w_{t+}^{\ i} = p(\mathbf{z}_t \mid \mathbf{x}_t = \mathbf{x}_{t-}^{\ i}) \cdot w_{t-}^{\ i}$。用粒子集合 $\{(x_{t+}^{\ i}, w_{t+}^{\ i}) \mid i = 1,...,N\}$ 表示后验分布 $p(\mathbf{x}_t \mid \mathbf{z}_t)$。

重采样：将粒子权重归一化，有 $\sum_i w_{t+}^{\ i} = 1$。根据公式(2-16)计算 $\hat{N}_{eff}$。

如果 有 $\hat{N}_{eff} < T$，则从 $\{(x_{t+}^{\ i}, w_{t+}^{\ i}) \mid i = 1,...,N\}$ 中进行 N 次有放回的采样，然后将所有粒子的权重设为 1/N。

End

图 2-2 基于重采样的标准序列蒙特卡洛滤波算法

Fig.2-2 Flowchart of standard sequential Monte Carlo filtering algorithm with particle resampling

然而，重采样算法在解决粒子集合退化问题的同时，也带来了新的问题。首先，重采样限制了算法的并行性；另外，如果一个粒子具有很高的权值，则其将被采样多次，使新粒子集合存在相同粒子的多个克隆版本，从而影响粒子集合的多样化问题。该问题被称为“采样贫瘠”(sample impoverishment)，尤其在过程噪声较小的情况下比较严重。第三个问题是当粒子多样性受到限制的时候，给予粒子的估测将发生退化。解决该问题的一

种方法是假设粒子的状态变化是已由前向滤波器确定，然后在采用后向滤波器重新计算粒子的权值。另一种方法是采用基于马尔科夫链的蒙特卡洛模拟算法(Markov Chain Monte Carlo, MCMC)。

近些年来，针对采样贫瘠问题提出了一些解决技术。一种方法是resample-move 算法[66]。另一种广泛应用的策略是基于正则化的方法[67]。

2.3.5 弥补非最优提议分布缺陷的技术

跟踪过程中更常见的情况是无法获得最优提议分布。例如当采用状态转移模型 $p(\mathbf{x}_i \mid \mathbf{x}_{i-1})$ 相对于状态观测关联模型 $p(\mathbf{z}_k \mid \mathbf{x}_k)$ 具有平坦得多的分布形式，势必造成大多数的粒子具有较低的权值分布。部分学者提出了能使粒子运动到“适当区域”的方法。Clapp 和 Oudjane 都是在先验分布和相似性分布之间引入中间分布[68, 69]。粒子根据在这些中间分布上的权值进行重采样，从而“驱赶”粒子运动到特征空间中正确的区域。当 $p(\mathbf{z}_k \mid \mathbf{x}_k)$ 呈尖峰状分布但能够分解为诸多较平坦分布的时候，分区采样是有效的解决粒子权值过小问题的办法[49]。

2.4 小结

序列蒙特卡洛滤波算法虽然具有在解决复杂视觉跟踪问题上的通用性，然而，这种通用性也限定了算法从根本上说只是一个理论框架，即只指给出了解决视觉跟踪问题的总体策略。如何有效地应用序列蒙特卡洛滤波算法解决具体的视觉目标跟踪问题，仍然有很多应用相关的问题需要探讨。

从本章的分析中可以得出，应用序列蒙特滤波算法到具体问题，主要需要根据问题定义两个模型：编码了目标物体运动规律的状态转移模型 $p(\mathbf{x}_i \mid \mathbf{x}_{i-1})$ 和目标状态及其观测间的关联模型 $p(\mathbf{z} \mid \mathbf{x})$，该两个模型构成了序列蒙特卡洛滤波算法的核心，同时又是问题相关的。在本论文后面的章节中，针对人脸和人体跟踪中的具体问题，分别探讨了如何构造两种模型的策略和具体方法。

第3章 可区分性目标模型的动态构建

3.1 引言

目标表观建模是视觉跟踪问题性能的决定因素之一。构建目标表观模型主要需要解决两方面的问题：选择什么样的图像特征和如何在图像特征取值分布的基础上描述被跟踪的目标。视觉跟踪研究的实践表明：图像特征选择和基于图像特征的目标描述从根本上决定了跟踪算法的鲁棒性和实时性。虽然这一问题得到了领域内学者的极大重视和不懈努力，但针对大规模应用需求，其仍是视觉目标跟踪研究领域中面临的最困难问题之一。本章针对该问题，提出了一种全新的目标表观建模方法。图像像素虽然被广泛应用，但其易受各种噪声的干扰，这里将最近在目标检测领域取得成功的 Haar 小波特征引入到视觉跟踪领域。基于 Haar 小波特征，采用分类器组合的方式构建目标表观的可区分性模型。不同于描述性模型建模了目标表观“什么样”，提出的基于分类器组合的可区分性模型建模了目标/背景之间的差异性信息，从而从根本上保证了跟踪算法能够将目标从纷繁复杂的背景中区分开来。另外，注意到在采用序列蒙特卡洛滤波算法进行目标跟踪的过程中，由于序列蒙特卡洛滤波算法本身的性质，从而有大量的“背景”粒子的存在。在以往工作中，这些粒子被认为对最终结果贡献很小而被忽视。提出的算法则利用了“背景”粒子中蕴含的背景分布信息，从而为目标表观模型随背景的变化而实时更新提供了有效的方法。在基于图像区域的跟踪问题中，相对于人脸跟踪，人体跟踪更具有挑战性。在人体跟踪问题上实验了提出的算法。实验结果表明，相比于目前最具代表性的跟踪算法之一：Mean Shift[15]，提出的算法在公开的测试序列上取得了更好的跟踪效果。

3.2 问题的提出和相关工作

在以往工作中，主要是依据相关经验或者有限实验结果来进行图像特征的选择，并在选择的图像特征之上进行目标表观模型的构建。得到的模型被期望在给定视觉跟踪问题上能够满足两点要求：

- 能够正确地刻画目标表观，从而使目标表观和背景在模型层面上对算法是可区分的，在根本上保证算法的鲁棒性；

- 具有尽可能低的计算复杂度，满足实时性计算的要求。

但是，这两种性能之间本身就存在着矛盾。复杂的目标表观模型往往刻画了更多的目标表观细节，客观上带来更好的可区分性，但同时造成计算量的增加。实时性计算的要求则限定了可选择的图像特征的种类和数量，从而趋向于选择简单的目标模型。所以，目前普遍采用的图像特征，如直方图[15]，轮廓[70]，模板[7]，直方图和轮廓的结合[13] 及其他基于多线索的方法[71]等都是在可区分性和计算低复杂性之间的折衷。

在这些工作中，目标模型的构建过程在目标跟踪任务开始之前，而且这些预设的图像特征在整个跟踪过程中往往固定不变，其中隐含的假设是预先定义好的目标表观模型除了能满足实时计算要求外，总能够在目标和背景之间具有足够的可区分度。然而，实际情况表明，目标的运动和外界成像因素因素等总是造成目标和背景的同时变化。在单一固定的目标模型下，无法保证在模型层面上**总是**能够很好的区分前景和背景。例如，选用颜色直方图作为图像特征刻画目标表观，当目标运动到一个颜色分布和目标模型很相近的背景区域时，则在模型层面上无法区分目标和该背景区域，从而容易导致系统受到背景的干扰而丢失目标，引起系统的失效。所以，鲁棒的跟踪系统应该具有自适应地改变其内部所采用的目标表观模型的能力，根据背景变化去动态地选择简单有效的图像特征从而始终保证目标模型的在可区分情况下具有尽可能低的复杂性。

提出的算法从时间轴上前景/背景分类的角度来看待和解决视觉跟踪中的目标表观建模问题。从构造分类器的角度，目标的图像表观(以下简称目标表观)是唯一的正例集合而背景的图像表观(以下简称背景表观)是唯一的反例集合。在有限时间内考察背景图像，可以得到相对稳定的反例集合。这样，只要在该正例/反例集合上构建分类器作为目标的刻画模型，并且由于训练集和测试集的一致性，算法就具有在该时间段内正确区分目标和背景的能力。另一方面，在时间段之外，目标表观模型需要具有根据背景的变化自适应地发生改变的能力，去除失效的特征，加入具有强分类能力的特征，从而维护目标模型的可区分性和计算低复杂性。

在最近几年的成果中，已有工作将自适应地目标表观模型构建思想应用到跟踪问题中。在[16]中，H. Stern和B. Efros提出一种自适应地选择颜色空间构造直方图的方法。系统可以从多个色彩空间中根据其对前景和背景的可区分性动态地选择当前最优的色彩空间来构建人脸的直方图表达。Collins进一步扩展该方法到一般的图像目标跟踪中[17]。他也同样采用了直方图作

为图像特征。通过对RGB色彩空间中的三个分量赋予不同的权重进行组合，从而可以根据背景的变化依据目标模型可区分度从49种不同的直方图中进行选择。上述两种方法都采用直方图作为被跟踪目标的图像特征并能够得到比采用单一直方图作为目标模型进行跟踪的更好效果。然而，由于失去了空间信息，直方图的区分能力是有限的，当背景中存在色彩分布上和目标近似的区域时，直方图方法往往会由于出现目标混淆而失效。在[72]的工作中，Gabor特征被用来同时进行目标和背景的建模，并采用基于Gabor特征的分类器构建目标模型。虽然Gabor特征具有很好的分类能力，但其高计算复杂度对于跟踪这样具有实时要求的应用来说是不适合的。

在[16, 17, 72]的工作中，背景信息的重要性已经得到了实验结果的证明。

3.3 方法概述

本节将简述通过选择图像特征动态构建和更新目标模型，从而同时保证目标模型具有可区分性和低复杂性的算法。由于过完备 Haar 小波特征集在物体识别领域的卓越表现[73]，在这里引入 Haar 小波特征到视觉跟踪领域作为目标的图像特征刻画方法。系统的框架如图 3-1 所示。

初始化阶段：对背景的图像表观(以下简称背景表观)进行随机采样，将过完备 Haar 特征集在背景表观采样点上的特征值作为反例，在目标的图像表观(以下简称目标表观)上的特征值作为正例，然后利用 Fisher 判别准则对每个 Haar 特征分量的分类能力进行评价，再采用前向特征选择算法从过完备 Haar 特征集合中选择当前分类能力最强的特征子集作为目标的初始模型。

目标跟踪阶段：对每帧新图像数据，利用当前目标位置及其以往运动轨迹作为卡尔曼滤波器的输入，从而预测目标在下一帧中可能的出现位置，并进一步利用这一位置信息在当前帧上对其周围背景进行随机采样。之后，比较当前目标模型和下一帧目标可能出现区域的背景窗口之间的相似度。如果高度相似，说明在下一帧中跟踪目标可能与背景混淆，则需要重新选择特征改变目标模型，以降低目标模型与该区域背景窗口的相似度。同时，还需要计算目标模型和当前目标表观之间的匹配度，如果匹配度过低，这表明该模型对跟踪对象的描述能力不足，从而需要更新目标模型提高目标模型和目标表观之间的匹配度。然后对下一帧图像的目标继续跟踪。

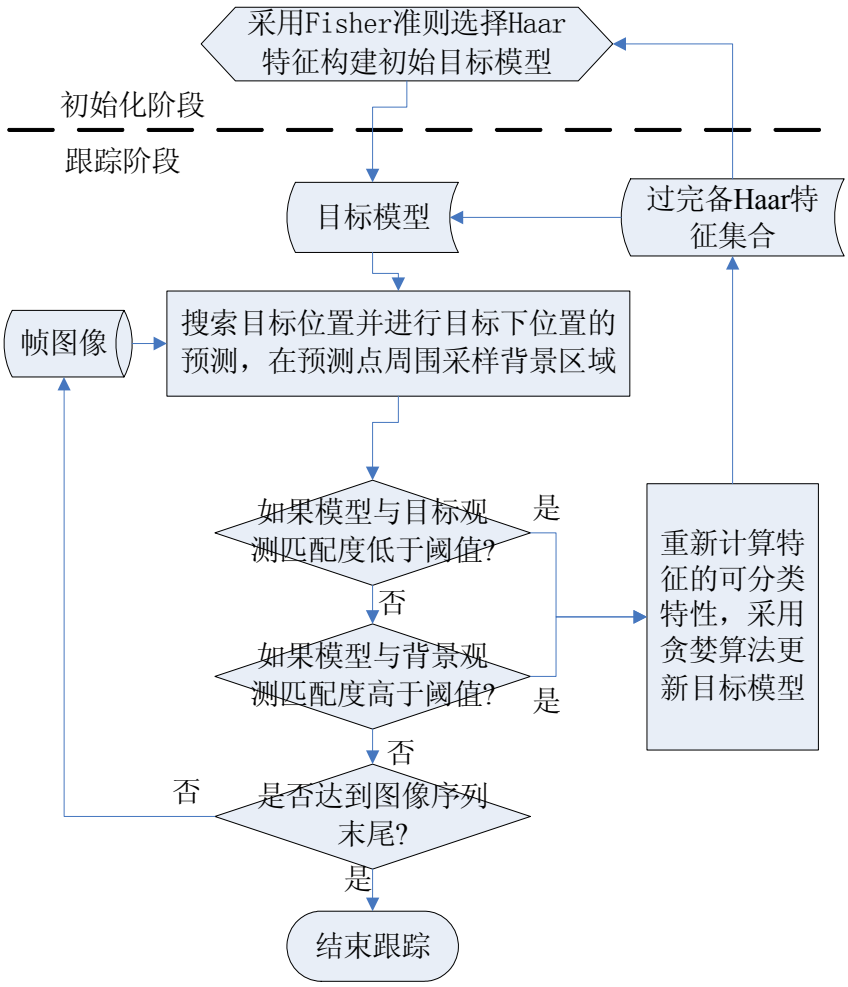


图 3-1 嵌入在线特征选择过程的目标跟踪方法流程图

Fig.3-1 Flowchart of object tracking method with online feature selection embeding

3.4 基于目标/背景差异信息的特征选择

本节首先描述所采用的图像特征，进而给出如何根据当前目标/背景差异信息选择特征来构建目标模型的方法。

3.4.1 特征集合

采用过完备 Haar 小波基特征集合作为图像特征来构建目标模型。在

Viola 的工作中[73]， 基于过完备 Haar 特征集合的分类器在物体识别问题上的优异性能得到了充分的表现，而且，积分图方法的引入使得 Haar 特征值的求取能够满足实时计算的要求。与跟踪研究中普遍采用的直方图，轮廓等特征相比，Haar 特征不仅有更强的表现能力，同时 Haar 特征还提供了多层次刻画目标的能力。

实现中采用了三种 Haar 特征(见图 3-2)。每一类 Haar 特征有四个控制参数：特征在所刻画图像区域中的位置(x, y)，特征的宽度和高度(w, h)。通过四个参数的不同组合，可以得到数以万计的特征。由于这些特征是对所刻画图像区域的过完备描述，所以所有不同特征的总集合称为过完备 Haar 特征集合。为了后面叙述的方便，用 $F=\{f_i \mid i=1,...,N\}$ 表示特征集合，用 $V(z)=\{v_i(z) \mid i=1,...,N\}$ 表示图像窗口 z 上所求得的特征向量。

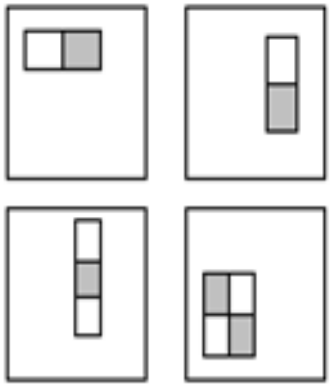


图 3-2 当前跟踪系统中所采用的 Haar 特征

Fig.3-2 Haar Features used in current system implementation

3.4.2 目标建模

对每个 Haar 特征，构造相应的弱分类器。考虑到计算复杂性的约束，采用如图 3-3 所示的弱分类器:

$$h_i(z)=\begin{cases}1 & if\ v_i(z)\in[\overline{v}_i^{pos}-\theta_i,\overline{v}_i^{pos}+\theta_i]\\ 0 & otherwise\end{cases} \tag{3-1}$$

其中， $\overline{v}_i^{pos}=\overline{v}_i(z^{pos})$ 是目标表观上 Haar 特征在特征分量 f_i 上的均值， θ_i 是分类器阈值， $v_i(z)$ 是图像窗口 z 上 Haar 特征在特征分量 f_i 上的特征值。

最后根据目标模型内所有弱分类器的投票输出，系统判断该窗口为目标表观或者背景表观。

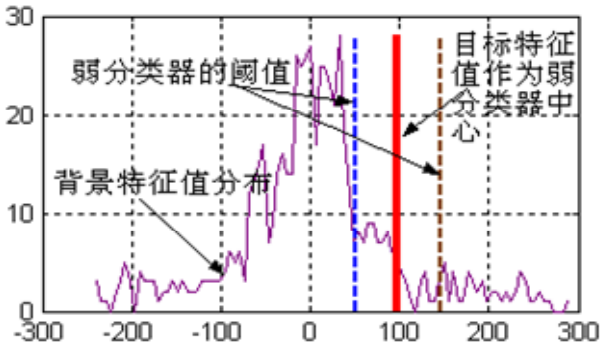


图 3-3 弱分类器及其与前景背景特征值分布的关系

Fig.3-3 Weak classifier and its relation to distrition of foreground/background feature values

虽然很多研究者的实验验证了 AdaBoost 是将若干弱分类器组合成一个强分类器的有效方法，但 AdaBoost 训练非常耗时，因而难以用于在线的特征选择，于是提出了一种结合贪心和 Fisher 准则的快速特征选择方法：

步骤 1：采用面积滤波和信息量度量初步选择特征。对一个 24x24 的基准窗口来说，可以利用的 Haar 特征数目有几万个，如此众多的特征数量保证了 Haar 特征集的刻画能力，但也同时存在特征冗余。首先删除那些计算面积小于一定阈值(在实验中，阈值等于 16)的特征。然后对特征进行空间上的等间距采样。在目标表观上求取的特征值，通常较大的特征值对应像素变化大的区域，包含了更多的纹理信息。据此，对当前余下的特征，以较大的概率对特征值比较大的特征进行采样进一步缩减原始特征集合中的特征个数。经过上面步骤，特征集中大概留有 4000 个左右的 Haar 特征。

步骤 2：通过对目标周围的背景进行采样选择可区分性好的特征。在目标周围的图像上通过均匀分布做 K 个采样，其中 K-L 个图像窗口对应背景表观，L 个图像窗口对应目标表观(图 3-4 演示了部分采样过程)。于是，对每一个特征，可以得到它的 L 个正例和 K-L 个反例。Fisher 线性判别准则被用来对每一个特征的分类性能进行评价[74, 75]：

$$R_i = \frac{|\overline{v}_i^{pos}(f_i) - \overline{v}_i^{neg}(f_i)|^2}{S^{pos}(f_i) + S^{neg}(f_i)} \tag{3-2}$$

其中，$\overline{v}_i^{pos}(f_i)$ 和 $\overline{v}_i^{neg}(f_i)$ 分别是正例集合和反例集合在特征分量 f_i 上的均值。$S^{neg}(f_i) = \sum_{m=1}^{K-p}((v_i^{neg}(f_i) - \overline{v}_i^{neg}(f_i))^2$ 和 $S^{pos}(f_i) = \sum_{m=1}^{p}((v_i^{pos}(f_i) - \overline{v}_i^{pos}(f_i))^2$ 分别是正例集合和反例集合的散度。

根据这些特征的分类能力由强到弱对特征分量降序排列，顺序选择前 M 个特征相对应的弱分类器加入目标模型，直到满足(3-3)式：

$$(\sum_{m=1}^{M} R_m) > T \tag{3-3}$$

其中 T 是保证弱分类器组合后分类能力的阈值(在实验中取 T=0.45)。通过这些弱分类器的线性组合，得到一个刻画目标的强分类器 $H(t,M) = \{h_m \mid m = 1,...,M\}$，即目标模型，其中 h_m 是公式 (3-1)中定义的弱分类器，t 表明这一模型从 t 时刻开始有效。

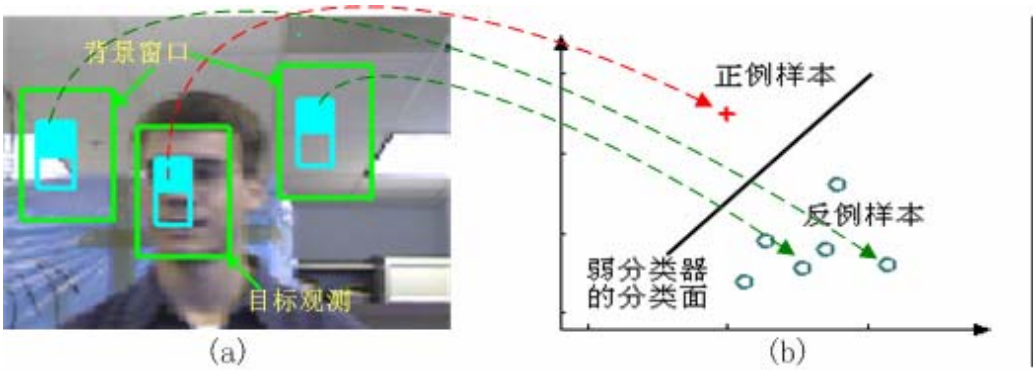
3.5 跟踪过程中目标模型的动态更新

在跟踪过程中，需要针对背景的变化和目标自身的变化不断调整目标模型。为了后面叙述方便，首先定义当前图像表观窗口和目标模型之间的相似性如(3-4)式：

$$w(z) = \frac{\sum_{m=1}^{M} h_m(z)}{M} \tag{3-4}$$

对目标模型 $H(t,M)$，可以将其分成两个子集合：$H_1(t,M_1)$ 和 $H_2(t,M_2)$。

其中 $H_2(t,M_2)=\{h_m(z^{pos})==0\,|\,m=1,...,M_2\}$ 包含那些在目标表观上输出为零的弱分类器， 而 $H_1(t,M_1)=H(t,M)/H_2(t,M_2)$ 。



(a)背景和目标表观采样过程及特征值求取示例　(b) 通过采样获得的正反例样本构造一个弱分类器

图 3-4 表观采样和弱分类器训练示例

Fig.3-4 Illustration of observation sampling and classifier training

3.5.1 维护目标/背景差异性的模型更新

跟踪过程中，如果在预测的下一帧候选目标位置附近的背景表观窗口和当前目标表观在当前目标模型上的相似性(利用公式(3-4)计算)很接近，则容易造成目标歧义而使系统丢失目标。因此，需要更新模型，以区别未来的目标和背景表观，从而避免目标歧义现象的发生。

首先找到目标在 t 时刻的位置 $\mathbf{x}(t)$。然后利用贝叶斯滤波的过程模型(Process Model)预测目标在 $t+1$ 时刻的候选位置 $\tilde{x}(t+1)$ 。之后，以 $\tilde{x}(t+1)$ 为中心，在其周围均匀采样背景表观中的 K 个窗口，对每一个采样窗口 z_k^{neg}，利用公式(4)计算其与当前目标模型的相似度 $w(z_k^{neg})$ 。进而计算

$$\Delta w_k = |\, w(z^{pos}) - w(z_k^{neg}) \,| \tag{3-5}$$

其中 $w(z^{pos})$ 表示目标模型在目标表观上的相似度。如果有 $\Delta w_k < T$ (实验中，定义 T=0.17)， 则 z_k^{neg} 所对应的背景表观可能造成目标的混淆，称这样的背景表观为“威胁表观”。

为避免当目标模型和目标表观的匹配值突然下降时，系统因为威胁表观的存在而丢失目标，收集所有“威胁表观”的特征作为反例来重新选择特征以改变目标模型，使改变后的目标模型与“威胁表观”之间的相似度降低，从而避免目标混淆的出现。

模型更新的步骤如下：首先从 $H(t,M)$ 中丢弃模型子集 $H_2(t,M_2)$。然后根据每个特征的可分类性从小到大地从模型中移走当前最弱的子模型，直到只剩下的子模型数量降至原来总数量的 70%。然后将前面收集到的“威胁表观”上得到的 Haar 特征值作为反例，根据第 4.4 节描述的特征选择方法进行重新选择特征更新目标模型。

3.5.2 维护目标描述一致性的更新

在某些情况下，仅仅更新目标模型的参数(如弱分类器的阈值)无法反映由于目标的运动而引起的目标表观的剧烈变化(比如，目标的拓扑结构发生改变，剧烈的光照变化等等)。另一方面随着时间的流逝，系统的参数会由于累计误差发生偏移，造成目标模型与目标表观间匹配度的下降。这时需要重新选择合适的特征对目标模型进行调整。

假设在时刻 t，如果目标模型在目标表观上的匹配度 $\bar{w}(z^{pos}) = \frac{1}{p}\sum w(z^{pos}) < T$ (实验中 T 设为 0.75)，表明当前的目标模型不能合理解释目标表观的变化从而引起目标模型与目标表观的偏离。于是，回溯到前一帧，从 $H(t-1,M)$ 中抛弃特征子集合 $H_2(t-1,M_2)$ 并从剩余的原始 Haar 特征集中重新选择 M_2 个最显著的特征加入目标模型中。

3.6 采用动态建模的目标跟踪

将前面提出的动态目标建模方法应用于目标跟踪，这里选择卡尔曼滤波作为跟踪的一般框架，构成一个完整的跟踪模块[76]。在实验中，定义卡尔曼滤波的状态为 $\mathbf{x}(t)=\begin{bmatrix} x \\ y \\ dx \\ dy \end{bmatrix}$，表观为 $\mathbf{y}(t)=\begin{bmatrix} x \\ y \end{bmatrix}$，其中 x 和 dx 分别是目标图像在水平方向的位置和运动速度，y 和 dy 分别是目标图像在垂直方向的位置和运动速度。跟踪的基本过程如下：

1. **初始化阶段：**在参考图像中定义目标表观 z^{pos} 以及目标的初始状态 $\mathbf{x}(0)$。采用卡尔曼滤波的过程模型预测目标的可能出现区域：

$$\tilde{\mathbf{x}}(t)=\mathbf{A}\mathbf{x}(t-1)+\mathbf{w}(t) \tag{3-6}$$

其中 $\mathbf{A}=\begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$，$P(\mathbf{w})\sim N(0,Q)$，$Q=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$，$\Delta t=1$。在 $\tilde{\mathbf{x}}(1)$ 所对应的背景表观周围，采用均匀分布进行背景采样，得到 K 个背景表观 $\{z_k^{neg}\mid k=1,...,K\}$。采用第 2 节描述的特征选择方法从过完备 Haar 特征集合中构建初始目标模型 $H(0,M)$ 。

2. **目标表观提取和特征更新过程：**采用当前目标模型 $H(t,M)$ 在卡尔曼滤波的预测区域内搜索目标的图像表观 $z^{pos}(t)$。根据 $z^{pos}(t)$ 更新目标状态的估计 $\mathbf{x}(t)=\tilde{\mathbf{x}}(t)+K(t)(\mathbf{y}(t)-\mathbf{H}\tilde{\mathbf{x}}(t))$，其中 $\mathbf{H}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$，$K(t)$是跟踪过程中动态变化的卡尔曼增益。
3. **目标位置预测过程：**利用方程(3-6)对下一帧的目标可能出现区域进行预测，得到 $\tilde{\mathbf{x}}(t+1)$。
4. **根据背景进行目标模型的更新：**采用 4.1 节所述的方法，搜索在 $\tilde{\mathbf{x}}(t+1)$ 附近的“威胁表观”。如果存在，则重新选择特征更新 $H(t,M)$，从而保证不会出现目标歧义的现象。
5. **根据目标表面呈现的变化进行目标模型的更新：**采用 4.2 节所述的

方法，如果目标模型和目标表观之间的相似度 $\overline{w}(z^{pos}) < T$,重新选择特征更新 $H(t,M)$ 保证目标模型和目标表观之间的相似度，在新目标模型的基础上进行后续帧的目标跟踪。

3.7 将模型更新融入序列蒙特卡洛滤波算法

3.7.1 背景粒子的存在

在序列蒙特卡洛滤波框架中，采用一组粒子及其相应权重的集合 $\{x_i^k, w_i^k \mid k = 1,...,N\}$ 来对目标状态的后验分布进行逼近。然后根据后验分布，求取某种意义下的最优跟踪解，即目标的运动状态。在绝大多数场合中，应用关心的是后验分布的期望。在用粒子逼近的后验分布中，期望表示为

$$E(p(\mathbf{x}_i \mid \mathbf{z}_{1:i})) = \frac{\sum_k w_i^k \mathbf{x}_i^k}{\sum_k w_i^k} \tag{3-7}$$

从公式(3-7)中可以看到，粒子对期望的贡献与其权重成正比。然而，在采用序列蒙特卡洛算法过程中，不可避免地要出现很多具有很小权重的粒子，这些粒子的出现主要来源于以下三方面的因素：

1) 采用非最优的运动模型：粒子的运动可以看作在状态空间中对目标状态在时间轴上变化的一种追随。由于目标未来运动的不可预知性，无法保证从无数的运动模型中找到最好的模型，来确保粒子传递的效率。这种非最优的运动模型客观上使部分粒子远离目标的真实状态，从而具有较小的权重。
2) 序列蒙特卡洛算法本身的随机特性：蒙特卡洛方法本身是随机模拟的意思。虽然这种随机模拟的特性使算法在求解过程中具有对局部最小等问题的鲁棒性。另一方面，也造成了很多粒子位于图像中的背景区域，从而具有较小的粒子权重。
3) 跟踪中为保证算法的鲁棒性，对粒子的多样性要求：由于序列蒙特卡洛滤波算法的基本策略是采用粒子群来模拟目标状态的后验概率而不是目标状态的期望。所以，保持粒子的多样性是应用序列蒙特卡洛算法的一个基本要求，从而要求有很多粒子位于背景区域。

将位于背景区域的粒子称为“背景粒子”，在视觉跟踪任务中，由于这些粒子不含有目标状态的信息，所以其权重都是比较小的(如果其权重比较大反而会造成较大的估计误差)。

在以往工作中，这些粒子由于对最终结果贡献很小而被认为用途而小。然而，这些粒子虽然不含有目标状态的信息，然而却含有关于背景的信息，实在序列蒙特卡洛算法跟踪过程中对背景的采样，并且这些采样早已存在，只是还没有被利用而已。

所以，将前面提出的目标模型自适应构建和更新过程有机地嵌入到序列蒙特卡洛算法中，由于背景粒子的存在，使得模型的更新过程只需要很少的额外计算量，从而大大提高了算法的效率。

3.7.2 融入自适应目标模型更新的跟踪算法

假设当前目标模型为 $H(t,M)=\{h_m \mid m=1,...,M\}$，其中 h_m 是公式(3-1)定义的弱分类器。

根据背景变化重新选择特征

在 t 时刻，采用序列蒙特卡洛滤波算法的相似性模型对采样点进行评估之后，得到权重更新的粒子集合 $PT(t)=\{x_t^k, w_t^k \mid k=0,...,K\}$。然后，根据预测状态 x_t^k 和其相应的权重 w_t^k，采用 K 均值算法对 $PT(t)$ 进行聚类分析。$\{cp_c \mid c=1,...,C\}$ 是聚类得到的结果，其中 cp_c 对应集合 $PT(t)$ 的一个子集，并且有 $\forall i,j\in[1,C], i\neq j, \exists cp_i \cap cp_j=\varnothing$，$cp_1 \cup ... \cup cp_C = PT(t)$ (C=1 表示后验概率单峰的特殊情况)。用 $x(cp_c)$ 表示子集合 cp_c 中具有最大权重的粒子，$w(cp_c)$ 表示相应的权重。

按照 $w(cp_c)\in\{w(cp_c) \mid c=1,...C\}$ 的降序排列，根据目标的历史运动轨

迹，依次考察 $x(cp_c)\in\{x(cp_c)\,|\,c=1,...,C\}$ 是否具有与预测运动相符合的运动形式。假设 $x(cp_{c1})$ 是第一个符合运动约束的粒子，则用 cp^{pos} 表示该粒子所在的子集合。于是，目标在时刻 t 的状态采用子集合 cp^{pos} 中的粒子根据公式(3-7)进行计算。而其他子集合中的粒子则认为其对应于背景区域。用 $CP^{neg}(t)=\{cp_c^{neg}\,|\,c=1,...,C-1\}$ 表示这些“背景”子集合。

对每一个背景子集合 cp_c^{neg}，计算

$$\Delta w_c=|\,w(cp^{pos})-w(cp_c^{neg})\,| \tag{3-8}$$

如果 Δw_c 低于某阈值 T(实验中 T=0.17)，则该 $x(cp_c)$ 所对应的背景区域被标记为“威胁”区域，即当目标模型在目标表观上的匹配度存在波动时可能造成目标混淆的背景区域。

所有从威胁区域收集的 Haar 特征值作为背景采样点成为重新选择特征调整目标模型的依据。首先，从目标模型 $H(t,M)$ 中去除弱分类器子集合 $H_2(t,M_2)$。然后，根据分类能力的降序排列，从弱分类器子集合 $H(t,M)$-$H_2(t,M_2)$ 中顺序剔除弱分类器直到满足目标模型中的弱分类器个数是原始目标模型 $H(t,M)$ 的 70%。再根据第二节中描述的基于背景信息的特征选择方法，从原始 Haar 特征集合中选择等同数量的特征来补充剔除的特征。不同之处在于，在跟踪过程中，目标模型的调整不是依据所有的粒子采样的背景区域，而是只将前面分辨出的威胁区域作为更新模型的依据和反例。

根据前景剧烈变化重新选择特征

在时刻 t，通过公式(3-4)计算目标模型和目标表观之间的匹配度，如果匹配度的均值 $\bar{w}(z^{pos})=\frac{1}{p}\sum w(z^{pos})$ 低于某阈值(实验中设为 0.75)，则表明由于目标运动而造成的目标表观变化已不能很好地用当前目标模型进行解释。于

是，首先从目标模型$H(t,M)$中去除弱分类器子集合$H_2(t,M_2)$。然后根据第二节中描述的特征选择策略从原始 Haar 特征集合中重新选择$\|H_2(t,M_2)\|$个新特征。

3.8 实验结果

在人体跟踪和汽车跟踪两类问题上测试了所提出的跟踪算法的性能。这是由于人体和汽车跟踪问题仍然没有完美的解决方案，并且该跟踪任务具有广阔的商业应用前景。在这里给出了一些具有代表性的跟踪结果并给出分析。

将提出的算法与两个代表性算法进行了比较：Mean Shift (MS)跟踪算法和 Mean Shift Ratio (MSR)跟踪算法[15, 17]。MS 跟踪算法是当前最好的图像块跟踪算法之一，其基于 mean shift 模式搜索算法，采用基于核函数的直方图进行前景匹配。MSR 跟踪算法则扩展了标准 mean shift 算法。通过对 RGB 颜色空间中三个色彩通道的线性组合，在算法中内嵌了 49 种不同的直方图。在跟踪过程中，算法根据每个直方图对目标和背景的区分度动态选择其一。

在当前算法的实现中，没有加入任何图像前处理的模块。通常构建目标模型的特征在 120-180 之间。在粒子滤波的实现中使用了 1089 个粒子。在未作特殊优化的情况下，系统在 PentiumIII 1.5GHz PC 上的运行速度大概为 10Hz。

算法的复杂度分析如下：算法的主要计算量发生在对图像子窗口和模型相似度的求取和搜索过程中。对每一个 Haar 特征的求取，平均需要计算 6 次加法，则对一个图像子窗口，平均需要的计算量为 6X(120+180)/2=900。对视频中的一针图像，搜索目标位置所需计算量为 1089X900=9.8e+5。而对 1.5GHz 的 PC 来说，假设机器每时钟周期可以计算一次加法，理论上其一秒钟大概进行 1.5e+9 次加法运算，所以，提出的算法在只占用部分 CPU 资源的情况下完全可以满足跟踪任务中实时计算的要求。

3.8.1 人体跟踪实验

首先在人体跟踪问题上检验算法性能。所采用的公开测试序列可以在

[77]得到。这里给出的两个测试序列在该视频监控视频序列集合中分别称为“OneStopNoEnter1cor.mpg”和“ThreePastShop2cor.mpg”(分别用序列 A 和 B 表示)。序列 A 有 725 帧，图像分辨率为 384×288。序列 A 中被跟踪人体本身具有较大的变化，如部分遮挡，视角、尺度的变化等。在原始手工标注文件中，共有人体区域、凝视方向、头、手、足和肩部等状态信息（见图 3-5）。根据算法测试的要求，采用了人体区域的状态信息：包括重心位置、高度和宽度共四维特征。

图 3-5 图像标注信息的示例

Fig.3-5 One example of image annotaton information

图 3-6 到 3-8 分别给出了序列 A 上 MS 跟踪算法、MSR 跟踪算法和提出的算法的部分示例跟踪结果。从图 3-6 中可以看到，MS 跟踪算法由于只使用了直方图刻画人体，最终被背景中的人体所吸引，丢失了跟踪目标。而在利用了背景信息之后，MSR 跟踪算法能够始终跟踪人体的大致位置，但由于 MSR 算法采用直方图进行目标建模，使算法对尺度变化的定位能力较弱，所以对目标区域的刻画并不正确(图 3-7)。而提出的算法则能够始终正确地找到目标区域(图 3-8)。

图 3-9 到 3-11 则分别给出了序列 A 上 MS 跟踪算法、MSR 跟踪算法和提出的算法的跟踪误差定量分析。由于是标准测试序列，序列中被跟踪对象的状态真值可以在[77]获得。在三幅图中，从图 3-9 中可以看到，由于背景

的色彩分布的影响，MS 跟踪算法很快开始偏离目标，定位在错误的背景区域上。由于 MSR 算法在背景采样的基础上，通过计算前景和背景在不同色彩空间上的相似性，从而在所有可能的色彩空间组合中可以选择最具区分性的表示方法，从而在实验结果上也表现出了更好的鲁棒性（见图 3-10）。但由于该算法采用直方图作为物体的图像特征描述方式，虽然在直方图的基础上，采用了核函数来增强对物理空间色彩变化的刻画程度，但从本质上说，改进的直方图仍然大量丢失了色彩分布在物理空间上的信息，所以在尺度定位上，MSR 算法的定位还是表现出了较大的跟踪误差。提出的算法在目标人体的重心定位和尺度定位上都比 MS 和 MSR 算法具有相对小的误差(图 3-11)。

图 3-6 MS 算法在人体序列 A 上的部分跟踪结果示例

Fig.3-6 Some tracking results on people sequence A by the MS tracker

图 3-7 MSR 算法在人体序列 A 上的部分跟踪结果示例

Fig.3-7 Some tracking results on people sequence A by the MSR tracker

图 3-8 提出的算法在人体序列 A 上的部分跟踪结果示例

Fig.3-8 Some results on people sequence A by the proposed tracker

图 3-9 MS 跟踪算法在人体序列 A 上的平均跟踪误差

Fig.3-9 Mean trakcing errors on sequence A by the MS tracker

图 3-10 MSR 跟踪算法在人体序列 A 上的平均跟踪误差

Fig.3-10 Mean tracking errors on sequence A by the MSR tracker

图 3-11 提出的算法在人体序列 A 上的平均跟踪误差

Fig.3-11 Mean tracking errors on sequence A by the proposed tracker

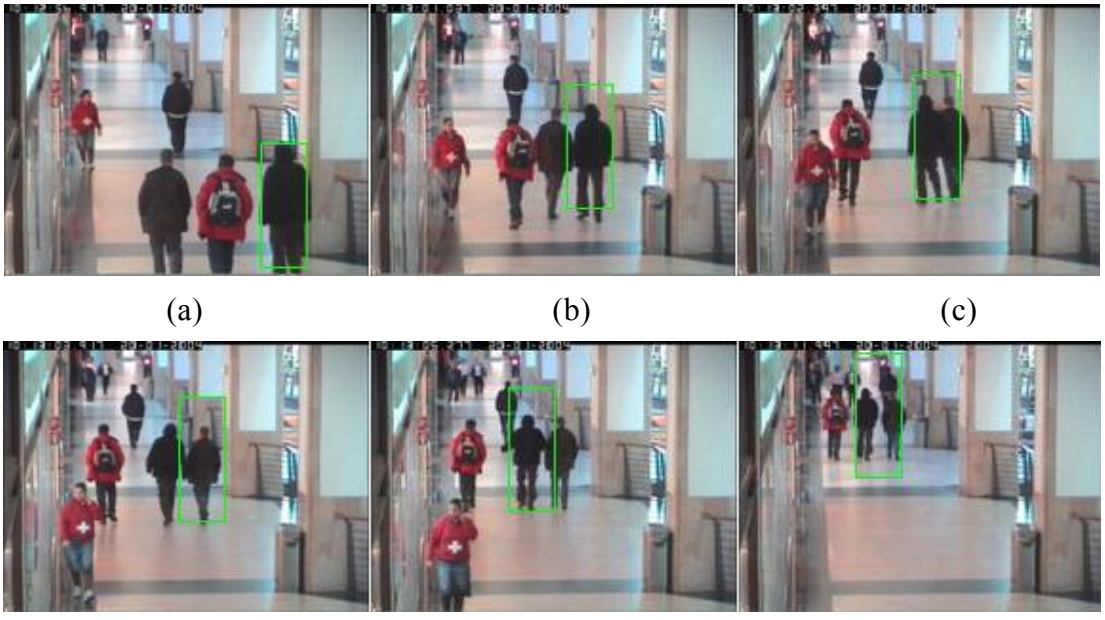
为了检验三种算法的鲁棒型，通过随机采样的方法改变算法的初始状态条件，令算法的初始输入在标注值的 2%增量之内变化。以变化的初始条件为输入，分别对三种算法进行了 100 次重复的实验，表 3-1 给出了三种算法的平均误差和平均方差。

表 3-1 MS, MSR 和提出的算法在平均跟踪误差上的比较

Table3-1 Comparing mean tracking errors among MS, MSR and proposed algorithms

	MS(单位: 像素)	MSR(单位: 像素)	提出的算法(单位: 像素)
水平方向平均跟踪误差	28.3	11.3	3.6
水平方向跟踪误差方差	9.2	4.3	1.3
垂直方向平均跟踪误差	30.2	4.6	2.4
垂直方向跟踪误差方差	7.7	2.9	1.2
水平尺度平均跟踪误差	9.7	7.3	3.3
水平尺度跟踪误差方差	4.8	4.1	2.4
垂直尺度平均跟踪误差	13.6	19.4	3.8
垂直尺度跟踪误差方差	5.7	6.2	1.9

序列 B 有 1527 帧和 384×288 的图像分辨率。相对于序列 A，在序列 B 中的目标人体周围始终有两个与目标非常相似的人和他同行(注意提出的算法的输入是灰度级图象)。而且在行进过程中，三人之间互相交换了彼此之间的位置（ http://www.jdl.ac.cn/user/jywang/projects/HaarTracking.htm 上有完整的视频演示)。在序列 B 上用相同的初始条件测试了 MS 跟踪算法、MSR 跟踪算法和提出的算法。部分示例跟踪结果分别显示在图 3-8 到 3-10 中。从图 3-8 中可以看到，当两个人接近的时候(图 3-8(b))，MS 算法无法判断那个是正确的目标从而将目标区域在两人之间不断切换（图 3-8(c)(d)(e))。图 3-9 给出了 MSR 算法的跟踪结果。可以看到，虽然 MSR 算法可以正确地找到目标位置(图 3-9(c)(d))，但仍然是尺度问题使其无法找到正确的目标区域(图 3-9(e))。图 3-10 给出了提出算法的跟踪结果，从图中可以看出，算法可以将目标和其他两个人区分开，很好地定位目标区域。通过图 3-11 可以看出在线特征选择过程在区分目标和背景方面所发挥的作用。图中给出了以目标为中心的 128X128 区域中目标模型和图像区域之间的相似度。图中 G 色彩通道的像素值大小和相似度成正比。

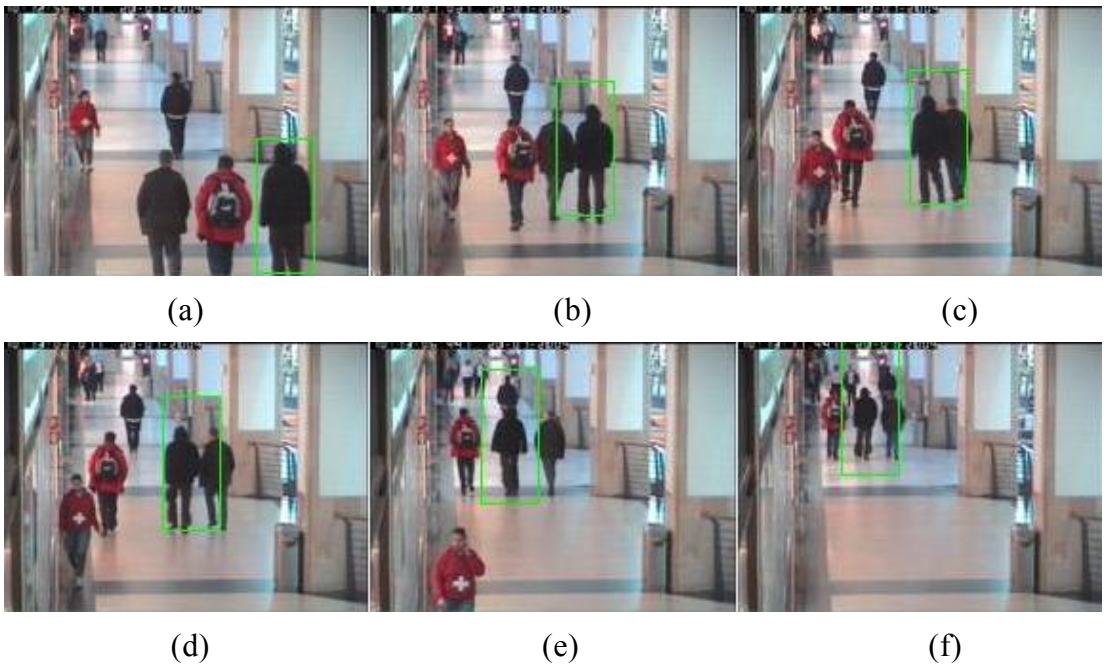


(a) (b) (c)

(d) (e) (f)

图 3-12 MS 算法在人体序列 B 上的部分跟踪结果示例

Fig.3-12 Some tracking results on people sequence B by the MS tracker

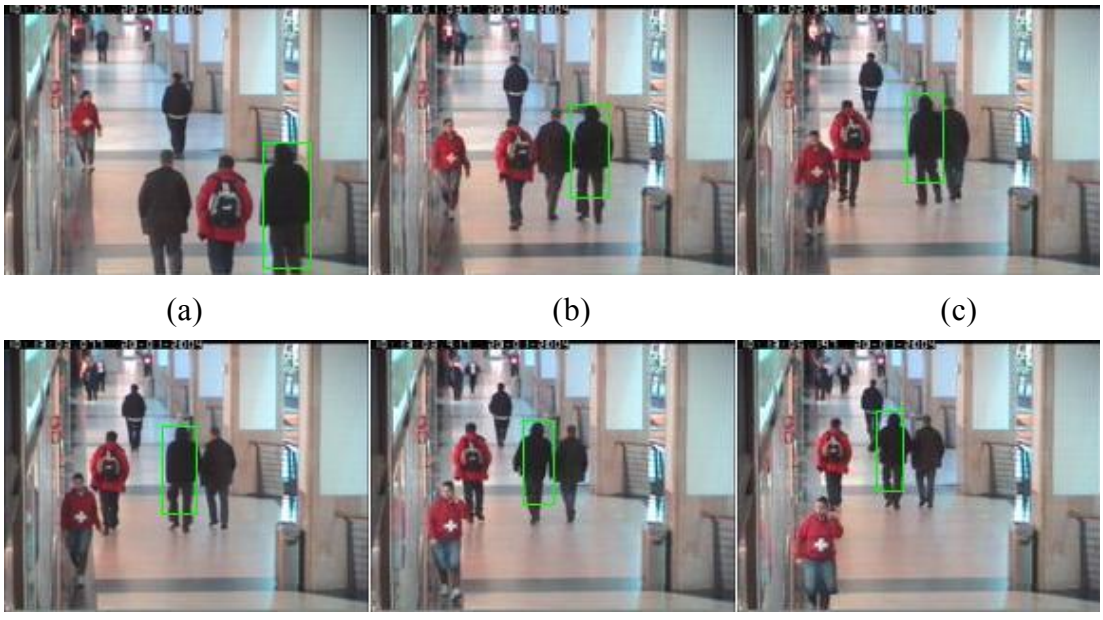


(a) (b) (c)

(d) (e) (f)

图 3-13 MSR 算法在人体序列 B 上的部分跟踪结果示例

Fig.3-13 Some tracking results on people sequence B by the MSR tracker



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

图 3-14 提出的算法在人体序列 B 上的部分跟踪结果示例

Fig.3-14 Some tracking results on people sequence B by the proposed tracker

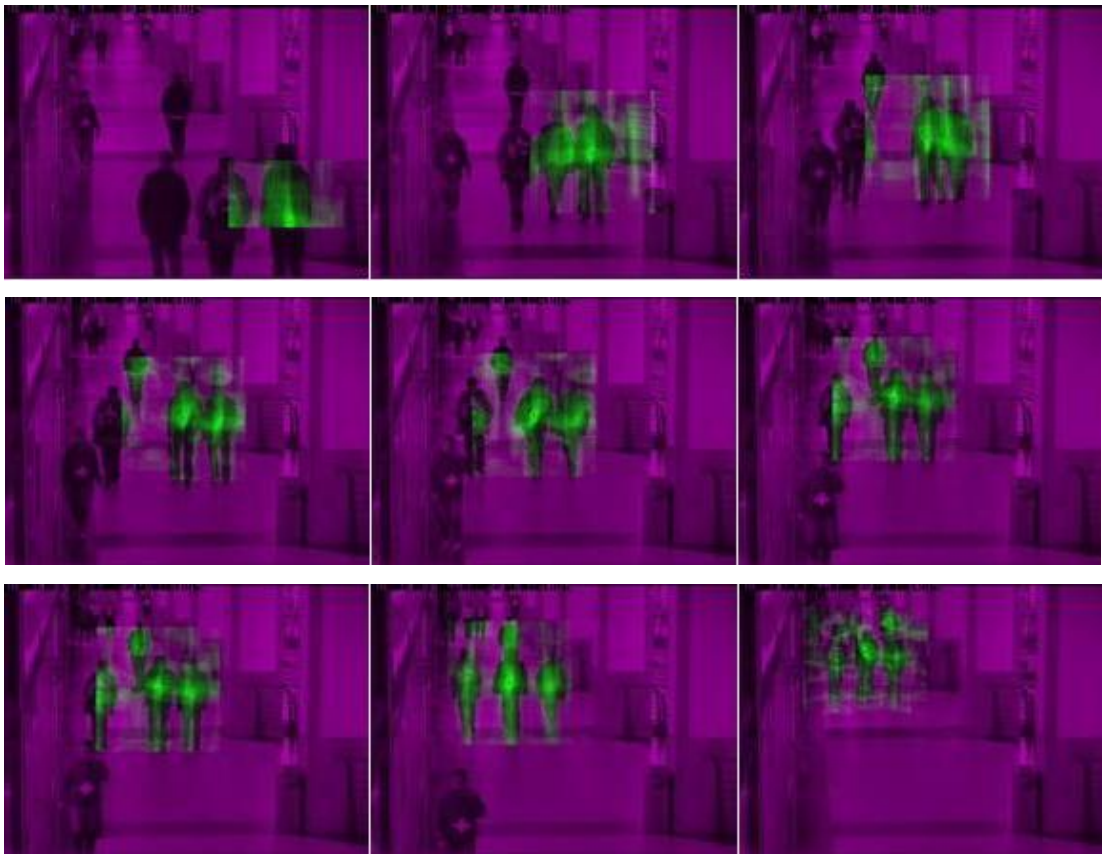


图 3-15 提出的算法在人体序列 B 上跟踪性能的定量分析

Fig.3-15 Quantitativly analyzing the performance of the proposed tracker on people sequence B

图 3-16，3-17 和 3-18 分别给出了MS跟踪算法，MSR跟踪算法和提出的算法在人体序列B上的跟踪误差定量分析。由于该序列同样是标准测试序列，序列B中被跟踪对象的状态真值可以通过http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/获得。相对于序列A，序列B的背景更为复杂多变。从图 3-16 中可以看到，由于目标人体周围行人的影响，使算法对目标区域的判断发生误差，从而在正确的人体目标和行人区域游移。由于MSR跟踪算法由于考虑了背景信息，从而在实验结果上表现出了更好的鲁棒性（见图 3-17）。但在尺度定位上，由于直方图的描述能力，MSR算法的定位还是表现出了较大的跟踪误差。提出的算法在目标人体的重心定位和尺度定位上都比MS和MSR算法具有相对小的误差。

为了检验三种算法的鲁棒型，同样通过随机采样的方法改变算法的初始状态条件，令算法的初始输入在标注值的 2%增量之内变化。以变化的初始条件为输入，分别对三种算法进行了 100 次重复的实验，表 3-2 给出了三种算法的平均误差和平均方差比较。

图 3-16 MS 跟踪算法在人体序列 B 上的平均跟踪误差

Fig.3-16 Mean tracking errors on sequence B by the MS tracker

图 3-17 MSR 跟踪算法在人体序列 B 上的平均跟踪误差

Fig.3-17 Mean tracking errors on sequence B by the MSR tracker

图 3-18 提出的算法在人体序列 B 上的平均跟踪误差

Fig.3-18 Mean tracking errors on sequence B by the proposed tracker

表 3-2 在序列 B 上 MS，MSR 和提出的算法平均跟踪误差的比较

Table3-2 Comparing mean tracking errors on sequence B among MS, MSR and proposed algrorithms

	MS(单位: 像素)	MSR(单位: 像素)	提出的算法(单位: 像素)
水平方向平均跟踪误差	4.7	4.6	1.8
水平方向跟踪误差方差	1.9	1.3	1.3
垂直方向平均跟踪误差	7.8	7.9	2.6
垂直方向跟踪误差方差	1.3	0.9	1.2
水平尺度平均跟踪误差	6.4	9.1	2.3
水平尺度跟踪误差方差	1.5	1.3	1.4
垂直尺度平均跟踪误差	14.7	13.4	3.1
垂直尺度跟踪误差方差	4.1	2.2	0.9

3.8.2 汽车跟踪实验

实验二针对复杂场景中的汽车跟踪任务。这里给出的测试序列是采用家用摄像机拍摄的。序列具有 3309 帧和 320×240 的图像分辨率。由于是在室外环境中，除了背景干扰外，光照，遮挡等很多不确定因素需要考虑。图 3-19 到 3-21 分别给出了 MS 算法、MSR 算法和提出算法的部分跟踪结果示例。MS 跟踪算法和 MSR 跟踪算法都很快丢失了目标，而提出的算法能完成该长序列跟踪任务。

图 3-19 MS 算法在汽车序列上的部分跟踪结果示例

Fig.3-19 Some tracking results by the MS tracker on car sequence

图 3-20 MSR 算法在汽车序列上的部分跟踪结果示例

Fig.3-20 Some tracking results by the MSR tracker on car sequence

图 3-21 提出算法在汽车序列上的部分跟踪结果示例

Fig.3-21 Some tracking results by the proposed tracker on car sequence

图 3-22，3-23 和 3-24 分别给出了 MS 跟踪算法，MSR 跟踪算法和提出的算法在汽车序列上的跟踪结果。序列中被跟踪对象的状态采用手工标注，共有重心位置、高度和宽度四维特征。相对于人体跟踪序列中相对简单的室内背景，汽车跟踪序列中的室外背景更为复杂。从图 3-22 可以得出结论，MS 跟踪算法很快丢失了跟踪目标。在本跟踪序列中，MSR 跟踪算法同样被背景吸引，丢失了目标。这是由于 MSR 利用背景信息的策略所致。MSR 算法对目标物体周围的区域色彩信息之建立一个直方图（见图 3-25），从而相当于对背景色彩进行平均化处理。当背景色彩比较单纯时，该方法能够有效地增加算法的鲁棒性，但若背景很复杂的情况下（如本序列），这

种宽泛的处理方法并不能将目标和特定的相似背景区域相区分，从而造成跟踪的失败。而提出的算法能够很好描述目标的纹理分布，从而在细节上将目标和背景进行区分，从而可以成功地跟踪该长序列。

为了检验三种算法的鲁棒性，同样通过随机采样的方法改变算法的初始状态条件，令算法的初始输入在标注值的 2%增量之内变化。以变化的初始条件为输入，分别对三种算法进行了 100 次重复的实验，表 5-3 给出了三种算法的平均跟踪误差和平均方差。

图 3-22 MS 算法在汽车序列上的跟踪误差曲线

Fig.3-22 Tracking error curve on car sequence by the MS traker

图 3-23 MSR 算法在汽车序列上的跟踪误差曲线

Fig.3-23 Tracking error curve on car sequence by the MSR traker

图 3-24 提出的算法在汽车序列上的跟踪误差曲线

Fig.3-24 Tracking error curve on car sequence by the proposed traker

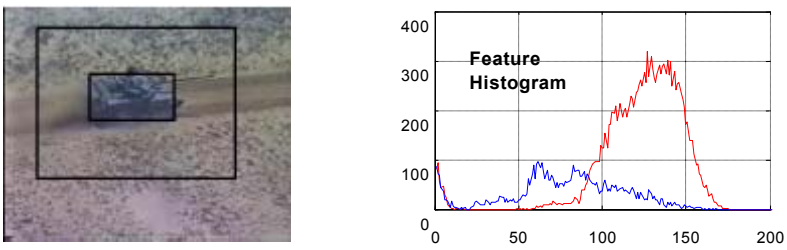


图 3-25 MSR 跟踪算法的背景采样策略。从前景和前景周围的背景分别建立直方图，然后将二者的相似性度量作为可区分性判据

Fig.3-25 Background sampling strategy used in MSR tracker. Constructing color histogram from foreground and its relevant background respectively and treat likelihood ratio between histograms as the discriminative measurement.

表 3-3 在汽车跟踪序列上 MS，MSR 和提出的算法的平均跟踪误差的比较

Table 3-3 Comparing mean tracking errors on car sequence among MS, MSR and proposed algorithms

	MS 算法	MSR 算法	提出的算法
水平方向平均跟踪误差	41.6	43.9	5.2
水平方向跟踪误差方差	19.2	11.3	1.7
垂直方向平均跟踪误差	5.2	11.6	4.1
垂直方向跟踪误差方差	2.7	3.9	1.2
水平尺度平均跟踪误差	46.1	39.2	6.3
水平尺度跟踪误差方差	24.8	14.1	2.2
垂直尺度平均跟踪误差	32.5	27.7	4.1
垂直尺度跟踪误差方差	25.7	16.2	1.9

3.9 小结

本章针对基于序列蒙特卡洛算法的目标跟踪中的核心问题——目标的建模与描述，提出了一种新的自适应动态建模方法，该方法同时维护了前景/背景的可区分性、前景描述的稳定性和计算的实时性。

该方法的创新点可以总结如下：

1) 首次将 Haar 小波特征作为刻画目标表观的图像特征引入视觉跟踪领域；
2) 提出基于目标/背景信息差别的目标模型构建方法；
3) 将模型构建和更新方法有机嵌入到序列蒙特卡洛滤波算法中，从而在跟踪过程中可以有效率地根据背景变化更新目标模型。

通过对比实验结果证明，提出的方法可以有效地进行复杂背景下的长序列跟踪，全面提升跟踪算法的性能。

第4章 基于多运动模型的复杂运动建模和推断

4.1 引言

由于计算复杂性的限制，视觉跟踪算法通常基于局部搜索的策略确定目标的运动状态。所以根据目标运动规律确定目标在当前时刻以较高概率出现的区域成为决定跟踪算法效率的关键问题。

现实世界中的很多物体，如原野上飞驰的汽车，晨练中的老者，小品演员的面部表情等，都具有复杂的运动模式。如何针对复杂运动建立模型，始终是解决这些目标跟踪问题的瓶颈所在。本章提出了采用多运动模型对目标复杂运动进行建模和估计的框架。在此基础上，针对具有多种运动模式和具有高维运动状态的两类常见的复杂目标运动，将多模型的估计框架融入到序列蒙特卡洛滤波算法中，从而针对两类复杂运动问题提出了标准序列蒙特卡洛滤波算法的两个改进：基于多模型切换的序列蒙特卡洛滤波算法(Multi-model Switching sequential Monte Carlo Filter, MSMCF)和基于多模型协同的序列蒙特卡洛滤波算法(Multi-model Cooperating sequential Monte Carlo Filter, MCMCF)。

4.2 问题的提出和相关工作

4.2.1 具有多种运动模式的复杂运动

设需跟踪目标物体在 t 时刻的状态向量为 $\mathbf{x}_t$。对时序跟踪问题，有 $\mathbf{x}_t = \mathbf{f}_d(\mathbf{x}_{1:t-1})$，其中 $\mathbf{x}_{1:t-1} = \{\mathbf{x}_1, ..., \mathbf{x}_{t-1}\}$ 是目标物体的状态时序，函数 $\mathbf{f}_d(\cdot)$ 则刻画了目标物体状态转移的轨迹，例如 $\mathbf{f}_d(\cdot)$ 可以是匀速直线运动模型，布朗运动模型等。

复杂运动模式的一种情况是：跟踪目标具有多种运动模式 $\{d_i \mid i = 1, ..., M\}$，并在某时间段内以集合 $\{d_i \mid i = 1, ..., M\}$ 中的一种模式 d_i 进行

运动，比如人具有走、跑、跳等诸多不同运动模式。相对于运动模式集合 $\{d_i \mid i=1,...,M\}$，可以构造状态转移模型的集合 $\{\mathbf{m}_i \mid i=1,...,M\}$，并使两集合间具有一一映射的关系。相对于标准序列蒙特卡洛滤波算法中只采用单一状态转移模型，本论文提出在序列蒙特卡洛滤波算法同时集成多个状态转移模型，并根据目标运动的轨迹做出判断，从而自适应地选择正确的状态转移模型，保证粒子传递和目标状态转移的一致性。

4.2.2 具有高维状态空间的复杂运动

另一种常见的复杂运动模式是由复杂的目标状态引起，如人体的多自由度造成人体姿态估计问题中的高维状态向量。与运动分析相关的研究领域如图形学中的运动合成等中的研究表明，绝大部分复杂的高维运动具有约束其可能运动组合的子空间，并且可以通过子空间的基来合成和分解复杂运动。

在本方法中，假设目标的复杂运动模式 $d_{complex}$ 可以由一组简单运动模式 $\{d_i \mid i=1,...,M\}$ 的线性组合来逼近。构造一组对应的状态转移模型 $\{\mathbf{m}_i \mid i=1,...,M\}$，则可以将原来对应 $d_{complex}$ 的复杂状态转移模型 $\mathbf{m}_{complex}$ 分解为一组简单状态转移模型的表示形式，$\mathbf{m}_{complex}=\sum_i c_i \cdot \mathbf{m}_i$，从而降低了跟踪具有复杂运动模式目标的困难。

针对具有复杂运动模式的目标跟踪算法研究，主要集中在人体姿态估计和手部姿态估计等研究领域。在该领域中，处理复杂运动模式的策略主要可分为三类：

融合强搜索算法到序列蒙特卡洛滤波算法中：在采用状态转移模型传递粒子并采用状态表观关联模型赋予粒子权重之后，再采用强搜索算法求精粒子在状态空间中的最终状态。Deutscher 将模拟退火搜索算法融入到序列蒙特卡洛滤波算法中来进行人体姿态的估计[56]。Choo 将每个粒子看作一个具有进化能力的马尔科夫链，并在单采样时刻上根据目标表观来递归传递粒子以确定其最终的位置[78]。

构造问题相关的复杂提议分布或者状态转移模型：复杂的提议分布或

者状态转移模型通常提供了更好的起始搜索位置，不但减少了计算强度，同时降低了被局部极小点吸引的概率。Leventon 和 Yacoob 借助专业的运动捕获设备采集训练数据，采用主成分分析技术获取目标物体运动的问题相关模型[79, 80]。Brand 和 Molina-Tanco 采用隐马尔科夫模型学习人运动的规律来构造运动状态转移模型[81, 82]。王采用基于无监督聚类的策略学习人体运动模型[83]。虽然问题相关的模型能够明显降低所需粒子的数目，获得更好的跟踪结果，学习和数据收集的过程却是冗长而充满挑战的。在[70]中, 最新的表观数据被用来作为传递粒子的依据之一。但是，该技术目前只针对二维跟踪问题有效。

借助运动约束或者限定目标的运动范围：为了简化具有复杂运动模式的目标跟踪问题，MacCormick 采用层次性搜索算法来跟踪具有关节性运动特点的物体[84]。Bregler 和 Deutscher 假定对运动物体的观察角度是固定的从而大大地简化了问题[85, 86]。Niyogi 和 Rohr 则仅仅针对人体运动中的步行情况进行跟踪[87, 88]。这些约束或者限制虽然极大地简化了问题，同时也降低了算法的通用性。

提出的算法可以看作通过构造复杂的状态转移模型来解决具有复杂运动模式的目标跟踪问题。在视觉跟踪领域，已经有一些开创性的工作涉及如何集成多个模型来完成跟踪任务[9, 89-91]。这些工作在模型的利用上都是排他性的，即某一时刻只有一个模型是有效的。而在提出的算法中，多模型是可以协同工作的。

4.3 多模型运动估计框架

多模型运动估计框架(以下简称多模型框架)的基本思想是通过贝叶斯网络来融合多个运动模型进行运动的估计。所有子模型在运动估计的过程中通过互相交换信息共同解释复杂的运动模式，最终估计结果基于每个子模型估计结果的统计融合。在某时间采样点，或许某子模型起到主导的作用，而在另一采样时刻，或许所有的子模型在运动解释的意义上是同等重要的。

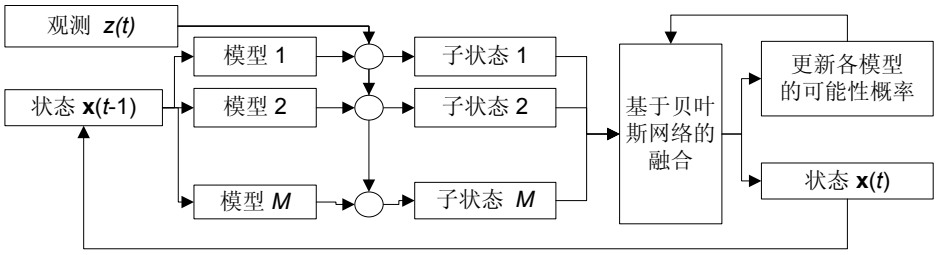


图 4-1 多模型运动估计框架的流程

Fig.4-1 flowchart of multi-model motion estimation framework

多模型框架在采样时刻 t 的计算流程见图 4-1。首先，用 M 个不同的状态转移模型分别作用于 $t-1$ 时刻的目标状态 $\mathbf{x}(t-1)$，得到对 $\mathbf{x}(t)$ 的 M 种预测 $\{\tilde{\mathbf{x}}_m(t)\,|\,m=1,...,M\}$。然后对子模型 m，根据提取的表观数据 $\mathbf{z}(t)$ 和相应的表观状态关联模型 $p_m(\mathbf{z}_t\,|\,\mathbf{x}_t)$ 来评估 $\tilde{\mathbf{x}}_m(t)$ 的信度。最后，根据子模型之间的马尔科夫转移特性假设，运用贝叶斯网络融合各子模型的状态估计结果得到 $\mathbf{x}(t)=\sum_m \lambda_m\cdot\tilde{\mathbf{x}}_m(t)$ 的估计。

4.3.1 定义

为对方法进行数学上的详尽描述，作如下定义：

$P_{mn}{}^{t-}$：在时刻 t 之前，被子模型 m 解释的部分目标运动将被子模型 n 解释的概率。这些概率预先根据经验或实验设定，并且满足 $\sum_{m=1}^{M} P_{mn}{}^{t-}=1$，将所有的 $P_{mn}{}^{t-}$ 归纳到一起写成矩阵的形式，称为子模型转移矩阵：

$$M_{mat}{}^{t-}=\begin{bmatrix} P_{11}{}^{t-} & \cdots & P_{1M}{}^{t-} \\ \cdots & \cdots & \cdots \\ P_{M1}{}^{t-} & \cdots & P_{MM}{}^{t-} \end{bmatrix} \tag{4-1}$$

$P_{mn}{}^{t+}$：在时刻 t 之后，曾被子模型 m 解释的部分目标运动被子模型 n 解

释的概率。

P_m^{t-}：目标运动在时间段$[t-1,t)$内被子模型 m 解释的概率；

P_m^{t+}：子模型通过贝叶斯网络交互之后，目标运动被子模型 m 解释的概率。

4.3.2 模型交互过程

于是，可以推导得到如下的关系：

$$P_{mn}^{t+} = P_{mn}^{t-} P_m^{t-} / P_n^{t+} \tag{4-2}$$

$$P_m^{t+} = \sum_{m=1}^{M} P_{mn}^{t-} P_m^{t-} \tag{4-3}$$

首先，对每个子模型 n 求出一个混合状态估计：

$$\tilde{\mathbf{x}}(t)^n = \sum_{m=1}^{M} P_{mn}^{t+} \hat{\mathbf{x}}(t)^m \tag{4-4}$$

则目标状态 $\mathbf{x}(t)$ 是所有子模型状态估计 $\tilde{\mathbf{x}}(t)^n$ 的概率叠加：

$$\mathbf{x}(t) = \sum_{n=1}^{M} P_n^{t+} \tilde{\mathbf{x}}(t)^n \tag{4-5}$$

对子模型 m 来说，根据 $\tilde{\mathbf{x}}_m(t)$ 从 $p_m(\mathbf{z}(t)|\mathbf{x}(t))$ 得到的期望表观 $\tilde{\mathbf{z}}_m(t)$ 和实际表观之间的差值用 $d_m(t) = \|\mathbf{z}_m(t) - \tilde{\mathbf{z}}_m(t)\|$ 表示，其中 $p_m(\mathbf{z}(t)|\mathbf{x}(t))$ 是对应子模型 m 的状态表观关联模型。设表观向量维数是 R 且差值 $d_m(t)$ 服从高斯分布，则子模型 m 能够解释表观的可能性函数为：

$$V_m(t) = \frac{\exp(-(d_m(t))^2 / 2\sigma_m(t)^2)}{\sqrt{(2\pi)^R \sigma_m(t)}} \tag{4-6}$$

其中，$\sigma_m(t)$ 是与 $d_m(t)$ 对应的方差。于是，最后更新子模型 m 的下一时刻先验概率为：

$$P_m^{(t+1)-} = V_m(t) P_m^{t+} / C \tag{4-7}$$

其中，$C = \sum_{m=1}^{M} V_m(t) P_m^{t+}$ 是归一化因子。

通过不断组合验证几种不同的假设，从而递归地得到目标的状态。多模型框架提供了解决复杂跟踪问题的一种方式：即通过分解复杂的运动，从而将复杂运动的跟踪问题转化为几个简单运动的跟踪问题。

4.4 多模型切换序列蒙特卡洛滤波算法

在本小节中，将多模型估计框架和序列蒙特卡洛滤波算法相结合，提出基于多模型切换的序列蒙特卡洛滤波算法来解决前面提到的第一类复杂运动跟踪问题。

算法的主要流程详述如下：用 $\mathbf{s}_t = \{\mathbf{x}_t^k, w_t^k, m_t^k \mid k = 0,...,N\}$ 表示 t 时刻的粒子集合，其中 m_t^k 表示集合中粒子 k 在时刻 t 按照状态转移模型 m 进行运动。对每一粒子，根据概率分布 P_m^{t-} (定义见 4.3.1 节)选择运动模型 m 作为其状态转移模型，并满足按照子模型 m 运动的粒子在粒子集合中所占的比例与 P_m^{t-} 成正比。也就是说，粒子集合被依概率划分为 M 个子集合，每个粒子子集合在估测子状态 $\tilde{\mathbf{x}}_m(t)$ 方面类似于单独的序列蒙特卡洛滤波器(序列蒙特卡洛滤波算法的跟踪流程请见第 3 章)。然后，依据 $\tilde{\mathbf{x}}_m(t)$ 对 $p(\mathbf{z}_m(t) \mid \mathbf{x}_m(t))$ 进行采样得到 $\tilde{\mathbf{z}}_m(t) \sim p(\mathbf{z}_m(t) \mid \mathbf{x}_m(t))$，计算提取的表观 $\mathbf{z}(t)$ 和 $\tilde{\mathbf{z}}_m(t)$ 之间的差异 $d_m(t)$。对 M 个子状态转移模型，可以得到 $\{d_m(t) \mid m = 1,...,M\}$。选择粒子子集合 $\{\mathbf{x}_t^k, w_t^k, m_{\min}\}$，其中 $m_{\min} = \arg\min_m(d_m(t))$。则目标的运动状态估计为

$\mathbf{x}(t)=\sum w_t^k \mathbf{x}_t^k \quad \mathbf{x}_t^k \in (\mathbf{x}_t^k, w_t^k, m_{\min})$。最后，用每个子模型的估计与最终结果的距离来更新该子模型的重要性度量，从而为下一步的估计做好准备。

算法的详细流程见图 4-2：

4.5 实验部分

本节中，对提出的多模型切换序列蒙特卡洛滤波算法的性能进行了实验验证。实验共分两个部分：首先采用一段参数可控的自拍视频序列来定量验证算法的正确性。然后，在公共测试序列上，针对人脸跟踪问题进行了实验研究。当前的算法实现只进行了简单的优化，可以在 Pentium III，667MHz 的个人计算机上以 5 帧/秒的速度处理视频序列。

在两类实验中，都选择目标在图像平面的像素位置和速度构成四维状态向量：

$$\mathbf{x}=(p_x, p_y, v_x, v_y)^T \tag{4-8}$$

采用基于 2 维色彩分布的非参数目标模型建模目标表观[2]。对每个采用状态转移模型 $p_m(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 生成的假设区域，通过求取该区域的色彩分布与目标模型的 Bhattacharyya 距离作为状态表观关联模型 $p(\mathbf{x}_t \mid \mathbf{z}_t)$ 的输出。

4.5.1 基于可控视频序列的算法验证

测试序列采用普通摄像机拍摄，具有 320x240 的像素分辨率和 3 帧/s 的采样率。序列中主体是一辆玩具汽车，其平均行进速度大概在 1 米/s 左右。在行进过程中，受控汽车两次改变运动模式，首先做直线行驶，接着作圆周旋转运动，最后又恢复直线行驶状态。该段视频可以检验基于多模型切换序列蒙特卡洛滤波算法根据目标运动模式自动选择正确运动模型传递粒子的能力和算法的正确性。图 4-3 给出部分图示跟踪结果。

根据该测试序列的特点，共采用了两种运动模型。一个是近似匀速直线运动模型(Nearly Constant Velocity Motion Model，NCVMM)，另一个是近似匀角速度平面旋转模型(Nearly Constant Horizontal Turn Motion Model，NHTMM)。关于该两个运动模型的详细定义请见[9]。根据经验设定联结两种

模型的马尔科夫转移矩阵在跟踪过程中不变化，该矩阵设置为：

$$M_T = \begin{Bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{Bmatrix} \tag{4-9}$$

循环

预测过程：对状态转移模型 $P_m(x_i \mid x_{i-1})$ 中采样 N_m 个粒子，满足 $N_m / N \approx P_m^{t-}$。则总粒子个数为 $\sum_m N_m = N$。

验证过程：通过观测状态关联模型 $w_t^k = p_m(\mathbf{z}_t \mid \mathbf{x}_t) w_{t-1}^k$ 评估粒子权重。

交互过程：1)．对每一个粒子子集合，计算估计状态值 $\tilde{\mathbf{x}}_t^m = \sum_{j=1}^{N_m} \frac{w_t^j \mathbf{x}_t^j}{w_t^j}$；

2)．根据公式(5-2)更新子模型 m 的概率；

3)．根据公式(5-3)更新子模型 m 的转移概率；

4)．根据公式(5-6)对子模型 m 计算 $d_m(t)$；

5)．最终状态估计结果为采用子集合 $m_{\min} = \arg\min_{d_m(t)}(m)$ 中的所有粒子所计算得到的目标状态 $\mathbf{x}_t = \sum_{j=1}^{N_t} \frac{w_t^j \mathbf{x}_t^j}{w_t^j}$；

6)．根据公式(5-7)，利用残差 d_t^m 更新子模型 m 的概率，为下时刻的计算做好准备。

重采样过程：对粒子集合，计算归一化的权重的方差。如果方差超过一定的阈值，则根据每个粒子的权重大小作为其出现的概率，有放回的从粒子集合中进行采样生成新的粒子集合。

图 4-2 MSMCF 算法的流程

Fig.4-2 Flowchart of MSMCF algorithm

实验结果表明，矩阵 M_T 中数项的微小调整对跟踪结果影响很小，说明算法对该矩阵的数项变化不敏感，从而假设其恒定是合理的。

为了对比实验结果，采用 NCVMM 作为状态转移模型实现了标准序列蒙特卡洛滤波算法。实验中，基于多模型切换的序列蒙特卡洛滤波算法和标准序列蒙特卡洛滤波算法都选用 1000 个粒子进行目标跟踪，部分跟踪结果如图 4-3 所示。

从图 4-4 可以看到，当汽车首次改变其运动模式的时候，标准序列蒙特卡洛滤波算法就丢失了跟踪目标，而提出的基于多模型切换的序列蒙特卡洛滤波算法却能够由始至终的完成目标跟踪任务。为了定量衡量两种算法之间的性能差别，定义了两个度量来比较算法间的效率：一个是粒子群的重心和目标物体真实状态之间的距离 D_{cgt}；另一个是粒子权重的方差 C_w。D_{cgt} 和 C_w 在跟踪过程中的变化分别图示于 4-5 和 4-6 中。若减少所用粒子数目，较大的 D_{cgt} 会造成算法性能的更严重下降。从图 4-5 中可以看到，基于多模型切换的序列蒙特卡洛算法比标准序列蒙特卡洛算法具有更小的 D_{cgt}，这也揭示了为什么在图 4-4 中，当粒子数目低于某阈值后，标准序列蒙特卡洛算法丢失了跟踪目标而提出的算法工作良好。另一个度量 C_w 从状态表观关联函数的角度诠释了粒子群对真实目标状态的逼近程度。权重集合的方差是粒子滤波算法中广泛用来度量算法效率的指标。从图 4-6 中可以看出，基于多模型切换的序列蒙特卡洛滤波算法比标准序列蒙特卡洛滤波算法具有更小的 C_w 均值，也证明了多模型切换序列蒙特卡洛滤波算法更具效率。

所提出算法的根据目标运动模式动态改变所采用运动模型的能力可以从图 4-7 中得到验证。同时，算法还具有判断目标运动模式的能力。

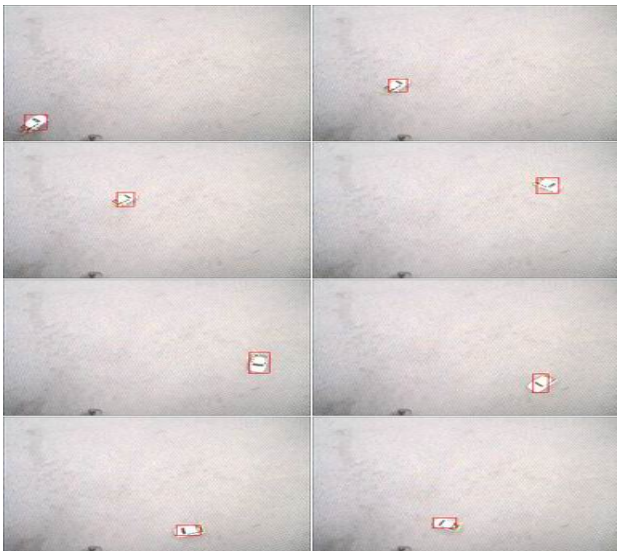


图 4-3 采用 MSMCF 算法在测试序列上获得的部分跟踪结果

Fig.4-3 Some tracking results obtained by MSMCF on test sequence

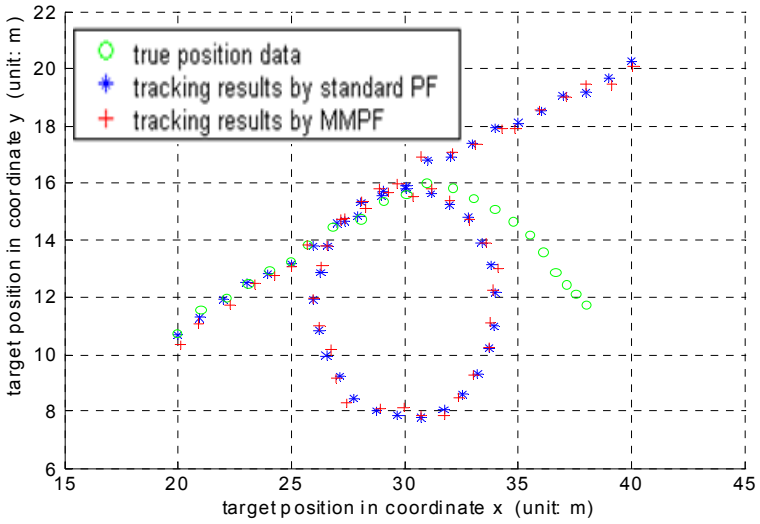


图 4-4 MSMCF，SMCF 的跟踪结果和目标真实状态之间的比较

Fig.4-4 Comparison of tracking results obtained from MSMCF, SMCF and target true states

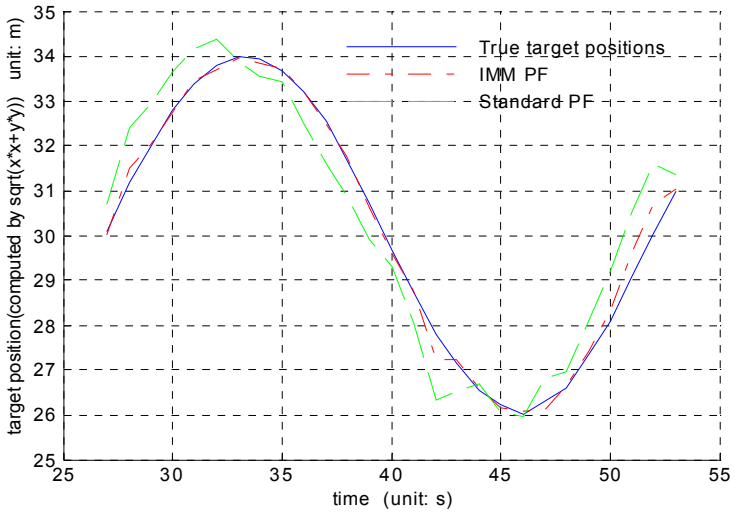


图 4-5 MSMCF,SMCF 的 D_{cgt} 变化和目标真实状态间的关系

Fig.4-5 Trajectories of D_{cgt} of MSMCF, SMCF and their relations to object true states

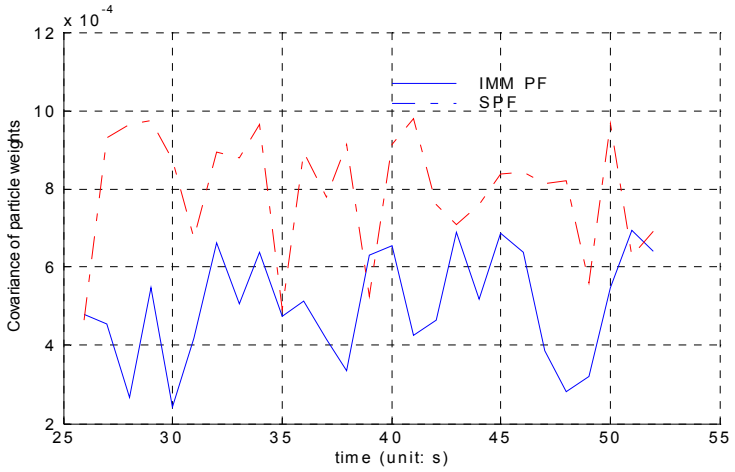


图 4-6 MSMCF 算法和 SMCF 算法的 C_w 指标变化

Fig.4-6 Trajectories of C_w of MSMCF and SMCF

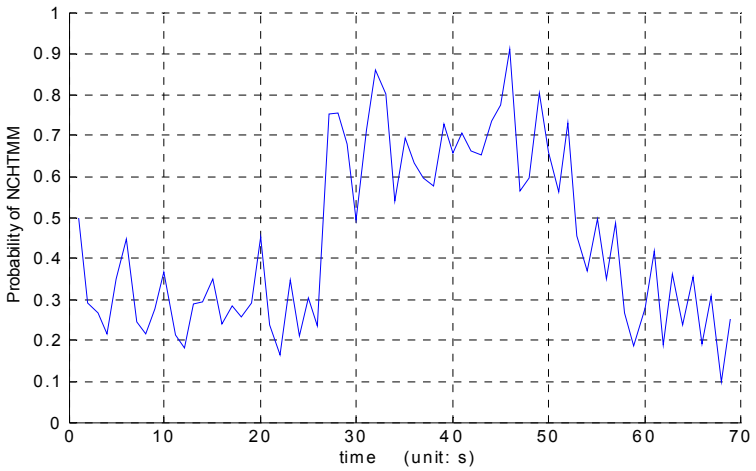


图 4-7 目标的运动模式可用 NCHTMM 运动模型解释的概率

Fig.4-7 Probability that the target resides on NCHTMM

4.5.2 基于公共测试序列的算法验证

以 http://www.ces.clemson.edu/~stb/research/headtracker/seq/ 上公布的人脸跟踪标准测试序列为测试样例集合，根据提出的算法的特点，选择了其中的“seq_fast.tar.gz”和“seq_jw.tar.gz”测试序列。

序列“seq_fast.tar.gz”共有 32 帧，由于目标在图像平面内具有较快的运动速度，同时具有急停急转等变向运动模式，从而为运动预测带来困难，尤其当目标突然改变运动方向的瞬间，如果只采用简单的单一运动模型很难保证算法的鲁帮性，而实验结果也证实了这一点。

根据该两测试序列的特点，共采用了两种运动模型。一个是 NCVMM, 另一个是反向近似直线运动模型。马尔科夫转移矩阵的设定同公式(4-9)。

在本序列上，同样将采用 NCVMM 作为状态转移模型的标准序列蒙特卡洛滤波算法（SMCF）与基于多模型切换的序列蒙特卡洛滤波算法(MSMCF)进行了比较。

在“seq_fast.tar.gz”序列上的对比结果可以看出(图 4-8)，当目标在由右向左突然改变方向时，采用单一运动模型会无法准确地对目标运动进行补

偿，从而造成目标的丢失。而基于多模型切换的序列蒙特卡洛滤波算法，系统可以正确地选择合适的反向近似匀速直线运动模型，从而能够正确地补偿目标运动。

该公共测试序列伴随有手工标注的状态文件。图 4-9 比较了提出的 MSMCF 算法和 SMCF 算法之间的跟踪误差曲线。图中的四条曲线分别代表在时间轴上算法跟踪误差的变化。从图中可以看出，SMCF 算法的水平跟踪误差要远大于 MSMCF 算法的水平跟踪误差，而二者在垂直跟踪误差上差别不大。这是由于该序列中目标人脸的运动主要是水平方向，并且运动模式的改变也主要发生在水平方向。从而证明了当目标运动模式改变的跟踪情境中，多模型自适应的 MSMCF 算法要优于单模型的 SMCF 算法。当然，当目标物体具有单一运动模式的情况下，如本测试序列中的目标人脸垂直方向运动，MSMCF 和 SMCF 具有相当的跟踪性能。

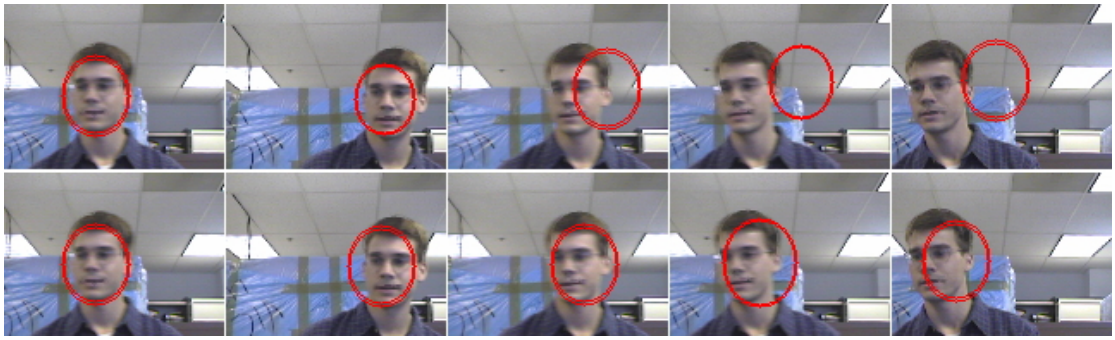


图 4-8 SMCF 算法(上行)和 MSMCF 算法(下行)在“seq_fast.tar.gz”序列上跟踪结果的部分比较

Fig.4-8 Comparison of tracking results obtained from SMCF (top row) and MSMCF (bottom row) on video “seq_fast.tar.gz”

图 4-9 在序列“seq_fast.tar.gz”上的 MSMCF 和 SMCF 算法的跟踪误差比较

Fig.4-9 Comparing tracing errors between MSMCF and SMCF on video “seq_ fast.tar.gz”

“seq_jw.tar.gz”序列相对于“seq_fast.tar.gz”序列要更为复杂。在该序列中有遮挡，光照等环境变化和人脸姿态等自身变化。该序列共有 100 帧。在“seq_jw.tar.gz”序列上的实验结果同样证明了算法的有效性，同样在目标突然改变运动方向的时候，相对于只采用单一模型的标准序列蒙特卡洛滤波算法(图 4-9)，基于多模型切换的序列蒙特卡洛滤波算法正确地跟踪了目标(图 4-10)。

图 4-10 在“seq_jw.tar.gz”序列上 SMCF 算法的部分跟踪结果

Fig.4-10 Some tracking results obtained by SMCF on video “seq_jw.tar.gz”

图 4-11 在“seq_jw.tar.gz”序列上 MSMCF 算法的部分跟踪结果

Fig.4-11 Some tracking results obtained by MSMCF on video “seq_jw.tar.gz”

图 4-12 在“seq_jw.tar.gz”序列上 MSMCF 和 SMCF 算法的跟踪误差比较

Fig.4-12 Comparing tracking errors between MSMCF and SMCF on video “seq_jw.tar.gz”

该公共测试序列同样伴随有手工标注的状态文件。图 4-12 比较了提出的 MSMCF 算法和 SMCF 算法之间的跟踪误差曲线。图中的四条曲线分别代表在时间轴上算法跟踪误差的变化。从图中可以看出，SMCF 算法的水平

平均跟踪误差要远大于 MSMCF 算法的平均水平跟踪误差，并且误差大小很不稳定。二算法在垂直跟踪误差上则差别不大。这也是由于序列中人脸的主运动方向和运动模式改变发生在水平方向，从而证明了当目标运动模式改变的跟踪情境中，MSMCF 算法要优于 SMCF 算法。

4.6 基于多模型协同的序列蒙特卡洛滤波算法

序列蒙特卡洛滤波算法的优势在于其在贝叶斯理论框架下对状态转移模型 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 进行带有随机性质的采样而产生假设集合，相对于单一假设机制如卡尔曼滤波中的预测策略，假设集合在一定程度上能够克服噪声的影响并且可以对多通道信息进行有效融合。一般意义上，生成的假设越多，得到精确跟踪结果的机会越大，但同时也带来计算量的增长。

如果目标状态具有比较高的维数或者在状态空间中具有很复杂的运动轨迹，采用序列蒙特卡洛滤波算法进行目标跟踪将由于粒子数目的指数级增长而造成计算能力超出负荷。MacCormick 定义了评估序列蒙特卡洛滤波算法中粒子集合 $\{\mathbf{x}_t^i, \pi_t^i \mid i = 1...N\}$ 效率的两个度量[84]，从而可以根据该度量估测所需计算量。一个度量是粒子生存能力指标 $D = (\sum_{i=1}^{n} (\pi_t^i)^2)^{-1}$，另一个是粒子生存率 $\alpha \approx \frac{D}{N}$，其中 N 是粒子的个数。在保证算法跟踪性能的前提下，可以推断出需要的粒子数目应该满足 $N \geq \frac{D_{\min}}{\alpha^d}$，其中 $D_{\min}$ 是在保证序列蒙特卡洛滤波算法性能基础上，可接受的生存能力指标最小值。从上面公式可以看出，所需粒子数目的决定因素是 α^d 中的目标维数 d。上述分析揭示了应用序列蒙特卡洛滤波算法到高维跟踪问题的困难所在。从滤波算法本身考虑，高维状态空间使生成有效假设的概率降低，因此，如何对目标的状态转移特性进行有效的建模，从而有效利用有限的粒子，成为解决高维状态空间中目标跟踪的关键问题。

序列蒙特卡洛滤波算法生成假设集合的特性使多模型状态跟踪框架可以自然地与其融合起来：相对于标准序列蒙特卡洛滤波算法将所有粒子用相同的状态转移模型进行粒子的传递从而生成同类假设，多模性跟踪框架具有将

粒子分布在不同的状态转移模型上从而生成不同种类的假设的能力，并且通过贝叶斯网络融合各子模型的估测结果可以得到目标的最终状态估计。

针对由高维状态空间造成的复杂运动模式的目标跟踪问题，将多模型运动估计框架和标准序列蒙特卡洛滤波算法融合，提出了基于多模型协同的序列蒙特卡洛滤波算法：

用 $s_t = \{\mathbf{x}_t^k, w_t^k, m_t^k \mid k = 0,...,N\}$ 表示时刻 t 的粒子集合，其中 m_t^k 表示集合中粒子 k 在时刻 t 按照子模型 m 进行状态转移。对每一个粒子，根据概率分布 P_m^{t-} 定义其状态转移模型，并且基本满足如下的关系：按照子模型 m 运动的粒子在粒子集合中所占的比例与 P_m^{t-} 成正比。依据以上原则，粒子集合被划分为 M 个子集，每个子集合中的粒子根据标准序列蒙特卡洛滤波算法求取对目标状态的部分估计。不同于前面的基于多模型切换的序列蒙特卡洛滤波算法中只采用一个子集进行最终目标状态的估计，最终的目标结果估计是 M 个子集合的标准序列蒙特卡洛滤波算法所得到的估计结果的基于贝叶斯网络的融合。详细的算法步骤见图 4-13：

4.7 头部运动估计

4.7.1 概述

本节在头部运动估计问题上检验了提出的基于多模型协同的序列蒙特卡洛滤波算法的性能。

头部运动可以分为刚性运动和非刚性运动。刚性运动中，头部被看作刚体从而具有六自由度运动模式。在非刚性运动中，比如表情变化等，通常假设头部姿态固定或者只有微小变化。绝大多数已有算法只针对刚性运动或者非刚性运动的单一情况进行头部运动的估计[92]。在现实应用中，如基于视频的节目主持人表情分析等，都需要同时对两类运动进行混合估计。

将基于多模型协同的序列蒙特卡洛滤波算法应用于头部运动估计问题，并同时考虑刚性运动和非刚性运动。通过这样一种基于贝叶斯理论的时序滤波框架，希望算法能够解决两个问题：一是算法能够在图像序列质量比较低

的情况下鲁棒工作；二是当发生跟踪误差的时候，算法能够不借助外部调整而以较大的概率从跟踪误差中恢复，从而避免跟踪误差随时间轴的传递并最终导致跟踪任务的失败。

共设计了两种试验来验证算法的性能。为了定量评估算法的性能，第一组实验使用状态值预知的合成视频数据。第二组实验则以摄像机获取的真实视频为测试序列来验证算法在真实场景下的跟踪性能。由于同时考虑了刚性运动和非刚性运动，表达头部运动的原始状态向量高达 66 维。通过利用基于多模型协同的序列蒙特卡洛滤波算法，原始的高维空间运动可以分解为 8 个一维的子运动模式从而大大降低了该问题的难度和所需计算量。

在相同测试序列中比较了标准序列蒙特卡洛滤波算法和基于多模型协同的序列蒙特卡洛滤波算法的性能。由于计算量随目标状态维数呈指数递增，标准序列蒙特卡洛滤波算法难以应用到这样的复杂跟踪问题中，而基于多模型协同的序列蒙特卡洛滤波算法则可以在合理的计算量需求下完成跟踪任务。

4.7.2 头部运动的表示

以合成真实感面部动画为目的，MPEG-4 标准中定义了采用两组参数控制的三维头部模型(图 4-14 简单描述了该模型的工作原理)，其中一组参数能够精确地定义头部形状、大小和纹理，称为人脸定义参数(Face Definition Parameter，FDP)。另外一组参数能够完全控制头部的刚性运动和非刚性运动，称为面部动画参数(Facial Animation Parameter，FAP)。通过调整三维面部线框模型上的关键特征控制点，FAP 可以重构头部的刚性运动和口腔、嘴、眼睛、面颊等部位的非刚性运动。根据面部区域的不同，标准中的 FAP 参数共有 68 个并被分为 10 组(在实验中，具有高层语义的发音和表情 FAP，由于没有明确定义而没有被采用)。实现了 MPEG-4 标准中定义的三维头部模型用来直观的表示头部运动状态和从假设状态生成虚拟表观 (在本工作中，不考虑 FDP 的获取，并假设 FDP 已经针对特定人被精确地调整好)。

循环

预测过程：从 $p_m(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 中采样 N_m 个粒子 $\{\mathbf{x}_t^k, w_t^k, m_t^k = m \mid k = 0,..., N_m\}$，满足 N_m / N 正比于 P_m^{t-} 且满足 $\sum_{m=1}^{M} N_m = N$；

验证过程：通过观测状态关联模型 $w_t^k = p_m(\mathbf{z}_i \mid \mathbf{x}_t^{k-}) w_{t-1}^k$ 评估粒子权重；

交互过程：1)．对每一子集合，计算状态估计值 $\tilde{\mathbf{x}}_t^m = \sum_{j=1}^{N_m} \frac{w_t^j \mathbf{x}_t^j}{w_t^j}$；

2)．计算子模型 m 的概率 $P_m^{t+} = \sum_{m=1}^{M} P_{mn}^{t-} P_m^{t-}$；

3)．计算子模型间的转换概率 $P_{mn}^{t+} = P_{mn}^{t-} P_m^{t-} / P_m^{t+}$；

4). 对每一粒子集合，计算滤波后的子状态估计 $\hat{\mathbf{x}}_t^n = \sum_{m=1}^{M} P_{mn}^{t+} \tilde{\mathbf{x}}_t^m$；

5). 最终状态估计结果是所有滤波后子状态的概率权重加和 $\mathbf{x}_t = \sum_{n=1}^{M} P_n^{t+} \hat{\mathbf{x}}_t^n$；

更新子模型概率：对每个子模型计算其估计残差 d_t^m 从而根据公式（5-7）计算概率 $P_t^{(t+1)-}$。

重采样：对粒子集合，计算归一化的权重方差。如果方差超过一定阈值，则根据每个粒子权重大小作为其出现概率，有放回的从粒子集合中进行采样生成新的粒子集合。新生成集合中的每个粒子的权重为 1/N。

图 4-13 MCMCF 算法的主要流程

Fig.4-13 Flowchart of MCMCF algorithm

实验中， 首先尝试直接采用序列蒙特卡洛滤波算法在 66 维的参数空间中进行头部运动状态的估计。虽然算法中采用的粒子数目已经达到10^8数量

级，但仍无法得到稳定收敛的跟踪结果，究其主要原因，是因为绝大多数粒子都生成了无用假设。该原因部分可用图 4-15 解释。在图 4-15 中，可以看到很多不真实的人脸图像，这些图像是根据序列蒙特卡罗滤波算法在状态空间中所做出的预测所生成的，由于原始的 FAP 共有 68 维。而人脸的表情由于受到物理因素的制约，因而并不是该高维空间中的所有采样点都对应合理的表情，如果只采用序列蒙特卡罗滤波算法不加约束地搜索状态空间，则会造成大量的粒子生成无用的假设，而这种无用假设，从表观上就表现为类似图 4-15 的形式。 这也解释了为什么算法不收敛的原因。正如 Forsyth 曾指出，对大于 10 维的状态空间，采用标准序列蒙特卡洛滤波算法进行运动估计很难获得收敛的跟踪结果[1]。

图 4-14 MPEG-4 标准中定义的三维头部模型的示例

Fig.4-14 Illustrating the 3D face model defined in MPEG-4

前面的实验基本否定了采用标准序列蒙特卡洛滤波算法在原始高维空间上直接进行状态估计的策略。注意到为合成真实感的人脸动画定义的 FAP 各元素之间并不统计独立。首先，采用主成分分析(Principle Component Analysis，PCA)方法对 FAP 训练数据集进行降维以得到头部非刚性运动的本质运动表达(假设头部的刚性运动和非刚性运动是互相独立的，在 FAP 数据训练集合中将与头部姿态相关的三个 FAP 元素设为常值，从而暂时屏蔽掉头部姿态的影响)。将特征值按降序排列，前 5 个特征值囊括了训练集合中 99%的变化，类似于 AAM 方法[33]，将前 5 个特征向量作为头部非刚性运动的子运动模型。针对头部的刚性运动，采用三个近似匀速运动模型分别对头部的深度旋转，平面旋转和俯仰旋转运动进行建模描述。 所以，最终的头部运动由 8 个一维子模型的系数所构成的 8 维向量进行描述。在基于多模型协同的序列蒙特卡洛滤波算法中，结合了两类共 8 个子模型来进行头部运动的估计。

4.7.3 运动模型

在前面的小节中，得到了 5 个特征向量构成的运动模型和 3 个近似匀速直线运动模型，头部的运动状态可以由这 8 个一维运动模型的系数来表示。从而，描述头部运动的原始复杂模型被分解为 8 个简单的运动子模型：

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \sum_{m=1}^{8} \lambda_m \cdot p_m(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \tag{4-10}$$

其中，$p_m(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 表示运动子模型 m。

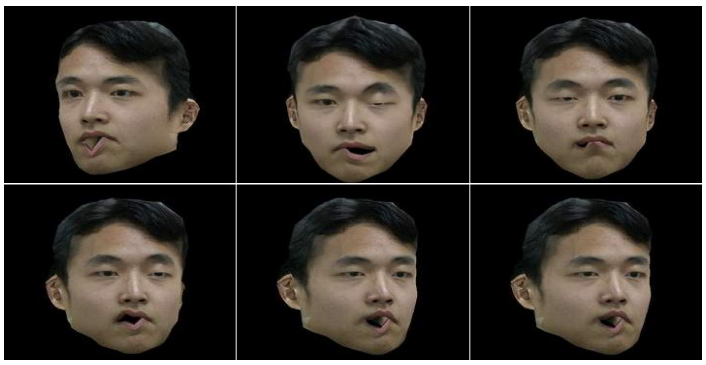


图 4-15 标准序列蒙特卡洛滤波算法产生的无用假设示例

Fig.4-15 Useless hypotheses generated by standard sequential Monte Carlo filter

头部运动估计的过程可以用图 4-16 表示。所有子模型在估计过程中构成串行结构，该结构构成基于多模型协同的序列蒙特卡洛滤波算法的串行估计模式。前一个子模型的估计权重(子状态)$\lambda_{\text{m-1}}$ 为后一个子模型的估计权重 λ_m 提供了搜索起点。

设定子模型转移矩阵在跟踪过程中保持不变。在实验中将对角线元素设为 0.72 ，非对角线元素设为 0.04 (实验结果表明，算法对子模型转移矩阵中元素的微小变化不敏感)。初始的子模型概率 p_m^{0-} 设为 $1/M$ 。跟踪过程中，子模型 m 在时刻 t 的距离残差为：

$$d_t^m = \frac{1}{|\lambda_m^t - \lambda_m^{t-1}|} \tag{4-11}$$

4.7.4 评估粒子权重

考虑 t 时刻的一帧输入 I_t，首先构造脸部区域的差图像 $\Delta I_t = I_t - I_{t-1}$，其中 I_{t-1} 是 t-1 时刻的输入帧 (所有的图像通过特征点进行匹配校准，在目前算法中采用手工标注特征点的方法)。然后，按照匹配滤波器(matched filter)的方法归一化 ΔI_t，使 $\|\Delta I_t\| = 0$ 并且 $\mathrm{var}(\Delta I_t) = 1$。之后,由粒子 $\mathbf{x}_{t-1}^{\ i}$ 和 $m_{t-1}^{\ i}$ 对状态转移模型 $p_{m_{t-1}^{\ i}}(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 进行采样生成 t 时刻的目标状态预测 $\tilde{\mathbf{x}}_t^{\ i}$，再根据三维头部模型和 $\tilde{\mathbf{x}}_t^{\ i}$ 生成表观图像 $\tilde{O}_t^i$。计算 $\tilde{O}_t^i$ 和 I_{t-1} 之间的差图像 $\Delta O_t^i = |\tilde{O}_t^i - I_{t-1}|$。最后，粒子的权重为

$$w_t^i = \Delta\tilde{O}_t^i \cdot \Delta I_t^{\ i} = <\Delta\tilde{O}_t^i, \Delta I_i^{\ t}> \tag{4-12}$$

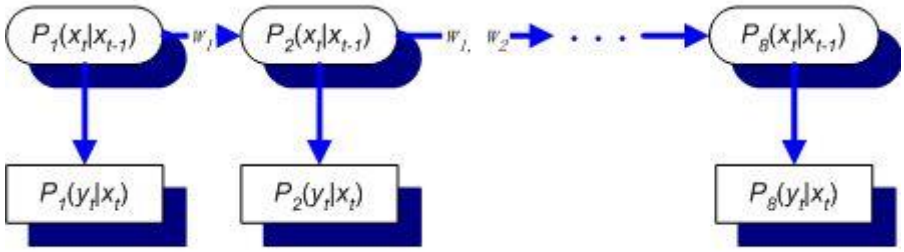


图 4-16 头部跟踪问题的多模型粒子滤波推断过程

Fig.4-16 The inference structure for head motion estimation

4.7.5 算法的定量性能评测

针对三个不同对象，采用三维头部模型合成了 3 段视频序列。其中一段视频序列的分辨率为 320x240，包含 317 帧图像。当算法中采用的粒子数目

达到 5000，得到了稳定收敛的跟踪结果。图 4-17 给出了部分估计结果，图中奇数行显示了用三维模型合成的实验图像，偶数行则是在算法估计的目标运动状态的基础上，采用三维模型所合成的头部图像。从图中可以看到，除了一些细节问题，比如皱眉，眼睛张合的大小等，估计的参数能够很好地解释原图像，从而说明算法估计的有效性。

图 4-17 原始图像帧和采用估计结果的合成帧之间的比较

Fig.4-17 Comparing original and synthesized frames using estimations

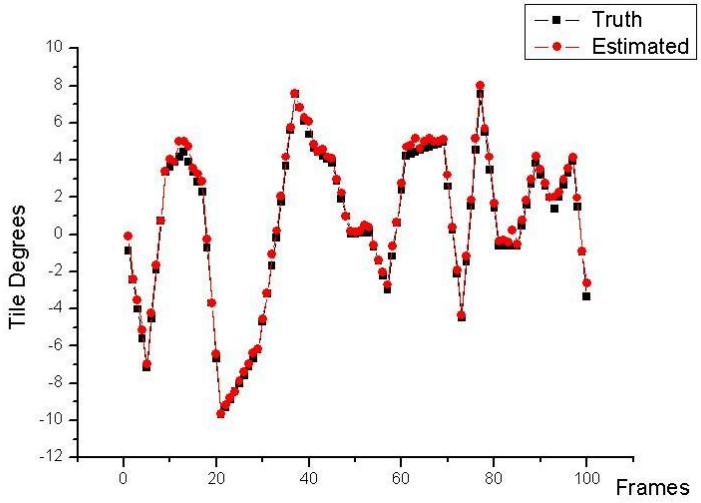


图 4-18 头部俯仰的估计值和真实值的比较

Fig.4-18 Comparing estimations and ground truth of head tilt

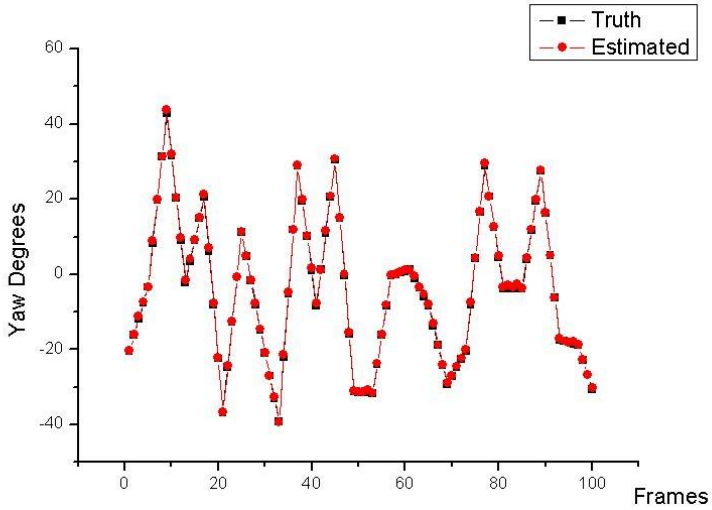


图 4-19 头部深度旋转的估计值和真实值的比较

Fig.4-19 Comparing estimations and the ground truth of head yaw

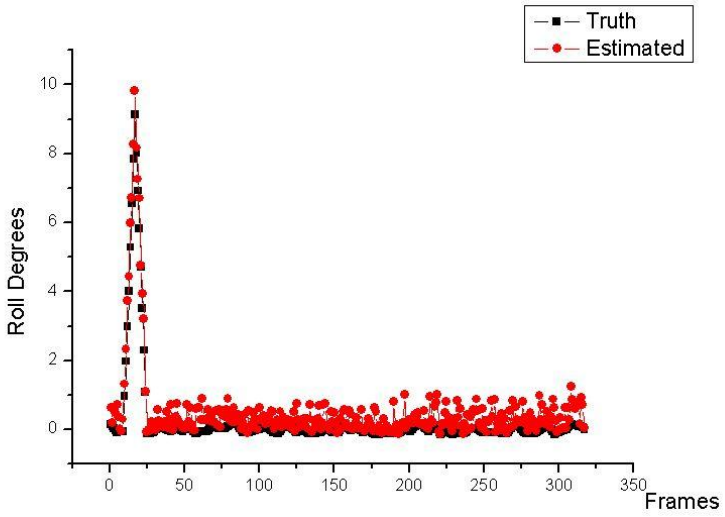


图 4-20 头部平面旋转的估计值和真实值的比较

Fig.4-20 Comparing estimations and the ground truth of head roll

图 4-18，19 和 20 对比了前 100 帧的目标运动状态跟踪结果和目标真实运动状态。由于表情等脸部运动形式的跟踪精确度不容易直观地表示出来，这里给出头部的深度旋转，平面旋转和俯仰的比较结果，图中红色的虚线表示跟踪结果，蓝色实线表示目标真实状态。对所有 3 个序列来说，头部深度旋转、平面旋转和俯仰的平均误差分别为 0.2908, 0.2973 和 0.3166 度。图 4-21 中给出了算法在不同子模型间分布粒子的能力 (这里随机选择了子模型 3)。从图 4-21 中看到，当子模型上的运动越剧烈，说明目标的运动能被该子运动模型描述的概率越高，相应地分配在该子模型上进行状态估计的粒子就越多。

在标准序列蒙特卡洛滤波算法中采用相同的目标状态进行头部运动跟踪。使用的运动模型为几乎匀速直线运动模型和布朗运动模型。要得到与基于多模型协同的序列蒙特卡洛滤波算法相当的跟踪结果，大概需要 2×10^6 数量级的粒子，如此高的计算资源要求使标准序列蒙特卡洛滤波算法在实际问题中很难应用。

4.7.6 算法性能的定性分析

录制了两段视频序列来检测算法在真实数据上的性能。视频序列的分辨率都是 320x320，分别有 132 和 189 帧。注意到重建的三维头部模型和真实人之间的差别所构成的噪声干扰。在广泛采用的梯度下降搜索的跟踪算法中，这种噪声往往容易造成误差的传递和累积，最终导致使跟踪算法失效的“漂移”现象。在算法中采用 6800 个粒子得到了稳定收敛的结果。图 4-22 和 4-23 中奇数行显示了部分原始视频帧，偶数行则显示了根据跟踪结果重构的对应视频帧。

为了检验算法对估计误差的鲁棒性，随机在算法的部分帧估计结果上加入噪声产生人工 “漂移”。从结果的统计平均，当加入 $0.28\times3\times\sigma$ 3 的偏移，算法在 4 帧之内可以 90.5%的概率从错误中恢复。

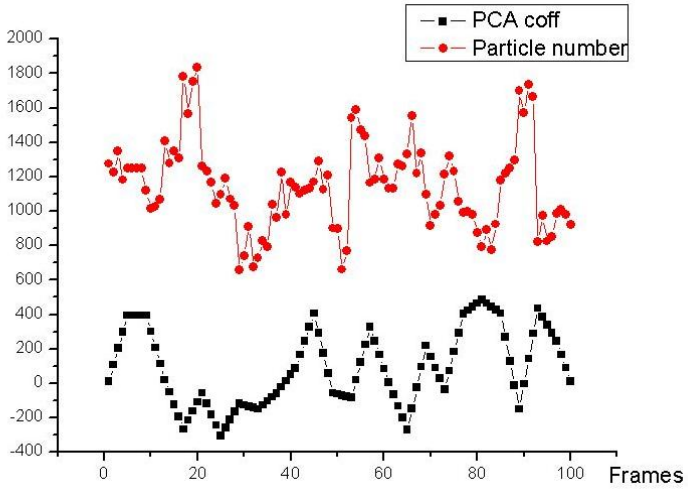


图 4-21 跟踪过程中粒子数在某子模型上的分布情况

Fig.4-21 Distribution of particles during tracking in one sub-model

图 4-22 原始视频帧与根据估计结果合成帧之间的对比

Fig.4-22 Comparing original frames with synthesized frames using estimated FAP values

图 4-23 原始视频帧与根据估计结果合成帧之间的对比

Fig.4-23 Comparing original frames with synthesized frames using estimated FAP values

4.8 小结

本章针对具有复杂运动模式的目标运动建模问题进行了研究。首先总结了具有复杂运动模式的两种目标跟踪问题，并提出了融合多模型的运动估计框架。在该框架的基础上，将其与标准序列蒙特卡洛滤波算法结合，从而提出了两种序列蒙特卡洛滤波算法的改进：基于多模型切换的序列蒙特卡洛滤波算法和基于多模型协同的序列蒙特卡洛滤波算法。实验证明，该两种算法在针对特定种类具有复杂运动模式的目标跟踪问题上相对于标准序列蒙特卡洛滤波算法具有明显的性能提升。

第5章 融合光流和模型的面部特征点跟踪算法

5.1 引言

很多视觉应用，如基于图像序列的三维人脸重建、面部表情分析、智能人机交互等，都倚重于准确的面部特征点跟踪技术。但是由于面部具有典型的非刚性形变等复杂运动模式，并且具有多以肤色为主调的平滑纹理，所以面部特征点跟踪是基于特征点运动感知任务中的困难问题。在本章中，在序列蒙特卡洛滤波算法的框架下融合了基于光流和基于特定模型的方法来解决面部特征点跟踪问题。

现有的面部特征点跟踪方法主要可以分为基于特定描述模型(以下简称模型)和基于光流的方法。基于模型的方法通常对特征点的变化进行描述性建模。在跟踪过程中，利用模型从每帧图像中选择具有物理对应关系的相同特征点集合，并在相邻帧的特征点集合之间建立一一映射来完成跟踪任务。McKenna 等人选择 Gabor 特征，采用点分布模型描述面部特征点的运动[93]。很多的唇动跟踪方法都是 Kass 和 Terzopoulos 提出的 Snake 方法或者 Yuille 提出的变形模板技术的改进[29, 94]。Coots 提出的主动形状模型(Active Shape Model)和主动表观模型(Active Appearance Model)在面部特征点跟踪问题上，尤其是当特征点具有复杂运动模式时，获得了优异的效果[32, 33]。基于模型的方法可以有效刻画具有复杂运动模式的面部特征点集，但模型的获取通常需要繁琐的示例学习，而且随着需要跟踪的面部特征点的增加，模型愈加复杂，对优化求解过程中的起始搜索位置也愈加敏感。

与基于模型的方法不同，基于光流的方法通常在参考帧中选择一组特征点，并假设特征点纹理在帧间保持不变或者变化微小，然后通过在局部图像窗口内进行匹配搜索完成跟踪任务[95]。在基于光流的特征点跟踪领域，Lucas 和 Kanade 采用平面位移运动模型来匹配立体视计算中的对应图像[96]。基于[96]的工作，Kanade-Lucas-Tomasi (KLT) 算法采用图像灰度差的平方和(SSD)作为特征点的匹配准则[97]。KLT 算法被广泛应用于特征点跟踪问题中，并且存在很多改进方案[98]。基于光流的方法虽然有较高的求解效率并适合跟踪稠密的特征点集合，但对纹理变化复杂的情况，常常由于误匹配而造成被跟踪点的丢失。在[99]的工作中，探讨了融合光流和 DAM 模型

的特征点跟踪方法。

提出的面部特征点跟踪算法通过结合基于光流和基于特定运动模型跟踪方法的优点来克服各自的不足，从而达到鲁棒跟踪面部特征点的目的。

5.2 算法的理论层面分析

序列蒙特卡洛滤波算法依据重要性采样定理对后验分布 $p(\mathbf{x}_t \mid \mathbf{z}_t)$ 进行带有随机性的采样来估计最优解。标准序列蒙特卡洛滤波算法的提议分布通过状态转移模型 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 构造(见第 3 章)。虽然 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 编码了目标运动状态在状态空间中的变化轨迹，但只采用该模型对未知状态进行预测是低效率的，尤其当目标具有复杂运动模式时，如人脸特征点跟踪问题，构造 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 首先就是困难的问题。低效率的提议分布降低了序列蒙特卡洛滤波算法的精度，同时增加了为保证算法收敛所需要的计算量。

另一方面，序列蒙特卡洛滤波算法是非参数的贝叶斯时序滤波器。基于非参数的策略虽然赋予了算法在概率上逼近任意形式后验分布的形式。但需要在高维状态空间中进行目标运动状态估计时，标准序列蒙特卡洛滤波算法仍无法逃脱非参数方法的固有弱点：为保证估计结果收敛到真实值，所需要的采样数量随目标状态的维数呈指数增长，从而造成计算上的不可解。

只采用 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ 在状态空间中传递粒子并没有利用隐含了 $\mathbf{x}_t$ 重要信息的最新表观信息 $\mathbf{z}_t$。所以摒弃标准序列蒙特卡洛滤波算法中将提议分布简单地定义为 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$，同时利用 $\mathbf{z}_t$ 和 $\mathbf{x}_{t-1}$ 来进行粒子传递，即将提议分布定义为 $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t)$。在提出的面部特征点跟踪算法中，虽然基于光流算法的帧间运动估计是不鲁棒的，但其计算简单，并能够有效提取人脸特征点的部分最新表观 $\mathbf{z}_t$，从而改善了粒子传递效率，降低了粒子位于状态关联函数“尾部”的概率，提升了算法总体性能。

在借助最新表观指导粒子传递方面，已经有部分学者进行了探讨。在[100]中，Isard 等人将基于颜色的分割算法和基于主动轮廓的跟踪算法在标准序列蒙特卡洛滤波跟踪框架下相结合进行手部运动的跟踪。算法首先采用分割算法找出图像中的肤色区域，然后以肤色区域为线索，进行粒子的传递和权重更新。Freitas 采用扩展卡尔曼滤波算法构建序列蒙特卡洛滤波算法中的提议分布[101]，扩展卡尔曼滤波算法本身包含了表观提取的过程。Julier 和 Ullman 提出了 Unscented Kalman Filter (UKF)算法[102]。算法认为相对于去逼近任意非线性函数，逼近高斯函数是更简单可行的策略，并采用离散的采样点集合来参数化后验概率的均值和方差。实验结果表明，UKF 算法相对于扩展卡尔曼滤波算法具有更好的估计精度和更准确的状态协方差估计。然而，UKF 仍然无法处理本质非线性问题。在[102]和[101]的基础上，Merwe 等人将扩展卡尔曼滤波算法替换为 UKF 算法[65]，采用 UKF 构造提议分布的序列蒙特卡洛滤波算法具有能够利用最新的表观信息 $\mathbf{z}_t$ 的优秀特性，并通过合成数据实验观察到相对于[101]中算法的性能提升。Rui 和 Chen 将[65]提出的改进序列蒙特卡洛滤波算法应用到基于听觉的话者定位和基于视觉的人体跟踪研究中，并获得了优于标准序列蒙特卡洛滤波算法的结果[70]。

5.3 结合光流和模型的面部特征点跟踪

KLT 光流算法是特征点跟踪领域的经典方法[60]，至今仍被广泛应用，但只有被跟踪的特征点具有丰富纹理且形变较小，算法才能取得理想的跟踪结果。对人脸跟踪问题而言，由于绝大多数面部是平滑的肤色，并且运动过程中面部含有非线性形变，所以直接采用 KLT 跟踪方法推断面部运动是很困难的问题。在本节中，将特定模型与 KLT 光流算法相结合，从而根据面部运动的特点提出面部特征点运动的跟踪方法。

5.3.1 KLT 光流跟踪算法

考虑图像序列 $I(\mathbf{x},t)$，$\mathbf{x}=(u,v)$ 是图像像素坐标值。在高采样帧率假设下，认为相邻帧间对应小区域的灰度值恒定，

$$I(\mathbf{x},t)=I(\mathbf{\Phi}(\mathbf{x}),t+\delta t) \tag{5-1}$$

其中 $\Phi(x)$ 是运动场。若有限时间间隔内特征点区域形变很小，定义 $\mathbf{\Phi}(\mathbf{x})$ 为位移运动模型,

$$\Phi(\mathbf{x}) = \mathbf{x} + \mathbf{d} \tag{5-2}$$

于是算法只需关心相邻帧间特征点位移运动$\mathbf{d}$。特征匹配度定义为图像窗口间的平方误差和(Sum of Squared Differences, SSD),

$$\varepsilon[\mathbf{d}(\mathbf{x})] = \sum_{R} [I(\mathbf{x}+\mathbf{d}(\mathbf{x}), t+\tau) - I(\mathbf{x}, t)]^2 w(\mathbf{x}) \tag{5-3}$$

其中$w(\mathbf{x})$是窗加权函数，通常取等权窗或者高斯窗(见图5-1)。

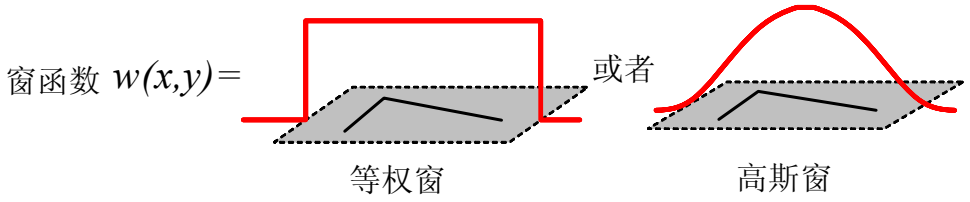


图 5-1 等权窗和高斯窗

Fig.5-1 Gaussian kernel and uniform kernel

对$I(\mathbf{x}+\mathbf{d}(\mathbf{x}), t+\tau)$进行泰勒展开并只保留一阶线性近似项，有

$$\begin{aligned} I(\mathbf{x}+\mathbf{d}(\mathbf{x}), t+\tau) &= I(\mathbf{x}, t) + \frac{\partial I}{\partial \mathbf{x}} \cdot \mathbf{d}(\mathbf{x}) + \frac{\partial I}{\partial t} \cdot \tau \\ &= I(\mathbf{x}, t) + \mathbf{g} \cdot \mathbf{d}(\mathbf{x}) + h \cdot \tau \end{aligned} \tag{5-4}$$

将(5-4)带入(5-3)可得

$$\varepsilon[\mathbf{d}(\mathbf{x})] = \sum_{R} (\mathbf{g} \cdot \mathbf{d} + h \cdot \tau)^2 w(\mathbf{x}) \tag{5-5}$$

令$\frac{\partial \varepsilon}{\partial \mathbf{d}} = 0$，有

$$\sum_{R} (\mathbf{g} \cdot \mathbf{d} + h \cdot \tau) \mathbf{g} w(\mathbf{x}) = 0 \tag{5-6}$$

由于$(\mathbf{g} \cdot \mathbf{d})\mathbf{g} = (\mathbf{g}\mathbf{g}^T)\mathbf{d}$，并假设$\mathbf{d}$在小区域$R$中一致，有

$$[\sum_{R} \mathbf{g}\mathbf{g}^T w(\mathbf{x})] \cdot \mathbf{d} = -\tau \sum_{R} \mathbf{g} h w(\mathbf{x}) \tag{5-7}$$

$$\mathbf{G} \cdot \mathbf{d} = \mathbf{e} \tag{5-8}$$

其中，$\mathbf{G}=\sum_{R}\mathbf{g}\mathbf{g}^{T}w(\mathbf{x})=\sum_{R}\begin{bmatrix} I_u^{\,2} & I_uI_v \\ I_uI_v & I_v^{\,2}\end{bmatrix}\cdot w(u,v)$ 是图像区域的二阶自回归矩阵，

$e=-\tau\sum_{R}\mathbf{g}hw(\mathbf{x})=-\tau\sum_{W}w(\mathbf{x})I_t[I_u \quad I_v]^{\mathrm{T}}$，$[I_u \quad I_v]=\frac{\partial I}{\partial\mathbf{x}}=[\frac{\partial I}{\partial u},\frac{\partial I}{\partial v}]^{T}$。

相邻帧间的特征点匹配变成解方程组 $\mathbf{d}=\mathbf{G}^{-1}\cdot\mathbf{e}$。根据 Newton–Raphson 梯度下降法迭代地求解 **d** 直到其收敛。当矩阵 G 奇异，无法确定两个唯一的 **d**，称为光流算法中“孔径问题”(aperture problem)。为避免该种情况，需要选择矩阵 G 非奇异的那些特征点。对矩阵 **G** 进行奇异值分解，得到特征值 $\{\lambda_1,\lambda_2\}$和特征向量$\{\Phi_1,\Phi_2\}$，选取的特征点应满足 $\min(\lambda_1,\lambda_2)>T$，阈值 T 与图像质量相关，图像噪声越大，需要越高的阈值。

5.3.2 基于尺度空间理论的特征点尺度自动选择

对 6.2 节给出的面部特征点跟踪算法，影响其性能的一个重要因素是特征点的选择。特征点的选择包含两个关键因素：特征点位置的选择和特征点尺度的选择。当预先设定的特征点尺度不同时，所选择的特征点有较大差别并对算法性能有较大影响。

在本节中，针对特征点选择的问题对算法进行改进，相对于根据经验或者实验的特征点选择方法，将图像匹配领域的尺度空间理论引入到特征点跟踪领域中，提出特征点的自动选择算法。

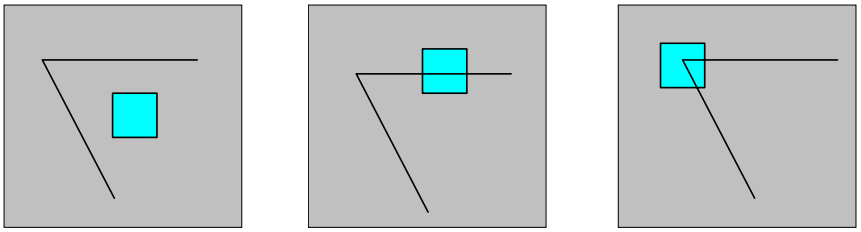
5.3.3 多尺度 Harris 特征点选择算法

首先给出特征点的选择标准。假设特征点的定位误差是 (u,v)，则由定位误差引起的匹配误差为，

$$E(u,v)=\sum_{R}[I(u+du,v+dv))-I(u,v)]^{2}w(u,v) \tag{5-9}$$

希望对选择的特征点有小的位移误差 (du,dv) 时，就会产生较大匹配误差 $E(u,v)$，从而使特征点对噪声有较好的鲁棒性。从图 5-2 可以直观地对此进行解释。在图 5-2(a)中，特征点位于平坦区域，当定位偏差 $du>0,dv>0$，可以看到由于 $E(u,v)$ 近似为零，所以算法很难准确定位该特征点位置，同时也容易受到噪声干扰。在图 5-2(b)中，当 $(du>0,dv=0)$，同样会出现 $E(u,v)$

近似为 0。但当$(du=0,dv>0)$，会有较大的匹配误差。说明位于边缘上的特征点只对垂直边缘方向的误差敏感，从而使此类特征点在垂直边缘方向比在水平边缘方向有相对高的定位精度和信度。图 5-2(c)中的特征点，无论特征点的位移发生在哪个方向，都会产生较大的$E(u,v)$，这也解释了角点在定位和跟踪上的优越性。



(a)　　(b)　　(c)

图 5-2 特征点纹理对其跟踪性能的影响

Fig.5-2 Characteristic texture and its relation to trackability

将公式(5-9)进行一阶泰勒展开

$$\begin{aligned} E(u,v) &= \sum_R [I(u,v)+\frac{\partial I}{\partial u}du+\frac{\partial I}{\partial v}dv-I(u,v)]^2 w(u,v) \\ &= \sum_R (\frac{\partial I}{\partial u}du+\frac{\partial I}{\partial v}dv)^2 w(u,v) \\ &= \sum_R \begin{bmatrix} du \\ dv \end{bmatrix} \begin{bmatrix} I_u^2 & I_u I_v \\ I_u I_v & I_v^2 \end{bmatrix} \begin{bmatrix} du & dv \end{bmatrix} w(u,v) \end{aligned} \tag{5-10}$$

注意到假定特征点区域 R 内所有像素点具有相同的误差$(du，dv)$，于是有

$$\begin{aligned} E(u,v) &= \begin{bmatrix} du \\ dv \end{bmatrix} \sum_R w(u,v) \begin{bmatrix} I_u^2 & I_u I_v \\ I_u I_v & I_v^2 \end{bmatrix} \begin{bmatrix} du & dv \end{bmatrix} \\ &= \begin{bmatrix} du \\ dv \end{bmatrix} \mathbf{G} \begin{bmatrix} du & dv \end{bmatrix} \end{aligned} \tag{5-11}$$

其中 G 是图像区域自回归矩阵(参见公式(5-8))。根据公式(5-11)可以得出结论：除 (du,dv)，矩阵 G 是影响 $E(u,v)$ 的唯一因素。

根据该原则，Harris 特征点选择算法为：首先对区域计算 **G**，然后计算

$$R(u,v)=\det(\mathbf{G})-k\cdot(\mathrm{trace}(\mathbf{G}))^2 \tag{5-12}$$

按照 $R(u，v)$ 的降序排列依次选择特征点，并且所选择的特征点应该满足 $\min(\lambda_1,\lambda_2)>T$。

求取自回归矩阵G，需要选择区域R的尺度(称为积分尺度)和求取 I_u，I_v的尺度(称为微分尺度)。已有工作证明[103]，当一幅图像发生尺度变化，特定的特征点的积分尺度和微分尺度能够准确地反应图像尺度的变化，该特定的微分尺度和积分尺度称为特征尺度。如果图像尺度变化未知，则需要在不同的积分和微分尺度上求取特征点，从而得到对同一图像特征点的多尺度表示。

多尺度Harris特征点提取的二阶矩表示为

$$\mathbf{G}(\mathbf{x},\sigma_I,\sigma_D)=K_g(\mathbf{x},\sigma_I)\otimes\begin{bmatrix} I_x^{\,2}(\mathbf{x},\sigma_D) & I_xI_y(\mathbf{x},\sigma_D)\\ I_xI_y(\mathbf{x},\sigma_D) & I_y^{\,2}(\mathbf{x},\sigma_D)\end{bmatrix} \tag{5-13}$$

其中 $K_g(\cdot)$ 是高斯核函数，σ_I 和 σ_D 分别是积分和微分尺度。将 σ_I 和 σ_D 在不同尺度上取值组合，便得到基于多尺度的Harris特征点。为减少搜索空间并不失一般性，约定 $\sigma_I=\alpha\sigma_D$，其中 α 是标量。矩阵 $\mathbf{G}(\mathbf{x},\sigma_I,\sigma_D)$ 的意义可用图5-3中的椭圆解释，设 $\mathbf{G}(\mathbf{x},\sigma_I,\sigma_D)$ 的特征值和特征向量分别为 $\{\lambda_1,\lambda_2\}$ 和 $\{\Phi_1,\Phi_2\}$。定义 $\lambda_1>\lambda_2$，则 Φ_1 对应该图像区域梯度变化最大的方向，Φ_2 和 λ_2 刻画了和 Φ_1 垂直方向的图像梯度变化情况。λ_1 和 λ_2 和特征点形式具有如表5-1的关系。

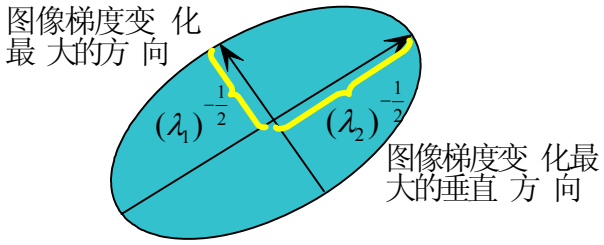


图 5-3 图像区域的自回归矩阵的物理意义

Fig.5-3 The vivid explanation of auto-correlatition matrix computed on image region

表 5-1 图像区域自回归矩阵和图像灰度值分布之间的关系

Table 5-1 Relations between auto-correlation matrix of image region and distribution of image pixel values

	λ_1和λ_2的关系	R
平坦区域	$\lambda_1 < T$ 并且 $\lambda_1 < T$	0<R<T
垂直边缘	$\lambda_1 \gg \lambda_2$	R<0
水平边缘	$\lambda_2 \gg \lambda_1$	R<0
角点	$\lambda_1 > T$ 并且 $\lambda_2 > T$	0<T<R

5.3.4 尺度空间理论

图像 $I(x,y)$ 的尺度空间 $L(x,y,\sigma)$ 定义为不同核宽 σ 的一组高斯核 $G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}\exp(-(x^2+y^2)/2\sigma^2)$ 对 $I(x,y)$ 进行卷积所得到的图像集合，二维图像的尺度空间是三维的[104]。

在图像配准研究领域，很多学者依据尺度空间理论提出了特征点尺度的确定准则，使特征点尺度只与图像内容有关，而与图像呈现尺度无关。在[105, 106]的工作中，采用高斯差(Difference-of-Gaussian，DoG)算子计算图像金字塔，选取金字塔三维局部邻域中的最值点为特征点。Linderberg在采用尺度归一化的微分算子所得到的三维尺度空间中选择局部极值做为特征点，并推荐使用拉普拉斯－高斯(Laplacian-of-Gaussian，LoG)算子[107]。

Lowe用DoG算子构造的图像金字塔来选择局部极值点进行物体描述和识别[108]。输入图像被具有不同核宽的高斯算子处理后得到平滑图像集合，对每两帧相邻图像做差得到图像的DoG算子卷积结果。DoG卷积结果金字塔中的三维极值被用来作为特征点。DoG算子是对LoG算子的一种良好逼近，同时相对于LoG算子能够大大降低运算复杂性。

通过在尺度空间中定义响应函数 $f_s(I,s)$，可以找到特征点的特征尺度，从而使特征点只于蕴含的图像内容有关，而与图像呈现形式无关。理想的响应函数应该具有如图5-4所示的性质：

1) 响应函数中极值的出现位置和尺度的变化成正比；
2) 在极值两端响应函数呈单调变化。

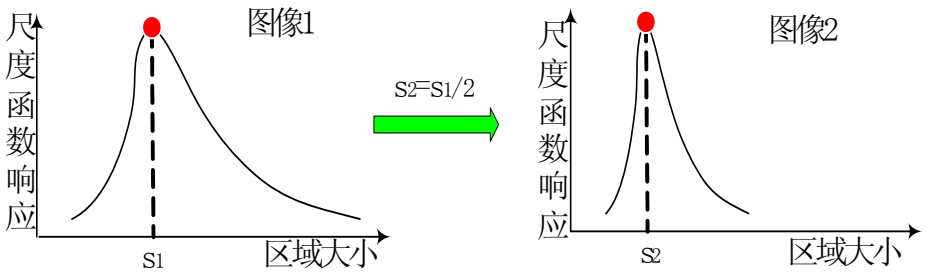


图 5-4 尺度响应函数示例

Fig.5-4 Illustration of scale response function

这里，采用[104]中定义的算子作为尺度响应函数，首先采用LoG算子与图像进行卷积，然后用 σ 对卷积结果进行归一化，

$$\left|\sigma^2(I_{xx}(\mathbf{x},\sigma)+I_{yy}(\mathbf{x},\sigma))\right| \tag{5-14}$$

通过调整核宽 σ，选择具有最大响应值的尺度 $\sigma_{\max}$ 作为特征点的特征尺度。

尺度选择的具体过程如图5-5：

基于尺度空间理论的特征点自动选择算法流程

输入：参考图像I。

输出：选择的特征点集合$\{(\mathbf{x}^k, s^k) \mid k = 1,...,K\}$，其中$s^k$是与特征点$k$相关联的特征点尺度。

初始化：定义初始的核函数尺度σ_0，并取$\sigma_n = \gamma^n \sigma_0$，并用$\sigma_n$对原始图像进行平滑得到尺度空间中的第$n$层表示(实验中，$\gamma = 1.4$)。在尺度空间中，根据公式(5-33)和(5-34)计算多尺度Harris特征点集合$\{\mathbf{x}^k \mid k = 1,...,K\}$。

For 特征点$\mathbf{x}^k$，

采用公式(5-34)在图像I的尺度空间L中分别计算对应$\sigma_0 \to \sigma_n$的尺度 上的响应值，选择响应最大记为σ_{ini}。

For $t = 0.7:0.1:1.4,$

定义 $\sigma_{tmp}{}^t = t \cdot \sigma_{ini}$ ，采用公式(5-34)在尺度$\sigma_{tmp}{}^t$上计算响应值v^t。

End

选择特征点的特征尺度为$s^k = \arg\max_{v^t}(\sigma_{tmp}{}^t)$。

End

图 5-5 基于尺度空间理论和多尺度 Harris 算法的特征点选择流程

Fig.5-5 Flowchart of interest point selection based on scale space theory and multi-scale Harris algorithm

5.3.5 嘴部特征点跟踪

由于嘴部特征点具有较大的非刚性形变，这里采用基于合成的分析策略，构造描述嘴部特征点运动特性的模型完成嘴部特征点的跟踪。

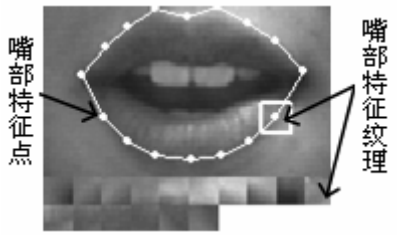


图 5-6 嘴部特征点的定义及其运动描述模型

Fig.5-6 Mouth feature points definition and its motion description model

以 AAM 方法中的表示策略为基础[33]，嘴部描述模型主要有两部分构成：点分布模型和纹理模型(见图 5-6)。对每幅训练图像 i，对嘴部特征点进行标注，然后将所有特征点的坐标值按顺序首尾连接构成向量

$$\{\mathbf{x}_i = \{(u_j, v_j) \mid j = 1,...,M\} \tag{5-15}$$

通过对所有训练图像的$\{\mathbf{x}_i \mid i = 1,...,N\}$进行主成分分析并将特征向量写为矩阵的形式，于是训练集合中的任一嘴部特征点分布可以写作

$$\mathbf{x}_i \approx \overline{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \tag{5-16}$$

其中，$\overline{\mathbf{x}}$为平均形状，$\mathbf{b}_s$为形状参数，$\mathbf{P}_s$为主成分特征向量构成的变换矩阵。上述点分布模型刻画了嘴部特征点的拓扑结构。然后，对$\mathbf{x}_i$中的每一特征点(u_j, v_j)，以其为中心定义图像子窗口$\Omega(u_j, v_j)$，将窗口内的像素按行优先展开成向量$g_j = \{I(u,v) \mid (u,v) \in \Omega(u_j, v_j)\}$，其中 I(u,v)表示(u,v)位置的像素值。再将 g_j 连接成向量作为嘴部的纹理描述集合$G_i = \{g_j \mid j = 1,...,M\}$。对所$G_i$进行主成分分析，从而得到如下统计模型：

$$\mathbf{g}_i \approx \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{5-17}$$

其中，$\bar{\mathbf{g}}$为平均纹理，$\mathbf{P}_g$为PCA 计算得到的纹理主成分特征向量形成的变换矩阵，$\mathbf{b}_g$为控制纹理变化的统计纹理参数。

进一步将公式(5-16)和(5-17)定义的形状和纹理模型融合起来。即：将$\mathbf{b}_s$和$\mathbf{b}_g$串接起来得到新的特征向量：

$$\mathbf{b}=\begin{pmatrix}\varphi_s\mathbf{b}_s\\ \mathbf{b}_g\end{pmatrix} \tag{5-18}$$

其中，对角阵φ_s用来调整$\mathbf{b}_s$和$\mathbf{b}_g$二者之间量纲的不同。对得到的$\mathbf{b}$进行主成分分析，进一步消除形状和纹理之间的相关性，从而得到表观模型：

$$\mathbf{b}=\bar{\mathbf{b}}+\mathbf{Q}\cdot\mathbf{c} \tag{5-19}$$

其中，$\bar{\mathbf{b}}$为平均表观向量，Q 为表观主成分特征向量形成的变换矩阵，c 为控制表观变化的统计表观参数。这样，给定表观模型参数 c 以及相应的相似变换参数$\Theta=(u_0,v_0,s,\theta)$，就可以合成一幅嘴部纹理：

$$I_m=T(warp^{-1}(\mathbf{s}_m,\mathbf{g}_m);\Theta) \tag{5-20}$$

其中：$warp^{-1}(\cdot)$表示将平均形状下的模型纹理$\mathbf{g}_m$变形为模型形状$\mathbf{s}_m$的操作，T 则表示对其进一步进行参数为Θ的相似变换，从而得到嘴部纹理I_m。而$\mathbf{g}_m$和$\mathbf{s}_m$则分别通过公式(5-16)和(5-17)得到。

根据基于合成的分析策略，输入一幅新的嘴部图像I_n，特征配准是一个表观模型参数 c 的优化过程，以期最终达到模型纹理与输入纹理的最佳匹配。匹配过程基于序列蒙特卡洛滤波算法。对特征点跟踪问题，假设已经知道上一时刻的特征点集合状态$\mathbf{x}_{t-1}$以及通过 KLT 光流算法的部分特征点跟踪结果$\tilde{\mathbf{z}}_t$，以$p(\mathbf{x}_t\mid\mathbf{x}_{t-1},\tilde{\mathbf{z}}_t)$为提议分布进行粒子的传递。由于人脸特征点跟踪

问题中，目标表观和目标状态的一致性， 并且将人脸特征点分为来能够个互不相交的集合，于是有

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \tilde{\mathbf{z}}_t) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_t \mid \tilde{\mathbf{z}}_t) \tag{5-21}$$

$p(\mathbf{x}_t \mid \tilde{\mathbf{z}}_t)$约束了嘴部模型的角度、尺度、重心位移等参数，从而大大降低了序列蒙特卡洛滤波算法中粒子的搜索空间。

融合帧间运动估计的面部特征点跟踪算法流程

算法输入：带有面部运动的视频序列

算法输出：面部特征点的跟踪结果

初始化： 对参考帧，采用 6.2.1 节中描述的特征点选择算法选择初始非嘴部面部特征点集合Φ_1，根据嘴部描述模型手工选择嘴部特征点Φ_2。

跟踪阶段：

For 每帧图像 t，

1) 根据尺度空间增强的 KLT 特征点跟踪算法得到面部特征点Φ_1^t的 跟 踪结果，根据Φ_1^t估计头部的旋转角度，尺度变化，位移等参数Θ_t。

2) 根据状态转移模型$P_m(\mathbf{x}_i \mid \mathbf{x}_{i-1}, \Theta_t)$中采样 N个粒子。

3) 根据 6.4.3 中定义的描述性嘴部模型估计每 个粒子$\mathbf{x}_t^j$的权重w_t^j。

4) 融合每个粒子的估计得到嘴部跟踪结果$\Phi_2 = \sum_{j=1}^{N} \frac{w_t^j \mathbf{x}_t^j}{w_t^j}$。

5) 对粒子集合，计算归一化权重的方差。如果方差超过阈值，则根据粒子权重大小作为其采样概率，有放回的从粒子集合中采样生成新集合。

End

图 5-7 融合帧间运动估计的面部特征点跟踪算法流程

Fig.5-7 Flowchart of facial feature tracking algorithm with inter-frame motion estimation

5.4 采用子空间约束的跟踪结果求精过程

5.4.1.1 特征点群运动的子空间约束

假设已经计算得到运动刚体上 P 个特征点在 F 帧中的位置，将所有位置信息写成观测矩阵的形式，

$$\mathbf{W}_{2F\times P}=\begin{bmatrix} u_{11} & \cdots & u_{1P} \\ \vdots & & \vdots \\ u_{F1} & \cdots & u_{FP} \\ v_{11} & \cdots & v_{1P} \\ \vdots & & \vdots \\ v_{F1} & \cdots & v_{FP} \end{bmatrix} \tag{5-22}$$

其中，(u_{fp},v_{fp}) 表示特征点 p 在帧 f 中的二维坐标。$\mathbf{W}$ 中一列表示同一特征点在 F 帧序列中的轨迹，$\mathbf{W}$ 中一行则表示同一帧中所有特征点的位置。用 $\mathbf{i}_f$，$\mathbf{j}_f$ 和 $\mathbf{k}_f$ 表示帧 f 中摄像机的方位，$\mathbf{i}_f$ 和 $\mathbf{j}_f$ 分别对应成像平面的 u 轴和 v 轴，则有 $\mathbf{k}_f=\mathbf{i}_f\times\mathbf{j}_f$。设在世界坐标系中选择目标物体上一点 $\mathbf{s}_p=(x_p,y_p,z_p)$ 在图像平面上投影为 (u_{fp},v_{fp})，在垂直投影关系假设下，可以得到

$$u_{fp}=\mathbf{i}_f{}^T(\mathbf{s}_p-\mathbf{t}_f) \tag{5-23}$$

$$v_{fp}=\mathbf{j}_f{}^T(\mathbf{s}_p-\mathbf{t}_f) \tag{5-24}$$

其中，$\mathbf{t}_f=(a_f,b_f,c_f)^T$ 是世界坐标系原点到目标特征点质心之间的位移。为简化问题，将世界坐标系的原心定义为特征点集合的质心，(5-23)和(5-24)转化为

$$u_{fp}=\mathbf{i}_f{}^T\mathbf{s}_p \tag{5-25}$$

$$v_{fp} = \mathbf{j}_f{}^T \mathbf{s}_p \tag{5-26}$$

将所有 F 帧的摄像机方向向量写成运动矩阵，

$$\mathbf{M}_{2F\times 3} = \left[\mathbf{i}_1, \ \ldots \ \mathbf{i}_F, \ \mathbf{j}_1, \ \ldots \ \mathbf{j}_F\right]^T \tag{5-27}$$

将刚体形状写为矩阵形式，

$$\mathbf{S}_{3\times P} = \left[\mathbf{s}_1 \ \ \ldots \ \ \mathbf{s}_P\right] \tag{5-28}$$

注意到有 $\sum_{p=1}^{P} \mathbf{s}_p = 0$，(5-22)可以表示为

$$\mathbf{W}_{2F\times P} = \mathbf{M}_{2F\times 3}\mathbf{S}_{3\times P} \tag{5-29}$$

根据矩阵性质可得，$rank(\mathbf{W}) \le \min(rank(\mathbf{M}), rank(\mathbf{S}))$，所以 $rank(\mathbf{W}) \le 3$。由于算法的结果误差和测量噪声等因素的存在，通常造成 $rank(\mathbf{W}) > 3$。将观测矩阵约束在维数低于 3 的子空间内，可以去除由于误差和噪声造成的矩阵升维，提高算法的跟踪精度。

应用空间秩约束将观测矩阵 W 投影到低维空间主要有两个步骤。首先，对矩阵 W 进行奇异值分解，

$$\mathbf{W}_{2F\times P} \to \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{5-30}$$

然后，令 $\mathbf{\Sigma}' = diag(\sigma_1, \sigma_2, \sigma_3)$ 是矩阵 $\mathbf{\Sigma}$ 左上角 3x3 子阵，并有 $\sigma_1 \ge \sigma_2 \ge \sigma_3 > 0$，并分别保留矩阵 U 和 V 中的对应的特征向量，得到 $\mathbf{U}'$ 和 $\mathbf{V}'$，重构观测矩阵 $\mathbf{W}'$，

$$\mathbf{U}'\mathbf{\Sigma}'\mathbf{V}'^T \to \mathbf{W}'_{2F\times P} \tag{5-31}$$

其中，$\mathbf{U}'$ 是运动子空间的一组基，而 $\mathbf{V}'$ 是形状子空间的一组基。

子空间约束实际上是对所有特征点属于同一刚性物体这一约束的数学化描述，相对于对每个特征点单独跟踪的策略或者使用平滑性约束等启发性约束，全局秩约束明确地限制了特征点运动之间的相互关系，从而可以得到更精确的跟踪结果。

然而人脸属于非刚体。Bregler 通过实验验证，具有非刚性形变的运动

物体，其形状可以用一组刚性形体的组合进行拟合[109]，

$$\mathbf{S}=\sum_{i=1}^{K} l_i\cdot\mathbf{S}_i \qquad \mathbf{S},\mathbf{S}_i\in IR^{3xP}, l_i\in IR \tag{5-32}$$

其中**S**是采用P个三维空间点进行描述的具有非刚性形变的物体形状，S_i是采用P个三维空间点描述的第i个基本刚体形状。在弱透视投影情况下，对第f帧，有成像关系：

$$\begin{bmatrix} u_{f1} & \cdots & u_{fP} \\ v_{f1} & \cdots & v_{fP} \end{bmatrix}=\mathbf{R}_f\cdot\left(\sum_{i=1}^{K} l_{fi}\cdot\mathbf{S}_i\right)+\mathbf{T}_f \tag{5-33}$$

其中，$\mathbf{R}_f^{2\times3}$是摄像机旋转矩阵的前两行。将世界坐标系原点定在物体三维形状点集的质心从而消去$\mathrm{T_f}$，可以有

$$\begin{bmatrix} x_{f1} & \cdots & x_{fP} \\ y_{f1} & \cdots & y_{fP} \end{bmatrix}=\begin{bmatrix} l_1\mathbf{R} & \cdots & l_K\mathbf{R} \end{bmatrix}\begin{bmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_K \end{bmatrix} \tag{5-34}$$

将序列中的特征点轨迹写成观测矩阵 **W**，有

$$\mathbf{W}^{2F\times P}=\mathbf{M}^{2F\times3K}\cdot\mathbf{S}^{3K\times P}=\begin{bmatrix} l_{11}\mathbf{R}_1 & \cdots & l_{1K}\mathbf{R}_1 \\ \vdots & & \vdots \\ l_{F1}\mathbf{R}_F & \cdots & l_{FK}\mathbf{R}_F \end{bmatrix}\begin{bmatrix} \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_K \end{bmatrix} \tag{5-35}$$

根据矩阵性质，对无噪声干扰的观测矩阵 **W** 应该有$rank(\mathbf{W})\le\min(3K,P,2F)$。由于通常有$\min(2F,P)>3K$，所以观测矩阵的子空间维数上限应为 $3K$。

5.4.1.2 基于子空间约束的求精算法流程

由 KLT 算法和嘴部描述模型得到面部特征点位置观测矩阵$W=\begin{bmatrix} U \\ V \end{bmatrix}$，通常有 rank($W$)>3K，这是由于跟踪误差的存在，使子空间的维数超出上界。对矩阵 W 应用子空间约束，将其投射到 $3K$ 维数的子空间中，从而达到消除孔径问题和去除噪声的目的。

算法的详细步骤如下所述：

1) 应用 KLT 跟踪算法，对嘴部区域以外的面部特征点进行跟踪(实验

中证明，KLT 算法能够跟踪具有较小非刚性形变的特征点)，并将结果写成观测矩阵 W_k。

2) 将 W_k 分成两个子阵 $W_k=[W_k^{rel},W_k^{unrel}]$，其中 W_k^{rel} 包含可以被 KLT 算法鲁棒跟踪的面部特征点。由于 W_k^{rel} 中特征点的位置仍可能受到噪声污染，先对 W_k^{rel} 进行奇异值分解，$W_k^{rel}\xrightarrow{SVD}L^{rel}M^{rel}R^{rel}$，其中，$L^{rel},R^{rel}$ 是特征向量矩阵，M^{rel} 是奇异值矩阵，通过保留 M^{rel} 中前 6 个最大奇异值得到 M^{rel_new} (由于面颊等非刚性运动，实验中令 K=2 能够比较好的逼近观测矩阵子空间真实维数)，重建 $W_k^{rel\_new}=L^{rel}M^{rel\_new}R^{rel}$，从而将 W_k^{rel} 投影到维数为 6 的子空间中并得到更精确的特征点位置 $W_k^{rel\_new}$。

3) 采用 $W_k^{rel\_new}$ 估计每帧图像中头部和摄像机之间的相对位移，平面内旋转等刚性运动，并据此初始化嘴部描述模型的起始搜索位置。采用第 6.2 节的方法得到嘴部特征点的跟踪结果 W_m。

4) 对于 W_k^{unrel}，利用人脸基本拓扑结构在运动过程中的不变性，根据其近邻特征点的位置对其当前位置进行插值估计，并认为估计偏差服从白噪声分布，则可以将嘴部特征点跟踪结果和 KLT 跟踪结果写成一个新的观测矩阵 $W=[W_k,W_m]$。然后通过奇异值分解将 W 投射到一个 9 维(K=3)的子空间中，从而得到新的特征点跟踪结果。

5) 上述采用子空间约束的求精过程迭代往复。在采用子空间约束求得新的特征点位置后，以该位置更新跟踪初值并输入 KLT 算法和嘴部描述模型，再次求取特征点的位置，然后再采用全局秩约束。直到在两次迭代之间的特征点位置变化小于某阈值(实验中，定为 0.1 像素)或者迭代超过一定的次数(实验中设为 5)。

5.5 实验部分

5.5.1 对 KLT 增强算法的实验验证

所用的测试序列为普通家用摄像机在两年前拍摄，具有 768x576 像素大小，10 Hz 采样频率，长度 256 帧。该序列中的年轻人具有相对光滑的皮肤，从而造成很多特征点周围具有贫乏的纹理。为了测试基于尺度空间的特征点尺度选择的效果，没有采用特定运动模型，只采用了基于尺度空间理论的 KLT 增强算法。

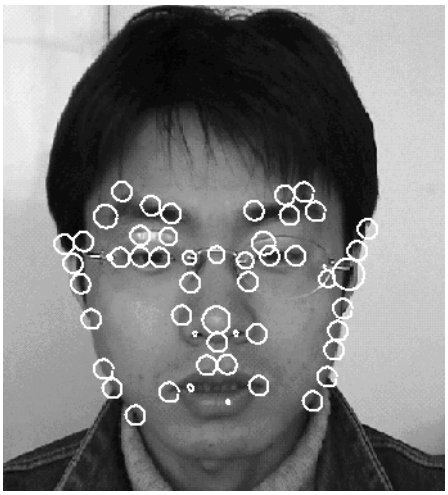


图 5-8 基于尺度空间和多尺度 Harris 特征点检测算法的结果

Fig.5-8 Feature selection results based on scale space theory and multi-scale Harris detector

由于面部的角点比较少，采用基于尺度空间和多尺度 Harris 特征点检测器的特征点选择方法选择了 50 个面部特征点(见图 5-8)。为了对比实验结果，首先在该序列上测试了标准 KLT 特征点跟踪方法的性能。从图 5-9 中下行的 KLT 算法跟踪结果可以看到，在全部 50 个特征点中只有 17 个特征点可以被完整地跟踪整个序列。从结果图像中分析可知，丢失的特征点主要集中于那些具有较大形变的特征点中(例如嘴部周围的特征点)。

图 5-9 中上行给出了基于尺度空间理论的增强特征点跟踪算法的部分实验结果，从图中可以得出结论，通过特征尺度的选择，可以使得特征点对形

变具有一定的鲁棒性，从而获得了更好的跟踪效果。

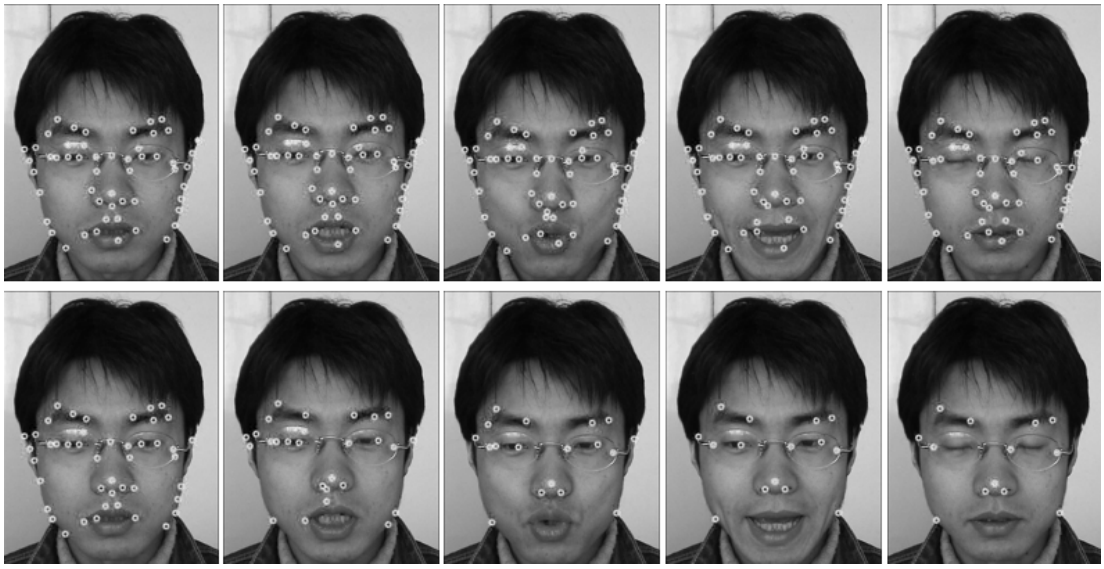


图 5-9 KLT 特征点跟踪算法(下行)和基于尺度空间理论增强的 KLT 特征点跟踪算法(上行)部分跟踪结果的比较。

Fig.5-9 Comparing results from scale-space theory enhanced KLT trakcer (top row) and original KLT tracker (bottom row)

图 5-10 和 5-11 分别给出了 KLT 跟踪算法和基于尺度理论空间增强的 KLT 跟踪算法的定量跟踪误差分析。从图 5-10 中可以看到，大部分的特征点发生了跟踪漂移从而具有较大的跟踪误差。而在图 5-11 中，则只有三个特征点跟踪丢失，从而说明根据尺度空间理论所选择的特征点尺度使特征点具有了一定的形变不变性，从而提高了算法的鲁帮性。

图 5-12 直接给出了根据公式（5-12）计算的特征点可跟踪性度量。从图中可以得出结论，经过自适应地特征点尺度选择之后，特征点的可跟踪性度量增大，从而增加了算法的鲁帮性。

图 5-10 在“wczhang.avi”序列上 KLT 算法的跟踪误差

Fig.5-10 Tracking errors on sequence “wczhang.avi” by KLT tracker

图 5-11 在“wczhang.avi”序列上基于尺度空间理论增强的 KLT 算法的跟踪误差

Fig.5-11 Tracking errors on sequence “wczhang.avi” by scale space enhanced KLT tracker

图 5-12 特征点可跟踪性度量在自适应特征尺度选择和经验设定之间的差异对比

Fig.5-12 Comparing features trackability with adaptive scale and fixed scale

5.5.2 子空间约束特征点跟踪算法的实验验证

采用具有丰富脸部形变的连续 100 帧图像作为测试样例集合验证算法。人脸在 139x177 像素尺寸的图像中大概占据 90x129 大小的像素区域。10Hz 的视频帧率使某些特征点具有较大的帧间位移，同时，对象人脸具有相对光滑的皮肤，造成某些特征点具有不丰富的纹理，这些都增加了跟踪任务的难度。

在实验中，采用基于尺度空间和多尺度 Harris 特征点检测器的特征点选择方法选择了嘴部区域之外的 63 个面部特征点。由于面部的大部分区域比较光滑，一些具有不丰富纹理的特征点被包含在跟踪任务集合中。另外，根据前面提到的嘴部描述模型，在参考帧中选取了 16 个嘴部特征点 (所有选择的特征点见图 5-13 (a)，构成模型训练集合的图像大概在 4000 幅左右)。

对具有丰富纹理，同时符合或者近似符合刚性运动的面部特征点，KLT 算法具有优秀的跟踪效果。但由于面部的很多特征点具有非刚性形变或者不丰富的纹理，从而只有 17 个特征点可以被完整地跟踪 100 帧 (被完整跟踪的特征点见图 5-13 (c))，尤其是具有较大形变的嘴部特征点几乎全部无法被鲁棒跟踪。

应用提出的特征点跟踪算法，全部选定的 79 个特征点中只有 7 个特征点跟踪丢失 (完整跟踪的特征点见图 5-13(b))。图 5-17 给出了一些跟踪结果的示例。在表 5-2 中给出了 KLT 算法和提出的算法的跟踪结果的定量比较(比较标准为手工标注结果)。从表中可以看出，两种算法的跟踪精度相当(提出的算法的跟踪误差稍有上升，这是由于需要跟踪很多具有不丰富纹理和非刚性形变的特征点)，而在可以完整跟踪的特征点数目上，提出的算法远远超过 KLT 算法，从而使人脸的三维重建，表情分析等高层任务成为可能。

图 5-14 和 5-15 分别给出了 KLT 算法和提出的算法在特征点跟踪误差上的定量分析。从结果中可得出结论，在该序列中，KLT 算法跟踪丢失了大量的特征点，而提出的算法则能够获得相对鲁帮的跟踪结果。

实验中，提出的算法实现在未作代码特殊优化的情况下，可在 PIII 766MHz 的 PC 上以平均 2-3 帧/秒的速度运行。

图 5-16 给出了子空间约束对跟踪结果的求精的作用。从该曲线可以看出，子空间约束能够在平均 3-4 次迭代中达到收敛，并且比较显著地降低跟踪误差。

表 5-2 KLT 跟踪算法和提出的跟踪算法的跟踪性能定量分析

Table5-2 Quantitative comparison of proposed method and KLT feture tracker

	原始点数	跟踪点数	水平跟踪精度	垂直跟踪精度
KLT 算法	79	17	0.0361 像素	0.0294 像素
提出的算法	79	72	0.0451 像素	0.0402 像素

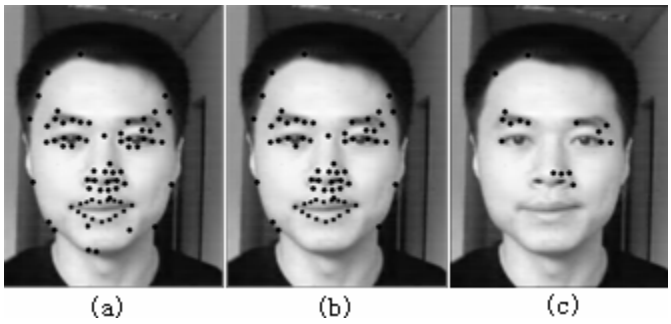


图 5-13 (a) 初始选定的人脸特征点集合 (b)可以被提出的算法完整跟踪的特征点集合；(c) 可以被 KLT 算法完整跟踪的特征点集合

Fig.5-13 (a)Initial feature set (b)Tracked feature by proposed tracker (c)Tracked features by KLT tracker

图 5-14 在“bcao.avi”序列上 KLT 算法的跟踪误差

Fig.5-14 Tracking errors on sequence “bcao.avi” by KLT tracker

图 5-15 在“bcao.avi”序列上提出的算法的跟踪误差

Fig.5-15 Tracking errors on sequence “bcao.avi” by proposed tracker

图 5-16 迭代使用子空间约束后平均误差递减的曲线

Fig.5-16 Error reduction curve by applying subspace constraints iteratively

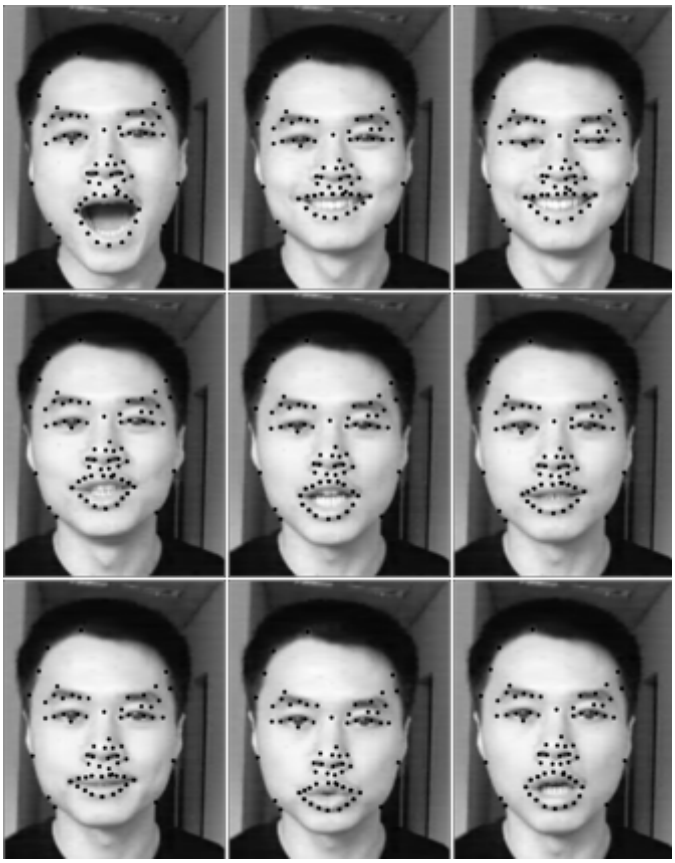


图 5-17 提出算法的部分跟踪结果图例

Fig.5-17 Some tracking results by proposed tracker

由于视觉跟踪中没有像人脸检测类任务那样建立了具有广泛共识和应用的公共测试序列，这里采用视频压缩测试的标准序列之一，“Foreman.mpeg”作为测试序列之一。

该序列中的人物具有夸张的表情，较大的面部形变和头部深度旋转，同时该视频的质量一般，所有这些因素都使面部特征点的跟踪成为困难的问题。

采用公式（5-12）的特征点可跟踪性度量准则和尺度空间理论选择了 24 个面部特征，同时根据预先标注的嘴部模型手工选择了 12 个嘴部特征点。

由于在该序列中具有夸张的嘴部运动和一定程度的头部深度旋转。对于

经典的 KLT 跟踪方法，只有两个点可以被完整跟踪。其中由于较大的形变，嘴部特征点全部无法鲁棒跟踪。当采用提出的方法进行面部特征点跟踪，全部 36 个特征只有 7 个跟踪丢失，表明采用尺度空间和子空间约束增强的 KLT 跟踪算法不仅能够处理一定的深度旋转问题，同时对形变也有一定的鲁帮性。

图 5-19，5-20 分别给出了 KLT 算法和提出的算法在特征点跟踪误差上的定量结果。从结果中可以得出，在该复杂的序列中，KLT 算法在序列跟踪的初期就丢失了大量的特征点，而提出的算法则能够获得相对鲁帮的跟踪结果。

图 5-18 提出的跟踪算法的部分跟踪结果（上行）和 KLT 跟踪算法的部分跟踪结果（下行）的比较

Fig.5-18 Tracking errors on sequence “foreman.mpeg” by the proposed tracker

图 5-19 在“foreman.mpeg”序列上 KLT 算法的跟踪误差

Fig.5-19 Tracking errors on sequence “foreman.mpeg” by the KLT tracker

图 5-20 在“foreman.mpeg”序列上提出的算法的跟踪误差

Fig.5-20 Tracking errors on sequence “foreman.mpeg” by the proposed tracker

5.6 小结

在本章中提出了通过融合低端运动信息改进序列蒙特卡洛滤波算法中粒子传递的策略，并将该策略应用于人脸特征点跟踪问题。提出的人脸特征点跟踪算法在序列蒙特卡洛滤波框架下结合了基于模型和基于光流的跟踪方法，在人脸具有夸张表情和不丰富纹理的情况下仍然可以有效地完成跟踪任务。实验结果也验证了算法在人脸特征点跟踪中的鲁棒性和准确性。另外，本章提出的算法具有一定的通用性，可以适用于其他具有非刚性形变的物体特征点跟踪问题。

结论

本文的目标是以序列蒙特卡洛滤波算法为基本框架，以人脸和人体跟踪问题为应用背景，研究视觉跟踪技术中面临的若干关键问题。基于此，本文研究了开发鲁棒实用的视觉跟踪系统所需要的核心技术和关键问题解决方案，重点探讨了与目标表观的建模，复杂运动的建模和推断，融合低端模型和高端模型的跟踪策略等相关的问题。所取得的主要研究成果和对基于视觉的目标跟踪研究的主要贡献如下：

1) 提出了可区分性目标表观模型的自适应建模和更新算法：目标表观建模是视觉跟踪算法性能的决定性因素之一。虽然这一问题得到了领域内学者的极大重视和不懈努力，其仍是阻碍视觉目标跟踪技术进入实际应用的最困难问题之一。论文中提出了一种新的自适应目标表观建模和更新方法，该方法在建模的过程中不仅仅考虑目标表观信息，同时对目标所处环境中的背景信息进行考察，可以对目标/背景差异信息进行有效建模，从根本上保证了模型的可区分性。在图像特征选择上，将来近在目标检测领域成功应用的 Haar 小波特征引入。基于该特征空间，采用分类器组合的方式构建目标表观的可区分性模型。另外，注意到在采用序列蒙特卡洛滤波算法进行目标跟踪的过程中，由于序列蒙特卡洛滤波算法本身的性质，从而有大量的“背景”粒子的存在。利用 “背景”粒子中蕴含的背景分布信息，为目标表观模型随背景的变化而实时更新提供了有效的方法。在人体跟踪问题上的实验结果表明，相比于目前最具代表性的跟踪算法之一 Mean Shift，提出的算法在公开的测试序列上取得了更好的跟踪效果。
2) 提出了集成多运动模型的复杂运动建模和推断方法：由于计算复杂性的限制，视觉跟踪算法通常基于局部搜索的策略确定目标的运动状态。所以，根据目标运动规律确定目标在当前时刻以较高概率出现的区域成为跟踪算法成败的关键问题。本论文提出了采用多运动模型对目标复杂运动进行建模和估计的框架。在此基础上，针对具有多种运动模式和具有高维运动状态的两类常见的复杂目标运动，将多模型的估计框架融入到序列蒙特卡洛滤波算法中，从而针对两类复杂运动问题提出了标准序列蒙特卡洛滤波算法的两个改进：基于多模型切换和基于多模型协同的序

列蒙特卡洛滤波算法。在面部跟踪和面部表情估计问题上分别验证了提出的基于多模型切换和多模型协同的序列蒙特卡洛滤波算法。实验表明，在降低计算复杂度的同时，算法相对比于标准序列蒙特卡洛滤波算法得到了较高的性能提升。

3) 提出了融合光流和特定模型的面部特征点跟踪算法：面部特征点跟踪是基于特征点的运动感知研究的典型应用。本论文在序列蒙特卡洛滤波算法的框架下融合了基于光流和基于特定模型的方法来解决面部特征点跟踪问题，来克服单独采用一类方法不足，从而达到鲁棒跟踪面部特征点的目的。基于尺度空间理论改进的 KLT 光流算法能够准确估计面部具有刚性或者近似刚性运动特征点的位移运动。光流估计结果为基于模型的形变特征点估计提供了更好的起始搜索位置，同时加速了基于序列蒙特卡洛滤波算法的搜索过程。由于基于光流的估计算法中存在跟踪误差，进一步采用特征点运动的子空间约束来求精跟踪结果。相比于广泛使用的 KLT 特征点跟踪算法，实验结果证实了提出的算法的有效性。

视觉目标跟踪是一项极具挑战性和涵盖广泛的研究课题，本论文仅针对其中的几个关键问题开展了研究，但即使针对这几个问题也还有很多工作需要进一步深入。在目标表观建模方面，当前的目标表观建模工作着重于解决目标表观和背景的区分程度问题，但对目标运动过程中其表观历史变化的信息建模却是有限的，只是更多地利用了当前时刻的目标表观信息。将目标的历史表观数据建模方法与已经提出的基于背景和前景差异信息的建模策略相融合，从而使目标表观模型包含两部分信息：目标/背景的差异信息和目标表观的稳定部分所构成的约束子空间。在多目标跟踪方面，当所有场景中的目标在图像区域中交集为零时，多目标跟踪跟踪问题退化为多个单目标跟踪问题。然而，如果目标之间存在互相遮挡的情况，则首先需要构造目标联合状态进行推断。在提出的方法中，基于 Haar 特征的模型和基于光流的方法由于采用了局部特征，都有处理目标局部遮挡的能力。后续工作将在现有方法中加入遮挡处理的策略，从而使算法能够适用于多目标跟踪问题；在系统实现方面，论文中提出的目标表观建模方法以 Haar 特征和分类器组合为框架，从而与目标检测算法共享相同的算法框架。由于目标跟踪和检测本来就是具有互补性质的视觉功能，跟踪问题通过利用帧间约束能够简化视频序列中的目标检测问题并且确定目标在时间轴上的对应，目标检测算法则为目标跟踪中的全遮挡、目标数目的变化等问题提供了可行的解决方案。

参考文献

1. Forsyth, D.A. and J. Ponce, *Computer Vision: A Modern Approach*. 2002: Prentice Hall Press.
2. Comaniciu, D., V. Ramesh, and P. Meer. *Real-time tracking of non-rigid objects using Mean Shift*. IEEE Proc. on Computer Vision and Pattern Recognition. 2000:142~149, Hilton Head Island, South Carolina.
3. Isard, M. and A. Blake, *CONDENSATION-Conditional density propagation for visual tracking.* International Journal of Computer Vision, 1998. **29**:5~28.
4. Wren, C., et al., *PFinder: Real-time tracking of the human body.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. **19**:780~785.
5. Stauffer, C. and W. Grimson, *Learning patterns of activity using real-time tracking.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**:747~57.
6. Rasmussen, C. and G. Hager, *Probabilistic Data Association Methods for Tracking Complex Visual Objects.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001. **23**(6):560~576.
7. Hager, G.D. and P.N. Belhumeur, *Efficient Region Tracking With Parametric Models of Geometry and Illumination.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(10):1125~1139.
8. Wu, Y. and T.S. Huang, *A Co-inference Approach to Robust Visual Tracking.* Proceedings of International Conference on Computer Vision, 2001. **2**:26~33.
9. Blackman, S. and R. Popoli, *Design and Analysis of Modern Tracking Systems*. 1999: Artech House.
10. Nummiaro, K., E. Koller-Meier, and L. Van Gool, *An Adaptive Color-based Particle Filter.* Image and Vision Computing, 2003. **21**(1):99~110.
11. Stricker, M. and M. Swan, *The capacity of color histogram indexing.* IEEE Proc. on Computer Vision and Pattern Recognition, 1994: 704~708.

12. Perez, P., et al., *Color-based probabilistic tracking.* European Conference on Computer Vision, 2002:661~675.
13. Birchfield, S., *Elliptical Head Tracking Using Intensity Gradients and Color Histograms.* IEEE Proc. on Computer Vision and Pattern Recognition, 1998:232~237.
14. Bradski, G.R., *Computer video face tracking for use in a perceptual user interface.* Intel Technology Journal, 1998. **Q2**.
15. Comaniciu, D., V. Ramesh, and P. Meer, *Kernel-based object tracking.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003. **25**(5):564~577.
16. Stern, H. and B. Efros, *Adaptive color space switching for face tracking in multi-colored lighting environments.* Int. Conf. On Automatic Face and Gesture Recognition, 2002:236~241.
17. Collins, R. and Y. Liu, *On-line selection of discriminative tracking features.* Proceedings of International Conference on Computer Vision, 2003:346~352.
18. 刘明宝, 姚鸿勋, 高文, 彩色图像的实时人脸跟踪方法. 计算机学报, 1998. **6**:527~532.
19. Yao, H. and W. Gao, *Face Detection and Location Based on Skin Chrominance and Lip Chrominance Transform from Color Images.* Pattern Recognition, 2001. **34**(8):1555~1564.
20. Baker, S. and I. Matthews, *Lucas-Kanade 20 Years On: A Unifying Framework.* International Journal of Computer Vision, 2004. **56**(3):221~255.
21. Bergen, J.R., et al., *Hierarchical model-based motion estimation.* European Conference on Computer Vision, 1992:237~252.
22. Matthews, I., T. Ishikawa, and S. Baker, *The Template Update Problem.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004. **26**(6):810~815.
23. Frey, B.J., *Filling In Scenes by Propagating Probabilities through Layers and Into Appearance Models.* IEEE Proc. on Computer Vision and Pattern Recognition, 2000:1185~1192.
24. Jojic, N. and B.J. Frey, *Learning Flexible Sprites in Video Layers.* IEEE

Proc. on Computer Vision and Pattern Recognition, 2001. **1**:199~206.

25. Rucklidge, W., *Efficient Guaranteed Search for Gray-level Patterns.* IEEE Proc. on Computer Vision and Pattern Recognition, 1997:717~723.
26. Olson, C.F., *Maximum-Likelihood Template Matching.* Proceedings of IEEE Proc. on Computer Vision and Pattern Recognition, 2000. **2**:52~57.
27. Morris, D.D. and J. Rehg, *Singularity Analysis for Articulated Object Tracking.* IEEE Proc. on Computer Vision and Pattern Recognition, 1998:289~296.
28. 高文, 金辉, 面部表情图像的分析与识别. 计算机学报, 1997. **9**:782~789.
29. Kass, M., A. Witkin, and D. Terzopoulos, *Snakes: active contour models.* International Journal of Computer Vision, 1988. **1**(4):321~331.
30. Leymarie, F. and M. Levine, *Tracking deformable objects in the plane using an active contour model.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993. **15**:617~634.
31. Paragios, N. and R. Deriche, *Geodesic Active Contours and Level Sets for the Detection and Tracking of Moving Objects.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000. **22**(3):266~280.
32. Cootes, T., et al., *Active Shape Models - Their Training and Application.* Int. J. of Computer Vision and Image Understanding, 1995. **61**(1):38~59.
33. Cootes, T., G. Edwards, and C. Taylor, *Active appearance models.* European Conference on Computer Vision, 1998. **2**:484~498.
34. Cootes, T.F., K. Walker, and C.J. Taylor, *View-based Active Appearance Models.* Proceeding of the 4th International Conference on Face and Gesture Recognition, 2000:227~232.
35. Heap, A.J. and D.C. Hogg, *Extending the Point Distribution Model using polar coordinates.* Image & Vision Computing, 1995. **14**(8):589~599.
36. Sonka, M., V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision(Second Edition)*. 2002: Brooks/Cole.
37. Black, M.J. and A.D. Jepson, *Eigentracking: Robust matching and tracking of articulated objects using view-based representation.* European Conference on Computer Vision, 1996:329~342.
38. Black, M.J., D.J. Fleet, and Y. Yacoob, *A framework for modeling*

appearance change in image sequence. Proceedings of International Conference on Computer Vision, 1998:660~667.

39. Torre, D.l., G. F., and M. S., S.J., *View-based adaptive affine tracking.* European Conference on Computer Vision, LNCS 1406, Springer Verlag, 1998:828~842.
40. Jepson, A.D., D.J. Fleet, and T.F. El-Maraghi, *Robust online appearance models for visual tracking.* IEEE Proc. on Computer Vision and Pattern Recognition, 2001. **1**:415~422.
41. Ross, D., J. Lim, and M.-H. Yang, *Adaptive Probabilistic Visual Tracking with Incremental Subspace Update.* European Conference on Computer Vision, 2004. **2**:470~482.
42. Koller, D., K. Daniilidis, and H.H. Nagel, *Model-Based Object Tracking in Monocular Image Sequences of Road Traffic Scenes.* International Journal of Computer Vision, 1993. **10**:257~281.
43. Gavrila, D.M. and L.S. Davis, *3-D model-based tracking of humans in action: A multi-view approach.* IEEE Proc. on Computer Vision and Pattern Recognition, 1996:73~80.
44. Sidenbladh, H., M.J. Black, and D.J. Fleet, *Stochastic tracking of 3D human figures using 2D image motion.* European Conference on Computer Vision, Springer Verlag, LNCS 1843, 2000:702~718.
45. Gao, X., et al., *Error Analysis of Background Subtraction.* IEEE Proc. on Computer Vision and Pattern Recognition, 2000.
46. Haritaoglu, I., D. Harwood, and L. Davis, *W4: real-time surveillance of people and their activities.* IEEE Transaction on Pattern Analysis and Machine Intelligence, 2000. **22**:809~830.
47. Elgammal, A., D. Harwood, and L. Davis, *Non-parametric model for background subtraction.* European Conference on Computer Vision, 2000. **2**:751~767.
48. Bar-Shalom, Y., *Extension of the probabilistic data association filter to multi-target environments.* Proc. 5th symposium on Nonlinear Estimation, 1974.
49. MacCormick, J. and A. Blake, *A probabilistic exclusion principle for tracking multiple objects.* Proceedings of International Conference on

Computer Vision, 1999. **1**:572~578.

50. Kalman, R., *A New Approach to Linear Filtering and Prediction Problems.* IEEE Transactions of the ASME--Basic Engineering, 1960. **82**(D):35~45.
51. Bar-Shalom, Y. and T.E. Foreman, *Tracking and Data Association*. 1988: Academic Press Inc.
52. Terzopoulos, D. and R. Szeliski, *Tracking with kalman snakes.* Active Vision, MIT Press, 1992:3~20.
53. Broida, T.J. and R. Chellappa, *Estimation of Object Motion Parameters from a Sequence of Noisy Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986. **8**(1).
54. Beymer, D. and K. Konolige, *Real-time tracking of multiple people using continuous detection.* Proceedings of International Conference on Computer Vision: Frame-Rate Workshop, 1999.
55. Rosales, R. and S. Sclaroff, *3D trajectory recovery for tracking multiple objects and trajectory guided recognition of actions.* IEEE Proc. on Computer Vision and Pattern Recognition, 1999. **2**:116~123.
56. Deutscher, J., A. Blake, and R. I. *Articulated body motion capture by annealed particle filtering*. IEEE Proc. on Computer Vision and Pattern Recognition. 2000, 2:1144~1149.
57. Sullivan, J., et al., *Object Localization by Bayesian Correlation.* Proceedings of International Conference on Computer Vision, 1999:1068~1075.
58. Li, B. and R. Chellappa, *Simultaneous Tracking and Verification via Sequential Posterior Estimation.* IEEE Proc. on Computer Vision and Pattern Recognition, 2000:2110~2117.
59. Liu, J.S. and R. Chen, *Sequential Monte Carlo methods for dynamic systems.* Journal of the American Statistical Association, 1998. **93**:1032~1044.
60. Tomasi, C. and T. Kanade, *Detection and tracking of feature points.* Carnegie Mellon University Technical Report CMU-CS-91-132, 1991.
61. Arulampalam, S., et al., *A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking.* IEEE Trans. Signal Processing, 2002. **50**(2):174~188.

62. Bergman, N., *Recursive Bayesian Estimation: Navigation and tracking applications.* PhD Thesis, Linkoping University, Sweden, 1999.
63. Doucet, A., *On sequential Monte Carlo Methods for Bayesian Filtering.* Technical Report, University of Cambridge, UK, Department of Engineering, 1998.
64. Lin, J.S. and R. Chen, *Sequential Monte Carlo Methods for Dynamical Systems.* Journal of the American Statistical Association, 1998. **93**:1032~1044.
65. Merwe, R.v.d., et al., *The unscented particle filter.* Advances in Neural Information Processing Systems, 2000. **12**.
66. Gilks, W.R. and C. Berzuini, *Following a Moving Target-Monte Carlo inference for Dynamic Bayesian Models.* Journal of the Royal Statistical Society, 2001. **63**(B):127~146.
67. Musso, C., N. Oudiane, and F. LeGland, *Improving Regularised Partile Filters.* In, eds A. Doucet, J. Freitas and N. Gordon, Sequential Monte Carlo Methods in Practice, Springer-Verlag, 2001.
68. Clapp, T. and S. Godsill, *Improving strategies for Monte Carlo particle filters.* Sequential Monte Carlo Methods in Practice, eds A. Doucet, J. Freitas and N. Gordon, Springer-Verlag, 2001.
69. Oudjane, N. and C. Musso, *Progressive Correction for Regularized Particle.* Proc. 3rd Int. Conf. on Information Fusion, 2000.
70. Rui, Y. and Y. Chen, *Better Proposal Distributions: Object Tracking Using Unscented Particle Filter.* IEEE Proc. on Computer Vision and Pattern Recognition, 2001:786~793.
71. 庄莉等, 视频中多线索的人脸特征检测与跟踪. 计算机学报, 2003. **26**(2):160~167.
72. Nguyen, H.T. and A.W.M. Smeulders, *Tracking Aspects of the Foreground against the Background.* European Conference on Computer Vision, 2004. **2**:446~456.
73. Viola, P. and M. Jones, *Rapid object detection using a boosted cascade of simple features.* IEEE Proc. on Computer Vision and Pattern Recognition, 2001:511~518.
74. Duda, R., P. Hart, and D.G. Stork, *Pattern Classification (2nd Edition).*

2000: Wiley-Interscience Press.
75. 边肇祺, 张学工, 模式识别. 北京:清华大学出版社, 2000.
76. Welch, G. and G. Bishop, *An Introduction to the Kalman Filter*. ACM SIGGRAPH, 2001. **Course 8**.
77. *CAVIAR Test Case Scenarios*. at: http://Homepages.inf.edac.uk/rbf/CAVIAR., 2004.
78. Choo, K. and F. D. *People tracking using hybrid monte carlo filtering*. Proceedings of International Conference on Computer Vision. 2001.
79. Leventon, M. and W. Freeman, *Bayesian estimation of 3D human motion from an image sequence*, *TR_98_06*. 1998, MERL Tech: Cambridge.
80. Yacoob, Y. and L. Davis, *Learned models for estimation of rigid and articulated human motion from stationary or moving camera*. International Journal of Computer Vision, 2000. **36**(1):5~30.
81. Molina-Tanco, L. and A. Hilton. *Realistic synthesis of novel human movements from a database of motion capture examples*. *IEEE Workshop on Human Motion*. 2000.
82. Brand, M. and A. Hertzmann. *Style machines*. *ACM SIGGRAPH*. 2000.
83. 王天树等, 人体运动非监督聚类分析. 软件学报, 2003. **14**(2):209~214.
84. MacCormick, J. and M. Isard. *Partitioned sampling, articulated objects, and interface-quality hand tracking*. *European Conf. on Computer Vision*. 2000.
85. Bregler, C. and J. Malik. *Tracking people with twists and exponential maps*. IEEE Proc. on Computer Vision and Pattern Recognition. 1998.
86. Deutscher, J., et al. *Tracking through singularities and discontinuities by random sampling*. Proceedings of International Conference on Computer Vision. 1999.
87. Niyogi, S. and E. Adelson. *Analysing and recognising walking figures in xyt*. IEEE Proc. on Computer Vision and Pattern Recognition. 1994.
88. Rohr, K., *Human movement analysis based on explicit motion models*, in *Motion-Based Recognition*, D. Boston, Editor. 1997, Kluwer Academic Publishers.171~198.
89. Isard, M. and A. Blake. *A mixed-state Condensation tracker with automatic model-switching*. Proceedings of International Conference on

Computer Vision. 1998.

90. Farmer, M., R. Hsu, and A. Jain. *Interacting Multiple Model (IMM) Kalman filters for robust high speed human motion tracking*. Int. Proc. on Pattern Recognition. 2002.
91. Wu, Y., G. Hua, and T. Yu. *Switching Observation Models for Contour Tracking in Clutter*. IEEE Proc. on Computer Vision and Pattern Recognition. 2003.
92. 梁国远, 查红彬, 刘宏, 基于三维模型和仿射对应原理的人脸姿态估计方法. 计算机学报, 2005. 28(5):792~800.
93. McKenna, S., et al., *Tracking facial feature points with Gabor wavelets and shape models.* Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication, 1997:35~42.
94. Yuille, A., *Feature extraction from faces using deformable templates.* International Journal of Computer Vision, 1992. **8**(2):99~111.
95. 高文, 陈熙霖, 计算机视觉-算法与系统原理. 清华大学出版社, 1999.
96. Lucas, B. and T. Kanade, *An iterative image registration technique with an application to stereo vision.* Int. Joint Conf. on Artificial Intelligence, 1981:674~679.
97. Tomasi, C. and T. Kanade, *Shape and Motion from Image Streams: A Factorization Method.* CMU-CS-TR-92-104, 1992.
98. Tommasini, T.e.a., *Making Good Features Track Better.* IEEE Proc. on Computer Vision and Pattern Recognition, 1998:178~183.
99. 宋刚, 艾海舟, 徐光祐, 纹理约束下的人脸特征点跟踪. 软件学报, 2004. 15(11):1607~1615.
100. Isard, I. and A. Blake, *ICondensation: Unifying low-level and highlevel tracking in a stochastic framework.* European Conference on Computer Vision, 1998.
101. de Freitas JF, et al., *Sequential monte carlo methods To train neural network models.* Neural Computation, 2000. **12**(4):955~93.
102. Julier, S.J. and J.K. Uhlmann, *A new extension of the Kalman filter to nonlinear systems.* Int. Symp Aerospace/Defense Sensing, Simul. and Controls, 1997.
103. Mikolajczyk, K. and C. Schmid, *Scale and affine invariant interest point*

detectors. International Journal of Computer Vision, 2004. **60**(1):63~86.

104. Lindeberg, T., *Detecting salient blob-like image structures and their scales with a scale-space primal sketch-A method for focus-of-attention.* International Journal of Computer Vision, 1993. **11**(3):283~318.
105. Crowley, J., *A representation for visual information.* Ph.D thesis, Carnegie Mellon University, 1981.
106. Crowley, J. and A. Parker, *A representation for shape based on peaks and ridges in the difference of low pass transform.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984. **6**(2):156~170.
107. Lindeberg, T., *Feature detection with automatic scale selection.* International Journal of Computer Vision, 1998. **30**(2):79~116.
108. Lowe, D.G., *Object recognition from local scale-invariant features.* Proceedings of International Conference on Computer Vision, 1999:1150~1157.
109. Bregler, C., A. Hertzmann, and H. Biermann, *Recovering non-rigid 3D shape from image streams.* IEEE Proc. on Computer Vision and Pattern Recognition, 2000. **2**:690~696.

攻读学位期间发表的学术论文

第一作者论文

1. 王建宇，陈熙霖，高文，赵德斌. 目标模型的动态构建. 软件学报，2006(5) [EI]
2. 王建宇，高文. 基于子空间约束的面部特征点跟踪算法. 高技术通讯，2005, Vol.9:24~28[EI]
3. Jianyu Wang, Xilin Chen and Wen Gao. Online Selecting Discriminative Tracking Features using Particle Filter, IEEE International Conference on Computer Vision and Pattern Recognition, Vol. 2, 1036~1041, San Diego, CA, 2005, USA[EI]
4. Jianyu Wang, Debin Zhao, Shiguang Shan, and Wen Gao, Approximating the Inference on Complex Motion Models using Multi-Model Particle Filter, Pacific-Rim Conference on Multimedia, Tokyo, Japan, 2004, 2:1011~1018 [SCI]
5. Jianyu Wang, DebinZhao, Wen Gao, Shiguang Shan, Interacting Multiple Model Particle Filter To Adaptive Visual Tracking, International Conference on Image and Graphics, Hong Kong, 2004:568~571[EI]
6. Jianyu Wang, Wen Gao, ShiGuang Shan, Xiaopeng Hu, Facial Feature Tracking Combined Model-Based and Model-Free Method, International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 2003, 3:205~208 [EI]
7. Jianyu Wang, Wen Gao and Xilin Chen, Approximating complex dynamic using set of simple models for motion inference, Pattern Recognition Letters (Submitted)
8. Jianyu Wang, Wen Gao and Xilin Chen, On-line Maintaining Discriminative Appearance Model using Particle Filter, Pattern Recognition Letters (Submitted)

合作论文

Wenchao Zhang, Shiguang Shan, Wen Gao, Jianyu Wang, Debin Zhao, Incremental Face-Specific Subspace for Online-Learning Face Recognition, Proc. of ACCV2004, Jeju, Korea, 2004, 2:1080~1084

专利

《身材信息辅助人脸信息的身份识别技术》，专利号：02153265.6。

哈尔滨工业大学博士学位论文原创性声明

本人郑重声明：此处所提交的博士学位论文《基于序列蒙特卡洛滤波算法的视觉目标跟踪》，是本人在导师指导下，在哈尔滨工业大学攻读博士学位期间独立进行研究工作所取得的成果。据本人所知，论文中除已注明部分外不包含他人已发表或撰写过的研究成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。本声明的法律结果将完全由本人承担。

作者签字：　　日期：　　年　月　　日

哈尔滨工业大学博士学位论文使用授权书

《基于序列蒙特卡洛滤波算法的视觉目标跟踪》系本人在哈尔滨工业大学攻读博士学位期间在导师指导下完成的博士学位论文。本论文的研究成果归哈尔滨工业大学所有，本论文的研究内容不得以其它单位的名义发表。本人完全了解哈尔滨工业大学关于保存、使用学位论文的规定，同意学校保留并向有关部门送交论文的复印件和电子版本，允许论文被查阅和借阅。本人授权哈尔滨工业大学，可以采用影印、缩印或其他复制手段保存论文，可以公布论文的全部或部分内容。

保密□，在　　年解密后适用本授权书。

本学位论文属于

不保密□。

作者签名：　　日期：　　年 月　日

导师签名：　　日期：　　年 月　日

致谢

值此论文即将完成之际，在此由衷感谢在本人攻读博士期间所有关心、帮助、支持我的老师、同学和家人们。正是老师们广博的实践经验、深厚的理论功底、严谨的治学态度和不倦的诲人精神，引导我从一个科研工作的门外汉，逐渐走向求知求真的最前线。正是同学之间相互鼓励和无私帮助的团队生活，使我在追求知识的道路上得到了志同道合者的有力支持与激励。正是家人们的真诚关心和鼎力支持，使得我在求学之路上获得了无尽的动力和温暖。

我首先要感谢我的导师高文教授。他是一位高瞻远瞩的领路者，在布满荆棘的道路上始终将正确的路标放在我前进的方向上。在我迷惘的时候，总是能够从他那里获得最具前瞻性的建议和帮助。高老师以他广博的知识、严谨的治学态度、高标准的研究水平、丰富的研究经验随时影响和教育着我。同时，他还竭尽所能，为我们提供了非常高标准、宽松、充满活力的研究环境。

感谢实验室的陈熙霖、赵德斌教授。两位老师严谨的科学作风、忘我的工作精神、无私的胸怀一直是我钦佩和学习的榜样。博士期间的工作一直受到两位老师的悉心关怀和指导，在跟两位老师的学习和交流中，使我终生受益。感谢实验室的姚鸿勋、刘岩老师。两位老师无私奉献、兢兢业业的工作态度和丰硕的科研成果，一直为我所钦佩和向往。在攻读博士期间，两位老师给我的无私帮助和热情关心，使我终生难忘。

感谢我的师兄山世光、曾炜、吕岩和苗军，他们从一个先行者的角度给了我无数的建议和关心，所有这些都使我有幸绕过了很多急流险滩，为我能顺利毕业提供了我莫大的帮助。感谢我的同学张鸿明、张文超、柴秀娟、曹波、杨澎、张晓华、闫胜业、陈杰、马丙鹏、唐杰和杜波，在于他们朝夕相处的日子里，一直给予我的热情帮助和支持。

感谢联合实验室其他老师和同学们，正是这个温暖和充满活力的集体，为我的博士生涯提供了一个良好的舞台和永久而美好的回忆。

感谢我的妻子陈思颖，她陪我度过了博士生涯这段最艰苦的岁月，使我一度失衡的心态恢复平静，勇于面对一切困难和挑战。感谢我的爸爸、妈妈和妹妹，他们都是这个国度中最普通的一员，但正是他们在我的求学道路上洒满了自己的汗水，在我最困难和最低落的时候，一个电话、一句问候都给

了我无尽的斗志。我谨将自己的论文献给他们，我最亲爱的家人们。

本工作受到参加国家 863 项目“生物特征识别核心技术与关键问题研究”(合同号：2001AA114190)，国家自然科学基金重点项目“基于生物特征的身份识别研究” (合同号：69789301)，国家 863 项目“多功能感知技术”(合同号：863-306-ZD03-01-2)，国家自然科学基金项目“人脸主动网格模型方法研究”(批准号:60473043)，中科院百人计划、上海市科委项目03DZ15013、以及上海银晨智能识别科技有限公司的大力资助。

个人简历

王建宇(1975.12)，男，辽宁营口人。主要研究领域为物体跟踪、物体检测、计算机视觉、图像处理、模式识别和机器学习。在攻读博士学位期间，发表和录用国内期刊论文 2 篇、国际会议论文 4 篇，另有与其他作者合作论文 1 篇、申请国内专利 1 项。

教育背景

1994.9~1998.7　哈尔滨工业大学流体传动控制系，工学学士。

1998.9~2000.7　哈尔滨工业大学机械电子工程系，工学硕士。

2000.9~至今　哈尔滨工业大学计算机科学与工程系，工学博士，论文题目：基于序列蒙特卡洛滤波算法的视觉目标跟踪。

攻读博士学位期间参与的科研项目

1　2002.7~现在 参加国家 863 项目“生物特征识别核心技术与关键问题研究”(合同号：2001AA114190)的研究工作。主要工作：人脸的检测和跟踪，基于视频序列的人脸识别，基于 3D 模型的人脸识别等工作；

2　2002.3~现在 参加国家自然科学基金(合同号为 69789301)重点项目“基于生物特征的身份识别研究”的研究工作。主要工作：人脸跟踪与检测，及基于特征点跟踪的视频序列理解等相关部分的研究；

3　2004.4~2004.10 上海银晨智能识别科技有限公司，开发智能视频监控系统。主要工作：基于背景建模的目标检测和跟踪模块；

4　2002.8~2002.11 参加了国家 863 项目“生物特征识别核心技术与关键问题研究—人脸检测与识别关键技术”的鉴定验收。主要工作：设计编写了基于照片的人脸检索系统；

5　2001.4~ 2002.7 参加国家 863 项目“多功能感知技术”(合同号：863-306-ZD03-01-2)的研究工作。主要工作：基于视频人脸 3D 模型重建，视频序列中的人脸特征点跟踪等研究工作。