密级：_____

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

硕士学位论文

基于多模态数据的目标跟踪算法研究

作者姓名：张学志

指导教师：卢孝强 研究员、 袁媛 研究员

中国科学院西安光学精密机械研究所

学位类别：工程硕士

学科专业：电子与通信工程

研 究 所：中国科学院西安光学精密机械研究所

二〇一七年五月

Research of Multi-modal Data Based Object Tracking

By

Xuezhi Zhang

A Thesis Submitted to

University of Chinese Academy of Sciences

In partial fulfillment of the requirement

For the degree of

Master of Engineering

Xi'an Institute of Optics & Precision Mechanics,

Chinese Academy of Sciences

May, 2017

科研道德声明

秉承研究所严谨的学风与优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中所引用的内容都已给予了明确的注释和致谢。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了致谢。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

作者签名：_____　　日　期：_____

知识产权声明

本人完全了解中科院西安光学精密机械研究所有关保护知识产权的规定，即：研究生在所攻读学位期间论文工作的知识产权单位系中科院西安光学精密机械研究所。本人保证离所后，发表基于研究生工作的论文或使用本论文工作成果时必须征得产权单位的同意，同意后发表的学术论文署名单位仍然为中科院西安光学精密机械研究所。产权单位有权保留送交论文的复印件，允许论文被查阅和借阅；产权单位可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。

（保密论文在解密后适用本授权书）

作者签名：_____　　导师签名：_____

日　期：_____　　日　期：_____

致　谢

时光匆匆，转眼三年的研究生求学生活即将结束。回首三年，我收获颇多，值此论文完成之际，谨向给予我关心和帮助的老师、同学和朋友表示衷心的感谢！

感谢我的导师卢孝强研究员和袁媛研究员。感谢卢老师对我科研工作的指导和日常生活的关心，他对工作的热情将积极地影响我的一生。感谢袁老师对我的关心和鼓励，从她身上学到了严谨治学和虚怀若谷的精神。

感谢李学龙研究员。他那一丝不苟和注重细节的做事风格，将使我受益终生。

感谢邱实老师，冯亚闯老师，刘康老师，黄举老师和彪延老师对我科研和生活上的帮助。感谢实验室的各位同学们的帮助与支持。

最后，祝愿光学影像分析与学习中心越来越好。

张学志

2017 年 5 月

摘 要

随着移动手机和监控摄像头等视频采集设备的广泛应用，每时每刻都会产生大量的视频数据。视频数据的智能化分析与挖掘能够从根本上解决海量视频数据的分析问题。而视觉目标跟踪作为计算机视觉领域的一个重要研究方向，它将在智能化视频分析与挖掘中起到十分重要的作用。视觉目标跟踪在现实生活中具有广泛的应用场景，如安防和无人驾驶等，具有十分重要的应用价值。视觉目标跟踪容易受到亮度变化，形变，运动模糊和尺寸变化等干扰因素的影响，仍然存在很多问题，具有十分重要的研究价值。

为了有效利用深度神经网络的强大特征表达能力和多模态数据的多源性与互补性，本文提出了两种目标跟踪算法：基于多模态数据和卷积神经网络的目标跟踪算法，以及基于多模态数据和全卷积双流网络的目标跟踪算法。为了验证本文算法的有效性，我们采集了一个多模态数据库 OptTrack，它同时包含可见光图像和红外图像, 总共包括 6 个视频序列。

基于多模态数据和卷积神经网络的目标跟踪算法采用了双融合策略，不仅融合了可见光图像的浅层特征图的空间信息和深层特征图的语义信息，而且也在算法级别上融合了可见光图像和红外图像。该算法可以分成两个步骤，首先在可见光图像的多层卷积特征图上使用平移相关滤波器预测目标位置；然后在红外图像的尺度金字塔上估计目标尺寸。我们在 OptTrack 中的 6 个视频序列上与 10 个目标跟踪算法做了对比实验，实验结果证明，该算法比其它算法的跟踪效果好。

为了解决基于深度神经网络的目标跟踪算法速度慢的问题，本文提出了基于多模态数据和全卷积双流网络的目标跟踪算法。该算法首先在可见光图像上使用全卷积双流网络，有效地利用了全卷积特性，只需前向传播一次，就能得到目标位置；然后在红外图像的尺度金字塔上估计目标尺寸。与 10 个目标跟踪算法的对比结果证明，该算法的跟踪效果很好。另外，该算法的跟踪速度比较快，平均跟踪速度约为 19 fps。

关键词：计算机视觉，视觉目标跟踪，多模态数据，卷积神经网络，双流网络

ABSTRACT

With the mobile phone and surveillance cameras and other video capture devices widely used, a lot of video data are produced. Intelligent analysis and mining of video data can fundamentally solve the problem of the massive video data analysis. As an important research direction in the field of computer vision, visual object tracking plays an important role in intelligent video analysis and mining. Visual object tracking has a wide range of applications in real life, such as security, driverless car, etc. So it has a very great application value. Visual object tracking is susceptible to interference factors such as brightness change, deformation, motion blur and size change, and there are still many problems in it. So it has very important research value.

Two object tracking algorithms are proposed in this thesis; they not only make full use of the powerful feature representation of deep neural network, but also take advantage of the multi-source and complementarity of multi-modal data. Two object tracking algorithms respectively are object tracking based on multi-modal data and convolutional neural network, and object tracking based on multi-modal data and fully-convolutional Siamese network. In order to verify the effectiveness of these algorithms, we collected a multi-modal database OptTrack, which also includes visible images and infrared images. There are six video sequences in OptTrack.

The object tracking algorithm based on multi-modal data and convolution neural network adopts the dual fusion strategy, it not only combines the spatial information of shallow feature maps and the semantic information of deep feature maps, but also fuses the visible image and the infrared image at the algorithm level. The algorithm can be divided into two steps. Firstly, the target position is predicted by applying the translation correlation filter to the multi-layer convolution feature maps from the visible image. Secondly, the target size is estimated on the scale pyramid of the infrared image. We compare the 10 target tracking algorithms on six video sequences in OptTrack. The experimental results show that the algorithm is robust and superior to all other algorithms.

In order to solve the problem that the traditional tracking algorithm based on deep neural network is slow, this paper proposes an object tracking algorithm based on multi-modal data and fully-convolutional Siamese network. Firstly, the fully-convolutional Siamese network is used on the visible image; the target position can be predicted with one forward propagation. Secondly, the target size is estimated on the scale pyramid of the infrared image. Compared with 10 object tracking algorithms, it is proved that the algorithm performance is good. In addition, its tracking speed is fast, the average tracking speed is about 19 fps.

Keywords: Computer vision; visual object tracking; multi-modal data; convolutional neural network; siamese network

目　录

第一章 绪论

1.1 研究背景和意义

随着移动手机和监控摄像头等视频采集设备的广泛应用，每时每刻都会产生大量的视频数据。海量的视频数据不仅使人们的生活更加便利和安全，而且也对现有的技术提出了两个挑战：如何有效地存储海量视频数据，以及如何智能化地分析与挖掘视频中的信息。相对而言，视频数据的智能化分析与挖掘更加重要。由于数据存储技术的快速发展，数据存储设备的价格下降很快，海量视频数据的存储问题可以通过横向增加廉价的存储设备来解决。在视频分析方面，目前主要以人工分析为主，计算机仅起到辅助作用。实现视频数据的智能化分析与挖掘不仅能够节省大量的人力和财力，而且能够从根本上解决海量视频数据的分析问题。在实现智能化视频分析与挖掘的过程中，计算机视觉必将起到举足轻重的作用。计算机视觉是一门研究如何让计算机理解图像或视频的交叉学科，涉及信号处理，数字图像处理，模式识别，机器学习和数学等学科。计算机视觉领域包含模式分类 [1,2]，目标检测，目标跟踪，图像分割和图像检索等研究方向。其中，视觉目标跟踪是计算机视觉 [3] 领域的一个重要研究方向，它在智能化视频分析与挖掘中起到十分重要的作用。视觉目标跟踪 [4] 是指，在给定第一帧图像中的目标位置和尺寸等状态的前提下，自动预测目标在后续帧图像中的状态。由于视觉目标跟踪是异常检测和行为分析等任务的基础 [5,6]，因此设计一种鲁棒的目标跟踪算法以精确地预测目标状态，对于后续的视频内容分析而言十分重要。

视觉目标跟踪在现实生活中具有广泛的应用场景 [7,8]，如图1.1所示。在交通监控方面，利用视觉目标跟踪算法获取车辆和行人的位置与速度等状态之后，不仅能够预测车辆和行人的运动趋势，而且能够对它们的异常行为进行分析。因此视觉目标跟踪对于实现智能化交通监控而言十分重要。在安防领域中，目前主要以人工监控为主，海量的视频数据为监控人员带来了巨大的工作量。利用视觉目标跟踪，目标检测和行为分析技术实现视频监控的智能化分析是解放监控人员的根本方法。在汽车自动驾驶领域，《2016 年度无人驾驶汽车领域创新报告》指出汽车无人驾驶技术已经到了决定性的转折点，苹果、谷歌、特斯拉和百度等科技和互联网公司正纷纷投身这一领域。实现汽

车无人驾驶的关键技术之一就是利用目标跟踪获取目标的运动信息。在人机交互领域，Kinect 和 VR 设备的广泛应用显著地提高了人们的生活质量，但如果没有目标跟踪技术的支撑，它们仅仅是一个数据采集或显示设备，无法实现人机交互功能。在机器人视觉导航方面，利用目标跟踪技术获取运动目标的状态对于提高机器人的智能化路线规划而言十分重要。在体育视频分析方面，对运动员的跟踪不仅可以帮助评委和解说员对其进行分析，而且可以为观众提供镜头特写等服务。上述这些应用场景的共同特点是需要借助摄像头捕捉视频数据，并利用智能化视频分析技术来实现自身功能。而视觉目标跟踪技术是视频数据智能化分析与挖掘的基础，因此目标跟踪技术在很多应用场景中发挥着重要作用，具有十分重要的应用价值。

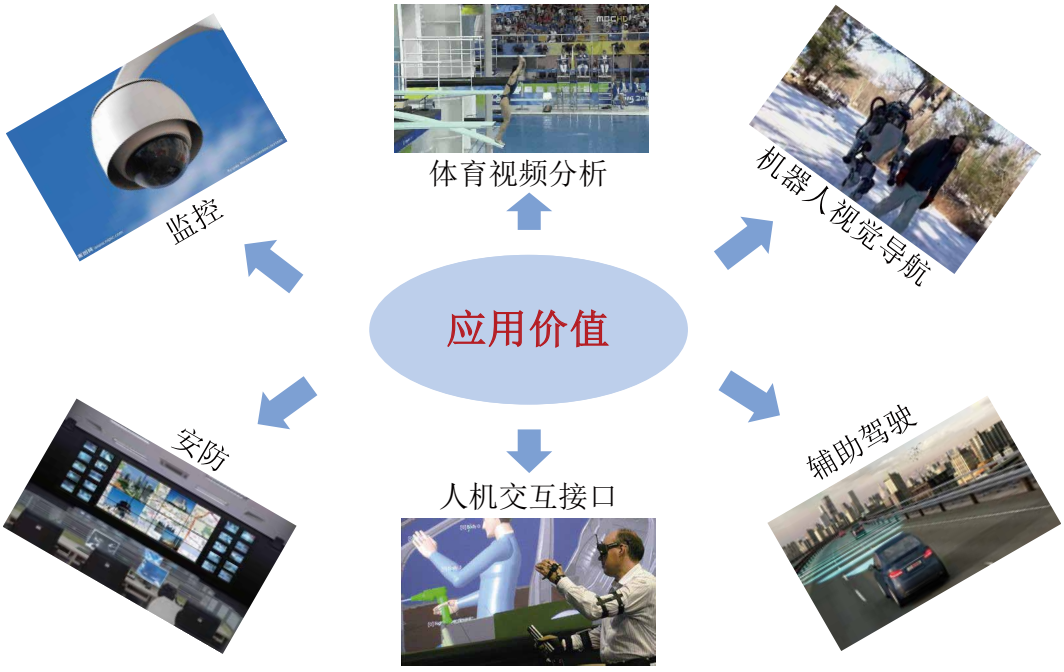


图 1.1　视觉目标跟踪的应用

尽管视觉目标跟踪方向已经研究了很多年，研究人员已经提出了大量的目标跟踪算法，但由于目标跟踪具有很多的挑战因素 [9–11]，如图1.2所示，其仍然存在很多问题。当目标在室内移动时，开灯或关灯将直接导致光照条件发生显著变化，如果提取的特征对光照变化敏感，将直接干扰目标跟踪算法的性能。目标在移动过程中容易受到其它物体的遮挡，当目标完全被遮挡时，对目标跟踪算法提出了巨大的挑战。非刚体目标在运动过程中可能发生姿态变化或形变，比如花样跳水运动员在下落过程中身体姿态一直在变化，这要求目标跟踪算法的跟踪框的大小和纵横比能够自适应改变。在图像采集过程

中，相机抖动会造成图像模糊，对于人眼都几乎无法辨别目标位置的模糊图像，利用目标跟踪算法实现智能化分析的难度非常大。当目标从远到近运动时，目标的尺寸将会发生显著变化，目标跟踪算法的跟踪框的尺寸应随目标的尺寸自适应变化。此外，目标跟踪算法还容易受到目标快速运动，旋转和背景混乱等因素的影响。由于目标跟踪容易受到上述挑战因素的干扰，因此设计一个鲁棒的目标跟踪算法仍然是一个难题，具有十分重要的科研价值。

(a) 亮度变化　(b) 形变　(c) 运动模糊　(d) 尺寸变化

图 1.2　视觉目标跟踪的挑战因素

海量视频数据的智能化分析与挖掘需要目标跟踪算法发挥基础性的作用，同时目标跟踪算法的广泛应用场景使其具有十分重要的应用价值，视觉目标跟踪算法的众多挑战因素使其具有很重要的科研价值，因此研究视觉目标跟踪算法具有十分重要的意义。本课题主要研究基于多模态数据和深度神经网络的视觉目标跟踪算法。

1.2　视觉目标跟踪算法的研究现状

由于广大科研工作者的不懈努力，已经提出了大量的视觉目标跟踪算法。但由于视觉目标跟踪算法容易受到光照变换，遮挡，形变，运动模糊，尺寸变化和背景混乱等因素的干扰，至今尚未有一种通用的、鲁棒的目标跟踪算法。为了提高视觉目标跟踪算法的性能，科研人员从数据源，特征和模型等不同的角度出发，提出许多改进算法。其中，近几年最显著的三个发展趋势是：基于深度学习的目标跟踪，基于多模态数据的目标跟踪和基于相关滤波器的目标跟踪。

绝大多数现有的目标跟踪算法仅仅利用了可见光图像。由于可见光固有物理性质的限制，基于可见光图像的目标跟踪算法无法在亮度极低，环境极其复杂的场景有效地跟踪目标。最近，研究人员提出了一些基于多模态数据的视觉目标跟踪算法 [12–14]，这些算法利用了多模态数据（比如可见光图像，深度图像和红外图像等）的互补性和多源性。基于多模态数据的目标跟踪算法的优点是，它可以利用多模态数据的互补性，设计出对复杂环境、光照变化与遮挡等干扰因素更加鲁棒的目标跟踪算法。

最近几年，深度学习在计算机视觉领域中得到了广泛应用 [2,15,16]，并取得了突破性的进展。深度学习成功的关键在于它具有多层网络结构，特征表达能力很强。科研人员利用多种深度神经网络，比如层叠降噪自编码网络 [17]，卷积网络 [18] 和循环网络 [19] 等，提出了一些基于深度学习的目标跟踪算法 [7,20]。基于深度学习的目标跟踪算法的优势在于，利用深度学习的强大特征表达能力，能够更好地表示目标和背景之间的差异。

基于相关滤波器的目标跟踪算法的速度很快，它已成为目标跟踪领域的研究热点。近年，研究人员提出了许多基于相关滤波器的目标跟踪算法。大部分目标跟踪算法都存在跟踪速度慢的问题，而基于相关滤波器的目标跟踪算法通过使用快速傅里叶变换，将空间域的相关运算转化成频域的乘积运算，极大地提高了跟踪速度。基于相关滤波器的目标跟踪方法的速度最快可达数百帧每秒，成功地解决了跟踪速度慢的问题。

1.2.1 视觉目标跟踪算法的分类

从特征的角度看，目标跟踪算法可以分成基于手工特征的目标跟踪算法 [21] 和基于深度特征的目标跟踪算法。基于手工特征的目标跟踪算法是指使用人工设计的特征来训练一个模型进行目标跟踪。与深度特征相比，人工设计的特征属于浅层特征，因此这类算法也可以称作基于浅层特征的目标跟踪算法。基于深度特征的目标跟踪算法 [22] 是指，利用深度神经网络（比如 SDAE，CNN 或 RNN）提取深度特征来训练一个模型进行目标跟踪，这类算法也可以称为基于深层特征的目标跟踪算法。由于深度神经网络具有多层网络架构，具有强大的特征表达能力，因此基于深度特征的目标跟踪算法可以有效地利用深度特征来改善跟踪效果。

从数据源角度看，目标跟踪算法可以分为基于单模态数据的目标跟踪算法和基于多模态数据的目标跟踪算法。基于单模态数据的目标跟踪算法是指仅使用单一数据源来

进行目标跟踪。按照数据源种类的不同，它又可以细分为基于可见光图像的目标跟踪方法，基于红外图像的目标跟踪和基于深度图像的目标跟踪。目前，绝大多数的目标跟踪算法都仅仅使用了单模态数据，尤其是可见光图像。基于多模态数据的目标跟踪算法是指通过融合多种不同的数据源来改善跟踪效果的跟踪算法。按照融合数据源的不同，又可以细分为两类：基于可见光图像与红外图像的目标跟踪算法，基于可见光图像和深度图像的目标跟踪算法。基于多模态数据的目标跟踪算法可以有效地利用多模态数据的互补性，因此它具有较好的鲁棒性。

从采样的角度看，目标跟踪算法可以分成基于稀疏采样的目标跟踪算法和基于密集采样的目标跟踪算法。基于稀疏采样的目标跟踪算法主要是指传统的判别式目标跟踪算法，它使用稀疏采样策略，在目标的邻近位置采集一些样本，根据距离目标的远近分成正负两类样本。由于目标跟踪具有实时性的要求，这类方法只能采集少量的样本。基于密集采样的目标跟踪算法主要是指基于相关滤波器的算法，它将目标跟踪任务看作回归问题。它使用密集采样策略，将每一个像素看作一个样本，样本的标签是通过以目标为中心的高斯函数计算得到的，即标签是连续的。相比较而言，基于稀疏采样的目标跟踪算法有两个缺陷：第一，仅使用少量样本训练模型，容易出现漂移现象；第二，样本之间的重叠区域较大，具有较大的冗余。相反，基于密集采样的目标跟踪算法使用了所有的样本，不容易出现漂移现象；同时可以利用快速傅里叶变换进行加速，跟踪速度很快。

1.2.2 基于深度学习的目标跟踪算法

近年，由于深度神经网络 [23] 具有强大的特征表达能力，它在计算机视觉领域的很多方向（比如目标检测和图文标注等）都取得了突破性的进展。但将深度神经网络用于视觉目标跟踪任务 [24,25]，需要解决两个问题：训练周期长和需要大量训练样本。众所周知，深度神经网络的训练周期很长，如果通过在线训练的方式，就不能满足目标跟踪的实时性要求。另外，深度神经网络具有多层网络结构，含有大量的参数，因此为了防止出现过拟合现象，需要大量的训练样本，然而目标跟踪具有样本量少的特点。如果使用深层神经网络，必然会发生过拟合，进而出现漂移现象；相反，如果使用浅层神经网络，学习到的特征的表达能力不强，不足以应对光照变化、遮挡和背景混乱等干扰因素。科研人员根据目标跟踪的自身特点，灵活地把各种深度神经网络应用在目标跟踪任务上，并得到了很好的结果。

根据深度神经网络架构的不同，基于深度神经网络的目标跟踪算法 [26,27] 可以进一步细分为基于降噪自编码网络的目标跟踪算法，基于卷积神经网络的目标跟踪算法和基于双流网络（Siamese Network）的目标跟踪算法。其中，基于卷积神经网络的目标跟踪算法最多。

DLT（Deep Learning Tracker）是一种最典型的基于降噪自编码网络（Stacked Denoising Autoencoder，SDAE）的目标跟踪算法 [17]。DLT 算法可以分成两个阶段：离线训练阶段和在线跟踪阶段。在离线训练阶段，DLT 算法使用 Tiny Images 数据集中的 8000 万张尺寸为 32x32 图像作为训练集，使用逐层训练方法训练一个层叠自编码网络。DLT 算法使用已训练网络的编码模块作为一个特征提取器，并在网络的最后添加一个逻辑回归层作为二分类器。在跟踪阶段，DLT 算法首先在目标附近采样，然后使用上述网络对样本分类以确定目标位置，最后再使用这些样本在线更新网络。DLT 是第一个使用自编码网络的目标跟踪算法，但由于网络结构简单，该算法的跟踪效果并不理想。

Fan 等 [28] 提出了一种使用卷积神经网络的行人跟踪算法。该方法首先构建了一个包含 2 万张行人图像的数据库，然后使用离线训练的方式训练一个浅层卷积神经网络，并将其用于在线目标跟踪。传统的卷积神经网络仅仅学习空域特征，并具有移不变的特性；而该算法可以同时学习空域特征和时域特征，时域特征有效表达了相邻帧图像之间的关系。此外，由于目标跟踪算法在目标附近采样，样本之间的距离比较小，因此移变特征更适合于目标跟踪任务。该算法使用的卷积神经网络架构具有移变特性，有助于消除漂移现象。

Tao 等 [29] 利用双流网络设计了 SINT（SIamese Instance search Tracker）算法，它是一种实时的目标跟踪算法。该算法在 ALOV 数据集上，通过离线训练双流网络，得到一个用于计算相似度的匹配函数。在跟踪阶段，首先在当前帧中的目标附近采样，然后利用学习到的匹配函数计算这些样本与第一帧中的目标之间的相似度，相似度最高的样本的位置作为目标位置。由于 SINT 算法在跟踪阶段不更新匹配函数，因此满足实时性的要求。

根据深度神经网络训练阶段的不同，基于深度神经网络的目标跟踪算法可以细分为以下三种方式：基于在线训练深度网络的目标跟踪算法，基于离线训练深度网络的目标跟踪算法和基于混合训练深度网络的目标跟踪算法。

基于在线训练深度网络的目标跟踪算法是指，不需要离线训练，仅在跟踪阶段初始化并训练深度神经网络的目标跟踪算法。Li 等 [30] 提出了第一个基于在线训练深度网络的目标跟踪算法，它使用了一个仅含 2 个卷积层的卷积神经网络。2015 年，Li 等提出了 DeepTrack 算法 [31]，它将多个卷积神经网络组成一个卷积神经网络池，并从中挑选最好的卷积神经网络来预测目标位置。由于目标跟踪具有实时性要求，因此这类算法的共同特点是网络层数少。压缩网络层数，可以减少在线训练时间，但由于网络结构简单，特征表达能力较弱，目标跟踪的效果并不理想。

基于离线训练深度网络的目标跟踪算法 [32] 是指，需要离线训练神经网络，在跟踪阶段不再需要更新网络的目标跟踪算法。由于基于在线训练的目标跟踪算法使用的网络架构相对简单，特征表达能力不强，因此大部分基于深度网络的目标跟踪算法都使用离线训练的方法。Wang 等 [33] 使用离线训练好的 VGG-Net [34] 设计了 FCNT 跟踪算法。该算法使用了多层卷积特征图，同时设计了一个特征图选择方法以去除冗余的特征图。siamFC 算法 [35] 通过离线训练方式得到一个全卷积双流网络，并使用该网络计算第一帧中的目标与当前搜索区域的相似性得分图，得分最高的位置作为目标的位置。由于基于离线训练的目标跟踪算法使用的网络层数较多，特征表达能力强，并且不需要在线训练，因此具有较高的性能。

基于混合训练深度网络的目标跟踪算法是指，既需要离线训练网络，在跟踪阶段也需要更新网络的目标跟踪算法。这类算法的典型代表是 DLT 算法。在离线训练阶段，DLT 算法使用 Tiny Images 数据集训练层叠自编码网络，在跟踪阶段，该算法在目标附近采样，并使用这些样本更新网络。

近年，深度学习在目标跟踪中得到了大量应用，并取得了较大的突破。从发展的角度看，基于深度神经网络的目标跟踪算法 [7,36] 经历了从在线训练方式到离线训练方式，从层叠自编码网络到卷积神经网络，再到双流网络 [37,38] 的历程。从跟踪效果的角度看，基于深度神经网络的目标跟踪算法的速度越来越快，精度越来越高。

1.2.3 基于多模态数据的目标跟踪算法

目前，绝大多数的目标跟踪算法仅使用了单模态数据 [39]，尤其是可见光图像。但基于可见光图像的目标跟踪算法受可见光物理性质的限制，应用场景有限。近年，科研人员提出了一些基于多模态数据的目标跟踪算法 [40–43]，充分利用了多模态数据的互

补性和多源性。基于多模态数据的目标跟踪算法又可以细分为基于可见光图像与红外图像的目标跟踪算法，以及基于可见光图像与深度图像的目标跟踪算法。

Wang 等 [3] 基于可见光图像和深度图像，设计了一种使用颜色、光流和深度信息的目标跟踪算法。Yuan 等 [44] 通过融合可见光图像和深度图像，提出了一种鲁棒的超像素目标跟踪算法 [44]。它不仅使用了超像素法获取中层特征的结构信息，而且也利用了深度图中目标和背景区别明显的特点。基于可见光图像和深度图像的目标跟踪算法同时使用可见光图像的颜色与纹理等特征，以及深度图像的深度信息，提升了算法的鲁棒性。

三星手机 Note7 同时配备了普通摄像头和红外摄像头，表明可见光图像和红外图像融合技术越来越重要。最大的视觉目标跟踪竞赛 VOT 指出，将会增加基于可见光图像与红外图像的目标跟踪比赛，一方面凸显了基于可见光图像与红外图像的目标跟踪算法的重要性，另一方面说明该类算法稀少，具有十分重要的研究价值。Kumar 等 [45] 利用低分辨率的红外图像和普通可见光图像，提高了行人跟踪算法的性能。基于可见光图像与红外图像的目标跟踪算法利用了红外图像具有不受光照条件影响和目标与背景差异较大的特点，有效地去除光照变化等因素的干扰。

1.2.4 基于相关滤波器的目标跟踪算法

最近几年，相关滤波器在目标跟踪方向得到了广泛应用 [46–50]。一方面，基于相关滤波器的目标跟踪算法使用密集采样方法，将每个像素看作一个样本，使用连续标签，因此不容易出现漂移现象。另一方面，这类算法可以通过快速傅里叶变换计算相关得分图，跟踪速度非常快，可以达到数百帧每秒。

Bolme 等 [51] 提出了基于 MOSSE 自适应相关滤波器的目标跟踪算法，它对光照变化，尺度变化和形变等具有鲁棒性，并且跟踪速度可以达到 669 帧/秒。该算法首先使用第一帧图像中的目标为模板训练一个相关滤波器；然后在下一帧图像中裁剪一个候选区域，并使用相关滤波器与该候选区域执行相关操作，得到相关响应得分图，得分最高的位置就是目标的位置；最后再更新相关滤波器。在计算相关响应得分图时，该算法使用了快速傅里叶变换进行加速。

CSK 算法 [52] 同时使用了线性相关滤波器和核技巧，并证明了核矩阵是一个循环矩阵。循环矩阵的特性使得每一个像素都可以作为一个样本，即密集采样，它具有抑

制漂移现象的优点。CSK 算法的跟踪速度可以达到 320 帧/秒。Henriques 等 [53] 提出了 DCF 算法，它将 CSK 从单通道推广到多通道。CN 算法 [50] 将 CSK 算法推广到多通道颜色特征，通过精心地选择颜色变换显著地提高了目标跟踪算法的性能。CN 算法的跟踪速度可以达到 100 帧/秒。

考虑到尺度变化对跟踪性能的影响，Danelljan 等提出了 DSST 算法 [54]。该算法可以分成两个阶段：预测目标位置和估计目标尺寸。在预测目标位置阶段，首先计算 HOG 特征图，然后计算 HOG 特征图与相关滤波器的响应以计算目标的位置。在估计目标尺寸阶段，首先以目标位置为中心裁剪不同尺寸的图像，构建一个尺度金字塔；然后在一个尺度金字塔上使用尺度相关滤波器计算目标的尺寸，有效地解决了基于相关滤波器的目标跟踪算法对尺度变化不鲁棒的问题。

1.3 现有方法的不足

传统的目标跟踪算法一般仅使用单模态数据，尤其是可见光图像。可见光图像包含丰富的颜色与纹理等特征，但当光照条件变化比较剧烈时，目标的表观将发生显著变化，基于可见光图像的目标跟踪算法受可见光物理性质的限制可能会跟踪失败。尽管红外图像不受光照变化的影响，并且目标与背景区别比较明显，但具有缺乏纹理和颜色等特征缺陷，基于红外图像的目标跟踪算法可以利用的特征比较少。因此，基于单一模态数据的目标跟踪算法没有使用多模态数据的互补性和多源性，算法的鲁棒性较差。

传统的目标跟踪算法都使用手工特征。手工特征与深度特征相比，属于浅层特征，对光照、遮挡与形变等因素不具有鲁棒性，因此严重影响了基于手工特征的目标跟踪算法的性能。深度神经网络，尤其是卷积神经网络的特征表达能力十分强大，且对微小的形变具有鲁棒性，因此基于卷积神经网络的目标跟踪算法的鲁棒性比传统的目标跟踪算法好。

基于深度神经网络的目标跟踪算法一般只能预测目标位置，对尺度变化不鲁棒，如 DLT，CNNTrack 和 FCNT 等。本文提出的算法通过尺度金字塔估计目标尺度，进一步提高了目标跟踪算法的性能。另外，基于深度神经网络的目标跟踪算法 [55] 一般不能满足实时性要求，比如，CNNTrack 需要在线训练卷积网络，而 FCNT 使用的 VGG-Net 的网络层数较深，计算量都比较大，跟踪速度都比较慢。本文提出的基于多模态数据和全卷

积双流网络的目标跟踪算法通过双流网络提高了跟踪速度，解决了跟踪速度慢的问题。

1.4 论文创新点

基于多模态数据和卷积神经网络的目标跟踪算法。 该算法采用了双融合策略，第一，融合了可见光图像的多层卷积特征图，即融合了浅层特征图的空间信息和深层卷积特征图的语义信息；第二，在算法级别上融合可见光图像和红外图像，先使用可见光图像预测目标位置，再使用红外图像估计目标尺寸，有效地利用了多模态数据的互补性。实验结果证明，该算法是一种鲁棒的目标跟踪算法，比其它算法的跟踪效果好。

基于多模态数据和全卷积双流网络的目标跟踪算法。 该算法解决了基于深度神经网络的目标跟踪算法速度慢的问题。与其它基于深度神经网络的目标跟踪算法相比，该算法跟踪速度快的原因包括以下几点。第一，该算法使用的全卷积双流网络不需要在线训练；第二，利用全卷积特性，该网络只需前向传播一次，就能输出目标位置；第三，该网络的网络层数适中，它比 Alex 的网络层数多，比 VGG-Net 的网络层数少，在特征表达能力和速度之间取了折中。该算法首先在可见光图像上使用全卷积双流网络预测目标位置，然后在红外图像上构建一个尺度金字塔模型，估计目标尺寸。

采集了一个多模态目标跟踪数据库。 目前，公开的视觉目标跟踪数据库都只包含单模态数据，尤其是可见光视频序列。而本文的研究方向是基于多模态数据的深度目标跟踪算法，需要使用同时包含可见光图像和红外图像的视频序列。为了评价本文提出的目标跟踪算法，我们构建了一个多模态目标跟踪数据集，名称是 OptTrack。该数据集共包含 6 个视频序列，每个视频都同时包含可见光图像和红外图像。

1.5 论文组织框架

本文的主要工作是研究如何有效地利用多模态数据和深度学习方法设计一种鲁棒的目标跟踪算法。各个章节的安排和主要内容如下：

第一章为绪论。本章首先介绍了单目标跟踪的研究背景和研究现状，并对现有的目标跟踪算法进行了分类。然后详细地描述了一些典型的目标跟踪算法，并对不同算法的优缺点进行了剖析。其次，针对于这些算法存在的缺点，说明了本文算法的出发点和创新点。最后，简单地介绍了本文的各章节内容。

第二章提出了基于多模态数据和卷积神经网络的目标跟踪算法。该算法融合了可见

光图像和红外图像，有效地利用了多模态数据的互补性。首先通过卷积神经网络提取可见光图像的多层卷积特征图，然后通过平移相关滤波器融合浅层的空间信息和深层的语义信息，以精确地预测目标的位置；最后在红外图像的尺度金字塔上高效地估计目标的尺度。本章首先描述了相关工作和算法概述，然后介绍了如何在可见光图像的多层卷积特征上使用平移相关滤波器预测目标位置，接着描述了如何在红外图像的尺度金字上使用尺度金字塔预测目标尺寸，最后，对实验结果进行了分析。

第三章提出了基于多模态数据和全卷积双流网络的目标跟踪算法。针对基于深度神经网络的目标跟踪算法存在速度慢的问题，该算法使用全卷积双流网络预测目标位置，该网络仅需前向传播一次，就能输出目标位置，提高了跟踪速度。另外，为了提高算法对尺度变化的鲁棒性，该算法使用红外图像的尺度金字塔模型估计目标的尺寸。

第四章为论文的总结和展望。

第二章 基于多模态数据和卷积神经网络的目标跟踪算法

本章利用多模态数据的互补性和卷积神经网络的强大特征表达能力，提出了一种基于多模态数据和卷积神经网络的目标跟踪算法。该算法使用了双融合策略，不仅融合了可见光图像的浅层特征图的空间信息和深层特征图的语义信息，而且也在算法级别上融合可见光图像和红外图像，先使用可见光图像预测目标位置，再使用红外图像估计目标尺寸。

2.1 相关工作

基于深度学习的目标跟踪算法。 近年，科研人员提出了一些基于深度神经网络的目标跟踪算法。Wang 和 Yeung [17] 使用一个离线训练的自编网络的编码模块作为特征提取器，使用传统的基于检测的跟踪框架，设计了一种基于层叠降噪自编码网络的目标跟踪算法（DLT）。Li 等 [31] 提出了 DeepTrack 算法，它使用多个卷积神经网络构建一个卷积神经网络池，从中挑选最好的卷积神经网络来预测目标位置。Wang 等 [33] 通过研究卷积神经网络的多层特征图，发现不同的网络层从不同的角度表达目标。根据这一发现，他们提出了全卷积神经网络跟踪算法（FCNT），使用两个卷积网络层的输出作为特征。Ma 等 [32] 利用卷积神经网络的多层卷积特征图，使用自适应相关滤波器预测目标的位置。上述这些目标跟踪算法都仅使用了可见光图像，而本章算法融合了可见光图像和红外图像。另外，它们仅预测目标位置，对尺寸变化不鲁棒，而本章算法使用红外图像的尺度金字塔模型计算目标尺寸。

基于多模态数据的目标跟踪算法。 为了充分利用多模态数据的互补性和多源性，科研人员提出了一些基于多模态数据的目标跟踪算法。Wang 等 [3] 基于可见光图像和深度图像，设计了一种使用颜色、光流和深度信息的鲁棒的目标跟踪算法。Yuan 等 [44] 通过融合可见光图像和深度图像，提出了一种鲁棒的超像素目标跟踪算法。Kumar 等 [45] 利用低分辨率的红外图像和普通可见光图像，提高了行人跟踪算法的性能。上述这些基于多模态数据的目标跟踪算法都仅使用了手工特征，没有使用深度特征，影响了目标跟踪算法的性能。

基于相关滤波器的目标跟踪算法。 基于相关滤波器的算法可以使用快速傅里叶变

换进行加速，具有跟踪速度快的特点，已成为目标跟踪方向的研究热点。Henriques 等[52] 使用线性相关滤波器和核技巧提出了 CSK 算法，它的跟踪速度很快。Danelljan 等[54] 提出了 DSST 算法，它使用一个尺度金字塔来估计目标尺寸。Henriques 等 [53] 提出了 DCF 算法，将 CSK 算法推广到多个通道。上述这些目标跟踪算法都仅使用了手工特征和单模态数据，而本章算法不仅使用深度特征，而且也融合了可见光图像和红外图像。

2.2 本章算法概述

针对上述现有方法的不足，本章提出了一个基于多模态数据和卷积神经网络的目标跟踪算法，它不仅融合了可见光图像和红外图像，而且融合了浅层卷积特征图的空间信息和深层卷积特征图的语义信息。该算法整体上可以分为两步：首先使用可见光图像的多层卷积特征图来预测目标位置，然后使用红外图像的尺度金字塔来进一步估计目标尺寸。算法流程图如图2.1所示。

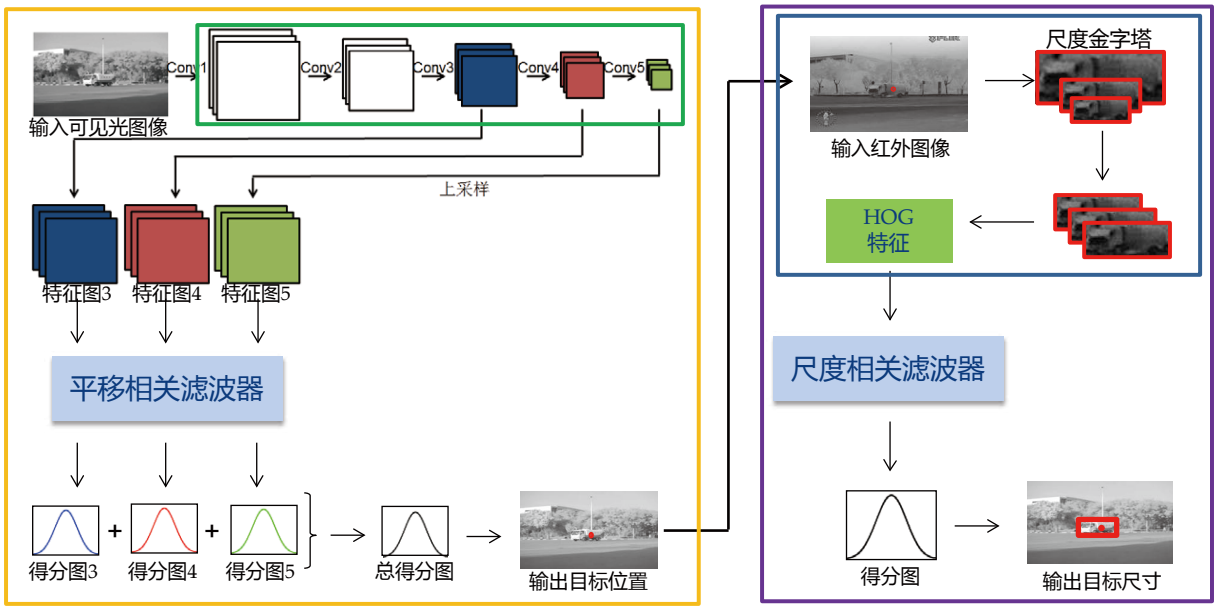


图 2.1 本章算法的整体流程图

该算法主要包括以下步骤：1）预处理可见光图像。由于我们采集的图像尺寸比较大，而本章算法使用的 VGG-Net-19 [34] 的输入尺寸为 $224 \times 224 \times 3$，因此首先需要对可见光图像进行尺寸变换，将可见光图像的维度变成 $224 \times 224 \times 3$。2）计算可见光图像的特征图。将预处理后的可见光图像输入到 VGG-Net-19，输出多层卷积特征图。3）预测目标的位置。首先在卷积特征图上应用平移相关滤波器，得到多层相关响应图；然后

对其进行加权求和，得到最终的相关得分图；最后将得分最高的位置作为目标位置。4）构建尺度金字塔模型。以上一步得到的目标位置为中心，裁剪不同尺寸的图像块，构建一个尺度金字塔。5）提取尺度金字塔模型的 HOG 特征。在不同尺度的图像块上使用 HOG 描述子，得到对应的 HOG 特征向量。6）估计目标的尺寸。在尺度金字塔的 HOG 特征上使用尺度相关滤波器，将得分最高的图像块的尺寸作为目标尺寸。

2.3 基于多层卷积特征图预测目标位置

本章算法首先提取可见光图像的多层卷积特征图，然后使用平移相关滤波器预测目标位置。

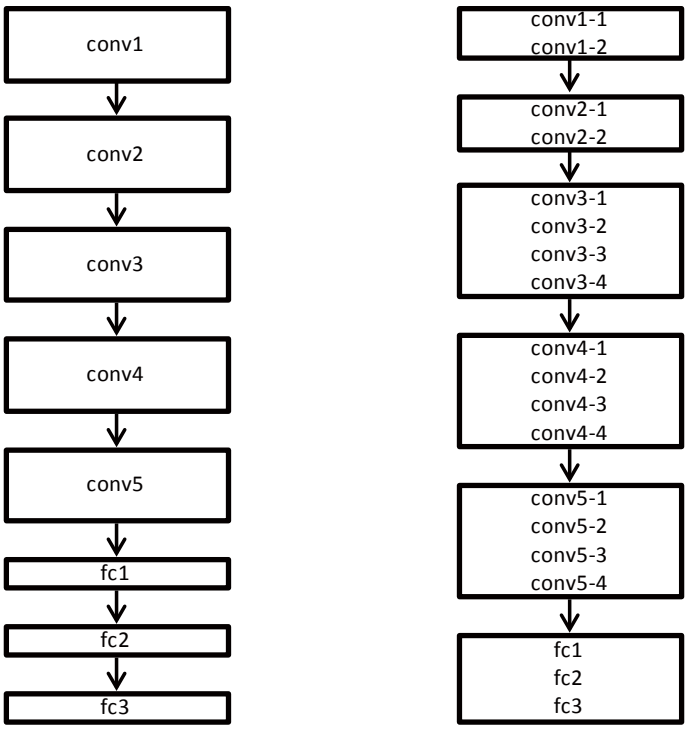
基于在线训练卷积网络的目标跟踪算法具有以下两个缺点。由于目标跟踪具有实时性要求，因此基于在线训练卷积网络的目标跟踪算法都使用浅层卷积网络，特征表达能力弱。此外，由于目标跟踪中样本量比较少，在线训练的方式很容易导致卷积网络过拟合，造成漂移现象。

本章算法属于基于离线训练卷积神经网络的目标跟踪算法。它有两个优点：第一，在跟踪阶段不需要对网络进行更新，因此可以节省训练网络的时间，提高跟踪速度；第二，与在线训练卷积神经网络方式相比，离线训练的方式使用更深的卷积网络，特征表达能力强，可以提高目标跟踪算法的性能。

离线训练卷积神经网络需要大量的训练样本，常用的数据集是 ImageNet。ImageNet 分类数据集包含 1000 个类别，每个类别大约有 1000 张图像，总共包含 120 万张训练图像，5 万张验证图像，15 万张测试图像。在 ImageNet 上训练好的卷积神经网络有许多，其中比较有代表性的是 AlexNet[56] 和 VGG-Net [34]，如图2.2所示。AlexNet 总共包含 8 个权重网络层，其中 5 个为卷积层，3 个为全连接层。按照网络层数的不同，VGG-Net 可以分成 3 个不同的版本。层数最多的 VGG-Net 总共包括 19 个权重网络层，其中 16 个为卷积层，3 个为全连接层，本文使用 VGG-Net-19 表示它。由于 VGG-Net-19 使用较小的卷积核（3×3），它比 AlexNet 包含更多的卷积层，能够提取更加鲁棒的语义特征，因此本章选择使用 VGG-Net-19 来做特征提取器。

全连接网络层的输出不适用于目标跟踪。全连接网络层的输出含有较高的语义信息，适用于目标分类，然而目标跟踪的首要任务是确定目标的位置。如果使用全连接层

的输出作为特征，虽然包含丰富的语义信息，但损失了大量的空间信息，不利于确定目标位置。为了更加精确地定位目标，本章提出的目标跟踪算法没有使用全连接层的输出作为特征。



(a) AlexNet　　(b) VGG-Net-19

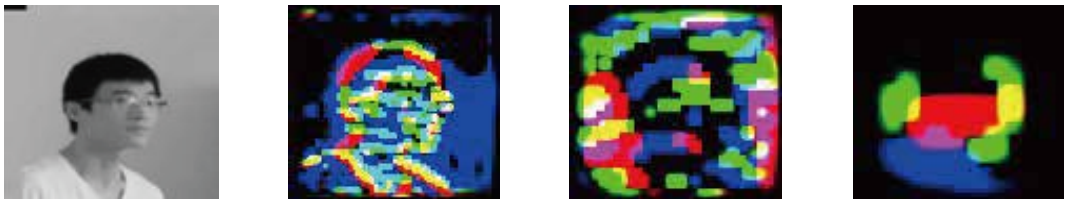
图 2.2　AlexNet 和 VGG-Net-19 的架构图

本章算法使用多层卷积特征图作为特征。在卷积神经网络中，由于卷积和池化操作，随着网络层数的加深，目标类别间的语义特征越来越强，而卷积特征图的尺寸越来越小。多层卷积特征图的特点如图2.3所示。浅层特征图属于低级特征，具有较低的判别能力；但它的尺寸比较大，包含较多的空间信息，这对于精确地定位目标而言非常重要。相反，深层特征图的尺寸比较小，只能粗略地估计目标的位置；然而深层网络层包含更多的语义信息，具有更高的判别能力。为了提高算法的鲁棒性，本章算法使用 $conv3-4$，$conv4-4$ 和 $conv5-4$ 三个卷积层的输出作为特征，融合了浅层特征图的空间信息和深层特征图的语义信息。图2.4展示了将一幅图像输入 VGG-Net-19 生成的 $conv3-4$，$conv4-4$ 和 $conv5-4$ 三个卷积层的特征图。

在将可见光图像输入到 VGG-Net-19 之前，需要进行一些必要的预处理操作。由于我们采集的可见光图像是灰度图像，且尺寸比较大，而 VGG-Net-19 的输入维度是

$224 \times 224 \times 3$，因此需要对可见光图像进行预处理，转换成 VGG-Net-19 指定的输入维度。

图 2.3 多层卷积特征图的特点



(a) 输入图像 (b) conv3-4 (c) conv4-4 (d) conv5-4

图 2.4 VGG-Net-19 的多层卷积特征图示例

将预处理之后的可见光图像输入到 VGG-Net-19，输出多层卷积特征图。由于卷积和池化操作，不同层的卷积特征图具有不同的尺寸。为了方便后续操作，需要将它们归一化成相同的尺寸。通过 VGG-Net-19 计算多层卷积特征图的过程如图2.5所示。

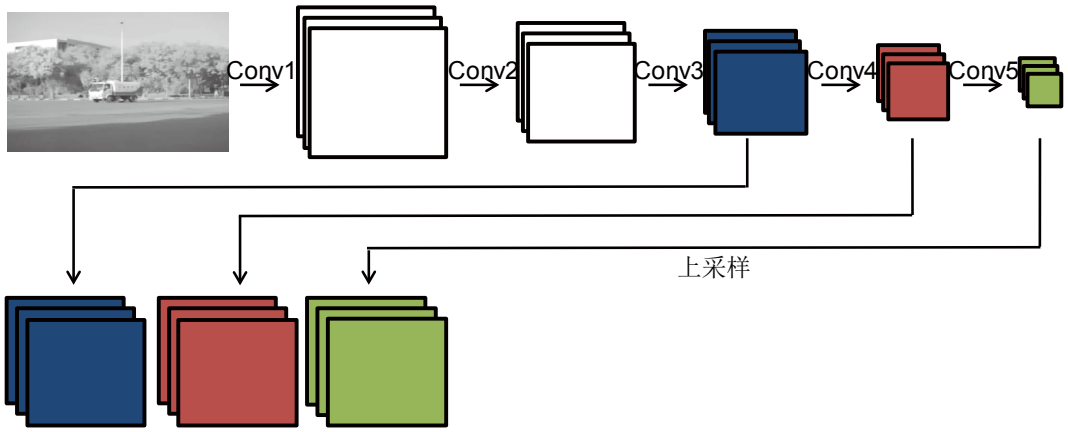
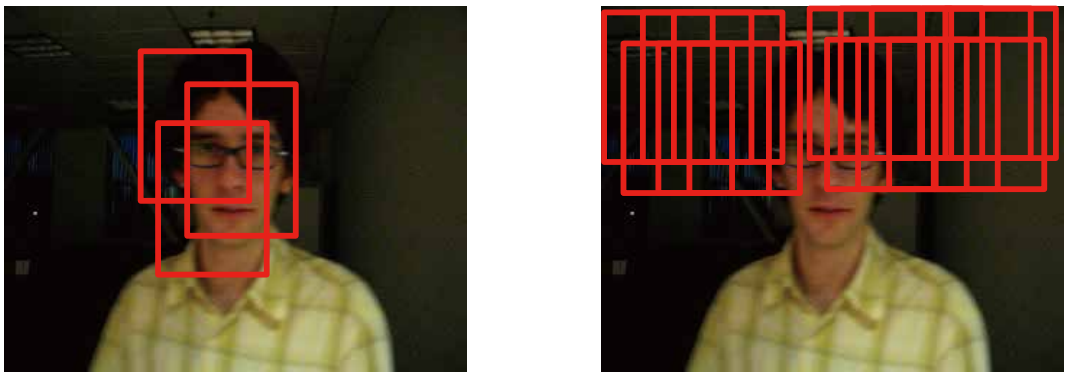


图 2.5 使用 VGG-Net-19 提取多层卷积特征图的流程图

本章算法通过在多层卷积特征图上使用平移相关滤波器来预测目标位置。基于相关滤波器的跟踪算法具有两个优点：密集采样和速度快。基于相关滤波器的跟踪算法采用密集采样的方法，如图2.6所示，将每个像素视作一个样本，通过以目标位置为中心的高斯函数为每个像素添加标签。相比于传统的目标跟踪算法使用的稀疏采样方法，密集采样方法使用所有的样本，不容易出现漂移现象。另外，由于可以使用快速傅里叶变换加

速计算相关得分图，因此这类算法的跟踪速度非常快。



(a) 随机采样　　(b) 密集采样

图 2.6　随机采样和密集采样的区别

本章算法将 $conv3-4$，$conv4-4$ 和 $conv5-4$ 三个网络层的输出作为多通道特征。使用 x_m 表示在第 m 层卷积特征图上以目标为中心截取的图像块。x_3，x_4 和 x_5 分别对应于 $conv3-4$，$conv4-4$ 和 $conv5-4$ 网络层的输出。x_{mnp} 表示 x_m 竖直移动 n 个像素和水平移动 p 个像素得到的图像块，$\{x_{mnp}\}$ 表示通过密集采样方法得到的样本集。y_{mnp} 表示 x_{mnp} 的标签，它是通过以目标为中心的高斯函数 $y_{mnp}=e^{-\frac{(n-n_0)^2+(p-p_0)^2}{2\sigma^2}}$ 计算得到的，其中 (n_0,p_0) 表示目标的中心位置，$\{y_{mnp}\}$ 表示所有样本的标签。下面将在每一层卷积特征图上训练一个平移相关滤波器。

平移相关滤波器的损失函数如下所示：

$$\varepsilon_m^d=\sum_{n,p}\left\|h_m^d x_{mnp}^d-y_{mnp}\right\|^2+\lambda\left\|h_m^d\right\|_2^2 \tag{2-1}$$

其中，h_m^d 表示第 m 层卷积特征图中第 d 个通道的平移相关滤波器，λ 表示正则项的参数，它可以避免分母是 0 的问题。为了得到最优的相关滤波器 h_m^{d*}，需要最小化损失函数：

$$h_m^{d*}=\arg\min_{h_m}\sum_{n,p}\left\|h_m^d\cdot x_{mnp}^d-y_{mnp}\right\|^2+\lambda\left\|h_m^d\right\|_2^2 \tag{2-2}$$

其中，$\cdot$ 表示内积运算。最优相关滤波器 h_m^{d*} 对应的频域解如下所示：

$$H_{mt}^d=\frac{Y\odot\overline{X_{mt}^d}}{\sum_{d=1}^D X_{mt}^d\odot\overline{X_{mt}^d}+\lambda}=\frac{A_{mt}^d}{B_{mt}^d+\lambda} \tag{2-3}$$

$$A_{mt}^d=Y\odot\overline{X_{mt}^d} \tag{2-4}$$

$$B_{mt}^{d} = \sum_{d=1}^{D} X_{mt}^{d} \odot \overline{X_{mt}^{d}} \tag{2-5}$$

大写字母表示对应小写字母的傅里叶变换解，比如，Y 表示 y 的快速傅里叶变换解。t 表示视频序列中的序号。H_{mt}^{d} 是第 t 帧图像的第 m 层特征图中的第 d 个通道得到的最优相关滤波器的傅里叶变换解。$\overline{X_{mt}^{d}}$ 表示 X_{mt}^{d} 的复共轭。

模型更新。为了加快求解最优相关滤波器，首先分别更新公式2-3的分子 A_{mt}^{d} 和分母 B_{mt}^{d}，然后再通过它们计算 H_{mt}^{d}，如下式所示。

$$A_{mt}^{d} = (1-\eta)A_{m(t-1)}^{d} + \eta Y \odot \overline{X_{mt}^{d}} \tag{2-6}$$

$$B_{mt}^{d} = (1-\eta)B_{m(t-1)}^{d} + \eta \sum_{d=1}^{D} X_{mt}^{d} \odot \overline{X_{mt}^{d}} \tag{2-7}$$

$$H_{mt}^{d} = \frac{A_{mt}^{d}}{B_{mt}^{d}+\lambda} \tag{2-8}$$

其中，η 表示学习率。第 t 帧图像中的第 m 层特征图的相关滤波器的响应，即相关得分图，可以通过公式 2-9计算得到：

$$f_{mt} = \mathcal{F}^{-1}(\sum_{d=1}^{D} H_{mt}^{d} \odot \overline{X_{mt}^{d}}) \tag{2-9}$$

其中，f_{mt} 表示第 t 帧图像中的第 m 层特征图的相关滤波器的响应，$\mathcal{F}^{-1}$ 表示逆傅里叶变换运算符。将不同网络层对应的相关得分图进行加权，可以融合多层卷积特征图，如公式2-10所示。

$$f_{t} = \sum_{m=3}^{5} f_{mt} \tag{2-10}$$

其中，f_t 表示加权之后的相关得分图。通过查找相关分数图 f_t 的最大响应所在的位置，就可以获得目标位置。

如图2.1所示，本节预测目标位置的方法可以分成三个步骤：

步骤 1，预处理可见光图像。由于我们采集的图像尺寸比较大，且为灰度图像；而 VGG-Net-19 的输入维度为 $224 \times 224 \times 3$，因此需要预处理操作。首先需要将原始图像的尺寸变换成 224×224；然后将灰度图像变成维度为 $224 \times 224 \times 3$ 的彩色图像；最后，使用 ImageNet 的均值图像对可见光图像进行均值化处理。

步骤 2，计算可见光图像的多层卷积特征图。将预处理后的可见光图像输入到 VGG-Net-19，前向传播一次，将 $conv3-4$，$conv4-4$ 和 $conv5-4$ 三个卷积层的输出作为特征图。由于三层卷积特征图的尺寸不相同，还需要使用双线性插值法将其变换成相同的尺寸，分别记为 x_3，x_4 和 x_5。

步骤 3，使用平移相关滤波器预测目标的位置。首先在不同卷积特征图上进行密集采样，并使用以目标位置为中心的高斯函数得到样本的标签。然后学习一个最优相关滤波器，并使用快速傅里叶变换计算不同层卷积特征图的相关得分图。接着为了融合多层卷积特征图的空间信息和语义信息，计算多个相关得分图的加权和，得到最终的相关得分图 f_t。最后通过查找 f_t 的最大值，获得目标的位置。

2.4 基于多尺度金字塔模型估计目标尺寸

现有的基于深度学习的目标跟踪算法只能预测目标的位置，对目标的尺度变化不鲁棒。本章提出的目标跟踪算法可以精确地估计目标尺寸。由于在目标跟踪任务中相邻两帧之间的尺度变化一般小于位置变化，因此我们首先使用一个平移相关滤波器预测目标位置，然后以目标位置为中心构建一个尺度金字塔，并使用尺度相关滤波器估计目标的尺寸。由于红外图像中目标和背景差异较大，分界线明显，更有利于估计目标尺寸，因此本章算法选择在红外图像上构建一个尺度金字塔。

计算目标尺寸的方法有 2 种，一种是遍历完整尺度空间的穷尽搜索方法，另一种是基于尺度金字塔的快速搜索方法。由于快速搜索方法可以显著地缩小搜索空间，比穷尽搜索方法更加高效，因此本章算法使用尺度金字塔来估计目标尺寸。

在构建尺度金字塔模型之后，需要提取图像块的特征。方向梯度直方图（Histogram of Oriented Gradient,HOG）特征，通过计算图像局部区域的梯度直方图来表示图像特征，在计算机视觉领域中应用十分广泛。由于红外图像中目标和背景的区分明显，包含显著的梯度信息，并且不同尺寸的图像块之间的 HOG 特征区别明显，因此本章算法通过提取红外图像的尺度金字塔的 HOG 特征向量作为判别目标尺寸的特征。

如图2.1所示，本节估计目标尺寸的方法可以分成三个步骤：

步骤 4，使用红外图像构建一个尺度金字塔。假设上一步得到的目标位置是 (n_0, p_0)，目标的尺寸是 $N \times P$，尺度相关滤波器的大小为 S，a 表示尺度金字塔中不同层之间的尺

度缩放因子。对于每一个 $l \in \left\{\left\lfloor -\frac{S-1}{2} \right\rfloor, \cdots, \left\lfloor \frac{S-1}{2} \right\rfloor\right\}$，以 (n_0, p_0) 为中心，裁剪尺寸为 $a^l N \times a^l P$ 的图像块 J_l，$\{J_l\}$ 表示构建的尺度金字塔。然后，使用双线性插值法将尺度金字塔中的图像块变换为相同尺寸，用于提取相同维度的特征向量。

步骤 5，提取尺度金字塔 $\{J_l\}$ 的 HOG 特征。为了降低光照变化和噪声的干扰，首先对 J_l 使用 Gamma 校正调节图像的对比度；其次计算每个像素的梯度大小和梯度方向；接着将 J_l 划分成小单元，并统计每个小单元的梯度直方图，得到每个小单元的特征描述子；然后将几个小单元组合成一个较大的块，串联其中每个小单元的特征描述子，得到块的描述子；最后将所有块的特征描述子串联起来，得到 J_l 的特征向量 k_l, 其中 k_l 向量的维度是 I。

步骤 6，使用尺度相关滤波器估计目标的尺寸。首先使用高斯函数生成 k_l 的标签 g_l。然后训练一个一维尺度相关滤波器，它的目标函数如公式2-11所示：

$$h^{i^*} = \arg\min_{h^i} \sum_l \left\| h^i k_l^i - g_l \right\|^2 + \lambda \left\| h^i \right\|_2^2 \tag{2-11}$$

其中，h^i 表示第 i 维特征对应的相关滤波器，h^{i^*} 是对应的最优尺度相关滤波器。λ 表示正则项的参数。$\cdot$ 表示内积运算。上面目标函数的频域解如式2-12所示：

$$H_t^i = \frac{G \odot \overline{K_t^i}}{\sum_{i=1}^I K_t^i \odot \overline{K_t^i} + \lambda} = \frac{A_t^i}{B_t^i + \lambda} \tag{2-12}$$

$$A_t^i = G \odot \overline{K_t^i} \tag{2-13}$$

$$B_t^i = \sum_{i=1}^I K_t^i \odot \overline{K_t^i} \tag{2-14}$$

大写字母表示对应小写字母的傅里叶变换解，t 表示视频序列中的序号。H_t^i 表示第 t 帧图像中第 i 维特征得到的最优相关滤波器的傅里叶变换形式。$\overline{K_t^i}$ 表示 K_t^i 的复共轭。

尺度相关滤波器的更新可以通过分别更新式2-12的分子 A_t^i 和分母 B_t^i 的方式实现，如下所示：

$$A_t^i = (1-\eta) A_{t-1}^i + \phi G \odot \overline{K_t^i} \tag{2-15}$$

$$B_t^i = (1-\eta) B_{t-1}^i + \phi \sum_{i=1}^I K_t^i \odot \overline{K_t^i} \tag{2-16}$$

$$H_t^i = \frac{A_t^i}{B_t^i + \lambda} \tag{2-17}$$

其中，ϕ 表示学习率。尺度金字塔和尺度相关滤波器的相关响应可以通过式2-18计算：

$$o_t = \mathcal{F}^{-1}\left(\sum_{i=1}^{I} H_t^i \odot \overline{K_t^i}\right) \tag{2-18}$$

尺度相关滤波器的响应 o_t 的最大值对应的尺寸，就是目标的尺寸。

另外，我们在 Algorithm 1中总结了本章提出的目标跟踪算法。

Algorithm 1 基于多模态数据和卷积神经网络的目标跟踪算法

输入: 视频序列和目标在第一帧图像中的位置和尺寸

输出: 目标在后续帧图像中的位置和尺寸

1. 预处理可见光图像。将输入图像的尺寸归一化为 $224 \times 224 \times 3$，并进行均值化处理。

2. 计算可见光图像的多层卷积特征图。将可见光图像输入到 VGG-Net-19，得到多层卷积特征图。

3. 使用平移相关滤波器预测目标位置。在多层卷积特征图上使用平移相关滤波器，以相关响应得分图的加权和的最大值对应的位置作为目标位置。

4. 使用红外图像构建一个尺度金字塔。以上一步得到的目标位置为中心，裁剪不同尺寸的图像块，构建一个尺度金字塔，并进行尺寸归一化。

5. 提取尺度金字塔的 HOG 特征.

6. 使用尺度相关滤波器估计目标尺寸。在尺度金字塔上使用尺度相关滤波器，以相关响应的最大值对应的尺寸作为目标尺寸。

2.5 实验结果分析

本节，首先分别介绍了实验中使用的数据集，评价方法和对比算法。然后对实验结果进行了定量和定性地评价。

2.5.1 数据集

目前，公开的视觉目标跟踪数据库 [4] 都只包含单模态数据，尤其是可见光视频序列。而本文的研究方向是基于多模态数据的深度目标跟踪算法，需要使用同时包含可见光图像和红外图像的视频序列，无法在现有的单模态目标跟踪数据库上进行实验。

为了评价本文提出的目标跟踪算法，我们构建了一个多模态目标跟踪数据库，名称是 OptTrack。该数据库总共包含 6 个视频序列，每个视频都同时包含可见光图像和红外图像。图2.7展示了 OptTrack 中的一些具有代表性的图像，左侧为可见光图像，右侧为对应的红外图像；从上往下依次是 OptCar，OptHead，OptHeadRotate，OptHeadCard，OptRoper 和 OptBook，序列长度分别是 81，464，296，462，241 和 353。

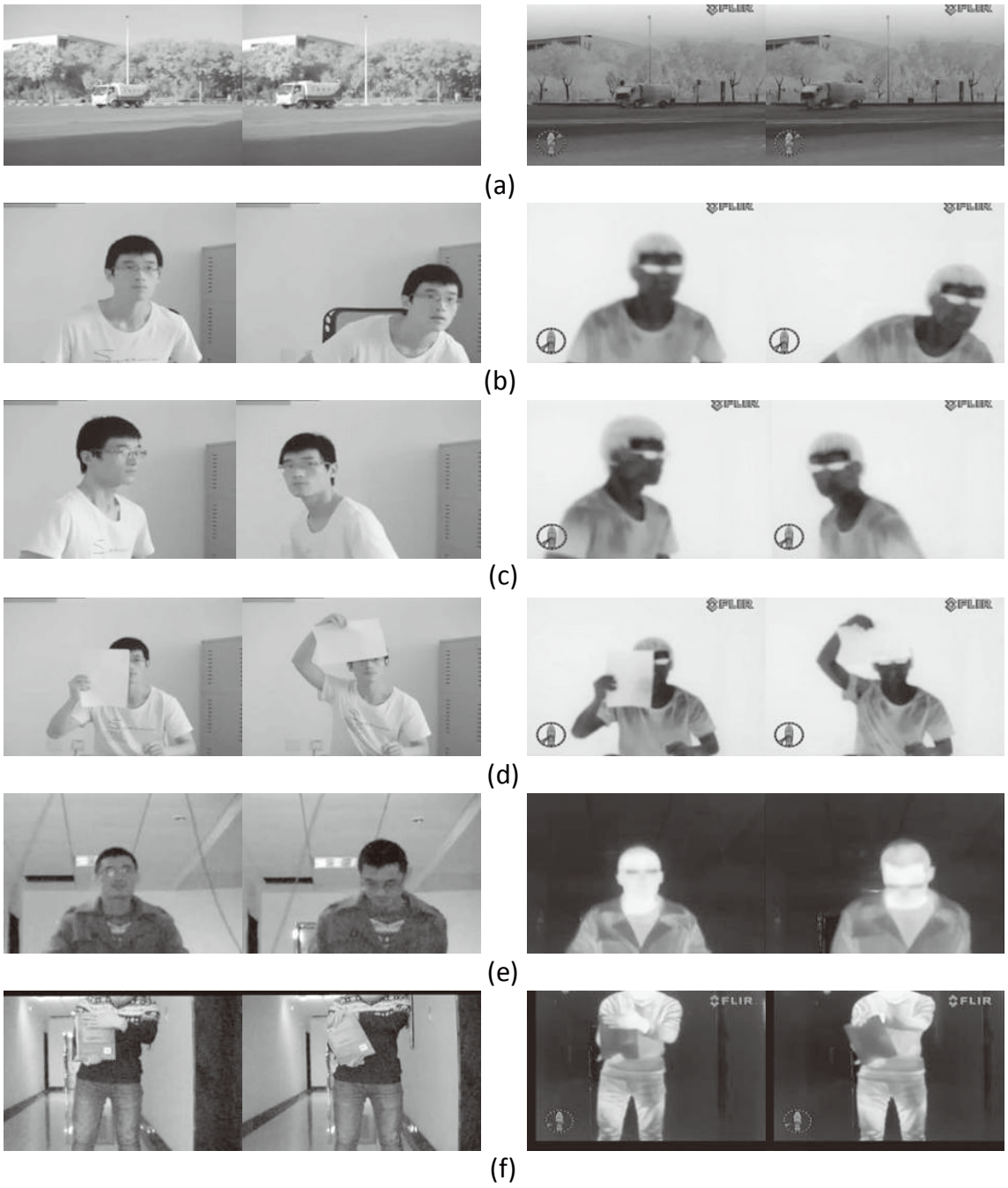


图 2.7 OptTrack 数据集的示例样本

所有视频序列的帧率都是 8 fps。它们是通过红外热成像系统在不同时间（白天或

晚上）和不同地点（室内或室外）采集的。OptRoper 和 OptBook 拍摄时间是晚上，其它均在白天拍摄。OptCar 拍摄地点是室外，其它序列均在室内拍摄。OptTrack 数据库考虑到了多种挑战因素，比如尺度变化，遮挡，平面内旋转，垂直于平面的旋转，快速运动和姿态变化等。表2.1展示了 OptTrack 数据库的详细信息。另外，由于在原始视频序列中，可见光图像和红外图像中的目标位置不同，因此我们对原始图像进行了手工配准和裁剪操作，造成各个视频序列的尺寸略有不同。

表 2.1　OptTrack 数据集的详细信息

名称	帧数	分辨率	时间	地点	挑战因素
OptCar	81	802x485	白天	室外	尺度变化
OptHead	464	802x485	白天	室内	快速运动，姿态变化
OptHeadRotate	296	802x485	白天	室内	旋转
OptHeadCard	462	802x497	白天	室内	遮挡，尺度变化，姿态变化
OptRoper	241	665x420	晚上	室内	快速运动，姿态变化
OptBook	353	835x545	晚上	室内	遮挡，旋转

2.5.2　评价指标

为了评价本文提出的目标跟踪算法，本文使用了以下的评价方法 [57]：平均中心误差，中心误差曲线，精度曲线，平均重合率和成功率曲线。

中心误差是指目标的实际位置和预测位置之间的欧氏距离。平均中心误差是指一个视频序列中所有帧的中心误差的平均值，它可以用来度量目标跟踪算法在一个视频序列上的整体性能。然而当跟踪失败时，目标跟踪算法的输出结果是任意的，平均中心误差可能会错误地度量跟踪性能。此时可以使用中心误差曲线和精度曲线来进行评价。中心误差曲线是指中心误差随时间的变化曲线。精度是指预测目标位置和实际目标位置之间的距离小于给定阈值的帧数占总帧数的比例，精度曲线表示精度随阈值的变化趋势，本文对算法的精度曲线排序时，使用的阈值是 20 个像素。

重合率是指实际目标框和预测目标框之间的交集与并集的比例，如公式2-19所示：

$$O = \frac{|s_p \cap s_r|}{|s_p \cup s_r|} \tag{2-19}$$

其中，s_r 表示实际目标框，s_p 表示预测目标框。分子表示实际目标框和预测目标框交集的像素数，分母表示它们并集的像素数。O 表示重合率。平均重合率是指一个视频

序列中所有帧的重合率的平均值，它可以用来度量跟踪算法在一个视频序列上的整体性能。成功率是指重合率 O 大于指定阈值的帧数占总帧数的比例，成功率曲线表示成功率随阈值的变化趋势。本文按曲线下面积（Area Under Curve，AUC）大小对成功率曲线进行排序。

2.5.3　对比方法

为了评价本章算法的性能，本章选择了 10 个目标跟踪算法作为对比算法。它们分别是 CT [58], ASLA [59], L1APG [60], LOT [61], ORIA [62], MTT [63], CSK [52], DSST [54], DLT [17] 和 CF2 [32]。其中，前七个目标跟踪算法的代码可以从 [4] 下载，其余算法的代码可以从作者主页下载。

2.5.4　实验结果分析

图2.8通过精度和成功率曲线展示了所有目标跟踪算法的整体性能。整体而言，本章提出的目标跟踪算法（Ours）的性能在所有目标跟踪算法中是最好的，CF2 算法排第二。一方面，因为本章算法通过组合可见光图像的多层卷积特征图，融合了空间信息和语义信息，所以本章算法能够更加精确地预测目标位置。尽管 CF2 算法使用了由粗到细的策略和多层卷积特征图，但它的精度比本章算法差。这是因为由粗到细的策略容易受到噪声的影响，而本章算法通过加权多层卷积特征图对应的相关得分图，降低了噪声的干扰。另一方面，本章算法的成功率曲线也比 CF2 算法好，这是因为它使用了红外图像的尺度金字塔的缘故。

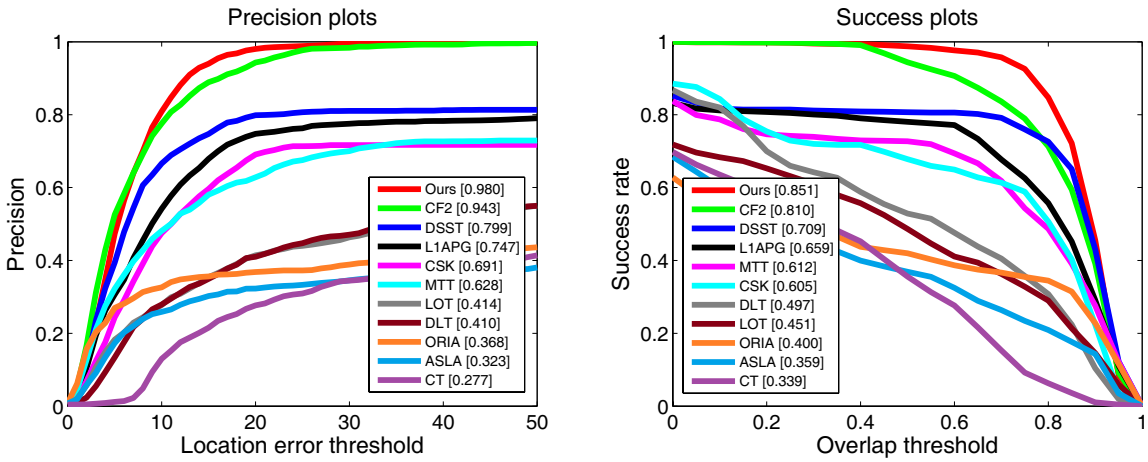


图 2.8　所有算法的整体精度曲线和成功率曲线

为了证明红外图像的尺度金字塔和可见光图像的多层卷积特征图的有效性，下面

将本章算法和其它使用卷积特征图的算法进行了对比。图2.9展示了这些目标跟踪算法的精度曲线和成功率曲线。通过比较本章算法（Ours）和不使用尺度金字塔的本章算法（Ours_c543_noscale）的精度曲线和成功率曲线，可以得到一个结论：红外图像的尺度金字塔的效果非常明显。尽管这些算法的精度曲线比较相似，但本章算法的成功率曲线比其它算法好，主要原因是它使用了红外图像的尺度金字塔模型，对尺度变化更加鲁棒。与仅使用单层卷积特征图且不使用尺度金字塔的本章算法（Ours_c3 _noscale，Ours_c4_noscale 和 Ours_c5_noscale）的精度曲线相比，使用多层卷积特征图的本章算法比使用单层卷积特征图的本章算法的精度更高。

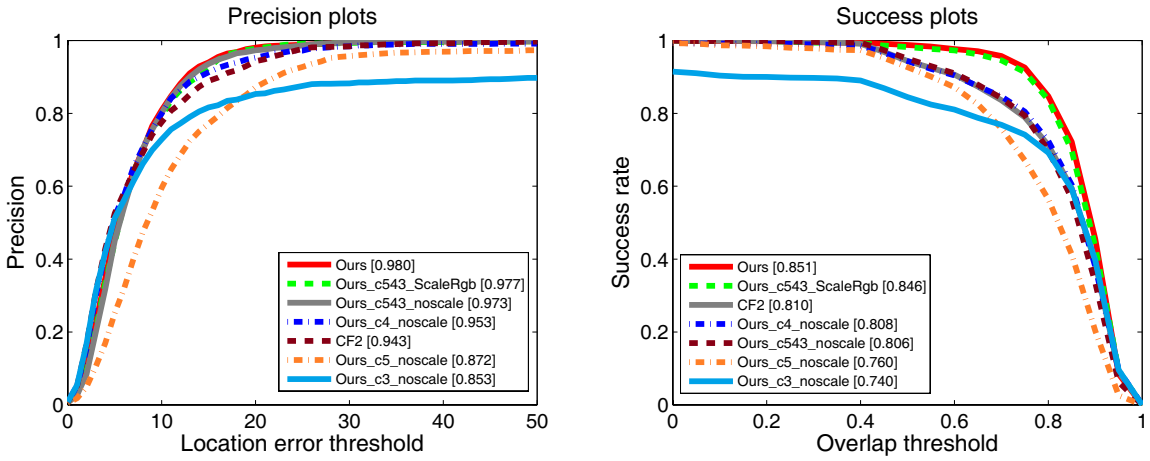


图 2.9　使用卷积特征图的算法的精度曲线和成功率曲线

图2.10展示了所有跟踪算法关于某个挑战因素的精度曲线和成功率曲线，分别用于评价尺度变化，遮挡，形变和快速运动对算法性能的影响。与其它算法相比，本章算法对尺度变化更加鲁棒，这是因为本章算法使用了红外图像的尺度金字塔来精确地估计目标的尺寸。在红外图像中，目标和背景的差别很大，有利于估计目标的尺寸；而尺度金字塔模型使得本章算法能够高效地估计目标尺寸。另外，在形变和快速运动方面，本章算法的精度曲线和成功率曲线也是最好的，主要原因是它融合了多层卷积特征图，同时使用了浅层卷积特征图的空间信息和深层卷积特征图的语义信息。在遮挡方面，尽管本章算法排第一，但相比于其它几个挑战因素，本章算法在遮挡方面略微差。这是因为本章算法使用了全局特征和相关滤波器，它们容易受到遮挡的干扰。

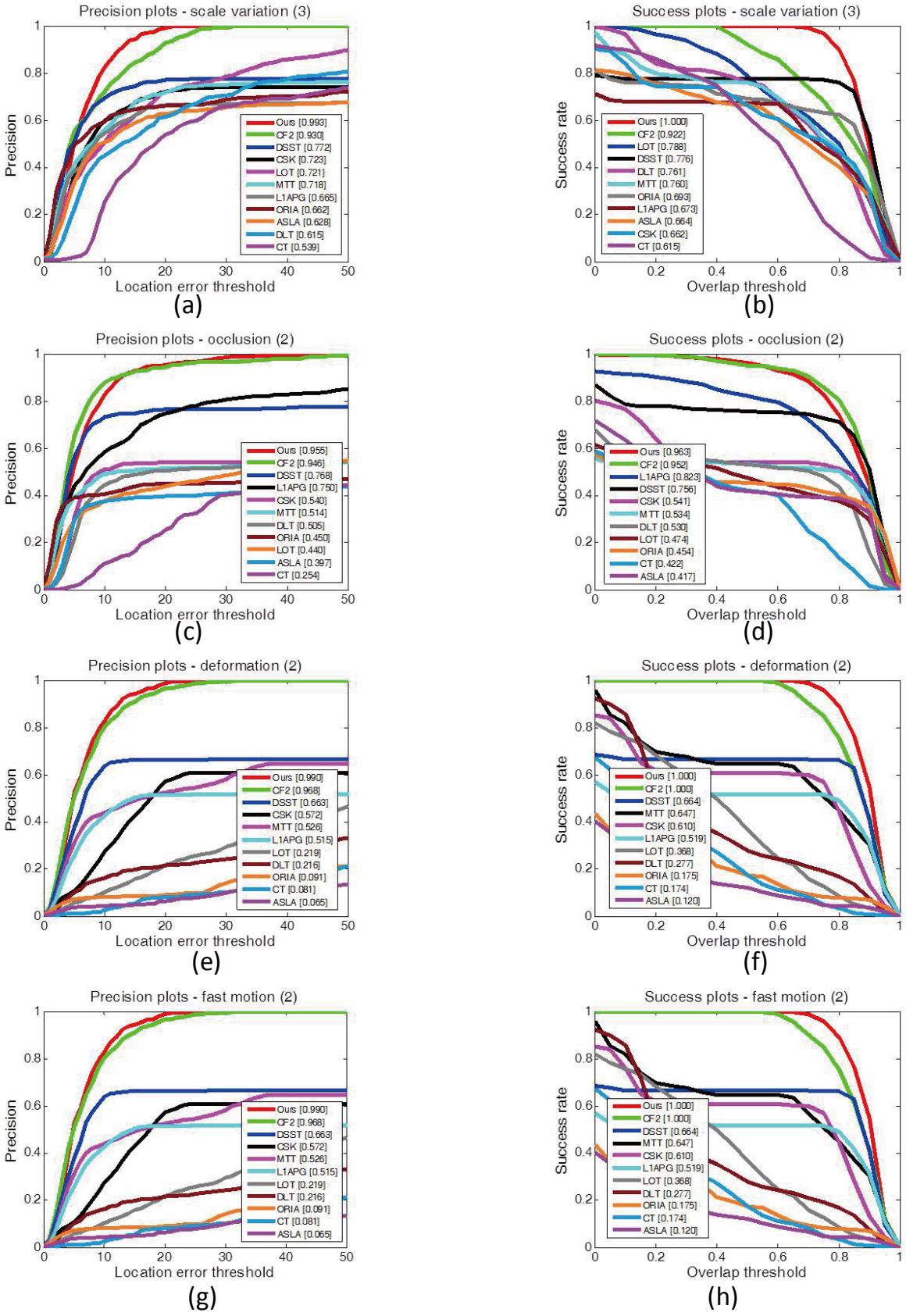


图 2.10 关于四种挑战因素的精度曲线和成功率曲线

图2.11展示了部分目标跟踪算法在 6 个视频序列上的中心误差曲线，横坐标是帧数，

纵坐标是中心误差。考虑到视觉效果，图2.11仅绘制了平均中心误差最小的 4 个对比算法和本章算法的中心误差曲线。从图2.11(a) 可以看出，所有对比算法的中心误差都随着时间推移而增长，而本章算法的中心误差在一个较小的区间内波动，证明了本章算法优于其它目标跟踪算法，具有较好的鲁棒性。从图2.11(e) 和图2.11(f) 可以看出，DSST，L1APG 和 CSK 算法的中心误差都很大，说明跟踪失败，而本文算法始终没有发生漂移现象。整体上，本文算法的中心误差在所有序列的所有帧上都小于 35 个像素，跟踪效果非常好。

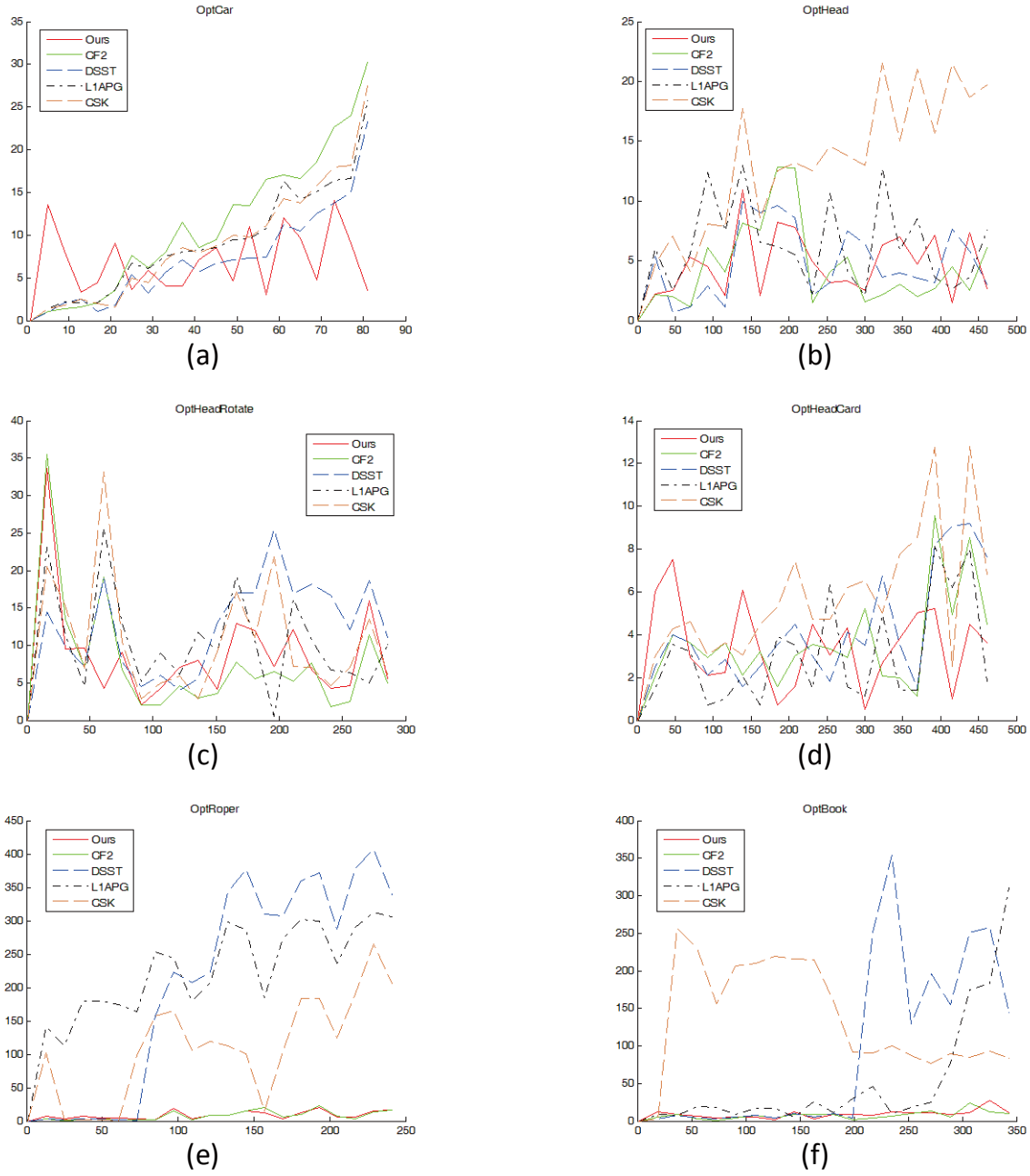


图 2.11　本章算法在六个视频序列上的中心误差曲线

图2.12展示了所有目标跟踪算法在 6 个视频序列上的实际跟踪结果，从上到下依次是 OptCar，OptHead，OptHeadRotate，OptHeadCard，OptRoper 和 OptBook 六个序列。本章算法的结果使用红色边框标记。从图2.12可以看出，本章提出的算法始终没有发生漂移现象，从定性的角度证明了本章算法在不同视频序列上都具有较好的跟踪效果。

图 2.12　所有算法在六个视频序列上的实际跟踪结果

表2.2定量地比较了平均中心误差，表2.3定量地比较了平均重合率，分别使用红色，蓝色和绿色表示排名第一，第二和第三。整体而言，除了本章算法和 CF2 之外，其它的目标跟踪算法都存在漂移现象。主要原因是 OptTrack 数据集中的视频序列包含了很多噪声和挑战因素，难度比较大。本章算法的平均重合率在所有的视频序列中都排前三名。尽管 DSST 算法中也使用尺度金字塔模型预测目标尺寸，但 DSST 的平均重合率比

本章算法的平均重合率低。原因包括两个方面。第一，DSST 算法使用手工特征，影响了跟踪性能，比如在 OptRoper 序列中出现漂移现象。而本章算法融合了 VGG-Net-19 的多层卷积特征图，对多种挑战因素具有鲁棒性。第二，DSST 使用可见光图像的尺度金字塔模型来估计目标的尺寸，但本章算法使用红外图像的尺度金字塔，红外图像中的目标和背景的外观差异比较大，有利于精确地估计目标尺寸。

表 2.2　平均中心误差

	OptCar	OptHead	OptHeadRotate	OptHeadCard	OptRoper	OptBook	平均值
CT	11.36	187.87	157.72	24.89	80.64	214.53	112.84
ASLA	7.23	228.54	131.05	22.86	172.13	176.39	123.03
L1APG	8.31	6.06	9.97	4.39	222.00	**61.51**	52.04
ORIA	**4.72**	267.75	147.31	22.61	100.98	225.60	128.16
MTT	9.62	11.43	11.97	5.04	83.74	444.62	94.40
CSK	8.66	12.71	**9.11**	5.06	106.39	138.52	**46.74**
DLT	16.50	119.89	44.94	8.34	63.13	165.51	69.72
DSST	**6.83**	**5.54**	11.02	**3.25**	202.84	95.18	54.11
CF2	10.81	**4.83**	**6.56**	**2.77**	**8.36**	**9.64**	**7.16**
LOT	8.39	177.81	207.67	9.32	**38.15**	260.50	116.97
Ours	**7.14**	**4.37**	**7.20**	**3.33**	**7.82**	**10.09**	**6.66**

表 2.3　平均重合率

	OptCar	OptHead	OptHeadRotate	OptHeadCard	OptRoper	OptBook	平均值
CT	0.60	0.18	0.23	0.68	0.31	0.03	0.34
ASLA	0.75	0.08	0.25	0.77	0.21	0.10	0.36
L1APG	0.75	0.90	0.86	0.92	0.04	**0.56**	0.67
ORIA	**0.90**	0.17	0.28	0.83	0.21	0.03	0.40
MTT	0.73	0.86	0.83	0.91	0.32	0.07	0.62
CSK	0.66	0.84	**0.87**	0.90	0.25	0.16	0.61
DLT	0.74	0.32	0.60	0.84	0.39	0.11	0.50
DSST	**0.89**	**0.90**	0.83	**0.93**	0.31	0.46	**0.72**
CF2	0.67	**0.90**	**0.87**	**0.92**	**0.81**	**0.78**	**0.83**
LOT	0.80	0.26	0.26	0.83	**0.51**	0.08	0.45
Ours	**0.88**	**0.92**	**0.88**	**0.92**	**0.86**	**0.76**	**0.87**

2.6 本章小结

本章针对于传统目标跟踪算法使用单模态数据和手工特征的不足，提出了一种基于多模态数据和卷积神经网络的目标跟踪算法。该算法采用了双融合策略，第一，融合了可见光图像的多层卷积特征图，同时利用了浅层卷积特征图的空间信息和深层卷积特征图的语义信息；第二，在算法级别上融合可见光图像和红外图像。在 OptTrack 数据库上与 11 个目标跟踪算法进行了对比实验，实验结果证明了本章算法优于其它算法，是一种对尺度变化和形变等因素具有鲁棒性的目标跟踪算法。

第三章 基于多模态数据和全卷积双流网络的目标跟踪算法

本章利用多模态数据的互补性和全卷积双流网络的全卷积特性，提出了一种基于多模态数据和全卷积双流网络的目标跟踪算法（FCST）。该算法在算法级别上融合了可见光图像和红外图像，首先在可见光图像上使用离线训练的全卷积双流网络预测目标位置，然后在红外图像上构建一个尺度金字塔来估计目标尺寸。由于全卷积双流网络仅需一次前向传播就可以预测目标的位置，因此本章算法的跟踪速度比较快，约为 19 fps。

3.1 相关工作

基于在线训练神经网络的目标跟踪算法。DLT 算法 [17] 是一种典型的基于降噪自编码网络的目标跟踪算法。在跟踪过程中，DLT 算法首先在目标附近采样，然后使用自编码网络对样本分类以确定目标位置，最后再使用这些样本在线更新网络。DeepTrack 算法 [31] 是一种典型的基于卷积神经网络的目标跟踪算法，它与 DLT 算法一样，也需要在线训练神经网络。在跟踪过程中，DeepTrack 也会在目标附近采样，然后训练一个卷积神经网络，并将其保存在一个卷积神经网络池中，最后从池中挑选最好的卷积神经网络来预测目标位置。虽然 DLT 和 DeepTrack 使用不同的网络架构，但是它们具有一个共同特点：在线训练网络。众所周知，神经网络的训练周期比较长，因此基于在线训练神经网络的目标跟踪算法具有跟踪速度慢的缺点。

基于离线训练神经网络的目标跟踪算法。 为了提高目标跟踪算法的性能，研究人员提出了一些基于离线训练神经网络的目标跟踪算法，比如 FCNT [33]，CF2 [32] 和本文第二章提出的基于多模态数据和卷积神经网络的目标跟踪算法等。这类算法的共同特点是使用离线训练的卷积神经网络（比如，AlexNet[56] 和 VGG-Net [34]）作为特征提取器。第二章提出的算法的跟踪速度较慢，平均速度约为 1 fps。主要原因是该算法使用的 VGG-Net 的网络层数高达 19 层，用它进行特征提取的时间比较长。FCNT 算法也使用 VGG-Net 作为特征提取器，但由于它使用了一种特征图选择方法来去除冗余的特征图，可以降低计算量，但跟踪速度也只有 3 fps。基于离线训练神经网络的目标跟踪算法解决了特征表达能力弱的问题，但也存在跟踪速度慢的问题。

基于双流网络的目标跟踪算法。 由于基于自编码网络和卷积网络的目标跟踪算法

都具有跟踪速度慢的缺陷，最近科研人员提出了基于双流网络的目标跟踪算法 [29,35,64]（比如，SINT [29],GOTURN [37] 和 siamFC [35]），解决了跟踪速度慢的问题。Tao 等 [29] 利用双流卷积网络设计了 SINT 算法，它是一种实时的目标跟踪算法。该算法在 ALOV 数据集上，基于双流卷积网络，通过离线训练的方式得到一个用于计算相似度的匹配函数。在跟踪阶段，首先在目标附近采样，然后利用训练好的匹配函数计算这些样本与第一帧中的目标之间的相似度，将相似度最高的样本位置作为目标位置。Held 等 [37] 基于双流网络设计了 GOTURN 算法，它的跟踪速度高达 100fps。该算法使用的双流网络仅需离线训练，在跟踪过程中不再需要更新。双流网络的两个输入分别是在 t-1 帧和 t 帧图像上以目标为中心裁剪的图像块，双流网络的输出是目标左上角坐标和右下角的坐标。该网络只需前向传播一次，即可输出目标的状态，因此目标跟踪的速度很快。siamFC 算法 [35] 通过离线训练的方式得到一个全卷积双流网络，并使用该网络计算第一帧图像中的目标与当前帧中的搜索区域的相似性得分图，将得分最高的位置作为目标位置。siamFC 的跟踪速度高达 58fps。SINT，GOTURN 和 SiamFC 算法的共同特点是网络只需前向传播一次，即可输出目标状态，因此跟踪速度比较快，解决了基于深度网络的目标跟踪算法的速度慢的问题。但是，上述三种方法仅使用了可见光图像，没有使用多模态数据的互补性和多源性。另外，SINT 和 GOTURN 算法都没有考虑目标尺寸的变化，对尺度变化不鲁棒；虽然 siamFC 算法使用了不同尺度的输入，但在尺度变化方面表现并不理想。

3.2 本章算法概述

针对上述现有方法的不足，本章提出了一个基于多模态数据和全卷积双流网络的目标跟踪算法。一方面，由于该算法使用的全卷积双流网络只需前向传播一次，即可输出目标位置，因此该算法的跟踪速度比较快，解决了基于神经网络的目标跟踪算法速度慢的问题；另一方面，该算法首先在可见光图像上应用全卷积双流网络预测目标位置，然后在红外图像上构建一个尺度金字塔来估计目标尺寸，在算法级别上融合了可见光图像和红外图像，有效地利用了多模态数据的互补性。本章算法整体上可以分为离线训练全卷积双流网络和在线跟踪两个阶段，其中在线跟踪阶段又可以细分为预测目标位置和估计目标尺寸两部分。算法流程图如图3.1所示。

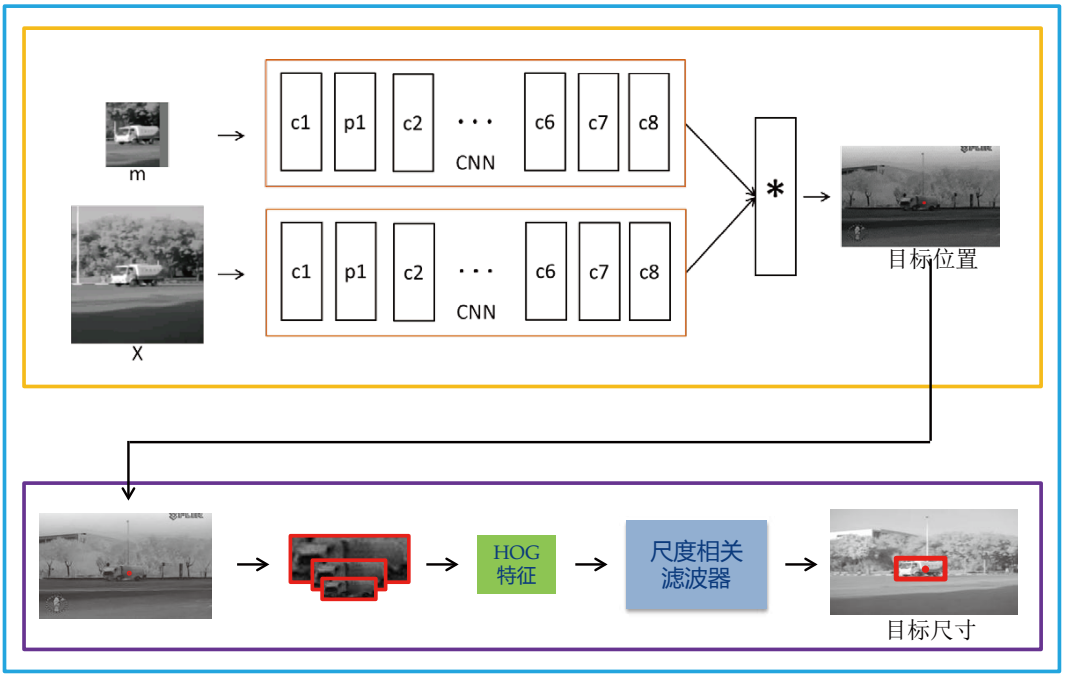


图 3.1　本章算法的整体流程图

本章算法将目标跟踪问题看作一个模板匹配问题，使用全卷积双流网络来度量相似性，实现对任意目标进行跟踪。在离线训练阶段，首先根据标签将 ImageNet Video 数据集转换成指定格式，然后使用随机梯度下降法训练一个全卷积双流网络，用于计算模板图像块和候选图像块之间的相似性。在跟踪阶段，首先将模板图像块和第 t 帧图像中搜索图像块输入到全卷积双流网络，相关网络层输出相似性得分图，相似性得分图中的最大得分的位置就是目标位置；然后，在红外图像上构建一个尺度金字塔，并使用尺度相关滤波器估计目标尺寸。

为了验证本章提出的算法的有效性，我们在 OptTrack 数据库上与 10 个目标跟踪算法进行了对比实验，实验结果证明本章算法的跟踪速度比较快，且性能优于其它目标跟踪算法。

3.3　离线训练全卷积双流网络

本节详细地介绍了离线训练全卷积双流网络的过程。首先介绍了全卷积双流网络的架构和性质；然后描述了如何从 ImageNet Video 数据集生成指定格式的训练数据集；最后介绍了如何训练全卷积双流网络。

3.3.1 全卷积双流网络的结构

本章使用的全卷积双流网络如图3.2所示。我们使用 $f(m,x)$ 表示整个网络。其中，m 表示模板图像块，x 表示候选图像块。双流网络中的 2 个橙色框区域表示卷积神经网络，二者完全相同，用字母 g 表示。该卷积网络包含 8 个卷积层和 3 个池化层，其中，池化层都是最大值池化层，表3.1列出了该卷积神经网络的参数。$*$ 表示互相关网络层。当 m 和 x 的尺寸相同时，该网络输出一个相关得分 s，计算公式如式3-1所示：

$$s = f(m,x) = g(m) * g(x) + f_0 \tag{3-1}$$

其中，f_0 表示一个常量，$g(m)$ 表示模板图像块的卷积特征图，$g(x)$ 表示候选图像块的卷积特征图，$*$ 表示互相关操作。

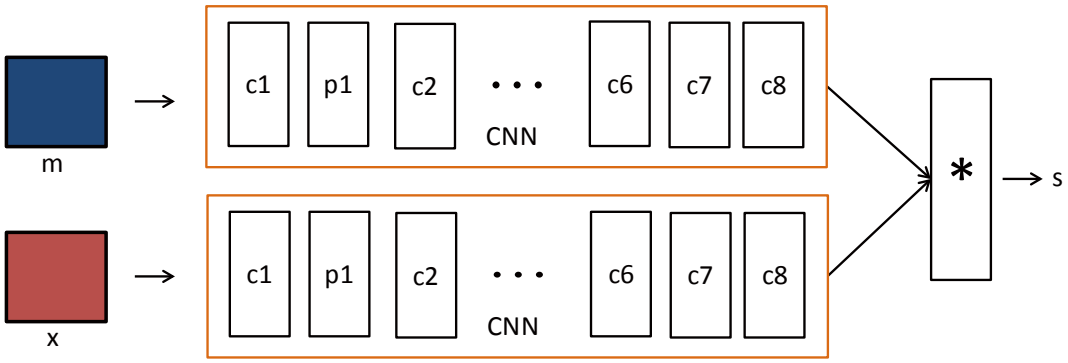


图 3.2 输入尺寸相同时的全卷积双流网络的架构

表 3.1 卷积网络模块的架构

网络层	参数	通道数	步长	模板图像块尺寸	搜索图像块尺寸
				127×127	255×255
conv1	3×3	64	1	125×125	253×253
pool1	2×2		2	62×62	126×126
conv2	3×3	128	1	60×60	124×124
conv3	3×3	128	1	58×58	122×122
pool2	2×2		2	29×29	61×61
conv4	3×3	256	1	27×27	59×59
conv5	3×3	256	1	25×25	57×57
pool3	2×2		2	12×12	28×28
conv6	3×3	256	1	10×10	26×26
conv7	3×3	512	1	8×8	24×24
conv8	3×3	512	1	6×6	22×22

本章算法综合考虑了特征表达能力和计算量两个方面，既保证了目标跟踪的实时性要求，又能够提取有效的特征。本章使用的卷积神经网络包含 8 个卷积层，层数介于 AlexNet 和 VGG-Net 之间。网络层数多时，特征表达能力强，但计算复杂度高；网络层数少时，计算复杂度低，但特征表达能力弱。

本章使用的双流网络具有全卷积特性。我们使用 L_τ 表示平移操作符，对于函数 h, 如果 $h(L_{k\tau}x) = L_\tau h(x)$ 对于任意的 τ 都成立，那么函数 h 对于步长 k 具有全卷积特性。为了保证双流网络的全卷积特性，双流网络的所有网络层均没有边缘填充操作。

按照传统的目标跟踪算法的思路，一般首先会在目标附近采样，作为候选图像块，然后使用全卷积双流网络计算模板图像块和候选图像块的相关得分，将得分最高的候选图像块的位置作为目标位置。由于这些样本存在大量的重叠区域，因此造成大量的重复计算，影响跟踪速度。

全卷积网络特性说明网络的输出与样本的位置无关，因此可以使用一个尺寸较大的搜索图像块代替候选图像块，然后通过一次前向传播计算出所有平移子窗口和模板图像块的相似性得分图，如图3.3所示。其中，X 表示尺寸较大的搜索图像块，S 表示网络输出的相似性得分图。S 中最大得分对应的位置就是目标的位置。由于节省了大量的重复计算，因此本章算法的跟踪速度比较快。

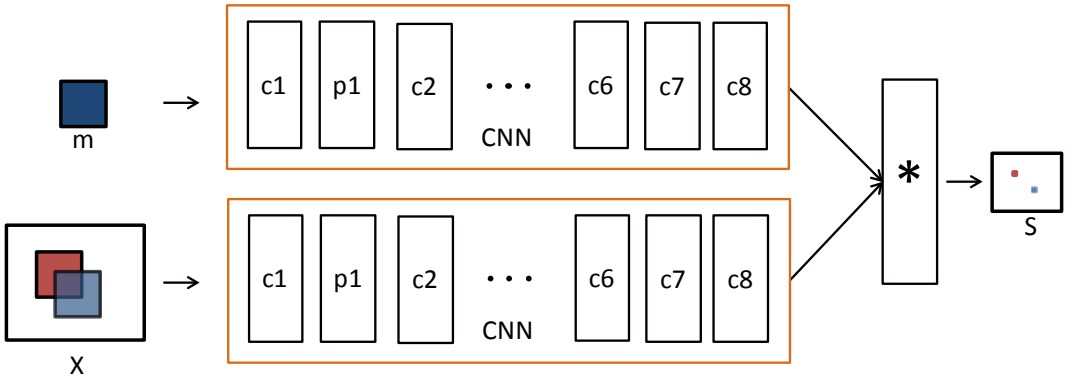


图 3.3　输入尺寸不同时的全卷积双流网络的架构

3.3.2　构建训练数据集

ILSVRC 2015 新增了 ImageNet Video 视频数据库。ImageNet Video 包括了 30 种不同的基本类别的视频序列，总共包括 4500 个视频序列，大约包含 100 多万张带标注的图像。与目标跟踪数据集 VOT，ALOV 和 OTB 相比，ImageNet Video 不仅具有视频数

量大的优点，而且包含不同的目标和场景，适用于训练深度神经网络。

为了使用 ImageNet Video 数据库训练全卷积双流网络，首先需要根据标注生成一对模板图像块和搜索图像块。本章使用的模板图像块和搜索图像块具有固定的尺寸，模板图像块的尺寸是 127×127，搜索图像块的尺寸是 225×225。模板图像块和搜索图像块既包含目标，又包含一定的背景区域。本文通过式3-2裁剪模板图像块：

$$s(w+2p) \times s(h+2p) = A \tag{3-2}$$

其中，$A = 127^2$ 表示模板图像块的面积，w 表示目标的宽度，h 表示目标的高度，$p = (w+h)/4$ 表示图像四周的填充尺寸，s 表示缩放因子。根据模板图像块的面积，目标的尺寸和填充尺寸计算图像的缩放因子 s。生成搜索图像块的方法与生成模板图像块的方法相似，此处省略。

3.3.3 离线训练全卷积双流网络

为了使用随机梯度下降法训练全卷积双流网络，首先需要定义一个损失函数：

$$l(y,s) = log(1 + exp(-ys)) \tag{3-3}$$

其中，s 表示一对模板图像块 m 和候选图像块 x 的相似性得分，$y \in \{+1,-1\}$ 是它们的标签。为了加快网络的训练速度，使用了一个较大的搜索图像块 X 代替候选图像块 x，得到一个相似性得分图 S。相似性得分图的损失函数定义为所有相似性得分的平均值：

$$L(y,s) = \frac{1}{|S|} \sum_{u \in S} l(y[u], s[u]) \tag{3-4}$$

其中，$y[u] \in \{+1,-1\}$ 表示 $s[u]$ 的标签。

全卷积双流网络的参数是通过随机梯度下降法求解下面的最优化问题得到的：

$$\arg\min_{\theta} \mathbb{E}_{(m,x,y)} L(y, f(m,x;\theta)) \tag{3-5}$$

其中，θ 表示全卷积双流网络的参数，$f(m,x;\theta)$ 表示模板图像块 m 和候选图像块 x 输入到全卷积双路网络后得到的相似性得分。

3.4 在线目标跟踪

3.4.1 基于全卷积双流网络预测目标位置

步骤 1，计算模板图像块的卷积特征图。首先按照公式3-2在第一帧图像中以目标为中心裁剪图像块，然后对它进行尺寸变换，将其尺寸变为 127×127，作为模板图像块 m。然后将模板图像块 m 输入到卷积网络模块 g 中，得到模板图像块的卷积特征图 $g(m)$, 尺寸为 6×6。由于本章使用的模板图像块没有更新，因此对一个特定的视频序列而言，模板图像的卷积特征图只需要计算一次。

步骤 2，计算第 t 帧图像中的搜索图像块的特征图。以第 $t-1$ 帧图像中的目标位置为中心，在第 t 帧图像上裁剪一个图像块，并将其尺寸变换为 255×255，作为搜索图像块 X。然后将搜索图像块 X 输入到卷积网络模块 g 中，得到搜索图像块的卷积特征图 $g(X)$, 尺寸为 22×22。每一帧图像都需要裁剪一个搜索图像块，并计算它的卷积特征图。

步骤 3，使用相关网络层计算相关得分图，并预测目标位置。将模板图像块的卷积特征图 $g(m)$ 和搜索图像块的卷积特征图 $g(X)$ 输入到相关网络层，得到一个尺寸是 17×17 的相关得分图。然后对相关得分图进行上采样操作，上采样后的相关得分图中的最大值的位置就是目标位置。

3.4.2 基于多尺度金字塔模型估计目标尺寸

通过全卷积双流网络只能预测目标的位置，不能准确地估计目标尺寸。为了提高本章算法对尺度变化的鲁棒性，本章通过在红外图像上构建一个尺度金字塔的方法，高效地计算目标的尺寸。

本章算法使用的尺度金字塔模型和2.4节中的方法完全相同，详细信息请参考2.4节。基于尺度金字塔模型估计目标尺寸的方法可以分为以下三个步骤：

步骤 4，使用红外图像构建一个尺度金字塔。以上一步得到的目标位置为中心，在红外图像上裁剪不同尺寸的图像块以构建一个尺度金字塔，并通过双线性插值法将这些图像块变换成相同的尺寸。

步骤 5，提取尺度金字塔中各个图像块的 HOG 特征向量。

步骤 6，使用尺度相关滤波器估计目标的尺寸。将上一步得到的 HOG 特征向量输入到一维尺度相关滤波器，最大相关响应对应的尺寸，就是目标的尺寸。

另外，我们在 Algorithm 2中总结了本章提出的目标跟踪算法。

Algorithm 2 基于多模态数据和全卷积双流网络的目标跟踪算法

输入: 视频序列和目标在第一帧图像中的位置和尺寸

输出: 目标在后续帧图像中的位置和尺寸

1. 计算模板图像块的卷积特征图。首先在第一帧图像中裁剪模板图像块 m，然后将其输入到卷积模块 g 中，得到卷积特征图 $g(m)$。

2. 计算第 t 帧图像中的搜索图像块的卷积特征图。首先在第 t 帧图像中裁剪搜索图像块 X，然后将其输入到卷积模块 g 中，得到卷积特征图 $g(X)$。

3. 使用相关网络层计算相关得分图，并预测第 t 帧中的目标位置。将 $g(m)$ 和 $g(X)$ 输入到相关网络层，得到相关得分图；相关得分图中的最大值对应的位置就是目标的位置。

4. 使用红外图像构建一个尺度金字塔。以上一步得到的目标位置为中心，裁剪不同尺寸的图像块，构建一个尺度金字塔，并进行尺寸归一化。

5. 提取尺度金字塔的 HOG 特征.

6. 使用尺度相关滤波器估计目标尺寸。在尺度金字塔上使用尺度相关滤波器，以相关响应的最大值对应的尺寸作为目标尺寸。

3.5 实验结果分析

本节，首先简单地介绍了实验中使用的数据集，评价方法和对比算法。然后对实验结果进行了定量和定性地评价。

3.5.1 数据集，评价指标和对比算法

由于本章算法需要同时使用可见光图像和红外图像，因此本章所有实验均在多模态数据库 OptTrack 上完成。有关 OptTrack 的更多信息请参考2.5.1。

为了评价本章算法，本章使用的评价方法包括：跟踪速度，平均中心误差，精度曲线，平均重合率和成功率曲线。有关评价指标的详细定义请参考2.5.2。

本章选择了 10 个目标跟踪算法作为对比算法，分别是 CT [58], ASLA [59], L1APG [60], LOT [61], ORIA [62], MTT [63], CSK [52], DSST [54], DLT [17] 和 siamFC_3s [35]。

3.5.2　实验结果分析

图3.4展示了所有目标跟踪算法在所有视频序列上的精度曲线和成功率曲线，可以用于评价目标跟踪算法的整体性能。从图3.4可以看出，本章提出的 FCST 算法的精度曲线和成功率曲线都明显优于其它目标跟踪算法。虽然 DSST 算法排名第二，但与 FCST 存在较大差距，原因有两点：第一，DSST 算法使用手工特征预测目标位置，而 FCST 使用表达能力更强大的深度特征预测目标位置；第二，尽管 DSST 和 FCST 都使用尺度金字塔模型估计目标尺寸，但 FCST 使用红外图像构建尺度金字塔，有效利用红外图像中目标和背景差异明显的特点。siamFC_3s 在精度曲线和成功率曲线中的排名分别是第六和第四。尽管 siamFC_3s 和本章算法 FCST 都在可见光图像上使用全卷积双流网络预测目标位置，但 siamFC_3s 仅包含 5 个卷积层，而 FCST 包含 8 个卷积层。由于 FCST 包含更多的卷积层，特征图包含更多的语义信息，因此 FCST 的精度曲线比 siamFC_3s 的精度曲线更好。另外，siamFC_3s 通过输入可见光图像的多尺度搜索图像块的方式对尺度进行估计，而 FCST 使用红外图像的尺度金字塔估计目标的尺寸。FCST 的成功率曲线比 siamFC_3s 好，证明了红外图像尺度金字塔模型在 FCST 中具有较大的作用。

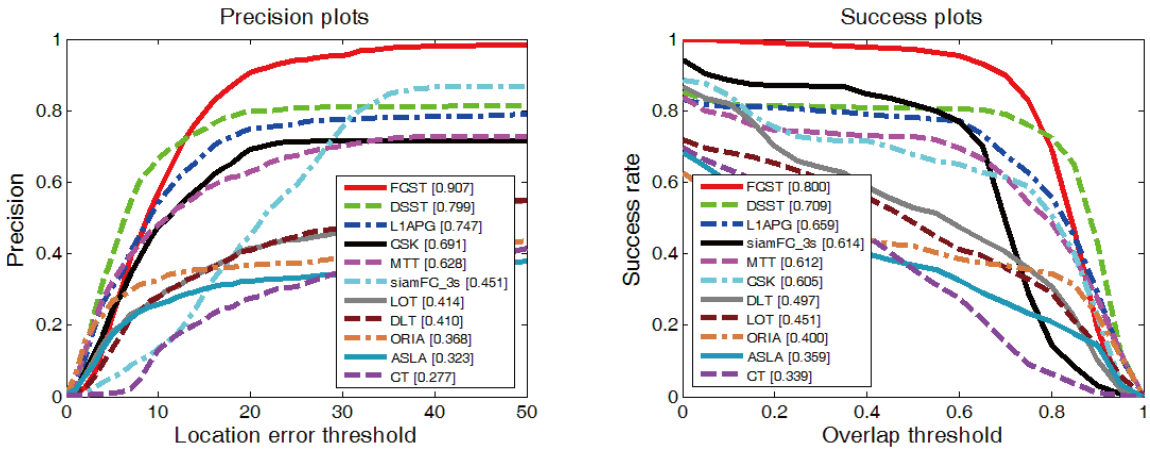


图 3.4　所有算法的整体精度曲线和成功率曲线

图3.5从不同的挑战因素方面展示了所有目标跟踪算法的精度曲线和成功率曲线，分别用于评价尺度变化，遮挡，形变和快速运动的影响。从图3.5可以看出，本章提出的 FCST 算法在所有精度曲线和成功率曲线中的排名都是第一，证明了 FCST 算法对尺度变化，遮挡，形变和快速运动均具有鲁棒性。相对而言，FCST 在遮挡方面的性能较差，这是因为 FCST 使用全卷积双流网络类似于模板匹配方法，对遮挡比较敏感。

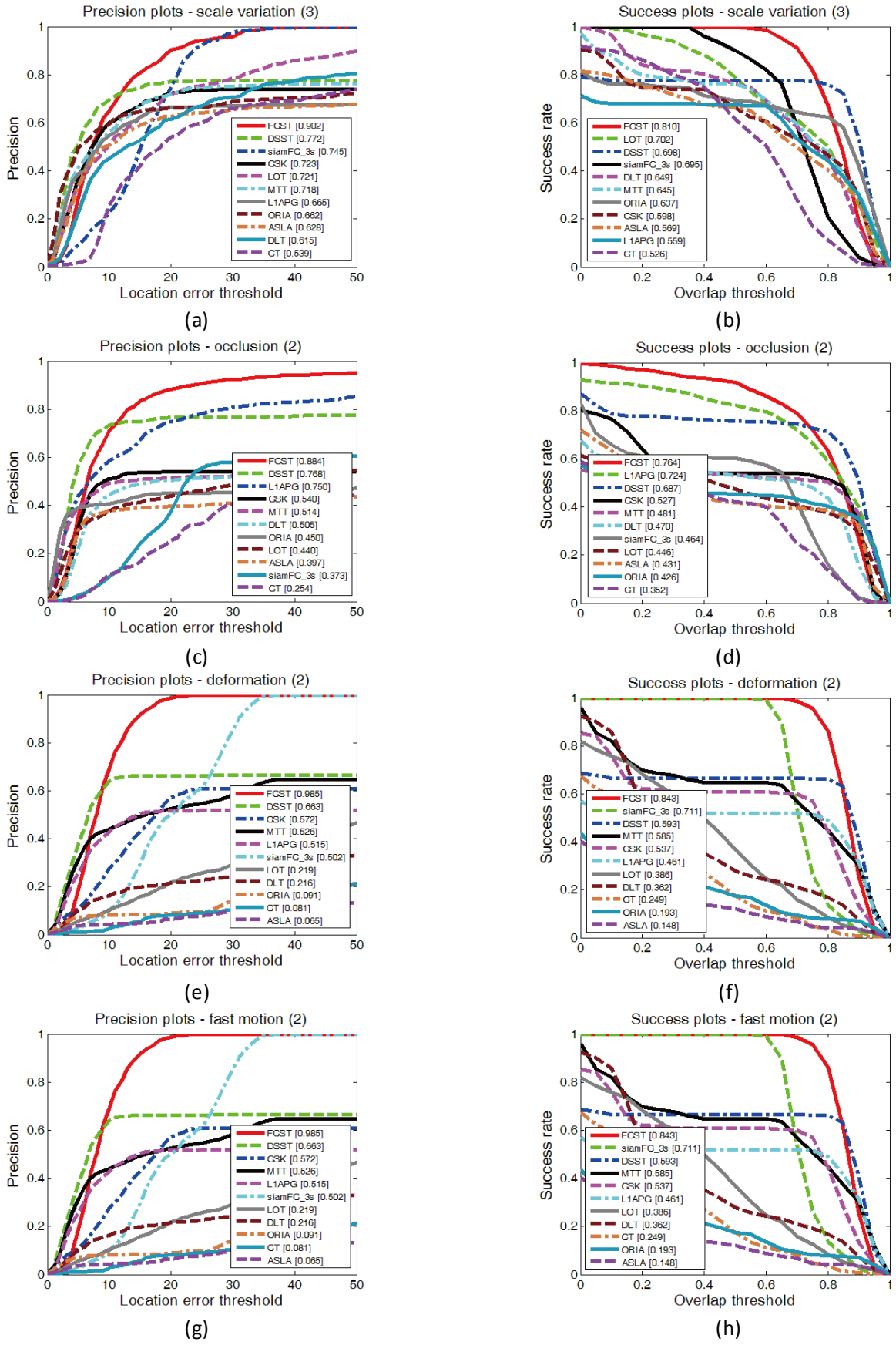


图 3.5　关于四种挑战因素的精度曲线和成功率曲线

表3.2和表3.3分别从平均中心误差和平均重合率的角度定量地对比了所有目标跟踪算法在 6 个视频序列上的性能。表3.2和表3.3分别使用红色，蓝色和绿色表示排名第一，第二和第三的算法。由于 OptTrack 中的视频序列包含了很多噪声和挑战因素，因此除了本文算法 FCST 之外，其它所有算法都出现了漂移现象，这证明了 FCST 算法具有较强的鲁棒性。FCST 算法在所有视频序列上的平均中心误差的最大值只有 14.5，具有较好的精度。FCST 的所有平均重合率的平均值为 0.82，是最大值；siamFC_3s 仅为 0.62，这证明了 FCST 使用的红外图像的尺度金字塔模型比 siamFC_3s 使用的多尺度搜索图像块的方法更加有效。

表 3.2　平均中心误差

	OptCar	OptHead	OptHeadRotate	OptHeadCard	OptRoper	OptBook	平均值
L1APG	8.31	**6.06**	**9.97**	**4.39**	222.00	**61.51**	52.04
CSK	8.66	12.71	**9.11**	**5.06**	106.39	138.52	**46.74**
MTT	9.62	11.43	11.97	5.04	83.74	444.62	94.40
LOT	8.39	177.81	207.67	9.32	**38.15**	260.50	116.97
DLT	16.50	119.89	44.94	8.34	63.13	165.51	69.72
ORIA	**4.72**	267.75	147.31	22.61	100.98	225.60	128.16
ASLA	**7.23**	228.54	131.05	22.86	172.13	176.39	123.03
DSST	**6.83**	**5.54**	**11.02**	**3.25**	202.84	**95.18**	54.11
CT	11.36	187.87	157.72	24.89	80.64	214.53	112.84
siamFC_3s	13.74	26.47	26.11	17.43	**14.49**	144.48	**40.45**
FCST	14.50	**8.34**	12.71	6.20	**8.79**	**18.60**	**11.52**

表 3.3　平均重合率

	OptCar	OptHead	OptHeadRotate	OptHeadCard	OptRoper	OptBook	平均值
L1APG	0.75	**0.90**	**0.86**	**0.92**	0.04	**0.56**	**0.67**
CSK	0.66	0.84	**0.87**	0.90	0.25	0.16	0.61
MTT	0.73	0.86	**0.83**	**0.91**	0.32	0.07	0.62
LOT	**0.80**	0.26	0.26	0.83	**0.51**	0.08	0.45
DLT	0.74	0.32	0.60	0.84	0.39	0.11	0.50
ORIA	**0.90**	0.17	0.28	0.83	0.21	0.03	0.40
ASLA	0.75	0.08	0.25	0.77	0.21	0.10	0.36
DSST	**0.89**	**0.90**	0.83	**0.93**	0.31	**0.46**	**0.72**
CT	0.60	0.18	0.23	0.68	0.31	0.03	0.34
siamFC_3s	0.63	0.70	0.72	0.75	**0.74**	0.19	0.62
FCST	0.78	**0.89**	**0.83**	0.86	**0.83**	**0.69**	**0.82**

表3.4对比了 11 种目标跟踪算法的跟踪速度。行标题表示不同的视频序列，列标题表示不同的跟踪算法。所有结果都是在没有使用 GPU 加速情形下获取的，实验的硬件环境包括，CPU 是 Intel i5-3470，内存大小为 8G。软件环境包括，Matlab 的版本是 R2014b。本章算法 FCST 使用 Matlab 语言和 MatConvNet 库实现。根据平均速度的大小，将 11 种目标跟踪算法从上到下进行排列。本章提出的 FCST 算法的平均跟踪速度是 18.90fps，排名第 4，基本满足了目标跟踪的实时性要求。第二章提出的基于多模态数据和卷积神经网络的目标跟踪算法的跟踪速度仅为 1 fps 左右，与它相比，FCST 算法的跟踪速度非常快。主要原因是：第一，FCST 采用全卷积双流网络，网络只需前向传播一次就能得到目标的位置；第二，基于多模态数据和卷积神经网络的目标跟踪算法包含 16 个卷积层，而 FCST 只包含 8 个卷积层，计算量更小。

表 3.4　各种目标跟踪算法在所有序列上的跟踪速度

	OptCar	OptHead	OptHeadRotate	OptHeadCard	OptRoper	OptBook	平均速度
CSK	126.96	119.62	162.68	220.20	209.13	185.26	170.64
CT	57.11	56.22	56.13	54.70	67.38	51.27	57.14
siamFC_3s	47.73	36.36	35.82	44.69	43.27	37.57	40.91
FCST	**23.43**	**13.42**	**12.66**	**23.99**	**21.20**	**18.69**	**18.90**
ORIA	6.96	7.77	7.60	4.79	6.81	5.45	6.56
DSST	11.80	3.42	3.18	5.85	6.32	4.56	5.86
DLT	4.08	2.00	2.43	4.20	1.63	0.95	2.55
ASLA	2.24	1.38	1.79	2.58	1.74	1.95	1.94
L1APG	2.05	1.50	1.65	1.61	1.40	0.94	1.53
MTT	0.59	0.50	0.47	0.50	0.48	0.76	0.55
LOT	0.34	0.14	0.17	0.26	0.32	0.22	0.24

图3.6展示了 11 个跟踪算法在 6 个视频序列上的实际跟踪框，从上到下依次是 OptCar，OptHead，OptHeadRotate，OptHeadCard，OptRoper 和 OptBook 六个序列。本章提出的 FCST 算法使用红色边框标记。由于 OptCar 序列相对简单，所有目标跟踪算法都没有发生漂移现象。在 OptHeadCard 序列上，少数目标跟踪算法跟踪失败。在剩余 4 个序列上，大量目标跟踪算法发生了漂移现象。FCST 算法在所有序列上都准确地预测了目标的位置和尺寸，定性的证明了 FCST 的有效性。

图 3.6　所有算法在六个视频序列上的实际跟踪结果

3.6　本章小结

本章针对基于深度神经网络的目标跟踪算法存在速度慢的问题，提出了一种基于多模态数据和全卷积双流网络的目标跟踪算法。该算法使用全卷积双流网络预测目标的位置。由于该网络前向传播一次就能输出目标位置，且网络层数适中，因此跟踪速度较快，约为 19 fps。此外，该算法在算法级别上融合了可见光图像和红外图像，利用可见光图像预测目标位置，利用红外图像估计目标尺寸。为了验证算法的有效性，我们在 OptTrack 数据库上与 10 个目标跟踪算法进行了对比实验。实验结果证明，该算法的跟踪速度比较快，且比其它算法的结果好。

第四章 总结与展望

本文主要研究了基于多模态数据和深度神经网络的视觉目标跟踪算法，提出了两种目标跟踪算法：基于多模态数据和卷积神经网络的目标跟踪算法，以及基于多模态数据和全卷积双流网络的目标跟踪算法。由于传统的目标跟踪数据库只包含可见光图像，因此我们采集了一个多模态目标跟踪数据库 OptTrack，它同时包含可见光图像和红外图像，总共包含 6 个视频序列。为了验证本文算法的有效性，在 OptTrack 数据库上与 10 个目标跟踪算法进行了对比实验，分别从定量和定性的角度证明了本文算法的有效性。本文总结如下：

1）本文提出的两种目标跟踪算法的共同特点是：在算法级别上融合了可见光图像和红外图像，有效地利用了多模态数据的互补性和多源性；借助卷积神经网络的强大的特征表达能力，增强了算法的鲁棒性；从整体上看，本文提出的两种目标跟踪算法都将目标位置和尺度分开处理，先预测目标的位置，然后再估计目标的尺寸。

2）基于多模态数据和卷积神经网络的目标跟踪算法主要解决了如何利用多模态数据和深度神经网络设计一个鲁棒的目标跟踪算法的问题。该算法使用了双融合的策略，不仅融合了可见光图像的多层卷积特征图的空间信息和语义信息，而且也在算法级别上融合了可见光图像和红外图像。

3）基于多模态数据和全卷积双流网络的目标跟踪算法主要解决了基于深度神经网络的目标跟踪算法的跟踪速度慢的问题。全卷积双流网络只需前向传播一次，就能输出目标的位置，因此该算法的跟踪速度比较快。

尽管本文提出的两种目标跟踪算法的性能较好，但仍然存在一些需要改进的地方：

1）在多模态数据融合方面，本文提出的两种目标跟踪算法采用了简单的算法级融合策略，使用可见光图像预测目标位置，使用红外图像预测目标的尺寸。为了更加有效地利用多模态数据的互补性，还需考虑更加复杂的融合策略。

2）本文提出的基于多模态数据和卷积神经网络的目标跟踪算法的网络层数多，目标跟踪效果好，但跟踪速度慢；而基于多模态数据和全卷积双流网络的目标跟踪算法的网络层数少，目标跟踪速度快，但跟踪效果不如第一种算法。因此，在兼顾跟踪效果和跟踪速度方面，仍然需要进一步研究网络架构和网络层数对目标跟踪算法的影响。

参考文献

[1] RANZATO M, HUANG F J, BOUREAU Y, et al. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2007:1–8.

[2] JARRETT K, KAVUKCUOGLU K, RANZATO M, et al. What is the Best Multi-stage Architecture for Object Recognition?[C] // Proc. IEEE International Conference on Computer Vision, 2009:2146–2153.

[3] WANG Q, FANG J, YUAN Y. Multi-cue Based Tracking[J]. Neurocomputing, 2014, 131:227–236.

[4] WU Y, LIM J, YANG M. Online Object Tracking: A Benchmark[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013:2411–2418.

[5] SEGUIN G, BOJANOWSKI P, LAJUGIE R, et al. Instance-Level Video Segmentation from Object Tracks[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:3678–3687.

[6] XIAO F, LEE Y J. Track and Segment: An Iterative Unsupervised Approach for Video Object Proposals[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:933–942.

[7] KRAFKA K, KHOSLA A, KELLNHOFER P, et al. Eye Tracking for Everyone[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:2176–2184.

[8] NING J, YANG J, JIANG S, et al. Object Tracking via Dual Linear Structured SVM and Explicit Feature Map[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:4266–4274.

[9] LU X, YUAN Y, YAN P. Robust Visual Tracking with Discriminative Sparse Learning[J]. Pattern Recognition, 2013, 46(7):1762–1771.

[10] DANELLJAN M, HÄGER G, KHAN F S, et al. Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1430–1438.

[11] ZHU G, PORIKLI F, LI H. Beyond Local Search: Tracking Objects Everywhere with Instance-Specific Proposals[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:943–951.

[12] ZHANG X, YUAN Y, LU X. Deep Object Tracking with Multi-modal Data[C] // Proc. International Conference on Computer, Information and Telecommunication Systems, 2016:1–5.

[13] NEWCOMBE R A, FOX D, SEITZ S M. DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-time[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015:343–352.

[14] MA L, LU J, FENG J, et al. Multiple Feature Fusion via Weighted Entropy for Visual Tracking[C] // Proc. IEEE International Conference on Computer Vision, 2015:3128–3136.

[15] MATHIEU M, HENAFF M, LECUN Y. Fast Training of Convolutional Networks through FFTs[C] // Proc. International Conference on Learning Representations, 2014:1–9.

[16] ZEILER M D, FERGUS R. Visualizing and Understanding Convolutional Networks[C] // Proc. European Conference on Computer Vision, 2014:818–833.

[17] WANG N, YEUNG D. Learning a Deep Compact Image Representation for Visual Tracking[C] // Proc. Annual Conference on Neural Information Processing Systems, 2013:809–817.

[18] WANG L, OUYANG W, WANG X, et al. STCT: Sequentially Training Convolutional Networks for Visual Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1373–1381.

[19] CUI Z, XIAO S, FENG J, et al. Recurrently Target-Attending Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1449–1458.

[20] ONDRUSKA P, POSNER I. Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks[C] // Proc. AAAI Conference on Artificial Intelligence, 2016:3361–3368.

[21] DEMI M. Contour Tracking with a Spatio-Temporal Intensity Moment[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2016, 38(6):1141–1154.

[22] ZHU G, PORIKLI F, LI H. Robust Visual Tracking with Deep Convolutional Neural Network Based Object Proposals on PETS[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1265–1272.

[23] GLOROT X, BENGIO Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks[C] // Proc. International Conference on Artificial Intelligence and Statistics, 2010:249–256.

[24] WANG L, PHAM N T, NG T, et al. Learning Deep Features for Multiple Object Tracking by Using a Multi-task Learning Strategy[C] // Proc. IEEE International Conference on Image Processing, 2014:838–842.

[25] SHRUTHI S. Vehicle Tracking Using Convolutional Neural Network[C] // Proceedings of the World Congress on Engineering, 2011:1–4.

[26] WANG L, LIU T, WANG G, et al. Video Tracking Using Learned Hierarchical Features[J]. IEEE Trans. Image Processing, 2015, 24(4):1424–1435.

[27] QI Y, ZHANG S, QIN L, et al. Hedged Deep Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:4303–4311.

[28] FAN J, XU W, WU Y, et al. Human Tracking Using Convolutional Neural Networks[J]. IEEE Trans. Neural Networks, 2010, 21(10):1610–1623.

[29] TAO R, GAVVES E, SMEULDERS A W M. Siamese Instance Search for Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1420–1429.

[30] LI H, LI Y, PORIKLI F M. Robust Online Visual Tracking with a Single Convolutional Neural Network[C] // Proc. Asian Conference on Computer Vision, 2014:194–209.

[31] LI H, LI Y, PORIKLI F. DeepTrack: Learning Discriminative Feature Representations by Convolutional Neural Networks for Visual Tracking[C] // Proc. British Machine Vision Conference, 2014:1834–1848.

[32] MA C, HUANG J, YANG X, et al. Hierarchical Convolutional Features for Visual Tracking[C] // Proc. IEEE International Conference on Computer Vision, 2015:3074–3082.

[33] WANG L, OUYANG W, WANG X, et al. Visual Tracking with Fully Convolutional Networks[C] // Proc. IEEE International Conference on Computer Vision, 2015:3119–3127.

[34] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C] // Proc. International Conference on Learning Representations, 2015:1–14.

[35] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-Convolutional Siamese Networks for Object Tracking[C] // Proc. European Conference on Computer Vision, 2016:850–865.

[36] HONG S, YOU T, KWAK S, et al. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network[C] // Proc. International Conference on Machine Learning, 2015:597–606.

[37] HELD D, THRUN S, SAVARESE S. Learning to Track at 100 FPS with Deep Regression Networks[C] // Proc. European Conference on Computer Vision, 2016:749–765.

[38] ZAGORUYKO S, KOMODAKIS N. Learning to Compare Image Patches via Convolutional Neural Networks[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015:4353–4361.

[39] BABENKO B, YANG M, BELONGIE S J. Robust Object Tracking with Online Multiple Instance Learning[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2011, 33(8):1619–1632.

[40] HUANG C, ALLAIN B, FRANCO J, et al. Volumetric 3D Tracking by Detection[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:3862–3870.

[41] BIBI A, ZHANG T, GHANEM B. 3D Part-Based Sparse Tracker with Automatic Synchronization and Registration[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1439–1448.

[42] GOUBET E, KATZ J, PORIKLI F. Pedestrian Tracking Using Thermal Infrared Imaging[C] // Proceedings of SPIE, 2006，6206:797–808.

[43] BORGHYS D, VERLINDE P, PERNEEL C, et al. Multilevel Data Fusion For The Detection of Targets Using Multispectral Image Sequences[J]. Optical Engineering, 1998, 37(2):477–484.

[44] YUAN Y, FANG J, WANG Q. Robust Superpixel Tracking via Depth Fusion[J]. IEEE Trans. Circuits Syst. Video Techn., 2014, 24(1):15–26.

[45] KUMAR S, MARKS T K, JONES M J. Improving Person Tracking Using an Inexpensive Thermal Infrared Sensor[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014:217–224.

[46] CHOI J, CHANG H J, JEONG J, et al. Visual Tracking Using Attention-Modulated Disintegration and Integration[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:4321–4330.

[47] LIU S, ZHANG T, CAO X, et al. Structural Correlation Filter for Robust Visual Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:4312–4320.

[48] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: Complementary Learners for Real-Time Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:1401–1409.

[49] ZHANG T, BIBI A, GHANEM B. In Defense of Sparse Tracking: Circulant Sparse Tracker[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:3880–3888.

[50] DANELLJAN M, KHAN F S, FELSBERG M, et al. Adaptive Color Attributes for Real-Time Visual Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014:1090–1097.

[51] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual Object Tracking Using Adaptive Correlation Filters[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010:2544–2550.

[52] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[C] // Proc. European Conference on Computer Vision, 2012:702–715.

[53] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2015, 37(3):583–596.

[54] DANELLJAN M, HÄGER G, KHAN F S, et al. Accurate Scale Estimation for Robust Visual Tracking[C] // Proc. British Machine Vision Conference, 2014:1–11.

[55] NAM H, HAN B. Learning Multi-domain Convolutional Neural Networks for Visual Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:4293–4302.

[56] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[C] // Proc. Annual Conference on Neural Information Processing Systems, 2012:1106–1114.

[57] CEHOVIN L, KRISTAN M, LEONARDIS A. Is My New Tracker Really Better Than Yours?[C] // Proc. IEEE Winter Conference on Applications of Computer Vision, 2014:540–547.

[58] ZHANG K, ZHANG L, YANG M. Real-Time Compressive Tracking[C] // Proc. European Conference on Computer Vision, 2012:864–877.

[59] JIA X, LU H, YANG M. Visual Tracking via Adaptive Structural Local Sparse Appearance Model[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012:1822–1829.

[60] BAO C, WU Y, LING H, et al. Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012:1830–1837.

[61] ORON S, BAR-HILLEL A, LEVI D, et al. Locally Orderless Tracking[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012:1940–1947.

[62] WU Y, SHEN B, LING H. Online Robust Image Alignment via Iterative Convex Optimization[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012:1808–1814.

[63] ZHANG T, GHANEM B, LIU S, et al. Robust Visual Tracking via Multi-task Sparse Learning[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012:2042–2049.

[64] LEAL-TAIXÉ L, CANTON-FERRER C, SCHINDLER K. Learning by Tracking: Siamese CNN for Robust Target Association[C] // Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016:418–425.

作者简介及在学期间发表的学术论文与研究成果

作者简介：

2014 年 9 月—2017 年 7 月　在中国科学院大学西安光学精密机械研究所获得硕士学位。

2009 年 9 月—2013 年 7 月　在曲阜师范大学物理工程学院获得学士学位

发表的学术论文：

[1] X. Zhang, Y. Yuan, and X. Lu,“Deep Object Tracking with Multi-modal Data,”in Proc. *International Conference on Computer, Information and Telecommunication Systems*, 2016, pp. 161-165. (EI 已检索)

奖励情况：

[1] 2016 年获 CITS 会议最佳论文奖

[2] 2016 年获中国科学院大学“三好学生”

[3] 2014 年获第四届济宁市自然科学学术创新奖论文和建议类一等奖