



华南理工大学  
South China University of Technology

# 硕士学位论文

基于生物启发模型的视觉跟踪

---

作者姓名	蔡博仑
学科专业	电子与通信工程
指导教师	毕淑娥 副教授
	徐向民 教授
	罗伟民 高级工程师
所在学院	电子与信息学院
论文提交日期	2016年4月20日

# **Object Tracking Based on Biologically Inspired Model**

A Dissertation Submitted for the Degree of Master

**Candidate: Cai Bolun**

**Supervisor: Prof. Bi Shue**

**Prof. Xu Xiangmin**

**Luo Weimin**

South China University of Technology

Guangzhou, China

分类号：TP391.4

学校代号：10561

学 号：201321009387

## 华南理工大学硕士学位论文

# 基于生物启发模型的视觉跟踪

作者姓名：蔡博仑

指导教师姓名、职称：毕淑娥 副教授、徐向民 教授

申请学位级别：工学硕士

学科专业名称：电子与通信工程

研究方向：智能视频信息处理

论文提交日期：2016年4月20日

论文答辩日期：2016年6月8日

学位授予单位：华南理工大学

学位授予日期： 年 月 日

答辩委员会成员：

主席：


殷瑞祥

委员：

林孔 姜士波 钟海波 晋建秀

# 华南理工大学 学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：

日期：2016 年 06 月 12 日

## 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华南理工大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅（除在保密期内的保密论文外）；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。本人电子文档的内容和纸质论文的内容相一致。

本学位论文属于：

保密，在 \_\_\_\_\_ 年解密后适用本授权书。

不保密，同意在校园网上发布，供校内师生和与学校有共享协议的单位浏览；同意将本人学位论文提交中国学术期刊（光盘版）电子杂志社全文出版和编入 CNKI《中国知识资源总库》，传播学位论文的全部或部分内容。

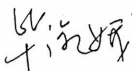
（请在以上相应方框内打“√”）

作者签名：



日期：2016.06.12

指导教师签名：



日期：2016.06.12

作者联系电话：13416186556

电子邮箱：

联系地址（含邮编）：

## 摘 要

目标跟踪是计算机视觉的核心问题，相关研究在人机交互、视频监控、自动驾驶等领域有广泛应用前景。近年来，尽管相关研究取得了众多进展，但在无约束环境中对通用目标实现可靠跟踪仍是难点，主要受到外观形变、光照变换、尺度变化、局部遮挡等问题影响。目前，研究者们从表观模型以及跟踪模型架两方面开展深入研究，试图解决上述跟踪难题。

由于多样的环境变化使得视觉跟踪仍是一个具有挑战的任务，人们提出一系列的目标跟踪算法各有优势和缺点，但仍然不能适应目标跟踪应用中的各种复杂场景。借鉴于人类视觉系统（HVS）优越的跟踪性能，设计一个基于生物启发模型的视觉跟踪将会获得更好的性能。然而受限于对于人类视觉系统的不完全理解，使其成为重要的挑战。本文通过视觉皮层中腹侧流通路的视觉认知机理，提出基于 HSV 的生物启发视觉跟踪器（BIT）。BIT 通过模拟浅层神经元（S1 单位和 C1 单位）提取生物启发表观特征；根据高层学习机制（S2 单位和 C2 单位）构建生物启发跟踪模型，结合生成模型和判别模型实现目标的位置。此外，快速 Gabor 近似（FGA）和快速傅里叶变换（FFT）的引进使得其成为一个实时学习和检测的跟踪算法。在两个基准数据集的实验表明，生物启发跟踪器在 TB50 数据库上取得 81.7% 的精度，优于现有的方法。特别地，加速技术使得生物启发跟踪算法获得了大约 45 帧每秒的跟踪速度。

此外，本文针对于生物启发表观模型（BIAM）和生物启发跟踪模型（BITM）分别提出了两个方面的改进，其中包括基于深度学习的改进与基于尺度自适应的改进。基于深度学习的改进采用深度卷积神经网络特征取代 BIAM 的 S1 和 C1 单元，实现对于跟踪目标更为鲁班的建模。基于尺度自适应的改进基于独立尺度估计方法，借鉴空间金字塔池化（SPP）提出一种快速的尺度估计特征。综合实验表明，以上两方面的改进能有效地提高跟踪器的综合性能，其中基于深度学习的 BIT 在 TB50 上取得 84.9% 预测率；基于尺度自适应的 BIT 在 ALOV300++ 上取得 0.724 的 F-score。

本文工作为物体跟踪有助于跟踪算法表观模型和跟踪模型的研究，以及推进生物视觉感知理论在机器视觉中的应用。

**关键词：** 生物启发模型；视觉跟踪；快速 Gabor 估计；深度学习特征；快速尺度特征

# Abstract

Visual tracking is an important issue in the field of computer vision, which has broader prospects for human Computer interaction, video surveillance, automatic driving. Although great progress has been made in recent years, it is still difficult to track a general object in wild environments, due to influences by object deformation, scale change, illumination change and occlusion. Researchers are trying to solve these problems from appearance model and tracking model.

A number of object tracker have been proposed, but most of them have their strengths and weaknesses, and could not handle all challenging situations. Given the superior tracking performance of human visual system (HVS), an ideal design of biologically inspired model is expected to improve computer visual tracking. This is however a difficult task due to the incomplete understanding of neurons' working mechanism in HVS. This paper aims to address this challenge based on the analysis of visual cognitive mechanism of the ventral stream in the visual cortex, which simulates shallow neurons (S1 units and C1 units) to extract low-level biologically inspired features for the target appearance and imitates an advanced learning mechanism (S2 units and C2 units) to combine generative and discriminative models for target location. In addition, fast Gabor approximation (FGA) and fast Fourier transform (FFT) are adopted for real-time learning and detection in this framework. Extensive experiments on large-scale benchmark datasets show that the proposed biologically inspired tracker (BIT) performs favorably (81.7%) against state-of-the-art methods on TB50. The acceleration technique in particular ensures that BIT maintains a speed of approximately 45 frames per second.

In addition, there are two aspects to improve bio-inspired appearance model (BIAM) and bio-inspired tracking model (BITM). First, a deep convolution feature is used to replace S1 and C1 units, and improves the tracking robustness. Second, a scale estimation method is only used to estimate bounding box size independently. On the other hand, a fast scale feature is designed based on Spatial Pyramid Pooling (SPP). Comprehensive experiments show that, two methods above can effectively improve the performance of BIT. The BIT based deep learning achieves 84.9% accuracy on TB50; the mean F-score of BIT based on scale estimation is 0.724 on ALOV300++.

This thesis exploits a novel research field for object tracking. BIT would be helpful for the research of appearance model and tracking model, and promote the employment and exploration of visual perception theory in computer vision.

**Keywords:** Biologically inspired model; visual tracking; fast Gabor approximation; fast scale feature

# 目 录

摘要 .....	I
Abstract .....	II
表格目录 .....	VII
插图目录 .....	VIII
第一章 绪论 .....	1
1.1 表观模型 (Appearance Model) .....	2
1.1.1 手工表观模型 (Handcrafted Appearance Model) .....	2
1.1.2 自动表观模型 (Automated Appearance Model) .....	4
1.2 跟踪模型 (Tracking Model) .....	5
1.2.1 生成跟踪模型 (Generative Model) .....	6
1.2.2 判别跟踪模型 (Discriminative Model) .....	6
1.3 生物启发模型 (Biologically Inspired Model) .....	7
1.3.1 灵长动物视觉皮层 .....	7
1.3.2 视觉通路的功能 .....	8
1.4 论文组织结构 .....	9
第二章 基于生物启发的跟踪器 .....	11
2.1 生物启发表观模型 (BIAM) .....	11
2.1.1 S1 单元: 初级简单细胞 .....	12
2.1.2 C1 单元: 皮层复杂细胞 .....	15
2.2 生物启发跟踪模型 (BITM) .....	16
2.2.1 S2 单元: 视觉调谐学习 .....	17
2.2.2 C2 单元: 独立任务学习 .....	18
2.3 实时生物启发跟踪器 (BIT) .....	19
2.3.1 快速 Gabor 近似 (FGA) .....	19
2.3.2 快速傅里叶变换 (FFT) .....	20
2.3.3 实时 BIT 算法总结 .....	22
2.4 本章小结 .....	22



<b>第三章 基于深度学习与尺度自适应的 BIT 改进</b>	24
3.1 基于深度学习的改进	24
3.1.1 生物启发特征与卷积神经网络	24
3.1.2 深度特征提取	25
3.1.3 基于深度学习的 BIT	26
3.2 基于尺度自适应的改进	28
3.2.1 独立尺度估计	29
3.2.2 快速尺度估计	31
3.2.3 尺度自适应的 BIT	33
3.3 本章小结	35
<b>第四章 实验分析对比</b>	36
4.1 对比实验数据库	36
4.1.1 TB50	36
4.1.2 ALOV300++	39
4.2 生物启发跟踪器对比实验	40
4.2.1 整体跟踪结果分析	40
4.2.2 详细跟踪结果分析	42
4.2.3 BIT 模型分析	48
4.2.4 实时性分析	49
4.3 基于深度学习的 BIT	49
4.3.1 TB50 对比实验	50
4.3.2 深度特征分析	51
4.4 自适应尺度的 BIT	53
4.4.1 TB50 对比实验	53
4.4.2 ALOV300++ 对比实验	54
4.5 本章小结	55
<b>第五章 总结与展望</b>	56
5.1 本文总结	56
5.2 未来展望	57
<b>参考文献</b>	59

攻读硕士学位期间取得的研究成果 .....	67
致谢 .....	68

# 表格目录

2-1	生物启发跟踪器中 S1 单元的参数选择 . . . . .	14
3-1	VGG-19 网络结构参数 . . . . .	27
4-1	TB50 跟踪数据库的视频属性分类 . . . . .	38
4-2	BIT 与其他 10 种跟踪算法在 TB50 数据库的预测精度 . . . . .	41
4-3	各类跟踪算法在 TB50 测试数据库 11 种挑战中的对比 . . . . .	42
4-4	TB50 数据库下各算法的跟踪速度与精度 . . . . .	50

# 插图目录

1-1	视觉信号处理流程 . . . . .	8
2-1	生物启发跟踪器 (BIT) 整体框架 . . . . .	11
2-2	Gabor 滤波器与生物电模型对比 . . . . .	12
2-3	多尺度 Gabor 滤波器卷积核 . . . . .	13
2-4	颜色名称 (CN) 特征空间映射图 . . . . .	14
3-1	LeNet 网络结构图 . . . . .	25
3-2	第一层卷积核可视化 . . . . .	28
3-3	VGG-19 各卷积层组的响应对比 . . . . .	28
3-4	DSST 尺度空间样本获取示例 . . . . .	30
3-5	空间金字塔池化与传统池化对比 . . . . .	32
3-6	基于 SPP 的快速尺度特征提取 . . . . .	33
4-1	TB50 跟踪数据库视频示例 . . . . .	37
4-2	ALOV300++ 跟踪数据库视频示例 . . . . .	39
4-3	BIT 与其他算法在 TB50 跟踪数据库上的预测精度图 . . . . .	42
4-4	TB50 跟踪数据库的可视化定性分析 . . . . .	43
4-5	TB50 中光照变化 (IV) 的预测曲线 . . . . .	44
4-6	TB50 中尺度变化 (SV) 的预测曲线 . . . . .	44
4-7	TB50 中局部遮挡 (OCC) 的预测曲线 . . . . .	45
4-8	TB50 中外观形变 (DEF) 的预测曲线 . . . . .	45
4-9	TB50 中运动模糊 (MB) 和快速运动 (FM) 的预测曲线 . . . . .	46
4-10	TB50 中旋转 (IPR/OPR) 的预测曲线 . . . . .	46
4-11	TB50 中超越视野 (OV) 的预测曲线 . . . . .	47
4-12	TB50 中背景干扰 (BC) 的预测曲线 . . . . .	47
4-13	TB50 中低分辨率 (LR) 的预测曲线 . . . . .	48
4-14	BIT 中 S2 单元和 C2 单元对比分析 . . . . .	49
4-15	TB50 跟踪数据库下各算法性价比 . . . . .	50
4-16	基于深度学习的 BIT 在 TB50 的预测精度对比 . . . . .	51

4-17 基于深度学习的 BIT 在低分辨率 (LR) 的预测曲线 . . . . .	52
4-18 基于深度学习的 BIT 在二维旋转 (IPR) 的预测曲线 . . . . .	52
4-19 不同 VGG 特征在 BIT 框架下的预测精度对比 . . . . .	53
4-20 BIT 与其他算法在 TB50 跟踪数据库上的成功率图 . . . . .	54
4-21 ALOV300++ 数据库下各个算法生存曲线 . . . . .	54

## 第一章 绪论

计算机视觉（Computer Vision, CV）是一门研究计算机通过成像设备来感知世界的科学，指具有计算能力的处理器（如个人电脑、移动设备、机器人等）利用成像设备（如网络摄像头、相机、视频监控等）获取的图像、视频等视觉信息，代替人眼对图像中的目标进行检测、识别和跟踪等高层次的理解。作为一个新兴学科，计算机视觉的相关理论研究和技术方案，试图建立从高维的低层图像数据中抽取低维的高层语义信息，以取代人类的视觉处理系统。因此，计算机视觉研究的首要任务包括：视觉信息的获取，视觉信息的处理分析和对视觉信息的感知。然而，计算机视觉的研究是交叉学科的研究，包含图像处理、模式识别、机器学习、甚至数据挖掘等知识，除此之外还需要借鉴生物学、心理学等学科的研究成果。

传统机器视觉研究主要关注于基于物体的研究，主要问题包括：物体检测、视觉跟踪和视觉理解。物体检测是在图像或图像序列中找到并定位给定的物体，其中需要应对尺度、光照、方向等的变化；视觉跟踪通过给定物体的初始位置，根据时空间相关性来快速精确地定位目标位置与运动轨迹；视觉理解是指建立在物体检测、视觉跟踪基础上，根据物体检测跟踪结果实现对图像内容的语义描述和更高层的理解。跟踪算法能够在视频序列中完成目标定位，获得目标的运动轨迹和形变参数，在人机交互、行为识别、自动驾驶、视频监控等领域有着广泛的应用前景。因此，本文将以视觉目标跟踪作为主要研究对象。

实际应用场景中的实时跟踪是一个有待深入研究的课题。一个理想的视觉跟踪算法<sup>[1]</sup>应具有以下特性：（1）实时性：实时性是指视觉跟踪算法能够在视频帧的采样周期内对视频中目标的位置作出精确且及时的估计；（2）鲁棒性：鲁棒性是指视觉跟踪算法可以应对目标运动过程中出现的光照、形变、遮挡等挑战；（3）稳定性：稳定性是指视觉跟踪算法具有较高的准确度和精确性以保证对连续跟踪的正确性；（3）唯一性：唯一性是指视觉跟踪算法应具有对环境中相似物体的抗干扰能力。如今，虽然已提出了各种各样的视频目标跟踪算法，但是大多数算法一般只适用于一些特定的目标、特定的环境，并且性能有待进一步优化和完善，因此在无约束环境下更加优秀的目标跟踪算法有待去进一步研究开发。

然而，由于受限于基础学科发展，纯粹的计算机理论和统计理论难以满足物体视觉跟踪的需求，无法很好地保证目标的表观模型和跟踪模型的不变性和区分性。生物

学研究表明，灵长动物具有很强的视觉模式识别功能。借鉴生物视觉系统对于视觉信息的抽象认知模型，提取不易受到客观环境影响的低层表观特征和设计多判别融合的高层跟踪模型为视觉跟踪提供了新的思路。在本文中将从生物启发模型的角度着重解决表观模型与跟踪模型的设计问题。

根据视觉跟踪算法<sup>[2]</sup>的模块分类，跟踪算法一般可分为表观模型（Appearance Model）和跟踪模型（Tracking Model）两个模块。在跟踪过程中，一般采用检测器实现对跟踪目标的初始定位，以初始目标为基础建立表观模型与跟踪模型实现对目标的跟踪。表观模型主要描述目标的视觉特征，其中包括颜色、纹理、边缘等，是实现鲁棒跟踪的先决条件。跟踪模型则是通过先验假设来估计目标的位置和状态，其中包括线性回归、Logistic 回归、支持向量机，随机森林等。

## 1.1 表观模型（Appearance Model）

表观模型是实现鲁棒跟踪的基础，优秀的表观模型有助于减轻后端跟踪模型的负担，使得算法可以更好地应对跟踪过程中出现的挑战，其中包括光照变化、尺度变化、外观形变、局部遮挡等。现有的表观模型大体可分为手工表观模型（Handcrafted Appearance Model, HAM）与自动表观模型（Automated Appearance Model, AAM）。

### 1.1.1 手工表观模型（Handcrafted Appearance Model）

手工表观模型主要依赖于特征设计者对跟踪目标的广泛观察和先验知识，通过先验知识和参数调节来设计鲁棒的手工特征。其中，手工表观模型可分为全局表观模型和局部表观模型：全局表观模型提取物体的整体统计特征，而局部表观模型融合局部特征表示估计整体统计分布。

#### 1.1.1.1 基于全局的手工表观模型

基于像素值的表观模型具有计算简单高效的优点，将原始像素值以向量或者张量形式作为特征描述，通过高级跟踪模型弥补特征本身的不足，以实现快速的视觉跟踪。但是，仅仅使用基于像素值的表观模型难以实现对视觉目标的鲁棒建模，难以应对多样化的跟踪环境。近年来，研究者尝试应用更为复杂的手工特征表观模型弥补目标表示的不足以提高跟踪的鲁棒性，其中包括纹理，颜色，轮廓，结构等。

直方图特征可以有效地获取目标区域中视觉特征的分布特性，且不受限于空间位置关系，具有对形变有较强鲁棒性。因此，Bradski 等人<sup>[3]</sup>提出的均值漂移跟踪算法

(Mean-shift Tracker, MST) 成为视觉跟踪领域的经典模型, 其采用 HSV 空间的色直方图作为目标特征, 并将其嵌入到 Mean-shift 跟踪模型中实现鲁棒的跟踪。然而, 直接色直方图容易造成空间位置信息的丢失, 导致在复杂环境中表现不佳。Laura 等人提出基于分布场特征的跟踪算法<sup>[4]</sup> (Distribution Fields Tracker, DFT), 分布场特征基于结构化直方图统计, 并通过空间和分布相关性增强直方图匹配的鲁棒性。相对经典的巴氏距离 (Bhattacharyya) 度量方法, 交叉点 (Cross-bin) 度量<sup>[5]</sup> 中采用堆土机距离 (Earth Mover Distance, EMD) 来避免巴氏距离对于光照变化容易出现的跟踪漂移现象。

基于色直方图方法受限于像素特征的单一性, 复杂的纹理特征被逐渐应用于视觉跟踪领域。梯度方向直方图<sup>[6]</sup> (Histogram of Oriented Gradients, HOG) 最早被提出于行人识别, 后逐渐应用于人脸识别、物体检测、场景识别等多种机器视觉任务。核相关滤波跟踪器<sup>[7]</sup> (Kernelized Correlation Filters, KCF) 应用 HOG 来提取局部边缘特征对跟踪目标建模, 基于局部的边缘统计信息可以应对微小的位置变化并且可以有效地抑制噪声的影响。局部二值模式<sup>[8]</sup> (Local Binary Pattern, LBP) 同样被广泛应用于视觉跟踪<sup>[9]</sup>, 因为 LBP 可以很好地解决光照的变化, 此外旋转不变的 LBP 特征还能对目标旋转鲁棒。小波变换作为近年来新提出的信号处理方法也逐渐被引进视觉跟踪领域, 其中基于 Harr 小波特征<sup>[10]</sup> 和压缩感知理论的压缩感知跟踪器<sup>[11]</sup> (Compressed Sensing Tracker, CST) 应用随机采样的方法提取 Harr 特征对跟踪目标进行建模。

### 1.1.1.2 基于局部特征的手工表观模型

全局的表观模型容易受到局部遮挡, 非刚性形变的影响, 采用局部特征的代表方法可以更好地应对以上挑战。局部特征以图像局部块 (Patch) 作为特征提取的基本单元, 融合对局部块的跟踪结果来决策判别目标的位置。其中常用的局部特征包括 SIFT<sup>[12]</sup>、SURF<sup>[13]</sup>、角点、超像素、显著性<sup>[14]</sup> 等。

局部特征的表观模型以 SIFT 特征为代表, Zhou 等人<sup>[15]</sup> 利用 SIFT 特征点匹配策略在跟踪视频中搜索目标对应的位置区域, 并通过局部色直方图度量进一步评估目标与模板间的相似性。Tang 和 Tao<sup>[16]</sup> 考虑了 SIFT 特征点之间的几何结构信息, 融合局部特征点匹配与全局拓扑结构来构建目标从局部到整体的描述。加速鲁棒特征<sup>[13]</sup> (Speed Up Robust Feature, SURF) 基于 Harr 小波变换的思想, 并利用积分图技术实现对 SIFT 的加速, 实现高效的对尺度和旋转不变的局部特征提取。Kim 等<sup>[17]</sup> 通过局部



SURF 特征的判别性与全局目标的相容性, 采用角点表观模型描述跟踪目标的外观变化。Grabner<sup>[18]</sup> 基于 Boosting 算法构建角点特征分类器, 将目标跟踪问题转换为背景和目标的二分类问题从而将目标从背景中分离出来。

基于超像素 (Super pixel) 分割方法<sup>[19,20]</sup> 将跟踪目标的原始像素映射到局部分组中, 目标的邻近区域被分割成一个或多个超像素集, 使用聚类的方法得到与每个超像素区域对应的局部模板字典, 并通过模板字典比较预测目标位置与状态。此外, Yang 等<sup>[21]</sup> 通过分类器判别性地搜索感兴趣区域 (Region Of Interest, ROI), 然后通过帧间的 ROI 匹配以获取跟踪目标的位置。然而, 基于显著性检测的跟踪器缺少目标和背景的判别机制, 容易受邻近相似物体的影响。

### 1.1.2 自动表观模型 (Automated Appearance Model)

自动表观模型主要关注如何依靠目标表观的原始数据自动学习适合于该目标跟踪的表观模型。自动表观模型可以避免手工特征的设计难题, 可以自动从目标中学习鲁棒的特征表示。然而, 需要在线更新的自动表观模型容易受到累积误差而导致跟踪漂移现象。

#### 1.1.2.1 基于子空间的自动表观模型

子空间的自动表观模型以主成分分析 (Principal Components Analysis, PCA), 线性判别分析 (Linear Discriminant Analysis, LDA) 和独立成分分析 (Independent Components Analysis, ICA) 等为代表。在视觉跟踪中, 由于目标原始表观存在高维信息的冗余, 因此可由一个低维子空间映射表示, 基于子空间的自动表观模型解决构建跟踪目标基模板的问题。子空间的自动表观模型以增量视觉跟踪<sup>[22]</sup> (Incremental Visual Tracking, IVT) 为代表。Skocaj 和 Leonardis<sup>[23]</sup> 为引入局部区域的相关性, 提出加权增量的跟踪算法, 在子空间更新中考虑跟踪目标不同区域的重要性。由于最小化重构误差有可能陷入局部最优解, Brand 等<sup>[24]</sup> 与 Levey 等<sup>[25]</sup> 提出在线奇异值分解的更新方法, 通过子空间的解析解实现更为精确的跟踪。为解决子空间样本均值更新问题, 增量 R-SVD 分解<sup>[26]</sup> 在实现子空间映射矩阵更新的同时也考虑均值的在线学习。此外, Li<sup>[27]</sup> 和 Wen 等人<sup>[28]</sup> 利用高维张量的子空间分解, 更好地保留目标的结构信息, 实现更为鲁棒的表观建模。然而, 基于子空间的表观模型需要实时更新以适应目标的变化, 无判别的子空间更新会引入跟踪误差, 不适应于长时的鲁棒跟踪。

### 1.1.2.2 基于字典表示的自动表观模型

稀疏约束由 Donoho<sup>[29]</sup> 引进机器视觉领域，越来越多基于字典表示的目标跟踪算法被提出。在目标跟踪中，根据候选目标与重构样本的误差估计目标位置，其中重构样本由字典稀疏地表示。Mei 和 Ling<sup>[30]</sup> 利用粒子滤波跟踪框架，通过重构系数的稀疏约束获得了目标样本的线性表示。多任务跟踪器<sup>[31]</sup> (Multi-task Tracker, MTT) 是一个基于多任务学习的稀疏字典优化框架，同时考虑样本相关性和稀疏系数约束。Zhong<sup>[32]</sup> 基于局部的稀疏字典避免局部遮挡的影响，并融合判别模型和生成模型取得了更为鲁棒的跟踪结果。为了提高字典表示的实时性，Li 等人<sup>[33]</sup> 基于压缩感知理论，通过求解正交匹配建立一个高效的子空间模型以实现实时跟踪。然而，基于字典表示的表观模型通过稀疏约束避免了子空间模型的累积误差问题，但字典优化和稀疏求解让基于字典的自动表观模型无法很好地满足实时性需求。

### 1.1.2.3 基于深度学习的自动表观模型

深度学习是具有多个隐层网络结构的学习方法的统称，其包括深度卷积神经网络，深度玻尔兹曼机，深度置信网等。深度学习算法广泛应用于计算机视觉领域，其中包括人脸识别，物体检测，场景分类等。视觉跟踪作为计算机视觉的传统领域，同样出现基于深度学习的表观模型的应用。Jin<sup>[34]</sup> 和 Naiyan Wang<sup>[35]</sup> 分别利用卷积神经网络 (CNN) 和层叠去噪自动编码器 (Stacked Denoising Auto-encoder, SDAE) 在目标跟踪任务中做了有效尝试与研究。首先与跟踪目标的样本局限，基于深度学习的表观模型普遍采用 ImageNet<sup>[36]</sup> 预训练深度表观模型，通过跟踪过程中对部分网络参数的动态调整来保证鲁棒的跟踪。基于深度学习的自动表观模型可以很好地刻画物体的本质属性，对于形变，光线等外界因素不敏感。然而，由于深度学习特征的高度抽象能力，使其对于同一类目标的表征过于相似，导致容易受场景中与跟踪目标相似物体的干扰。此外，计算瓶颈是深度学习表观模型无法真正实时应用的根本原因。

## 1.2 跟踪模型 (Tracking Model)

面对跟踪过程中的种种挑战，简单依赖于表观模型是远远不够的，需要依靠跟踪模型的统计建模。一般我们把跟踪模型分为两大类，一是生成模型 (Generative Model)，一是判别模型 (Discriminative Model)。生成模型无判别地在感兴趣区域内搜索并定位与跟踪模板最相似的目标位置。相反地，判别模型把目标跟踪问题当作一个监督学习

的过程，通过学习一个目标与背景的分类器，判别性地定位目标位置。

### 1.2.1 生成跟踪模型 (Generative Model)

生成模型 (generative model) 是基于后验概率的跟踪建模，根据统计分析挖掘跟踪目标的样本分布，通过似然概率定位期望最大的跟踪框位置。

经典的生成跟踪模型直接对跟踪目标的特征进行建模，如核空间跟踪算法<sup>[37]</sup>、Mean-Shift 跟踪算法<sup>[3]</sup> 等。增量视觉跟踪<sup>[22]</sup> (Incremental Visual Tracking, IVT) 以 PCA 为基础，通过在线的子空间学习和增量更新实现对于目标的鲁棒定位，但无判别的模型更新过程中不可避免地引入背景噪声。多事例学习 (Multiple Instance Learning, MIL)<sup>[38]</sup> 算法允许跟踪器利用一系列目标图像事例进行整合更新，提高跟踪器的鲁棒性。TLD<sup>[39]</sup> (Tracking Learning Detection) 算法创新性地融合跟踪器与检测器 (Tracking-by-detection)，避免在线更新中的误差累积，对于复杂背景中的目标形变具备了较强的鲁棒性。Wang<sup>[40]</sup> 基于主成分分析，通过维持一个整体性的目标外观概率模信息来表示跟踪的目标。生成跟踪模型基于目标本身进行建模，更好地关注物体的不变特征，然而其忽略了背景信息判别性地对目标的学习，因此容易受背景的干扰。

### 1.2.2 判别跟踪模型 (Discriminative Model)

判别模型 (discriminative model) 是寻找判别不同样本的分类平面，学习不同标签数据之间的差异性。其核心是把目标跟踪问题转化为分类器的学习问题，更加关注目标区域与背景图像的区分性。因此，判别跟踪模型的优化实际上是分类器选择、训练、更新问题。

为解决目标跟踪过程中的判别问题，机器学习经典的分类器方法被应用于视觉跟踪领域，其中包括支持向量机<sup>[41]</sup> (Support Vector Machine, SVM)、多视图 SVM<sup>[42]</sup> (Multi-view SVM)、结构化 SVM<sup>[41]</sup> (Structured SVM)、自适应增强<sup>[43]</sup> (Ada-boosting)、半监督增强 (Semi-boosting)<sup>[44]</sup> 等。Avidan 等<sup>[45]</sup> 基于光流估计方法，提出利用 SVM 分类器的离线训练实现物体的鲁棒跟踪。为解决目标姿态的变化和遮挡问题，基于碎片跟踪方法<sup>[46]</sup> (Fragments-based Tracking, Frag) 通过局部直方图作为跟踪目标的各个碎片特征，融合局部到整体的判别应对目标跟踪中形变、遮挡、光照等挑战。然而，静态的分块方法难以适应跟踪过程中复杂背景和外观旋转的情况。Kwon<sup>[47]</sup> 改进目标固定碎片的不足，提出一种碎片拓扑结构更新方法以处理目标姿态变化情况。Li 等人<sup>[48]</sup> 通过结合多种目标表观模型和粒子滤波框架，解决姿态变化和遮挡问题以实

现对于低帧率视频的鲁棒跟踪。

### 1.3 生物启发模型 (Biologically Inspired Model)

随着生物学领域对于生物神经研究的进展,灵长动物视觉皮层信息处理机制的研究已初见成果。建立于神经生物学和心理物理学的生物的理论基础上,生物视觉皮层的数学模型已初步形成。1855年, Panizza<sup>[49]</sup>通过生物学实验发现大脑内部存在专司与视觉信号处理的视觉皮层。二十世纪50年代, Barlow<sup>[50]</sup>通过青蛙视网膜的微电极研究,发现了视网膜神经节细胞的放电现象;60年代初, Roddick提出视觉皮层感知的数学模型,用于模拟视网膜对于外界视觉信号的感知。随后, Mcilwain<sup>[51]</sup>发现了猫视网膜细胞的感受野结构; Hubel等<sup>[52]</sup>发现大脑视皮层中对各个感受野的综合感知,不同视觉感受区对于神经激励的调谐联合处理。80年代, Mishkin和 Ungerleider<sup>[53]</sup>对猴脑视觉皮层进行深入研究,发现灵长类动物视觉皮层存在两条视觉通路,分别为负责腹侧流(形状特征提取)和背侧流(运动信息分析)。1999年, Riesenhuber和 Poggio<sup>[54]</sup>发现腹侧流和背侧流并不是简单的各司其职,两条视觉通路会在中高层的视觉皮层发生信息融合和相互作用,共同完成对物体和行为的识别。与此同时,中科院李朝义院士<sup>[55]</sup>提出整合野概念,初级视觉皮层神经元通过整合野与感受野的相互作用对图像进行高层特征提取。生物学研究的成熟和基本数学模型的提出,促使国内外各大科研机构将生物学启发模型应用于机器视觉领域。

#### 1.3.1 灵长动物视觉皮层

灵长类动物的视觉通路是一个跨学科的复杂的系统,它不仅涉及到眼球运动,而且还与下丘脑,大脑皮层等其他神经活动相关。从神经解剖学的角度来看,灵长类动物的视觉系统是由大量的有序排列、层次化排列的神经元细胞组成,其中包括眼球光学系统,视网膜,外侧膝状体(Lateral Geniculate Nucleus, LGN)和视觉皮层等。图1-1示出从视觉光刺激到神经信号的转换过程<sup>[56]</sup>。

通过的瞳孔和眼睛晶状体的调整,自然界的光信号以视觉刺激的形式投射到视网膜。在视网膜上,光信号转变为对视网膜神经节细胞的神经脉冲信号,然后发送到位于丘脑处的LGN。神经脉冲信号通过视觉LGN的初级处理,经神经元被传送到视觉皮层。视觉皮层<sup>[57]</sup>是指大脑皮层中主要负责处理视觉信息的一种典型的感觉型粒状神经皮(Koniocortex cortex),其位于大脑后部的枕叶的距状裂周围,包括初级视皮层(Primary Visual Cortex)、纹前皮层和纹外皮层。视觉皮层存在两种不同的视觉通路:

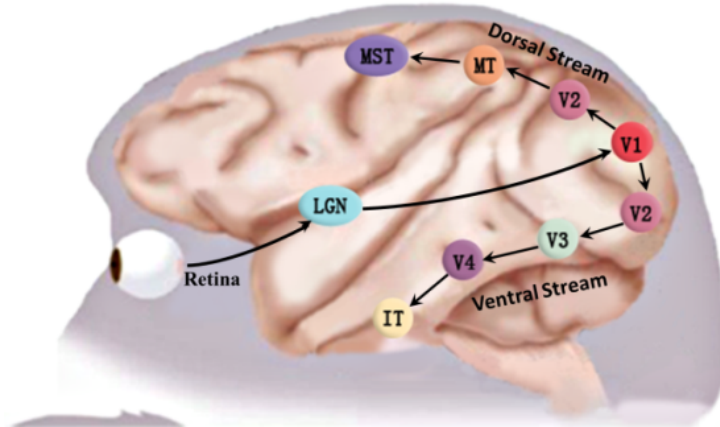


图 1-1 视觉信号处理流程

腹侧流（Ventral Stream）和背侧流（Dorsal Stream）。这两条通路均起源于初级视皮层（V1 区），亦称为纹状皮层，负责提取多尺度，多方向的局部感受野特征。初级视皮层处理结果通过未予纹前皮层的复杂视觉细胞（V2 区）进行最大池（Maximum Pool）操作，起中继区作用，负责存储视觉信息并确定信息流方向。其中一部分视觉信息通过 IT 区和 PFC 区，流入对物体空间信息敏感的腹侧流；另外一部分视觉信息通过 MT 区和 MST 流入对物体运动方向敏感的背侧流。

### 1.3.2 视觉通路的功能

生物研究表明，灵长动物视觉通路的神经细胞的层次化处理模型可以通过逐层视觉特征提取，逐步获取具有“不变性”与“区分性”的高层视觉特征。本论文旨在模拟视觉通路中腹侧流通路的信息处理过程实现对物体的鲁棒跟踪，现简要介绍腹侧流通路中各层次皮层的功能。

眼球光线系统由巩膜、角膜及其内容物等组成，其主要部分大体上像球状体。视网膜（Retina）由色素上皮层和视网膜感觉层组成，居于眼球壁的内层。色素上皮层由色素上皮细胞组成，与脉络膜紧密相连，具有支持和营养光感受器细胞、遮光、散热以及再生和修复等作用。光学信息通过眼球聚焦投影到视网膜上，形成视觉神经冲动，沿视觉通路将视信息传递到视觉中枢形成视觉，从而在大脑中建立起图像。

LGN 位于视束的后端、丘脑枕的外侧，其形如马鞍状。McAlonan 和 Cavanaugh 等<sup>[58]</sup> 通过对猕猴的视觉皮层进行研究表明 LGN 起到空间注意力调控作用。初级视皮层位于 Brodmann 17 区，也称为 V1 区，其输入主要来自位于丘脑的 LGN。生物研究表明，位于 V1 区的神经元细胞具有两个特性：一、神经元细胞仅对处于其感受野中的刺激产生响应，即单个神经元仅对与其相关的局部空间敏感；二、神经元细胞在感受野范

围内对纹理方向特征敏感，即单个神经元仅对某一频段呈现较强的响应。因此，初级视皮层的神经元细胞感受野可以被描述为具有局部性、方向性和带通特性的一系列信号编码滤波器，即提取某一频段的，某一方向的边缘、线段、条纹的卷积模板。

视皮层复杂细胞位于 V2 区，相比于简单视皮层细胞其感受野面积较大，接受来自 V1 区细胞的视觉信息，对其刺激呈非线性响应。由于复杂细胞结构复杂且各功能区域重叠分化不严格，其在感受视野内对具体位置信息的灵敏度相对简单细胞较低。同时，复杂细胞当有视觉信号经过视野区域的时间段内能对适当方向向量信号刺激产生一个持续性的响应。从视觉细胞对外界信号响应上讲，复杂细胞提供了一种不受位置局限的抽象化刺激的方向向量信息。

大脑皮层颞下（IT 区）的神经元参与视觉中的物体和人脸识别，属于纹外皮层区，实现对于视觉的高层感知，在将局部视觉信号整合到全局视觉起重要的作用。对 IT 区电生理特性的神经元属性的研究表明，其神经元表现对应视觉信息的调谐作用和分辨功能，如空间定位、立体视觉、距离估测及色调感受等。脑前额叶（PFC 区）接受和综合由脑的各部位传入的来自机体内外的各种信息，其不仅参与视觉客体的工作记忆加工过程，同时参与了视觉空间的工作记忆加工。生物研究表明，前额叶背外侧损伤的患者在视觉客体 and 视觉空间工作记忆均有明显损伤。

## 1.4 论文组织结构

本文基于生物启发模型提出一种新颖的生物启发跟踪器，并应用快速 Gabor 近似和应快速傅里叶变换，使其成为一个实用的实时算法。此外，针对 BIT 中表观模型和跟踪模型，分别提出基于深度学习和基于尺度自适应的改进。在 TB50 和 ALOV300++ 两个数据库上，相比于现有的跟踪算法，本文提出的方法均在性能上有较大的提升。本文组织架构如下：

第二章中基于 HMAX 模型提出一种新颖的生物启发跟踪器（BIT），模仿灵长动物视皮层中腹侧流通路的感知机制。其中包括生物启发表观模型与生物启发跟踪模型。此外，为了保证跟踪算法的实时性，本文提出了一种快速 Gabor 近似方法，和应用快速傅里叶变换，使 BIT 成为一个可以真正实用的实时视觉跟踪器。

第三章中以 BIT 为基础对其中尚存在的部分问题作进一步研究。基于深度学习的改进，迁移深度卷积特征优化 BIT 的表观模型；基于尺度自适应的改进，解决 BIT 跟踪模型无法自适应地调整跟踪框大小。此外，本章中基于空间金字塔池化，还提出一种

快速尺度估计的特征。

第四章中对本文提出的算法在 TB50 和 ALOV300++ 两个数据库上与现有经典跟踪算法做全面的分析对比，实验结果表明本文方法具有很好的跟踪性能，在两个数据库上均取得极具竞争性的结果。

第五章对本文的研究内容做了总结和分析，并对今后的目标跟踪的研究方向和工业化应用提出未来展望。

## 第二章 基于生物启发的跟踪器

生物启发跟踪器（Biologically inspired tracker, BIT）是基于层级最大模型<sup>[59]</sup>（Hierarchical Max Model, HMAX），模仿灵长动物视皮层中腹侧通路<sup>[60]</sup>（Ventral stream）的感知机制设计的视觉跟踪器。借鉴于 HMAX 模型在机器视觉领域中的成功应用，BIT 以 HMAX 模型为基础进行针对于视觉跟踪任务的设计和改造。如图2-1所示，BIT 主要包括两个模块：生物启发表观模型（Bio-inspired Appearance Model, BIAM）与生物启发跟踪模型（Bio-inspired Tracking Model, BITM）。其中，BIAM 主要模仿视皮层初级简单细胞（S1 单元）和视皮层复杂细胞（C1 单元），实现对物体外观特征的鲁棒提取；BITM 主要模仿高级视皮层中的视觉调谐学习（S2 单元）和独立任务学习（C2 单元），实现对物体位置的精确估计。此外，为了保证跟踪算法的实时性，本文提出了一种快速 Gabor 近似（Fast Gabor Approximation, FGA）方法优化 BIAM 模块的特征提取效率，并应用快速傅里叶变化（Fast Fourier Transform, FFT）实现 BITM 模块的算法加速，使 BIT 成为一个可以真正实用的实时视觉跟踪器。

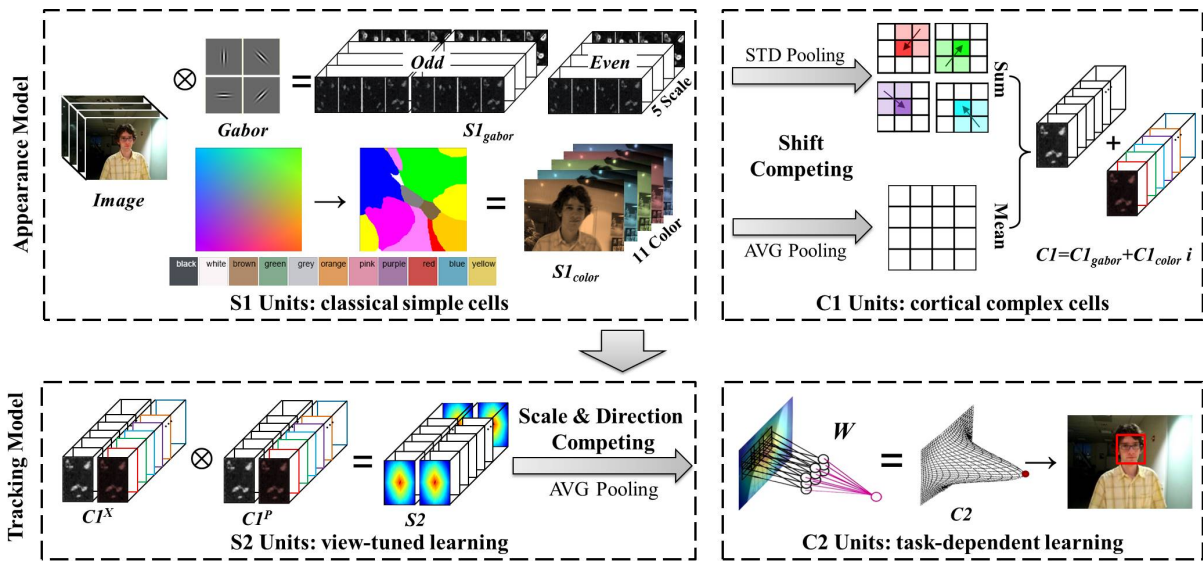


图 2-1 生物启发跟踪器（BIT）整体框架

### 2.1 生物启发表观模型（BIAM）

表观模型在视觉系统中扮演着重要的角色，因为其鲁棒性很大程度上决定了跟踪算法的抗干扰能力。在本论文中提出一种基于生物启发的表观模型（Bio-inspired Appearance Model, BIAM），实现对于跟踪物体外观的鲁棒建模。根据对视皮层纹状皮层（Striate cortex）和纹外皮层（Extrastriate cortex）的生物学研究<sup>[56]</sup>，视皮层中的初级



简单细胞（S1 单元）通过对于物体细节的刻画表现出对特征的选择性；而皮层复杂细胞（C1 单元）通过联合局部低层特征保证了模型的不变性。BIAM 模仿低层次视皮层活动，通过两层级联的结构，融合 S1 单元的细节特征提取与 C1 单元的外观特征抽象，保证表观模型在“区分性”和“不变性”上的统一，最终获得对跟踪目标的鲁棒外观建模。

### 2.1.1 S1 单元：初级简单细胞

生物研究表明，视皮层中的初级简单细胞（V1 区）表现出局部性、多尺度性和多方向性等，因此其感受野可以由一系列的尺度不一的二维 Gabor 滤波函数来逼近，并满足时频联合测不准原则。二维 Gabor 小波变换函数由 Granlund<sup>[61]</sup> 提出，并将其成功应用于图像处理领域。后续地，Knutsson<sup>[62]</sup> 和 Daugman<sup>[63]</sup> 等人从生物学角度对二维 Gabor 滤波器进行更深入的研究，结果表明灵长动物初级视觉皮层的感受野响应能够被二维 Gabor 函数很好的拟合。如图2-2 所示，从空域分布特性、方向选择特性、频域覆盖范围上，二维 Gabor 函数都能很好地模拟简单细胞感受野特性。其中，Gabor 滤波器的卷积操作对应于初级简单视觉细胞的局部敏感性，滤波器尺度对应于尺度选择性，梯度方向对应方向选择性，主瓣宽度对应于频率选择性。生物学的研究成果推动了二维 Gabor 小波分析方法在机器视觉领域的高速发展，其已被广泛应用于图像的边缘检测和人脸识别中。

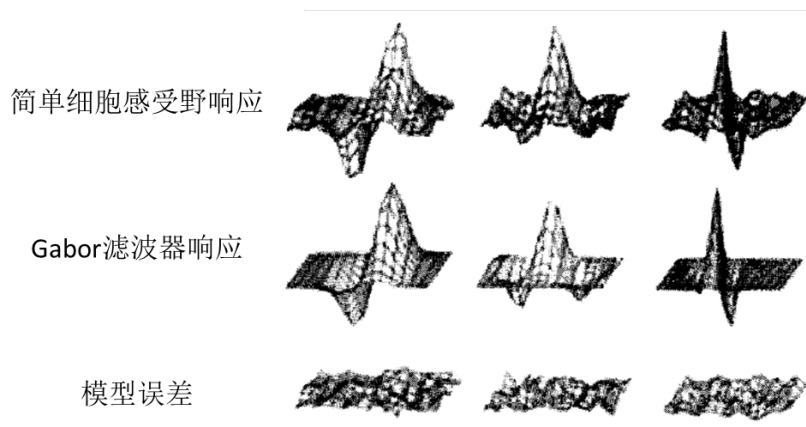


图 2-2 Gabor 滤波器与生物电模型对比

Gabor 变换<sup>[64]</sup> 属于加窗傅里叶变换，可以在不同频域、不同尺度、不同方向上提取相关的特征。因此，S1 单元可由一系列的 Gabor 滤波器进行表示，并提供近似于初级视觉细胞属性的感受野。相比于经典 HMAX 模型<sup>[59]</sup> 中的生物启发特征采用单一的偶 Gabor 滤波器，本文引入奇 Gabor 滤波器增强目标表观建模。因为，偶 Gabor 滤波器

只能提取对比度不敏感的纹理信息，而奇 Gabor 滤波器在提取纹理强度的同时也能提取纹理的梯度方向，能更好地刻画纹理外观。因此，本文中采用一系列的奇、偶 Gabor 滤波器联合对输入的原始灰度图像进行卷积得到 S1 单元的细胞响应，两类 Gabor 滤波器的表达式如下：

$$\begin{cases} G_{even}(x, y, \theta, s(\sigma, \lambda)) = \exp\left(-\frac{X^2+Y^2}{2\sigma^2}\right) \cos\left(\frac{2\pi}{\lambda}X\right) \\ G_{odd}(x, y, \theta, s(\sigma, \lambda)) = \exp\left(-\frac{X^2+Y^2}{2\sigma^2}\right) \sin\left(\frac{2\pi}{\lambda}X\right) \end{cases} \quad (2-1)$$

其中， $X = x \cos \theta + y \sin \theta$ ， $Y = -x \sin \theta + y \cos \theta$ ， $(x, y)$  为 Gabor 滤波器卷积核的坐标索引， $\theta$  为方向角度， $s$  为尺度空间大小（其中  $\delta$  对应于 Gabor 带宽， $\lambda$  为 Gabor 的波长）。根据灵长动物初级视觉细胞的感受野尺度<sup>[65]</sup>，在此选取尺寸空间从  $7 \times 7$  到  $15 \times 15$  的 Gabor 滤波器对 S1 单元的细胞感受野（Receptive Field, RF） $\xi$  进行建模，参数选择如表2-1。其中 Gabor 滤波器共在偶函数上选取 4 个方向角度（ $\theta = 0, \pi/4, \pi/2, 3\pi/4$ ），在奇函数上选取 8 个方向（ $\theta = 0, \pm\pi/4, \pm\pi/2, \pm3\pi/4, \pi$ ）。联合从  $7 \times 7$  到  $15 \times 15$  的 5 个尺度空间  $s$ ，共产生  $5 \times (4 + 8) = 60$  组 Gabor 滤波器（如图2-3）。对应的 S1 单元响应可以表示如下：

$$S1_{gabor}(x, y, \theta, s) = I(x, y) \otimes G_{even/odd}(x, y, \theta, s), \quad (2-2)$$

其中， $S1_{gabor}(x, y, \theta, s)$  是 V1 区的输出结果， $I(x, y)$  是输入的原始灰度图像。

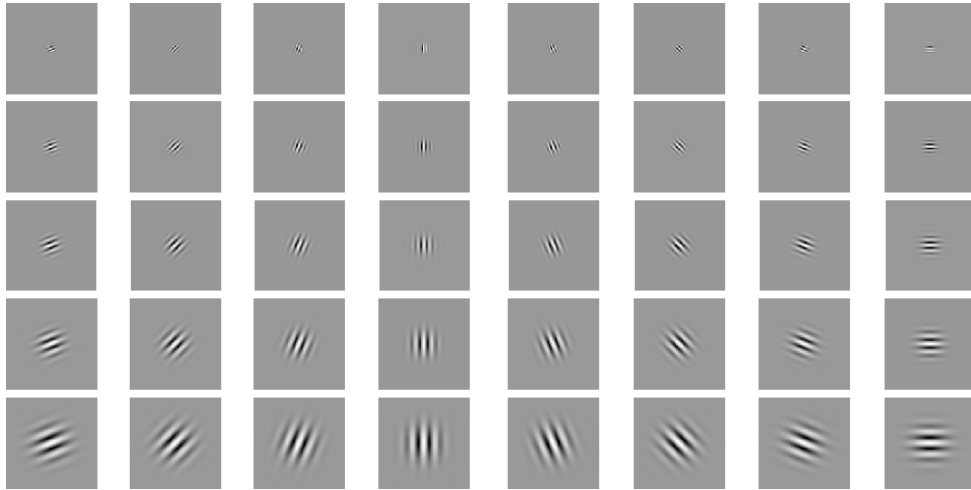


图 2-3 多尺度 Gabor 滤波器卷积核

生物启发模型已被应用于场景识别<sup>[66]</sup>和显著性检测<sup>[14]</sup>，已有的研究表明联合颜色和纹理信息的表观模型可以更为有效地对彩色目标进行建模。此外，在视觉跟踪中也证明了多不变特征<sup>[67]</sup>对于跟踪性能的提升起到了重要作用。在相邻的视觉感受野内，神经元中某一颜色（如蓝色）神经元的兴奋会对其他颜色（如黄色）神经元产生抑制作

表 2-1 生物启发跟踪器中 S1 单元的参数选择

尺度序号 $s$	感受野 $\xi$	带宽 $\sigma$	波长 $\lambda$
1	$7 \times 7$	2.8	3.5
2	$9 \times 9$	3.6	4.6
3	$11 \times 11$	4.5	5.6
4	$13 \times 13$	5.4	6.8
5	$15 \times 15$	6.3	7.9

用，这表现为视皮层中的颜色互对性<sup>[68]</sup>（Color Double-opponent）。因此，本文采用人类语言描述对 S1 单元的颜色信息进行刻画，其符合人类视觉皮层颜色单元对于彩色空间的描述。本文将基于人类语言描述训练的颜色名称<sup>[69]</sup>（Color Names, CN）特征作为物体的颜色特征描述。CN 特征通过 Google 图像搜索数据库训练得到的 CN 映射矩阵，将机器视觉中常用的 RGB 颜色空间被映射到 11 维的颜色概率空间，其中包括 11 种基本颜色：黑、棕、绿、粉、红、黄、蓝、灰、橙、紫、白（如图 2-4 所示）。CN 的概率特征可以表示如下：

$$S1_{color}(x, y, c) = Map(R(x, y), G(x, y), B(x, y), c) \quad (2-3)$$

其中  $R(x, y)$ ,  $G(x, y)$ ,  $B(x, y)$  对应 RGB 颜色空间的像素值， $c$  为 CN 特征维度索引， $Map()$  是 RGB 空间到 11 维 CN 颜色概率的映射矩阵。此外，为了保持颜色特征  $S1_{color}$  与纹理特征  $S1_{gabor}$  在单一尺度空间的维度一致性，将  $S1_{color}$  扩展为 12 维特征向量并把  $S1_{color}(x, y, 12)$  设为 0。

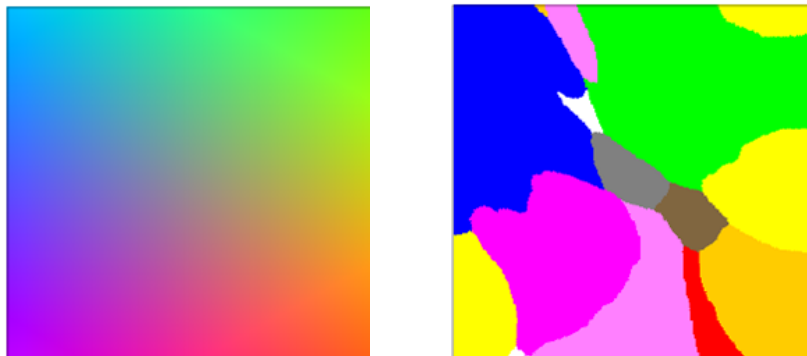


图 2-4 颜色名称（CN）特征空间映射图

本文采用复数的表示形式融合纹理特征  $S1_{color}$  和颜色特征  $S1_{gabor}$ ，跟踪目标被表示为 60 维的复数特征图：60 维的纹理实部特征包含 5 个尺度的 Gabor 滤波器输出响应；12 维的颜色虚部特征被复制 5 次以对应纹理特征的尺度空间。复数特征表示方法

可以有效的解决多视图 (Multi-view) 特征的权重平衡问题, 并能充分利用快速傅里叶变换的复频分析优势, 使其减少接近一半的计算复杂度。综上, S1 单元的复数响应可表示如下:

$$S1(x, y) = S1_{gabor}(x, y) + S1_{color}(x, y) i \quad (2-4)$$

### 2.1.2 C1 单元: 皮层复杂细胞

根据大量的解剖学和生理学知识<sup>[70]</sup>, 动物的视觉系统中有大量的视皮层区域直接的或间接的与视觉信息的处理相关, 通过神经元之间的级联处理获得各种不同的高层抽象。其中, 视皮层 V2 区的复杂视觉神经细胞的感受野范围大于 V1 区简单细胞, 接受简单神经元的响应输入, 综合局部线性特征, 起到范围检测器的作用。其表现为利用神经元间的竞争机制对于 V1 区的输出进行特征筛选。生物学研究表明, V2 区复杂视觉神经细胞的竞争机制可分为尺度竞争, 方向竞争和位置竞争:

- (1) 位置竞争: 复杂细胞具有比简单细胞更大的感受野, 复杂细胞在其感受野内接受简单细胞刺激, 刺激之间相互抑制, 形成位置竞争;
- (2) 方向竞争: 相同方向, 相同视野位置的不同尺度神经细胞之间的竞争, 旨在得到更具有尺度不变性的更高层特征;
- (3) 尺度竞争: 相同视野位置, 不同方向的神经细胞之间的竞争, 旨在保证对于旋转的鲁棒性。

神经元间的相互作用可分为兴奋性和抑制性: 当某一神经元处于兴奋的同时将对周围神经元释放抑制性递质。在视觉神经竞争过程中, 对于视觉刺激响应较弱的神经元细胞将被抑制, 其响应信息将不被传递至下后续各视皮层; 对于视觉刺激响应较强的神经元细胞将保持原有响应状态, 并将神经冲动经突触传入下一层次处理。基于生物神经元间抑制机理, 其数学模型可表示为: 对 V1 区简单视觉细胞的输出响应做数值比较, 被抑制的神经元赋值为零, 兴奋的神经元保持原有响应。因此可有效地去除冗余信息, 保留有效信息, 避免信息爆炸, 提取更具表征能力的高层的信息。

Lamp 等<sup>[71]</sup>对复杂的细胞空间整合特性的研究, 表明皮层复杂细胞可通过一系列的池化操作 (Pooling Operation) 来刻画。Riesenhuber 和 Poggio<sup>[54]</sup>认为复杂细胞有更大的感受野, 对应于图像特征表现为位置不变性, 因此相比线性的平均 (AVG) 池化, 非线性的最大 (MAX) 池化能更好地近似于 V2 区复杂细胞的竞争效应。Guo 等<sup>[72]</sup>基

于 MAX 池化提出修正的标准 (STD) 池化用于人脸年龄估计的生物启发特征提取。STD 池化避免了 MAX 池化带来的信息丢失, 更加充分地利用上层的有效信息, 在局部特征细节刻画方面明显优于 MAX 池化。此外, Dalal 和 Triggs<sup>[6]</sup> 提出采用四种不同的正则化因子用于梯度特征的归一化。为了保证 C1 单元特征对于局部偏移的鲁棒性和信息感知的有效性, 本文借鉴 STD 池化<sup>[72]</sup> 与梯度特征归一化方法<sup>[6]</sup>, 提出一种新的 C2 单元池化操作方法, 其表达式如下:

$$C1_{gabor}(x, y) = \sum_{\substack{(x,y) \in \Sigma \\ \delta_x, \delta_y \in \pm 1}} \frac{S1_{gabor}(x, y)}{N_{\delta_x, \delta_y}(x, y)} \quad (2-5)$$

$$N_{\delta_x, \delta_y}(x, y) = (S1_{gabor}^2(x, y) + S1_{gabor}^2(x + \delta_x, y + \delta_y) + S1_{gabor}^2(x + \delta_x, y) + S1_{gabor}^2(x, y + \delta_y))^{0.5} \quad (2-6)$$

其中, 池化单元的尺度  $n_s \times n_s = 4 \times 4$ 。

同样地, 由于 S1 单元的颜色特征  $S1_{color}$  基于 RGB 到 CN 的点对点 (point-to-point) 映射, 因此  $S1_{color}$  对于局部噪声比较敏感。在本文中采取平均 (AVG) 池化提取鲁棒的 C1 单元颜色特征。C1 单元的颜色特征  $C1_{color}(x, y)$  可表示为对 S1 单元的颜色特征  $S1_{color}(x, y)$  在池化单元  $\Sigma$  内 (尺度空间为  $n_s \times n_s = 4 \times 4$ ) 的子采样。C1 单元的神经元响应为在该神经元对应的 S1 单元中 16 个邻域神经元的平均激励, 其可表示为:

$$C1_{color}(x, y) = \frac{1}{n_s \times n_s} \sum_{(x,y) \in \Sigma} S1_{color}(x, y) \quad (2-7)$$

## 2.2 生物启发跟踪模型 (BITM)

跟踪模型用来预测跟踪目标的位置与所处的状态, 其可分为生成模型 (Generative modle) 和判别模型 (Discriminative modle)。基于高级视皮层的学习机制, 本文将颞下<sup>[73]</sup> (Infero-Temporal, IT) 皮层和前额叶皮层<sup>[74]</sup> (Pre-Frontal cortex, PFC) 的响应机制应用于生物启发跟踪模型 (BITM) 的设计。在本文中, 提出基于高级视皮层学习机制的跟踪模型, 其中包含生成模型和判别模型, 分别对应于视觉调谐学习<sup>[56]</sup> (View-tuned learning) 的 S2 单元与独立任务学习<sup>[56]</sup> (Task-dependent learning) 的 C2 单元。生成模型和判别模型的融合有效地提高了对跟踪目标的定位的精度, 并有效地避免了周围相似物体的干扰, 实现有效的鲁棒跟踪。

### 2.2.1 S2 单元：视觉调谐学习

灵长动物需要对外界环境的物体具有强大的感知能力<sup>[56]</sup>，因此要求其视觉系统的神经感应机制具有极强的组合综合能力。灵长动物的高级视皮层可以把当前与过去经过视皮层处理过的颜色、形状、纹理等各种不同的视觉信息进行高层整合，形成更为完备的视觉感知。在一定范围内，高级视觉皮层按照其所在的信息流的性质，对低层视皮层各个区域神经元间同种类型的视觉激励信号进行调控。在腹侧流视觉通路中，神经激励信号从复杂视皮层（V2 区）传输到颞下皮层<sup>[73]</sup>（IT 区），视皮层神经元表现出视觉感知的调谐特性，对目标识别发挥了关键作用。视觉调谐特性的学习过程可以被视为一个生成模型，S2 单位在其感受野内接收与综合 C1 单位的神经激励。其在视觉皮层的作用是并行地提高视觉感知的有效性，将记忆的特征模板和当前的 C1 特征做匹配操作，得到一个代表 C1 单元相似度的高层激励响应。

有别于经典 HMAX 模型中将随机选取的特征块（Patch）与整个 C1 特征图做稠密特征匹配。为了减少不必要的匹配计算与更多地关注跟踪目标本身，本文中的 S2 单元采用将 C1 单元中提取到整个跟踪目标的特征与目标邻近的 ROI 区域做模板匹配，从而很大程度上地减少了需要匹配的计算空间。每个 S2 单元响应通过计算 C1 单元记忆池中的存储模型  $P$  和新输入  $X$  欧式距离之间的径向基函数<sup>[65]</sup>（Radial Basis Function, RBF）。对于每个跟踪目标的 C1 特征，匹配响应  $r_{S2}$  可表示如下：

$$r_{S2} = \exp(-\beta \|X - P\|^2) \quad (2-8)$$

其中， $\beta$  是视觉调谐系数。根据上式，S2 的响应图通过计算 C2 单元中记忆模板与当前各个位置的匹配响应值。

根据式(2-8)，S2 单元的响应对应于 RBF 核方法，因此当 RBF 是一个标准正态函数（ $\beta = 1/2\sigma^2 = 1/2$ ）时，其可以近似地用线性函数表示：

$$\begin{aligned} r_{S2} &= \exp\left(-\frac{1}{2\sigma^2} \|X - P\|^2\right) \\ &= \exp\left(-\frac{1}{2} (X^T X + P^T P - 2X^T P)\right) \\ &\sim \exp(X^T P) \sim X^T P \end{aligned} \quad (2-9)$$

其中， $X^T X$  和  $P^T P$  作为自相关操作基本保持不变且接近于常数， $X^T P$  作为互相关操作基本落于 RBF 函数的线性区。此外，由于线性核函数<sup>[7]</sup>可以减少大量的运算量，被广泛应用于实时性要求高的应用中。因此，本文采用简单的线性核取代 RBF 核来计算 S2 单元的稠密响应图。在 C1 单元，对于新输入  $X$  的响应表示为  $C1^X(x, y, k)$ ，存储模

板  $P$  的响应为  $C1^P(x, y, k)$ ，其中  $k$  对应于 60 个特征图的索引（包括 12 个方向和 5 个尺度）。为了获得尺度和旋转的不变性，本文还采用 AVG 池化操作对多个特征图间进行池化，实现神经元的尺度竞争与方向竞争。综上所述，S2 单元对应的响应可表示为：

$$S2(x, y) = \frac{1}{K} \sum_{k=1}^K C1^X(x, y, k) \otimes C1^P(x, y, k) \quad (2-10)$$

## 2.2.2 C2 单元：独立任务学习

视觉神经信息处理是通过基于前馈/反馈和皮层内交错的水平连接实现神经递质多层次的复杂相互作用。针对视觉跟踪这种特定任务，神经激励沿着神经环路从 IT 视皮层传到前额叶<sup>[74]</sup>（PFC 区），用于学习目标对象和背景集群之间的差异性。根据生物学研究表明<sup>[56]</sup>，PFC 区是视皮层中与视觉任务相关的、带有反馈学习的视觉感知区域。因此，神经科学认为 PFC 区可以作为一个针对独立任务，以监督学习方式训练的分类器，接收并反馈底层神经元的响应。PFC 区的分类机制通过 100 万个神经突触接收 IT 区的神经激励，在极短的时间反馈出对象身份和信息（如目标的位置和大小）。监督学习在这个阶段需要调整突触权重并最小化训练集的误差来对特定任务进行拟合。因此，本文使用单层的卷积神经网络<sup>[75]</sup>（Convolutional Neural Networks, CNN）对应于 PFC 区中独立任务学习（C2 单位）的神经活动。

卷积神经网络由多层感知器<sup>[76]</sup>发展而来，被越来越广泛地应用于信号处理领域。多层感知器的网络结构忽略了图像数据的空间结构信息，把图像的输入像素看作是孤立的响应点，未考虑邻近像素之间的上下文信息。受生物视觉中局部感受野的启发，CNN 将多层感知器中的网络结构改进为卷积网络结构。相较于多层感知器，CNN 具有以下三个优点：神经元间连接、局部感受野、权值共享。这些属性使得 CNN 在图像处理上取得了比多层感知器更为优秀的表现。本文采用的卷积神经网络是一个单层的网络结构，每一个下层神经元与上一层的二维平面中固定区域的突触相关联，这对应于生物神经学中的感受野的思想，可表示为：

$$C2(x, y) = W(x, y) \otimes S2(x, y) \quad (2-11)$$

其中， $W$  是神经网络中突触的权值。下一小节中将详细介绍具体方法实现  $W$  的快速估算。

## 2.3 实时生物启发跟踪器 (BIT)

实时性是视觉跟踪的重要指标。实时性是指视觉跟踪算法能够在视频帧的采样周期内完成对视频中目标的位置估计, 并做出及时的响应。现有的很多跟踪算法过度地追求跟踪的精度而忽略了对实时性的约束。随着智能视频监控的发展, 多目标跟踪的需求逐步增加, 这也为目标跟踪的效率提出了更高的要求。以往基于生物启发模型的视觉跟踪器由于复杂的神经元模拟机制都无法真正提供实时有效的视觉跟踪, 如简单生物启发特征<sup>[77]</sup> (Simple Biologically Inspired Feature, SBIF) 跟踪方法, 显著性判别跟踪器<sup>[78]</sup> (Disciminant Saliency Tracker, DST) 等。在本文中, 快速 Gabor 近似 (Fast Gabor Approximation, FGA) 和快速傅里叶变化 (Fast Fourier Transform, FFT) 被应用于层级生物启发模型的加速。

### 2.3.1 快速 Gabor 近似 (FGA)

在节2.1.2中, 为了获得多样鲁棒的生物启发特征, 在 S1 单元中应用了多达 60 组的多尺度、多方向、多频率 Gabor 滤波器对灰度图像进行卷积。因此, 60 次的卷积操作成为了 BIAM 中主要的计算瓶颈, 严重地影响了生物启发跟踪器的实时性能。受到梯度直方图<sup>[6]</sup> (Histogram of Oriented Gradient, HOG) 的启发, 本文提出了一种快速 Gabor 估计 (Fast Gabor Approximation, FGA) 的方法, 其将极大地减少了卷积操作的计算量。

HOG 特征<sup>[6]</sup> 是一种在计算机视觉和图像处理中重要的纹理描述子, 被广泛应用于物体检测、目标跟踪、人脸识别等机器视觉领域。在图像中, HOG 通过边缘梯度和方向信息的密度分布来刻画局部目标的表观和形状。首先, HOG 将图像切割为细胞单元 (Cell), 在各个 Cell 统计各个方向梯度的直方图, 并把这些直方图组合起来构成特征描述子。与 S1 单元的局部感受野相似, HOG 也是在图像的局部区域上操作以保持几何和光照不变性, 因为这几何和光照具有局部一致性, 其变化只会出现在更大的感受野上。其次, 在对大感受野的直方图抽样及局部归一化等作用下, HOG 是具备更鲁棒和区分性的描述特征。其核心思想是, 采用两个正交的梯度卷积去估计各个方向纹理的梯度强度。

借鉴于 HOG 的核心思想, 本文采用 5 组正交的一维 Gabor 滤波器  $G_x(x, s(\sigma, \lambda))$  和  $G_y(x, s(\sigma, \lambda))$  对原始灰度做卷积操作。通过正交的一维的 Gabor 滤波器联合估计二维中各个方向的 Gabor 方向和强度, 因此可以极大的减少卷积操作的计算量。根据



表2-1的参数，5组共10个正交的Gabor响应图可表示如下：

$$\begin{cases} D_x(x, y, s(\sigma, \lambda)) = I(x, y) \otimes G_x(x, s(\sigma, \lambda)) \\ D_y(x, y, s(\sigma, \lambda)) = I(x, y) \otimes G_y(y, s(\sigma, \lambda)) \end{cases} \quad (2-12)$$

$$\text{where } \begin{cases} G_x(x, s(\sigma, \lambda)) = \exp\left(-\frac{x^2}{2\sigma^2}\right) \times \sin\left(\frac{2\pi}{\lambda}x\right) \\ G_y(y, s(\sigma, \lambda)) = \exp\left(-\frac{y^2}{2\sigma^2}\right) \times \sin\left(\frac{2\pi}{\lambda}y\right) \end{cases}$$

S1 单元的响应由式2-12中一维 Gabor 滤波器的梯度进行估算。令  $\Theta(x, y, s(\sigma, \lambda))$  和  $A(x, y, s(\sigma, \lambda))$  分别表示一维 Gabor 滤波器在  $(x, y)$  像素位置的梯度方向和强度，如下：

$$\begin{cases} \Theta(x, y, s(\sigma, \lambda)) = \tan^{-1}\left(\frac{D_y(x, y, s(\sigma, \lambda))}{D_x(x, y, s(\sigma, \lambda))}\right) \\ A(x, y, s(\sigma, \lambda)) = \sqrt{D_x^2(x, y, s) + D_y^2(x, y, s)} \end{cases} \quad (2-13)$$

我们假设 S1 单元响应为像素水平 (Pixel-level) 是稀疏的，因此可以通过一维 Gabor 响应的区域稀疏统计去估计多方向的 Gabor 梯度强度。当  $\Theta(x, y, s(\sigma, \lambda))$  属于对应 Gabor 梯度方向所在区间内，强度响应  $A(x, y, \theta, s(\sigma, \lambda))$  被定义为该方向上的 S1 单元响应：

$$S1_{\text{odd}}(\cdot) = \begin{cases} A(\cdot), \text{ if } \Theta(\cdot) \in [\theta - \pi/8, \theta + \pi/8) \\ 0, \text{ otherwise} \end{cases} \quad (2-14)$$

$$S1_{\text{even}}(\cdot) = \begin{cases} A(\cdot), \text{ if } \Theta(\cdot) \in [\theta - \pi/8, \theta + \pi/8) \cup \\ \quad [\theta + 7\pi/8, \theta + 9\pi/8) \\ 0, \text{ otherwise} \end{cases} \quad (2-15)$$

通过以上 FGA 算法，在保证极少计算量的同时有效地估算二维 Gabor 滤波器的响应，使得 S1 单元的计算基本满足实时需求。

### 2.3.2 快速傅里叶变换 (FFT)

随着机器学习中判别模型的发展，越来越多基于检测的跟踪方法 (Tracking by detection) 被提出。然而，基本所有基于检测的跟踪方法<sup>[32,39,79]</sup>都采用稀疏采样的策略：在每一帧中，根据粒子滤波算法在上一帧结果附近随机提取一定数量的样本，并通过概率判别找到最相似的目标。基于稀疏采样的方法可以有效的较少搜索空间的，但只能近似的逼近函数鞍点，并不能真正找到目标的最优解。在本文中，在 S2 和 C2 单元采用密集采样的方法，稠密地计算目标与 ROI 区域的判别距离，所以能更精确地确定目标位置。因此，加速 S2 和 C2 单元的稠密采样计算是 BIT 实时跟踪的重要问题。

根据卷积定理<sup>[80]</sup>，卷积核与原图像在二维空域中的卷积可由求其相应的傅里叶变换乘积的反变换而得。反之，在频域中的卷积可用在空域中乘积的傅里叶变换而得。按照卷积的定义计算，对于一个长度为  $N$  的序列，需要做  $(2N-1)$  组对位乘法和加法，其计算复杂度为  $O(N \times N)$ ；而利用傅里叶变换将序列从空域变换到频域上后，只需要一组对位乘法，其技术复杂度为  $O(N)$ 。因此，利用频域分析方法可以有效地简少卷积的运算量，本文将其应用于 S2 和 C2 单元的稠密采样。

在  $t$  时刻，S2 单元的稠密响应图的卷积操作重写如下：

$$S2_{t+1}(x, y) = \frac{1}{K} \sum_{k=1}^K C1_{t+1}^X(x, y, k) \otimes C1_t^P(x, y, k) \quad (2-16)$$

根据卷积定理可以将其转换为频域点乘的操作，表示如下：

$$\mathcal{F}[S2_{t+1}(\cdot)] = \frac{1}{K} \sum_{k=1}^K \mathcal{F}[C1_{t+1}^X(\cdot, k)] \odot \mathcal{F}[C1_t^P(\cdot, k)], \quad (2-17)$$

其中， $\mathcal{F}[\cdot]$  表示频谱傅里叶变换，表示元素间的点乘操作。频域分析方法的引进使得主要计算瓶颈成为分析域转换的计算量。因此，我们利用快速傅里叶变换<sup>[81]</sup>（FFT），可将整体计算复杂度降低为  $O(N \times \log N)$ 。FFT 是离散频域变换的加速计算方法的统称。1965 年，FFT<sup>[82]</sup> 的提出极大地减少了离散傅里叶变换的运算量。FFT 的基本思想是利用短序列离散傅里叶变换时指数因子周期性来避免不必要的重复计算。

与 S2 单元加速方法相仿，C2 单元的卷积和解卷操作同样可以由频谱分析方法快速实现。首先，定义 C2 单元的独立任务学习输出为类高斯（Gaussian-like）传输函数，作为 C2 单元的监督信号。通过监督信号的反向传导调整得到 C2 单元各个神经元的权值。C2 单元的监督信号表示如下：

$$\widetilde{C2}(x, y) = \exp\left(-\frac{1}{2\sigma_s^2}((x - x_o)^2 + (y - y_o)^2)\right), \quad (2-18)$$

其中， $\sigma_s$  是尺度参数， $(x_o, y_o)$  是跟踪目标的中心位置。同样根据卷积定理，神经元权值  $W$  在频域可表示为频域 C2 单元的监督信号与 S2 单元的响应点除。

$$\mathcal{F}[W(x, y)] = \frac{\mathcal{F}[\widetilde{C2}(x, y)]}{\mathcal{F}[S2(x, y)]} \quad (2-19)$$

在目标定位阶段过程表示形式与此相反，在  $t+1$  帧中，目标的位置  $(\hat{x}, \hat{y})$  定义为 C2 响应图的最大值所在的位置：

$$(\hat{x}, \hat{y}) = \arg \max_{(x, y)} C2_{t+1}(x, y), \quad (2-20)$$

其中， $C2_{t+1}(x, y) = \mathcal{F}^{-1}[\mathcal{F}[W_t(x, y)] \odot \mathcal{F}[S2_{t+1}(x, y)]]$ ， $\mathcal{F}^{-1}[\cdot]$  是反傅里叶变换。

为适应跟踪过程中目标的变化，本文应用最简单的更新方法对模型进行更新。在

第  $t$  帧，BIT 模型通过以下公式更新：

$$\begin{cases} C1_{t+1}^P(x, y, k) = \rho C1(\hat{x}, \hat{y}, k) + (1 - \rho) C1_t^P(x, y, k) \\ \mathcal{F}[W_{t+1}(x, y)] = \rho \mathcal{F}[W(\hat{x}, \hat{y})] + (1 - \rho) \mathcal{F}[W_t(x, y)] \end{cases} \quad (2-21)$$

其中， $\rho$  模型的学习率， $C1(\hat{x}, \hat{y}, k)$  是 C1 单元的空域模型， $\mathcal{F}[W(\hat{x}, \hat{y})]$  是神经元权值的频域模型。

### 2.3.3 实时 BIT 算法总结

综上所述，联合 FGA 和 FFT 两项加速方法，使得 BIT 成为真正实用的实时跟踪器。其中，在 S1 和 C1 单元，FGA 实现快速的特征提取，减少表观模型的计算量；在 S2 和 C2 单元，FFT 通过频谱分析方法将卷积/解卷操作转换为点乘/点除操作，减少跟踪模型的时间复杂度。本小节总结实时 BIT 的实现流程，如算法 2-1。

## 2.4 本章小结

本章提出了一个新的视觉跟踪框架 BIT，其模拟了灵长动物初级视皮层的神经活动。BIT 是一个四层的跟踪器，其中包括对于初级视皮层中的简单细胞模拟的 S1 单元，对应于皮层复杂细胞的 C1 单元；模拟 IT 区视觉调谐学习的 S2 单元和 PFC 区独立任务学习的 C2 单元。通过 S1 和 C1 单元的特征提取，构建生物启发表观模型，再经过 S2 和 C2 单元的学习得到生物启发的跟踪模型，融合生成模型和判别模型的优点，实现对于目标的精确跟踪。此外，本章还提出了通过 FGA 和 FFT 对于生物启发跟踪器的加速方案，使得其成为真正实用的跟踪方法。

**算法 2-1:** 实时生物启发跟踪器**Input:** 灰度图像  $I(x, y)$ , 彩色图像  $RGB(x, y)$ **Output:** 跟踪结果  $(\hat{x}, \hat{y})$ 设  $\widetilde{C2} = \mathcal{F} \left( \exp \left( -\frac{1}{2\sigma_s^2} ((x - x_o)^2 + (y - y_o)^2) \right) \right)$ 设  $C1^P(x, y, k) = C1(x, y, k)$  when  $t = 1$ **for**  $t = 1, 2, \dots$  **do****生物启发表观模型 (BIAM)**

S1 单元:

通过  $I(x, y)$  根据(2-14)和(2-15)计算  $S1_{gabor}$ 通过  $RGB(x, y)$  根据(2-3)计算  $S1_{color}$ 

C1 单元:

对  $S1_{gabor}$  根据(2-5) STD 池化得  $C1_{gabor}(x, y, k)$ 对  $S1_{color}$  根据(2-7) AVG 池化得  $C1_{color}(x, y, k)$  $C1(x, y, k) = C1_{gabor}(x, y, k) + C1_{color}(x, y, k)$ **生物启发表观模型 (BITM)**

S2 单元:

$$\mathcal{F}[S2(x, y)] = \frac{1}{K} \sum_{k=1}^K \mathcal{F}[C1(x, y, k)] \mathcal{F}[C1^P(x, y, k)]$$

C2 单元:

$$C2(x, y) = \mathcal{F}^{-1}[\mathcal{F}[W(x, y)] \odot \mathcal{F}[S2(x, y)]]$$

定位目标:  $(\hat{x}, \hat{y}) = \arg \max_{(x, y)} C2(x, y)$ 

模型更新:

$$C1^P(x, y, k) = \rho C1(\hat{x}, \hat{y}, k) + (1 - \rho) C1^P(x, y, k)$$

$$\mathcal{F}[W(x, y)] = \rho \mathcal{F}[W(\hat{x}, \hat{y})] + (1 - \rho) \mathcal{F}[W(x, y)]$$

**end for**

## 第三章 基于深度学习与尺度自适应的 BIT 改进

第2章中重点介绍了本文核心提出的生物启发跟踪器 (BIT)，本章将以 BIT 为基础对其中尚存在的部分问题作进一步研究。其中主要包括深度学习特征的应用与精确的尺度估计。深度学习特征是针对 BIT 表观模型的改进，迁移图像分类任务的卷积特征取代 BIAM，实现更加鲁棒的物体跟踪；尺度自适应估计是针对 BIT 跟踪模型的改进，解决 BITM 仅仅对于目标位置进行估计，无法自适应地调整跟踪框大小。

### 3.1 基于深度学习的改进

第2章中提出的生物启发跟踪器主要基于多尺度 Gabor 滤波器和 CN 颜色特征，其作为类手工特征提取方法具有速度快，实现简单的优点。同时，深度学习和卷积神经网络的兴起，为生物启发跟踪器中特征提取提供了新的思路和解决方案。相比起手工的生物启发特征，深度学习可以更为有效得融合纹理信息与颜色信息，在同一框架下实现对不变性特征的提取，很好地避免了不同类型的特征融合问题。本节将介绍如何利用深度学习特征取代生物启发特征，实现对物体的鲁棒跟踪。

#### 3.1.1 生物启发特征与卷积神经网络

随着大数据和人工智能的迅速发展，人工神经网络<sup>[83]</sup> (Artificial Neural Network, ANN) 作为机器学习中的重要分支，其重要性日益突出。ANN 作为仿生的算法，模拟大脑的处理机制，是目前最接近人脑的人工智能解决方案。ANN 这种模拟是具有生物学依据的。1977 年，Torsten Wiesel 和 David Hubel<sup>[84]</sup> 发现视皮层的分层结构，灵长动物的视皮层包含了不同类型的视觉神经元，其中视网膜细胞接收外界的光学刺激对应于系统输入，神经元之间通过神经递质传递神经冲动对应于权值连接，通过多层神经元的融合处理得到高层的语义表示对应于系统输出，这证明了 ANN 与视皮层感知的原理的一致性。1996 年，Bruno Olshausen 和 David Field<sup>[85]</sup> 从大量的黑白风景照中学习到了 400 个  $16 \times 16$  的基本碎片，这些基本碎片表现出类 Gabor 的特性，与生物学理论很好地对应。通过迭代学习和最小化残差，不同的物体图像均可以通过基本碎片的线性组合来表示。1980 年，Fukushima<sup>[86]</sup> 基于视觉感受野的概念，在机器学习领域提出卷积神经网络的概念。

深度学习是传统浅层模型的扩展。计算机和信息处理技术的发展推动了一系列机器学习模型，其中应用最广的包括支持向量机 (SVM)<sup>[41]</sup> 和 Logistic 回归<sup>[87]</sup> 等。浅层

模型的训练可以利用反向传播算法<sup>[88]</sup>（Back Propagation, BP）计算梯度，再利用下层的参数梯度可由上层模型的残差导出的规律在参数空间中寻找最优解。然而，浅层模型有限参数的局限性，对非线性问题的表示能力不足，针对复杂分类问题其泛化能力差，深度学习通过深度的非线性拟合能力克服浅层模型的这一弱点。

LeCun 等人<sup>[89]</sup> 基于视皮层感知机制提出通用的深度卷积神经网络（LeNet），如图3-1，并采用随机梯度下降（Stochastic Gradient Descent, SGD）<sup>[90]</sup> 的训练方法解决深度学习的梯度弥散问题，在手写体字符的识别取得非常优秀的性能。LeNet 是一个多层的网络结构，上层神经元与下层输出的局部区域信息关联，这种局部区域信息相关联的结构相当于生物神经学中的感受野的思想。卷积层对输入图像通过若干个可训练的特征提取器进行卷积，对应于视皮层的初级简单细胞（S1 单元）。对卷积特征图进行非线性映射和降维，映射的过程包括池化、加权、偏置和激活函数等，对应于视皮层复杂细胞（C1 单元）。通过多组的卷积层和池化层级联，最后经过全连接层和损失函数实现对于网络的训练。

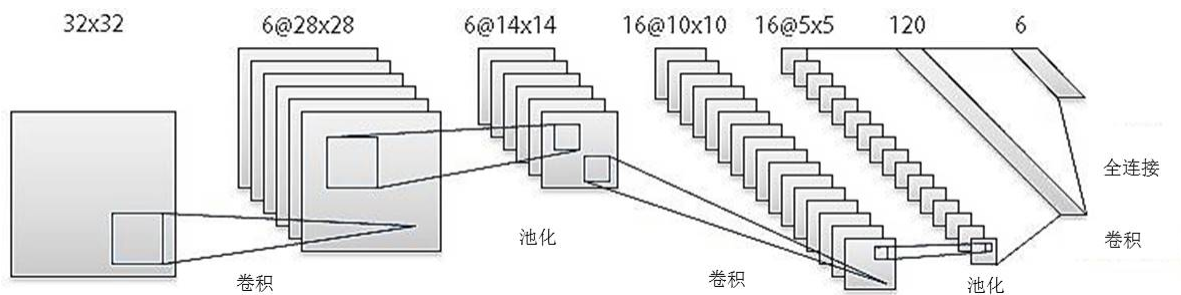


图 3-1 LeNet 网络结构图

### 3.1.2 深度特征提取

本文所采用的是牛津大学视觉几何组（Visual Geometry Group, VGG）在 ILSVRC 2012 数据库<sup>[36]</sup> 中预训练的 19 层卷积网络模型<sup>[91]</sup>（VGG-19），模型的结构如表3-1所示。在训练前，对训练图像进行预处理，首先将 ILSVRC 2012 数据库的所有训练样本归一化到  $256 \times 256$  的尺度空间，计算训练集上各个像素和通道的均值。在训练样本上减去均值，并随机切割成  $224 \times 224$  的样本块来保证平移不变性。通过前期的预处理操作，将最后的得到的训练样本输入到 19 层卷积的卷积网络中。

VGG-19 是多个  $3 \times 3$  小卷积核构成的深度神经网络，其分为 5 组，分别是 Conv1、Conv2、Conv3、Conv4、Conv5。卷积层对应于 BIT 框架中的 S1 单元，实现对于简单视皮层细胞的模型。每组卷积层由多个级联的小感受野的卷积核构成，小感受野卷积

核的级联相比于大感受野卷积核具有参数空间小，容易训练等优点，此外通过小感受野的级联同样可以获得大尺度感受野的特征刻画，与 S1 单元的多尺度性相对应。各个卷积层步幅和填充均为 1 像素，以此避免通过多层卷积操作而导致输入特征图的尺度缩小而带来的信息丢失。跟随着各个卷积层的是一个非线性激励函数<sup>[92]</sup>（Rectified Linear Unit, ReLU）。非线性激励函数（ReLU）的应用为卷积操作的线性变换引入非线性，Hard-zero 截断操作可以避免正负激励的中和现象。相比于同样非线性 Sigmoid 函数，ReLU 的局部线性有利于梯度的方向传导，使得网络更容易收敛。

在各卷积组间增加特征图的池化层，其对应于 BIT 框架中的 C1 单元，实现对视皮层复杂细胞的模拟。在 VGG-19 网络中采用最大（MAX）池化操作，模拟视觉细胞间的激励和抑制作用，采用尺度是  $2 \times 2$ ，步幅为 2，不重叠池化的方法。池化操作可以大量的减少特征图的空间维度，实现各组卷积层提取的特征逐步抽象。通过 5 组的卷积和池化层构成深度的特征提取栈，其后面跟着三个全连接层用于网络训练中梯度的 BP 传导。前两个全链接层包含 4096 维隐层节点，用于特征空间映射和降维。第三层是 1000 维的输出层，其对应于数据库的标签数量。最后通过 Softmax 损失函数对于网络进行梯度计算和反向传导。测试时，Softmax 层得到此样本属于每类的概率矩阵，通过概率矩阵中的最大概率值得到该测试图像类别。

图3-2中可视化深度卷积神经网络的第一层卷积层学习到的滤波器特征，从图中可以看出 CNN 可以自动从图片中学习类 Gabor 的滤波器。这与第 2 章中采用 Gabor 滤波器模拟初级视皮层的对应，说明了卷积神经网络模拟灵长动物视皮层的合理性。本文分别采用 Conv3, Conv4, Conv5 不同神经卷积层特征的输出取代 BIT 模型的表现模型，特征抽取结果如图3-3所示。由图中可以看出，对于目标的特征刻画，从底层到高层不同卷积层组提取的特征从具体到抽象。其中，Conv3 中可以明显的看到物体的轮廓信息，主要表现为底层纹理特征；Conv4 中的特征进一步抽象，物体本身集中为一团块，表现出对物体（Objectness）的检测能力；Conv5 是高层的语义表达，对于不同类型物体提取出不同的特征响应。

### 3.1.3 基于深度学习的 BIT

本文以生物启发跟踪器为基础框架，采用 VGG-19<sup>[91]</sup> 在 ILSVRC 2012<sup>[36]</sup> 中学习到卷积特征取代生物启发表观模型中的 S1 单元和 C1 单元。其中，卷积层对应于视皮层初级视觉细胞（S1 单元），池化层对应于视皮层复杂细胞（C1 单元），通过多层的 S1 和 C1 单元级联，提取与视皮层响应相仿的深度卷积特征作为表现模型。本文中

表 3-1 VGG-19 网络结构参数

类型	名称	输入	大小	数目	扩边	步长
输入	Input	$3 \times 256 \times 256$	–	–	–	–
卷积	Conv1-1	$3 \times 224 \times 224$	$3 \times 3$	64	1	1
卷积	Conv1-2	$64 \times 224 \times 224$	$3 \times 3$	64	1	1
池化	Pool1	$64 \times 224 \times 224$	$2 \times 2$	–	0	2
卷积	Conv2-1	$64 \times 112 \times 112$	$3 \times 3$	128	1	1
卷积	Conv2-2	$128 \times 112 \times 112$	$3 \times 3$	128	1	1
池化	Pool2	$128 \times 112 \times 112$	$2 \times 2$	–	0	2
卷积	Conv3-1	$128 \times 56 \times 56$	$3 \times 3$	256	1	1
卷积	Conv3-2	$256 \times 56 \times 56$	$3 \times 3$	256	1	1
卷积	Conv3-3	$256 \times 56 \times 56$	$3 \times 3$	256	1	1
卷积	Conv3-4	$256 \times 56 \times 56$	$3 \times 3$	256	1	1
池化	Pool3	$256 \times 56 \times 56$	$2 \times 2$	–	0	2
卷积	Conv4-1	$256 \times 28 \times 28$	$3 \times 3$	512	1	1
卷积	Conv4-2	$512 \times 28 \times 28$	$3 \times 3$	512	1	1
卷积	Conv4-3	$512 \times 28 \times 28$	$3 \times 3$	512	1	1
卷积	Conv4-4	$512 \times 28 \times 28$	$3 \times 3$	512	1	1
池化	Pool4	$512 \times 28 \times 28$	$2 \times 2$	–	0	2
卷积	Conv4-1	$512 \times 14 \times 14$	$3 \times 3$	512	1	1
卷积	Conv4-2	$512 \times 14 \times 14$	$3 \times 3$	512	1	1
卷积	Conv4-3	$512 \times 14 \times 14$	$3 \times 3$	512	1	1
卷积	Conv4-4	$512 \times 14 \times 14$	$3 \times 3$	512	1	1
池化	Pool4	$512 \times 14 \times 14$	$2 \times 2$	–	0	2
全连接	Fc1	$512 \times 7 \times 7$	4096	–	–	–
全连接	Fc2	4096	4096	–	–	–
全连接	Fc3	4096	1000	–	–	–

分别采用三种深度学习特征取代 BIT 中 BIAM，分别是（1）Conv3；（2）Conv4；（3）Conv5；算法框架如下图。

本文采用 VGG-19 网络在 ImageNet 预训练的网络参数取代 BIT 的表观模型。首先将跟踪目标归一化到  $256 \times 256$  的尺度大小，减去 ImageNet 数据库预保存的均值矩阵，



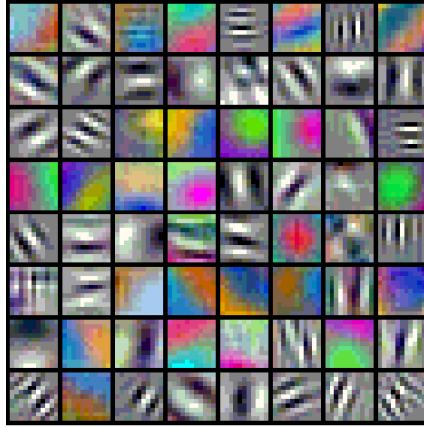


图 3-2 第一层卷积核可视化

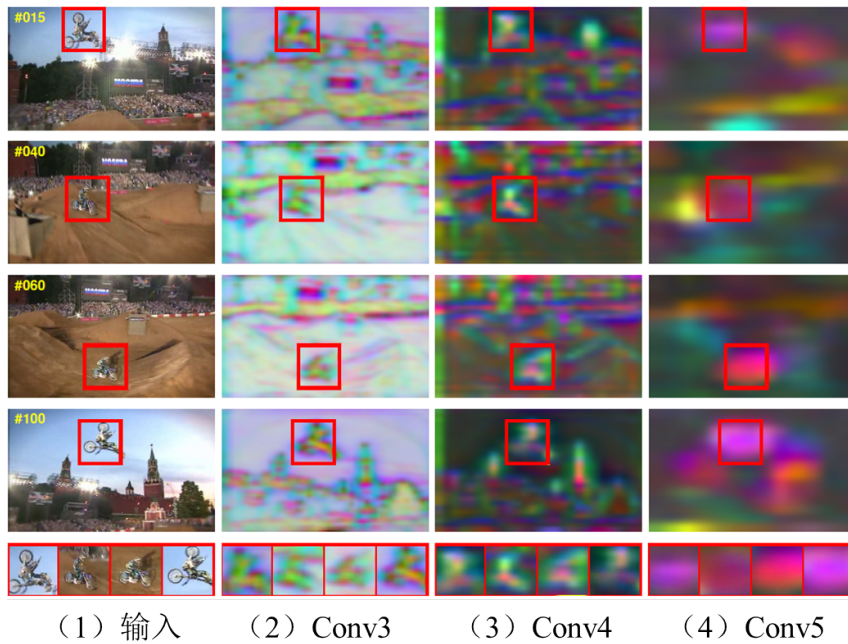


图 3-3 VGG-19 各卷积层组的响应对比

得到对于亮度不敏感的归一化图像。然后通过 VGG-19 多组的卷积层和池化层，从原始图像逐步抽象得到鲁棒的跟踪目标特征。最后，将得到的特征图重新归一化得到与第 2 章中 C1 单元对应的尺度大小（跟踪目标在  $n_s \times n_s = 4 \times 4$  下采样的尺度）。使用 VGG-19 提取出来的特征图取代 BIT 框架中 C1 单元的反应图，具体流程见算法 3-1。

### 3.2 基于尺度自适应的改进

基于生物启发跟踪器的跟踪性能受限于仅能为目标提供固定尺度的目标框，这大大影响了其在被跟踪目标尺度发生变化情况下的跟踪表现。在对于跟踪精度有较高要求的应用场合中，这样的跟踪性能是不足够的。针对 BIT 在尺度估计上的不足，本文引进独立尺度估计的思想并提出了一种快速的尺度估计特征解决这一难点。

**算法 3-1: 基于深度学习的 BIT****Input:** 彩色图像  $I(x, y)$ **Output:** 跟踪结果  $(\hat{x}, \hat{y})$ 设  $\widetilde{C2} = \mathcal{F} \left( \exp \left( -\frac{1}{2\sigma_s^2} ((x - x_o)^2 + (y - y_o)^2) \right) \right)$ 设  $C1^P(x, y, k) = C1(x, y, k)$  when  $t = 1$ **for**  $t = 1, 2, \dots$  **do****深度学习表观模型 (VGG-19)**

预处理:

将跟踪目标  $I(x, y)$  尺度缩放到  $256 \times 256$  尺度空间

对缩放后图像矩阵减去亮度均值矩阵, 剔除光照变化

VGG-19:

通过 VGG-19 各组卷积池化单元 (S1、C1 单元) 得到  $VGG(x, y, k)$ 

归一化:

根据  $n_s \times n_s = 4 \times 4$  将  $VGG(x, y, k)$  归一化得  $C1(x, y, k)$ **生物启发表观模型 (BITM)**

S2 单元:

$$\mathcal{F}[S2(x, y)] = \frac{1}{K} \sum_{k=1}^K \mathcal{F}[C1(x, y, k)] \mathcal{F}[C1^P(x, y, k)]$$

C2 单元:

$$C2(x, y) = \mathcal{F}^{-1}[\mathcal{F}[W(x, y)] \odot \mathcal{F}[S2(x, y)]]$$

定位目标:  $(\hat{x}, \hat{y}) = \arg \max_{(x, y)} C2(x, y)$ 

模型更新:

$$C1^P(x, y, k) = \rho C1(\hat{x}, \hat{y}, k) + (1 - \rho) C1^P(x, y, k)$$

$$\mathcal{F}[W(x, y)] = \rho \mathcal{F}[W(\hat{x}, \hat{y})] + (1 - \rho) \mathcal{F}[W(x, y)]$$

**end for****3.2.1 独立尺度估计**

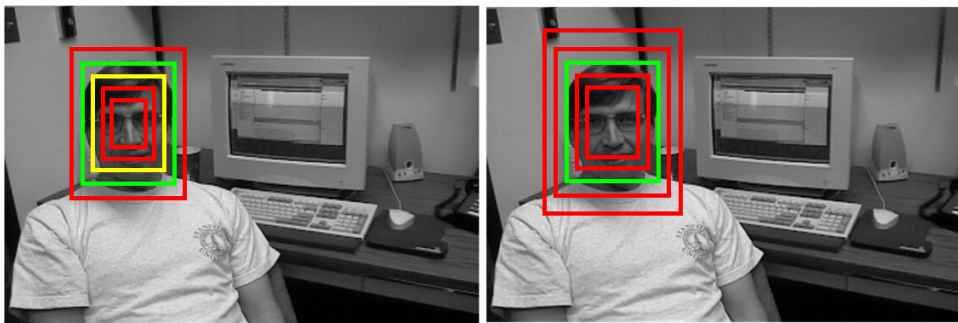
尺度估计是目标跟踪中普遍面临的挑战。第2章中从表观模型设计上解决对于尺度变化目标的特征表示, 提出的多尺度 Gabor 滤波器可以有效地提取具有尺度不变性的特征 (详细的实验分析将在第4章中展示)。然而, 通过尺度不变特征构造的方法只能解决尺度变化下的目标位置估计问题, 并不能自适应地调整目标跟踪框的大小。

Zhang 等人提出的时空上下文跟踪器<sup>[93]</sup> (Spatio Temporal Context, STC) 采用基于置信图 (Confidence Map) 的尺度估计算法。文中通过对比连续几帧跟踪框内的置信图方差来估计当前目标的尺度, 其能够一定程度上适应目标尺度的变化, 但并不能精确地调整跟踪框大小。此外, Martin Danelljan 等人通过运用最小方差和<sup>[94]</sup> (Minimum Output Sum of Squared Error, MOSSE) 建立目标的尺度空间模型, 提出了基于尺度空间的目标自适应估计算法<sup>[95]</sup> (Discrete Scale Space Tracker, DSST)。Hong 等以 DSST 为基础, 引入核相关滤波器<sup>[7]</sup> (Kernelized Correlation Filters, KCF) 提出独立相关滤波器<sup>[96]</sup> (Independent Correlation Filters, ICF) 用于精确尺度估计。以上算法都将目标定位与尺度估计分为两个独立的任务分别解决, 这与视皮层感知中 C2 单元的独立任务学习一致。因此, 本文引进 DSST 独立对目标跟踪框进行精确的尺度调整, 目标定位任务和尺度自适应任务分别进行学习。以下对 BIT 框架引进 DSST 算法做简要介绍。

DSST<sup>[95]</sup> 借鉴 MOSSE<sup>[94]</sup> 跟踪器的解决思路, 为目标外观建立了尺度空间的判别分析。在每一帧视频图像中, DSST 在已知的目标运动状态 (位置及尺度) 的基础上获取当前位置目标的多尺度样本。如图3-4所示, 黄色框为上一帧目标的状态估计结果, 称其为基准跟踪框 (Base Bounding Box) 的尺度  $\hat{s}_{t-1}$ 。通过基准跟踪框放大/缩小一定倍数, 获得不同尺度的尺度样本框集合  $S_t = \{s_t^n\}_{n=1}^{N_s}$ , 尺度空间集合  $S_t$  表示如下:

$$s_t^n = \alpha^n \hat{s}_{t-1}, n \in \left\{ \left\lfloor -\frac{N_s - 1}{2} \right\rfloor, \dots, \left\lfloor \frac{N_s - 1}{2} \right\rfloor \right\} \quad (3-1)$$

其中  $s_t^n$  为尺度空间样本框,  $\alpha$  为尺度缩放放大因子,  $N_s$  为尺度空间样本数, 本文采用与 DSST 一致的参数设置  $N_s = 33$ 。



(a) 尺度空间匹配。

(b) 尺度空间建立。

图 3-4 DSST 尺度空间样本获取示例

根据  $S_t$  在不同的尺度框在跟踪视频帧中截取不同大小的尺度样本, 并对尺度框内的图像样本分别进行尺度归一化, 以确保尺度样本框采集特征维度一致。然后, 通过

HOG 获取不同尺度样本框内图像的尺度特征  $f_t^n$ ，从而得到描述当前位置目标大小的尺度空间特征  $F_t = \{f_t^n\}_{n=1}^{N_s}$ 。

根据 MOSSE 训练一个尺度空间滤波器使得该滤波器能够满足不同尺度样本通过 MOSSE 滤波器的输出  $F_t \otimes H_t^*$  与期望输出  $G$  距离平方的总和最小。为达到这个目的，需解决下式所列最小化问题：

$$\min_{H^*} \sum_t |F_t \otimes H^* - G|^2 \quad (3-2)$$

根据卷积定理，对于滤波器  $H^*$  的最小化求解可以转换为频域解卷，可获得滤波器  $H^*$  的闭合解。由于尺度特征在跟踪过程中会因为外界环境扰动发生变化，因此滤波器需要实时地做调整更新以适应目标外观的变化。综上所述，MOSSE 的更新公式如下所示：

$$H_t^* = \frac{A_t}{B_t} \quad (3-3)$$

$$\begin{cases} A_t = \eta \mathcal{F}[G] \odot \mathcal{F}[F_t] + (1 - \eta)A_{t-1} \\ B_t = \eta \mathcal{F}[F_t] \odot \mathcal{F}[F_t^*] + (1 - \eta)B_{t-1} \end{cases} \quad (3-4)$$

其中  $A_t$  为滤波器第  $t$  帧对应的分子项， $B_t$  为第  $t$  帧滤波器的分母项， $\eta$  表示滤波器更新速率。当再次出现新一帧视频图像时，通过下式获取尺度空间滤波器响应：

$$y_s = \mathcal{F}^{-1} \left[ \frac{A_{t-1} \odot \mathcal{F}[F_t]}{B_{t-1} + \lambda} \right] \quad (3-5)$$

其中  $y_s$  为滤波器响应， $y_s$  的峰值对应尺度空间中当前目标尺度， $\lambda > 0$  用于控制滤波器目标方程标准化项的影响。详细请见 [52]。

### 3.2.2 快速尺度估计

DSST 中采用 HOG<sup>[6]</sup> 作为尺度特征对不同尺度样本提取对于尺度敏感的描述符。然而由于 HOG 特征的空间维度与图像大小成正比，因此需要对于尺度样本进行  $N_s$  次尺度归一化操作。实验证明，多达 33 次的尺度归一化占据了特征计算的大部分计算量。本文仿照第 2 章中的生物启发表现模型中的 S1 和 C1 单元，基于空间金字塔池化<sup>[97]</sup> (Spatial Pyramid Pooling, SPP) 的思想，提出一种快速的尺度特征提取方法。

#### 3.2.2.1 空间金字塔池化 (SPP)

SPP<sup>[97]</sup> 是空间金字塔匹配<sup>[98]</sup> (Spatial Pyramid Matching, SPM) 在卷积神经网络上的扩展。词袋<sup>[99]</sup> (Bag of Words, BoW) 模型在图像检索、行为识别、物体分类等领域得到了成功应用，然而由于 BoW 仅对于局部空间特征建模，忽略了大尺度空间的结构

信息。空间金字塔匹配<sup>[98]</sup>（SPM）是通过不同尺度空间的特征表示以有效地提取局部特征的空间位置关系。随着 SPM 在物体匹配的成功应用，其思想也被扩展到颜色直方图特征中，通过不同尺度空间的直方图统计来保持不同尺度下的维度不变性。分层梯度方向直方图<sup>[100]</sup>（Pyramid Histogram of Oriented Gradients, PHOG）是一种基于 SPM 改进的空间形状描述符，在不同尺度空间上统计目标边缘的梯度方向直方图，在物体识别、图片检索等机器视觉领域取得成功应用。

SPP 的提出是为了解决区域卷积神经网络<sup>[101]</sup>（Region-based Convolutional Neural Networks, RCNN）提取特征比较耗时的问题。SPP 由 He 等人在 SPP-Net<sup>[97]</sup> 论文中提出，随后被 RGB, Ren 等人扩展为 ROI Pooling 方法应用在 Fast-RCNN<sup>[102]</sup> 和 Faster-RCNN<sup>[103]</sup> 中。如图3-5，RCNN 需要对每个目标进行裁剪（Crop）并且缩放（Warp）后的区域进行卷积特征提取，当裁剪区域有部分重叠时会导致重复的卷积操作。而应用 SPP 之后只需对图像进行一次的卷积，通过金字塔的池化操作提取不同位置和尺度下对应的目标特征。SPP 相比于正常的池化操作的区别是，普通池化操作采用固定步长（Step）对目标进行池化降维，而 SPP 采用固定格子（Grid）数对目标进行池化降维。因此，在不同的尺度空间下能一致获得相同维度的特征表示，避免了裁剪和缩放带来的额外计算量。

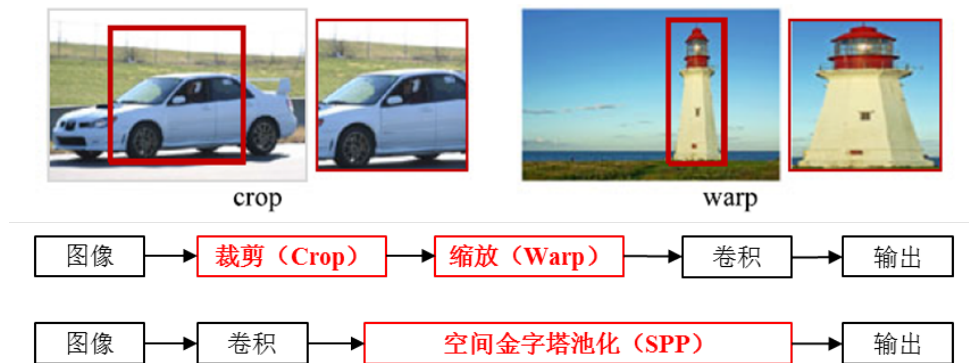


图 3-5 空间金字塔池化与传统池化对比

### 3.2.2.2 快速尺度特征（FSF）

本文借鉴 SPP 的思想，提出一种对于尺度敏感的快速特征提取方法，以取代 DSST 中采用的 HOG 特征。解决 DSST 尺度估计需要对不同尺度大小的样本进行多次缩放来保证特征维度的一致性。与第二章中 BIAM 特征提取过程类似，本章提出的快速尺度特征也由卷积层（S1 单元）和池化层（C1 单元）构成。

不同于目标定位对于尺度和方向不变性的要求，尺度估计需要提取的是对尺度变化敏感的表现特征。相比于目标定位采用多尺度 Gabor 滤波器，本文采用 2 组正交的单尺度一维边缘检测滤波器  $G_x = [+1, 0, -1]$  和  $G_y = [+1, 0, -1]^T$  对原始灰度  $I(x, y)$  做卷积操作。单尺度滤波器可以保证特征对于尺度的敏感性，同时因为不需要估计纹理的方向，因此通过正交的一维梯度滤波器计算二维纹理的强度。因此可以极大的减少卷积操作的计算量。快速尺度特征的卷积响应  $A_s$  可由公式(3-6)表示。

$$A_s = \sqrt{(I(x, y) \otimes G_x)^2 + (I(x, y) \otimes G_y)^2} \quad (3-6)$$

在池化单元中，本文利用 SPP 的空间金字塔的思想，对卷积响应  $A_s$  进行最大池化。即通过变步长的 MAX 池化对卷积响应  $A_s$  进行下采样，保证不同尺度下得到相同维度的尺度特征，如图3-6所示。因此，SPP 方法的应用可以避免 DSST 尺度估计方法中尺度特征提取时多次裁剪和缩放操作，有效提高了特征提取的效率。尺度特征可以用公式表示如下：

$$f^n(x, y) = \max_{(x', y') \in \Omega_r(x, y)} A_s(x', y'), r = s^n / \nu \quad (3-7)$$

其中， $s$  对应于公式(3-1)中的尺度空间样本框， $n$  是尺度空间的索引， $\nu$  是快速尺度特征的维度。

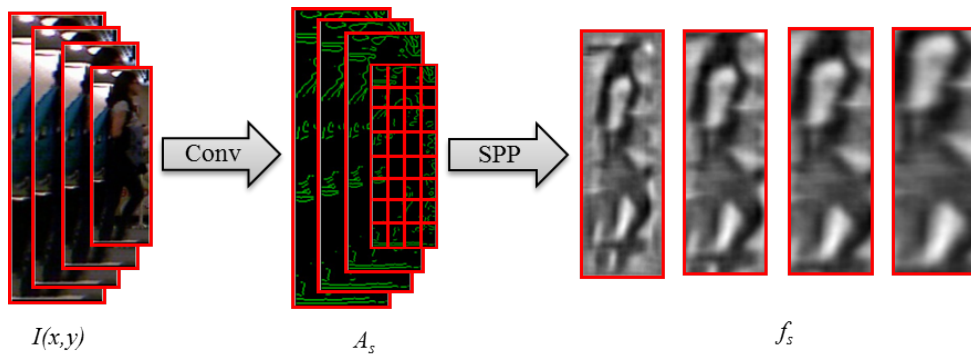


图 3-6 基于 SPP 的快速尺度特征提取

### 3.2.3 尺度自适应的 BIT

综上所述，本小节中通过 DSST 改进了 BIT 跟踪器的尺度估计问题，并提出了一种快速尺度特征提取方法。快速特征提取方法结合卷积操作和池化操作，与目标定位的 BIAM 中 S1 单元和 C1 单元相对应，解决了原来 DSST 方法中特征提取复杂的问题，保证了算法的实时性。本小节总结尺度自适应的 BIT 实现流程如算法 3-2。

**算法 3-2: 尺度自适应的 BIT**

**Input:** 灰度图像  $I(x, y)$ , 彩色图像  $RGB(x, y)$

**Output:** 跟踪结果  $(\hat{x}, \hat{y})$

设  $\widetilde{C2} = \mathcal{F} \left( \exp \left( -\frac{1}{2\sigma_s^2} ((x - x_o)^2 + (y - y_o)^2) \right) \right)$

设  $C1^P(x, y, k) = C1(x, y, k)$  when  $t = 1$

**for**  $t = 1, 2, \dots$  **do**

**生物启发表观模型 (BIAM)**

S1 单元:

通过  $I(x, y)$  根据(2-14)和(2-15)计算  $S1_{gabor}$

通过  $RGB(x, y)$  根据(2-3)计算  $S1_{color}$

C1 单元:

对  $S1_{gabor}$  根据(2-5) STD 池化得  $C1_{gabor}(x, y, k)$

对  $S1_{color}$  根据(2-7) AVG 池化得  $C1_{color}(x, y, k)$

$C1(x, y, k) = C1_{gabor}(x, y, k) + C1_{color}(x, y, k)$

**生物启发表观模型 (BITM)**

S2 单元:

$$\mathcal{F}[S2(x, y)] = \frac{1}{K} \sum_{k=1}^K \mathcal{F}[C1(x, y, k)] \mathcal{F}[C1^P(x, y, k)]$$

C2 单元:

$$C2(x, y) = \mathcal{F}^{-1} [\mathcal{F}[W(x, y)] \odot \mathcal{F}[S2(x, y)]]$$

定位目标:  $(\hat{x}, \hat{y}) = \arg \max_{(x, y)} C2(x, y)$

**快速尺度估计**

根据  $s^n$  和(3-6)通过  $I(x, y)$  提取计算  $A_s$

基于 SPP 根据(3-7)提取快速尺度特征  $f^n$

尺度估计:  $\hat{s}_t = \alpha^{\arg \max_n y_s(n)} s_{t-1}$

模型更新:

$$C1^P(x, y, k) = \rho C1(\hat{x}, \hat{y}, k) + (1 - \rho) C1^P(x, y, k)$$

$$\mathcal{F}[W(x, y)] = \rho \mathcal{F}[W(\hat{x}, \hat{y})] + (1 - \rho) \mathcal{F}[W(x, y)]$$

$$A_t = \eta \mathcal{F}[G] \odot \mathcal{F}[F_t] + (1 - \eta) A_{t-1}$$

$$B_t = \eta \mathcal{F}[F_t] \odot \mathcal{F}[F_t^*] + (1 - \eta) B_{t-1}$$

**end for**

### 3.3 本章小结

本章提出了生物启发跟踪器两个改进方案。其中引进了深度学习特征取代生物启发跟踪器的表观模型，本文迁移 VGG19 网络在 ILSVRC 2012 上预训练得到的网络参数。抽取其中前 5 组卷积和池化层，对应于原来 BIAM 中的 S1 单元与 C1 单元。另外，针对于 BIT 跟踪算法并不能产生尺度自适应变化的跟踪框，引进了 DSST 为跟踪框提供尺度估计，并在此基础上提出一种快速尺度特征。借鉴 SPP 池化的思想，同样采用卷积和池化的操作，实现尺度特征的快速提取。



## 第四章 实验分析对比

本文提出一种基于生物启发的目标跟踪算法（BIT），其中主要包括两个模块：生物启发表观模型与生物启发跟踪模型。本章在 Tracking Benchmark 50<sup>[104]</sup>（TB50）和 Amsterdam Library of Ordinary Videos 300++<sup>[105]</sup>（ALOV300++）两个大型的目标跟踪数据库上对比。实验分析主要包括跟踪器跟踪精度、跟踪模型分析、跟踪器效率分析等。此外，同时也对第3章中提出的基于深度特征和尺度自适应改进的BIT进行对比分析。

在本章中的所有实验中采用固定的参数设置，其中学习率  $\rho$  被设置为 0.02，根据前 5 帧中 C2 响应的变化趋势，C2 的尺度参数  $\sigma_s$  被设置 0.1 或 0.08（当 C2 的平均值递增时， $\sigma_s$  被设置为 0.1，否则设置为 0.08）。实验采用一台配置 Intel i7 3770（3.4GHz）的个人电脑，在 MALTLAB 2014A 中实现对 BIT 的对比测试。

### 4.1 对比实验数据库

#### 4.1.1 TB50

TB50<sup>[104]</sup> 是目标视觉跟踪算法近几年来最为主要的评测数据库，成为视觉跟踪对比实验的基准。其中总共包含 50 段测试视频（见图4-1），51 个跟踪任务，一共 23000 多帧。TB50 视觉跟踪测试集将 50 段测试视频分为 11 种挑战（具体见表4-1），其中包括：

- （1）光照变化（Illumination Variation, IV）：在目标区域中光照变化剧烈；
- （2）尺度变化（Scale Variation, SV）：目标大小长宽比等发生变化；
- （3）局部遮挡（Occlusion, OCC）：目标被部分或者完全遮挡；
- （4）目标形变（Deformation, DEF）：非刚性物体发生形变；
- （5）运动模糊（Motion Blur, MB）：由于目标运动或者摄像机抖动引起的目标区域模糊；
- （6）快速运动（Fast Motion, FB）：目标无规律的快速运动；
- （7）二维旋转（In-Plane-Rotation, IPR）：目标在同一图像平面内旋转；
- （8）三维旋转（Out-Plane-Rotation, OPR）：目标不在同一平面内旋转；

- (9) 视野超越 (Out-of-View, OV): 目标部分区域超出摄像机视野范围;
- (10) 背景干扰 (Background Clusters, BC): 背景中有与目标相似的物体;
- (11) 低分辨率 (Low Resolution, LR): 视频分辨率过低或跟踪目标过小。

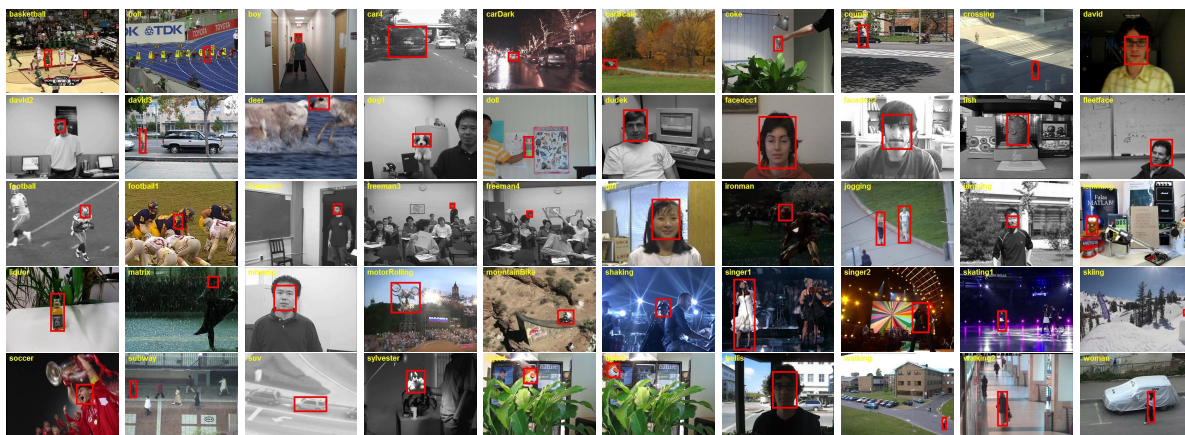


图 4-1 TB50 跟踪数据库视频示例

TB50 中提出了三种视觉跟踪算法测试方法, 常规测试 (One Pass Evaluation, OPE)、时间鲁棒性测试 (Temporal Robustness Evaluation, TRE) 以及空间鲁棒性测试 (Spatial Robustness Evaluation, SRE)。常规测试 (OPE) 是指每一段测试视频只跑一次结果评价, 从测试视频第一帧开始以标注数据为起始位置对算法进行测试; 时序鲁棒性测试 (TRE) 主要测试跟踪算法在不同起始时段下跟踪目标的稳定性, 算法分别从测试视频不同帧的标注数据开始测试; 空间稳定性测试 (SPE) 评测跟踪目标的尺度大小对于跟踪结果的影响, 算法从 8 个不同的起始位置及 4 个不同的尺度进行测试。本文以标准的 OPE 最为测试方法。

TB50 提出并采用两项基本指标对跟踪算法的跟踪精度进行评估, 其分别为预测精度和成功率。预测精度根据每一帧中心位置误差 (Center Location Error, CLE) 为一帧图像中跟踪算法跟踪结果目标框中心与标注目标框 (Ground Truth) 中心欧式距离, 当中心位置误差小于某一个阈值时则认为跟踪算法预测正确 (一般阈值选为 20 像素)。跟踪成功率是一项与尺度估计有关的的评测指标, 主要通过交叠面积参数 (Overlap Score) 定义一帧图像中跟踪算法跟踪结果目标框与标注目标框交叠面积占两目标框合面积的百分比。一段视频中交叠面积参数小于特定阈值的帧数占整段视频的百分比表示跟踪的成功率 (一般阈值选为 50%)。

表 4-1 TB50 跟踪数据库的视频属性分类

视频类别	视频名称
光照变化 (IV)	Basketball, Car4, CarDark, Coke, David, Doll, FaceOcc2, Fish, Ironman, Lemming, Liquor, Matrix, Mhyang, MotorRolling, Shaking, Singer1, Singer2, Skating1, Skiing, Soccer, Sylvester, Tiger1, Tiger2, Trellis, Woman
尺度变化 (SV)	Boy, Car4, CarScale, Couple, Crossing, David, Dog1, Doll, Dudek, FleetFace, Freeman1, Freeman3, Freeman4, Girl, Ironman, Lemming, Liquor, Matrix, MotorRolling, Shaking, Singer1, Skating1, Skiing, Soccer, Trellis, Walking, Walking2, Woman
局部遮挡 (OCC)	Basketball, Bolt, CarScale, Coke, David, David3, Doll, Dudek, FaceOcc1, FaceOcc2, Football, Freeman4, Girl, Ironman, Jogging.1, Jogging.2, Lemming, Liquor, Matrix, Singer1, Skating1, Soccer, Subway, Suv, Tiger1, Tiger2, Walking, Walking2, Woman
目标形变 (DEF)	Basketball, Bolt, Couple, Crossing, David, David3, Dudek, FleetFace, Jogging.1, Jogging.2, Mhyang, Singer2, Skating1, Skiing, Subway, Tiger1, Tiger2, Walking, Woman
运动模糊 (MB)	Boy, David, Deer, FleetFace, Ironman, Jumping, Liquor, MotorRolling, Soccer, Tiger1, Tiger2, Woman
快速运动 (FB)	Boy, CarScale, Coke, Couple, Deer, Dudek, FleetFace, Ironman, Jumping, Lemming, Liquor, Matrix, MotorRolling, Soccer, Tiger1, Tiger2, Woman
二维旋转 (IPR)	Bolt, Boy, CarScale, Coke, David, David2, Deer, Dog1, Doll, Dudek, FaceOcc2, FleetFace, Football, Football1, Freeman1, Freeman3, Freeman4, Girl, Ironman, Matrix, MotorRolling, MountainBike, Shaking, Singer2, Skiing, Soccer, Suv, Sylvester, Tiger1, Tiger2, Trellis
三维旋转 (OPR)	Basketball, Bolt, Boy, CarScale, Coke, Couple, David, David2, David3, Dog1, Doll, Dudek, FaceOcc2, FleetFace, Football, Football1, Freeman1, Freeman3, Freeman4, Girl, Ironman, Jogging.1, Jogging.2, Lemming, Liquor, Matrix, Mhyang, MountainBike, Shaking, Singer1, Singer2, Skating1, Skiing, Soccer, Sylvester, Tiger1, Tiger2, Trellis, Woman
视野超越 (OV)	Dudek, Ironman, Lemming, Liquor, Suv, Tiger2
背景干扰 (BC)	Basketball, CarDark, Couple, Crossing, David3, Deer, Dudek, Football, Football1, Ironman, Liquor, Matrix, Mhyang, MotorRolling, MountainBike, Shaking, Singer2, Skating1, Soccer, Subway, Trellis
低分辨率 (LR)	Deer, Ironman, MotorRolling, Walking2

### 4.1.2 ALOV300++

ALOV300++<sup>[105]</sup> 是目前视频数量最大的跟踪测试数据集，逐步被应用到视觉跟踪的综合性评测中。该数据集中包含了 314 段测试视频，共计 89364 帧。ALOV300+ 视觉跟踪测试集将其中包含的 314 段测试视频（如图4-2）分为 13 种挑战，其中包括：

- (1) 光照 (Light)：在目标区域中光照变化剧烈；
- (2) 覆盖 (Surface Cover)：跟踪物体外观被覆盖
- (3) 镜面 (Specularity)：跟踪目标会倒影周围环境；
- (4) 透明 (Transparency)：跟踪目标为透明物体；
- (5) 形变 (Shape)：非刚性物体发生形变；
- (6) 运动模糊 (Motion Smoothness)：由于目标运动引起的目标区域模糊；
- (7) 运动相关 (Motion Coherence)：跟踪目标与周围环境运动一致；
- (8) 杂乱 (Clutter)：跟踪环境复杂；
- (9) 混淆 (Confusion)：背景中有与目标相似的物体；
- (10) 遮挡 (Occlusion)：目标被部分遮挡；
- (11) 摄像头运动 (Moving Camera)：拍摄的摄像头发生运动；
- (12) 摄像头变焦 (Zooming Camera)：拍摄的摄像头发生焦距变化；
- (13) 长时 (Long Duration)：长时间的跟踪序列。



图 4-2 ALOV300++ 跟踪数据库视频示例

ALOV300+ 数据库根据三个基本准则评估跟踪算法的性能：（1）偏差（Deviation）：跟踪算法估计值和真实值之间的误差；（2）错跟（False Positive）：跟踪算法错误地判定背景区域为跟踪目标；（3）漏跟（False Negative）：跟踪算法无法跟踪到目标。ALOV300+ 数据库上通过跟踪结果与标定结果的交叠面积为基准，当交叠面积大于 50% 为正确跟踪。通过所有视频帧中的准确跟踪数  $n_{tp}$ 、错跟数  $n_{fp}$  和漏跟数  $n_{fn}$ ，计算跟踪算法的查准率（Precision） $precision = n_{tp}/(n_{tp} + n_{fp})$  和召回率（Recall） $recall = n_{tp}/(n_{tp} + n_{fn})$ ，最后计算 F-Score 指标如下：

$$F = 2 \times \frac{precision \times recall}{precision + recall} \quad (4-1)$$

## 4.2 生物启发跟踪器对比实验

本文选取 10 个代表性的跟踪算法作为对比，其中 RPT<sup>[106]</sup>，TGPR<sup>[107]</sup>，ICF<sup>[96]</sup> 和 KCF<sup>[7]</sup> 为近一年来优秀的跟踪算法；Struck<sup>[108]</sup>，SCM<sup>[32]</sup>，TLD<sup>[39]</sup>，VTS<sup>[109]</sup> 是 TB50 测试中表现最突出的跟踪器；IVT<sup>[22]</sup> 和 MIL<sup>[38]</sup> 作为经典的跟踪算法成为参照的基准。除了 50 段视频的综合测试外，还分属性进行测试，并对跟踪模型和实时性能作了详细的分析。

### 4.2.1 整体跟踪结果分析

首先，本文定性分析 BIT 与性能最高的 5 种算法（RPT，TGPR，ICF，KCF 和 Struck）在不同视频的关键帧中的跟踪结果。图4-4中共列举了以上算法在 TB50 数据库中所有 50 段跟踪视频段视频的跟踪结果。由图中可见，本文算法提出的 BIT 在所有视频中都基本能较好地跟踪，表现极好的鲁棒性。

图4-3为 BIT 与其他 10 个跟踪算法在 50 段视频下的测试对比结果。本文算法 BIT 在中心位置误差 20 阈值下的预测精度为 81.7%，高于最好的 state-of-art 方法（RPT，81.1%）0.6 个百分点。此外，与 TB50 标准测试中的最好结果 Struck 高出 15.4%。表4-2中展示了各有方法在中心位置误差 20 阈值下的跟踪精度。本文提出的 BIT 在 23 段视频中取得了最高的跟踪精度，相比于其他最高的 4 个算法 RPT，TGPR，ICF，KCF 分别是 23,11,20 和 15。此外，定义各段视频最高预测精度的 80% 为优秀结果，BIT 在 50 段测试视频中只有 8 段未达到优秀标准线，相比于其他的是个算法 RPT，TGPR，ICF，KCF 分别是 10,14,12 和 15。综上所述，BIT 基本取得了目前跟踪算法中最高的测试结果。

表 4-2 BIT 与其他 10 种跟踪算法在 TB50 数据库的预测精度（20 像素阈值）

	BIT	RPT <sup>[106]</sup>	TGPR <sup>[107]</sup>	ICF <sup>[96]</sup>	KCF <sup>[7]</sup>	Struck <sup>[108]</sup>	SCM <sup>[32]</sup>	TLD <sup>[39]</sup>	VTS <sup>[109]</sup>	MIL <sup>[38]</sup>	IVT <sup>[22]</sup>
Basketball	1.000	0.924	0.994	1.000	0.923	0.120	0.661	0.028	1.000	0.284	0.497
Bolt	1.000	0.017	0.017	1.000	0.989	0.020	0.031	0.306	0.089	0.014	0.014
Boy	1.000	1.000	0.987	1.000	1.000	1.000	0.440	1.000	0.980	0.846	0.332
Car4	0.973	0.980	1.000	1.000	0.950	0.992	0.974	0.874	0.363	0.354	1.000
CarDark	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.639	1.000	0.379	0.807
CarScale	0.718	0.806	0.790	0.806	0.806	0.647	0.647	0.853	0.544	0.627	0.782
Coke	0.931	0.962	0.945	0.887	0.838	0.948	0.430	0.684	0.189	0.151	0.131
Couple	0.607	0.679	0.600	0.107	0.257	0.736	0.114	1.000	0.100	0.679	0.086
Crossing	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.617	0.417	1.000	1.000
David	1.000	1.000	0.977	1.000	1.000	0.329	1.000	1.000	0.962	0.699	1.000
David2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.978	1.000
David3	1.000	1.000	0.996	1.000	1.000	0.337	0.496	0.111	0.742	0.738	0.754
Deer	0.831	1.000	0.859	0.817	0.817	1.000	0.028	0.732	0.042	0.127	0.028
Dog1	1.000	1.000	1.000	0.994	1.000	0.996	0.976	1.000	0.811	0.919	0.980
Doll	0.986	0.987	0.943	0.947	0.967	0.919	0.978	0.983	0.946	0.732	0.757
Dudek	0.862	0.849	0.751	0.899	0.859	0.897	0.883	0.597	0.871	0.688	0.886
FaceOcc1	0.877	0.663	0.664	0.855	0.878	0.575	0.933	0.203	0.485	0.221	0.645
FaceOcc2	0.933	0.990	0.468	0.968	0.972	1.000	0.860	0.856	0.936	0.740	0.993
Fish	1.000	1.000	0.975	1.000	1.000	1.000	0.863	1.000	0.992	0.387	1.000
FleetFace	0.581	0.562	0.453	0.627	0.556	0.639	0.529	0.506	0.642	0.358	0.264
Football	0.798	0.801	0.997	0.801	0.796	0.751	0.765	0.804	0.796	0.790	0.793
Football1	0.973	0.932	0.986	0.986	0.959	1.000	0.568	0.554	0.892	1.000	0.811
Freeman1	1.000	0.972	0.933	0.393	0.393	0.801	0.982	0.540	0.969	0.939	0.807
Freeman3	0.817	0.996	0.774	0.896	0.911	0.789	1.000	0.767	0.702	0.048	0.761
Freeman4	0.993	0.880	0.580	0.951	0.530	0.375	0.509	0.410	0.219	0.201	0.346
Girl	1.000	0.924	0.918	0.916	0.864	1.000	1.000	0.918	0.874	0.714	0.444
Ironman	0.157	0.181	0.217	0.199	0.217	0.114	0.157	0.120	0.247	0.108	0.054
Jogging.1	0.977	0.228	0.993	0.977	0.235	0.241	0.228	0.974	0.225	0.231	0.225
Jogging.2	1.000	0.179	0.997	0.186	0.163	0.254	1.000	0.857	0.186	0.186	0.199
Jumping	0.093	1.000	0.946	0.383	0.339	1.000	0.153	1.000	0.236	0.997	0.208
Lemming	0.491	0.537	0.349	0.509	0.495	0.628	0.166	0.859	0.554	0.823	0.167
Liquor	0.986	0.937	0.271	0.431	0.423	0.390	0.276	0.588	0.364	0.199	0.207
Matrix	0.360	0.440	0.390	0.350	0.170	0.120	0.350	0.160	0.200	0.180	0.020
Mhyang	1.000	1.000	0.947	1.000	1.000	1.000	1.000	0.978	1.000	0.460	1.000
M.Rolling	0.049	0.049	0.091	0.049	0.043	0.085	0.037	0.116	0.049	0.043	0.030
M.Bike	0.987	1.000	1.000	1.000	1.000	0.921	0.969	0.259	0.996	0.667	0.996
Shaking	0.970	0.995	0.970	0.025	0.025	0.192	0.814	0.405	0.921	0.282	0.011
Singer1	1.000	0.986	0.684	0.689	0.980	0.641	1.000	1.000	1.000	0.501	0.963
Singer2	0.036	0.913	0.970	0.038	0.945	0.036	0.112	0.071	0.358	0.404	0.036
Skating1	1.000	1.000	0.805	1.000	1.000	0.465	0.768	0.318	0.890	0.130	0.108
Skiing	0.136	0.136	0.123	0.111	0.074	0.037	0.136	0.123	0.062	0.074	0.111
Soccer	0.949	0.944	0.158	0.967	0.793	0.253	0.268	0.115	0.505	0.191	0.173
Subway	1.000	1.000	1.000	1.000	1.000	0.983	1.000	0.251	0.240	0.994	0.223
Suv	0.979	0.529	0.658	0.979	0.979	0.572	0.978	0.909	0.535	0.123	0.447
Sylvester	0.839	0.979	0.955	0.851	0.843	0.995	0.946	0.949	0.820	0.651	0.680
Tiger1	0.927	0.977	0.284	0.958	0.975	0.175	0.126	0.456	0.117	0.095	0.080
Tiger2	0.449	0.814	0.723	0.485	0.356	0.630	0.112	0.386	0.162	0.414	0.082
Trellis	1.000	1.000	0.979	1.000	1.000	0.877	0.873	0.529	0.503	0.230	0.332
Walking	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.964	1.000	1.000	1.000
Walking2	0.440	0.684	0.988	1.000	0.440	0.982	1.000	0.426	0.408	0.406	1.000
Woman	0.940	0.938	0.968	0.938	0.938	1.000	0.940	0.191	0.198	0.206	0.201
ALL	0.817	0.811	0.766	0.764	0.739	0.656	0.649	0.608	0.575	0.475	0.499
No. Best	23	23	11	20	15	14	14	11	8	3	8
No. Worst	8	10	14	12	15	26	23	28	29	41	33

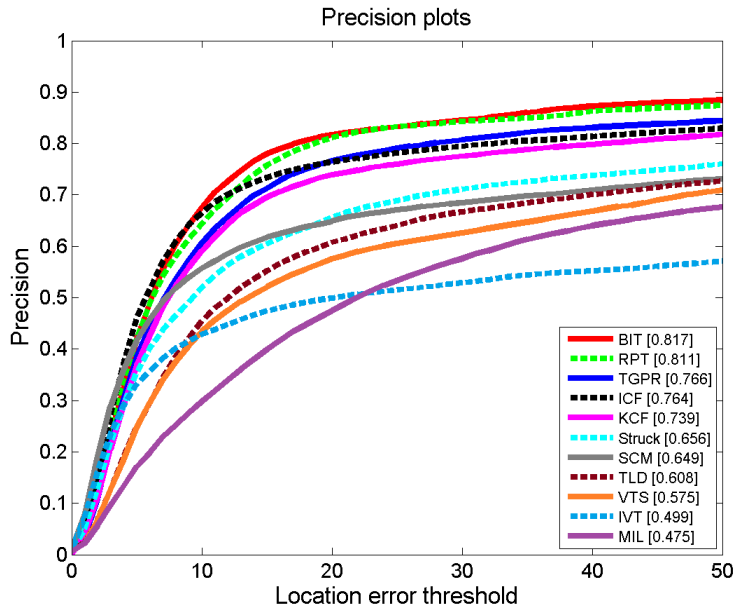


图 4-3 BIT 与其他算法在 TB50 跟踪数据库上的预测精度图

### 4.2.2 详细跟踪结果分析

以下图 5-3 至图 5-13 为不同视频分类下各跟踪算法的性能对比。下面分别对光照变化 (IV)，尺度变化 (SV)，局部遮挡 (OCC)，目标变形 (DEF)，运动模糊 (MB)，快速运动 (FM)，二维旋转 (IPR)，三维旋转 (OPR)，视野超越 (OV)，背景干扰 (BC) 及低像素图像 (LR) 等 11 类挑战进行对比分析。

表 4-3 各类跟踪算法在 TB50 测试数据库 11 种挑战中的对比

	BIT	RPT <sup>[106]</sup>	TGPR <sup>[107]</sup>	ICF <sup>[96]</sup>	KCF <sup>[7]</sup>	Struck <sup>[108]</sup>	SCM <sup>[32]</sup>	TLD <sup>[39]</sup>	VTS <sup>[109]</sup>	MIL <sup>[38]</sup>	IVT <sup>[22]</sup>
IV	<a href="#">0.764</a>	<b>0.827</b>	0.687	0.696	0.717	0.558	0.594	0.537	0.573	0.349	0.418
SV	<a href="#">0.786</a>	<b>0.802</b>	0.703	0.707	0.667	0.639	0.672	0.606	0.582	0.471	0.494
OCC	<b>0.854</b>	0.765	0.708	<a href="#">0.817</a>	0.744	0.564	0.640	0.563	0.534	0.427	0.455
DEF	<b>0.817</b>	0.748	<a href="#">0.768</a>	0.754	0.751	0.521	0.586	0.512	0.487	0.455	0.409
MB	<a href="#">0.663</a>	<b>0.783</b>	0.578	0.654	0.621	0.551	0.339	0.518	0.375	0.357	0.222
FM	<a href="#">0.643</a>	<b>0.745</b>	0.575	0.612	0.581	0.604	0.333	0.551	0.353	0.396	0.220
IPR	<a href="#">0.783</a>	<b>0.795</b>	0.706	0.739	0.731	0.617	0.597	0.584	0.579	0.453	0.457
OPR	<b>0.831</b>	<a href="#">0.807</a>	0.741	0.741	0.724	0.597	0.618	0.596	0.604	0.466	0.464
OV	<b>0.654</b>	<a href="#">0.641</a>	0.495	0.584	0.555	0.539	0.429	0.576	0.455	0.393	0.307
BC	<a href="#">0.789</a>	<b>0.840</b>	0.761	0.698	0.725	0.585	0.578	0.428	0.578	0.456	0.421
LR	0.369	0.478	<a href="#">0.539</a>	0.516	0.379	<b>0.545</b>	0.305	0.349	0.187	0.171	0.278

图4-5为光照变化类别跟踪任务的对比结果，BIT 取得 76.4% 的预测准确率，只低于 RPT 的 82.7%，位列所有跟踪算法第二位。这得益于 BIT 在 S1 单元引入的 Gabor 滤波器。Gabor 滤波器的局部感受野和边缘敏感性能很好地应对光照变化，提取对光照不



图 4-4 TB50 跟踪数据库的可视化定性分析 (BIT, RPT<sup>[106]</sup>, TGPR<sup>[107]</sup>, ICF<sup>[96]</sup>, KCF<sup>[7]</sup>, Struck<sup>[108]</sup>)



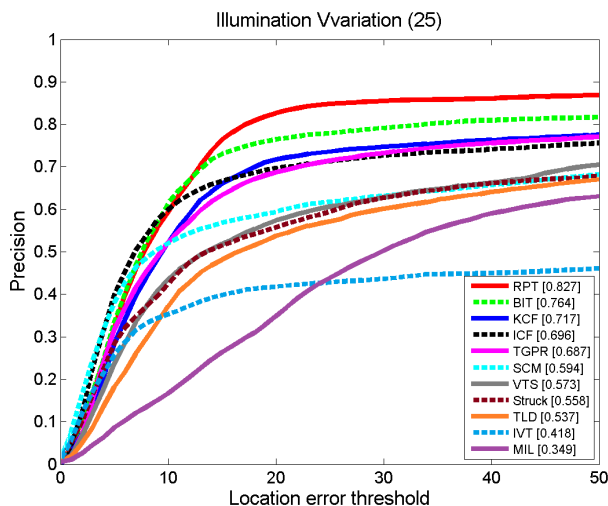


图 4-5 TB50 中光照变化 (IV) 的预测曲线

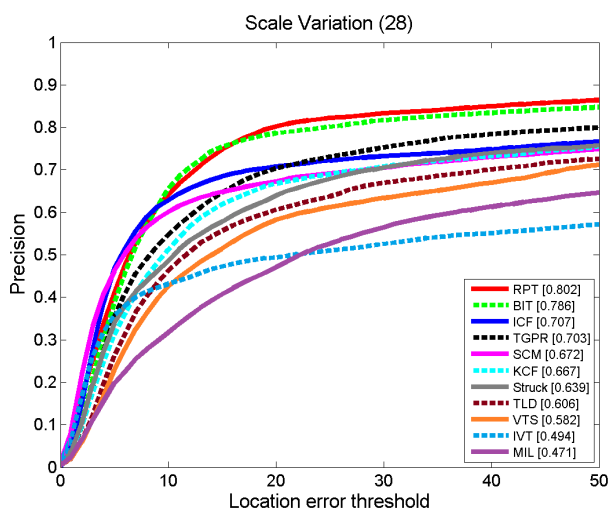


图 4-6 TB50 中尺度变化 (SV) 的预测曲线

敏感的鲁棒特征。RPT 在光照变化中也表现出鲁棒性，来自其局部碎片跟踪机制，局部块的特征提取与跟踪可以很好地应对光照不均和光照变化。

图4-6为尺度变化类别的对比结果，BIT 取得 78.6% 的预测准确率，仅略低于 RPT 的 80.2%。由于 BIT 算法的 S1 单元采用了多尺度的 Gabor 滤波器，并在 C1 单元中对多尺度空间采用 AVG 池化，保证了对于尺度变化的鲁棒性。RPT 方法有别于本文简单地采用尺度鲁棒的特征，而是采用了复杂的局部跟踪来估计整体尺度状态，因此能更好地应对尺度变化。

图4-7为局部遮挡类别任务的测试对比，BIT 取得 85.4% 的预测准确率，远高于其他所有跟踪算法。对于局部遮挡的鲁棒性得益于 BIT 算法中跟踪模型的 C2 单元。C2 单元采用一个全连接的卷积神经层，当目标出现遮挡时只能抑制部分神经元的响应，此时没受到遮挡的神经元仍然可以激活以实现目标的精确跟踪。因此，本文提出的 BIT

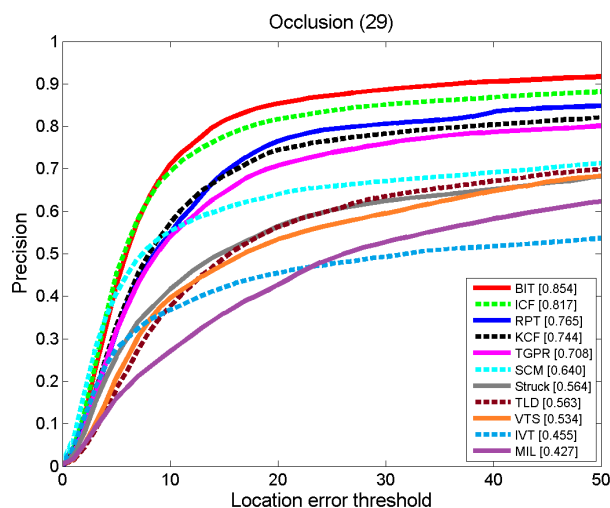


图 4-7 TB50 中局部遮挡 (OCC) 的预测曲线

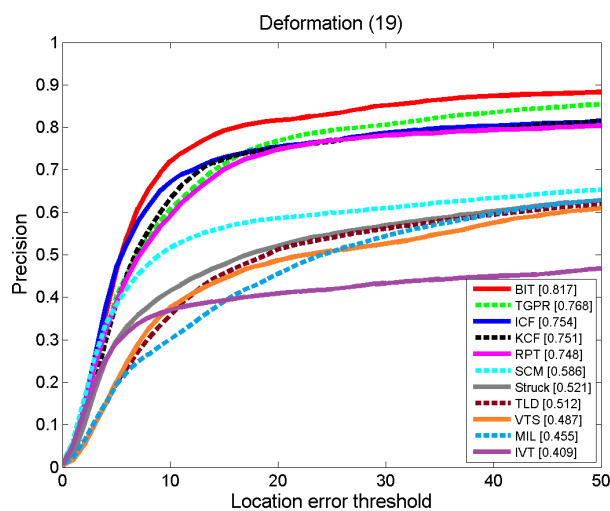


图 4-8 TB50 中外观形变 (DEF) 的预测曲线

取得了比第二名算法 (ICF, 81.7%) 高出 3.7% 的预测精度。

图4-8为外观形变类别视频的视频测试对比结果, BIT 取得 81.7% 的预测准确率, 也远高于其他所有跟踪算法。由于 BIT 算法中外观模型中的 S1 单元采用的 Gabor 滤波器和 C1 单元中的 STD 池化, 提供了对于形变鲁棒的特征表达。其中 Gabor 滤波器形变所造成的纹理变化不敏感, STD 池化获得了更大的特征感受野对于小范围形变鲁棒。排名第二的 TGPR 通过迁移长短时的模板记忆, 获得更能刻画物体本质的空间映射, 取得了 76.8% 的预测精度, 比 BIT 低了 4.9%。

图4-9分别为运动模糊和快速运动类别视频的视频测试对比结果, BIT 在这两类测评中均取得第二的预测精度 (分别为 66.3% 和 64.3%), 低于第一位的 RPT (78.5% 和 74.5%)。BIT 算法本身并没有特别针对于快速运动或者运动模糊的处理模块, 能取得较好的跟踪精度得益于 BIT 的两个实时加速方法使其成为一个可以密集采样的跟踪算

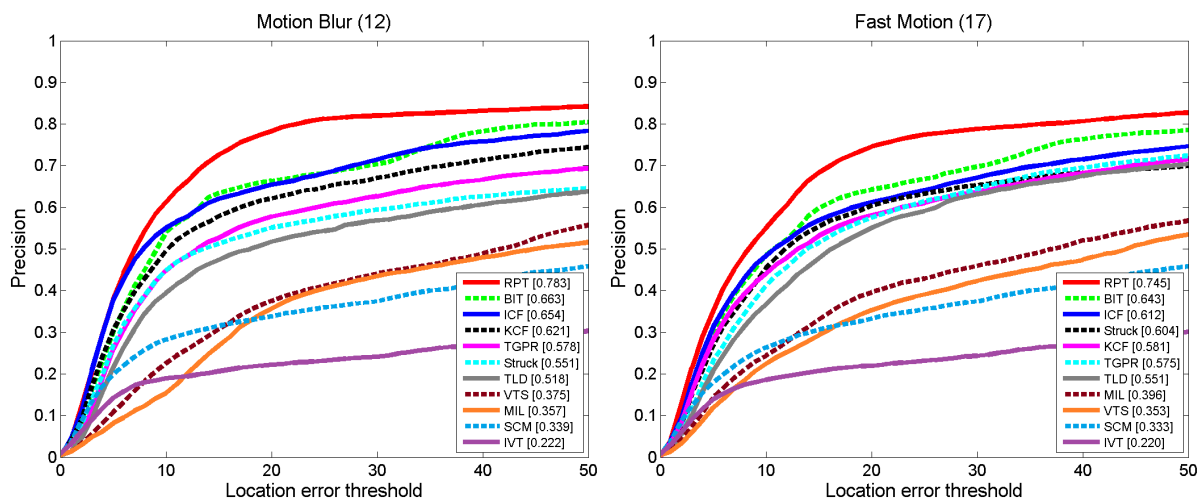


图 4-9 TB50 中运动模糊 (MB) 和快速运动 (FM) 的预测曲线

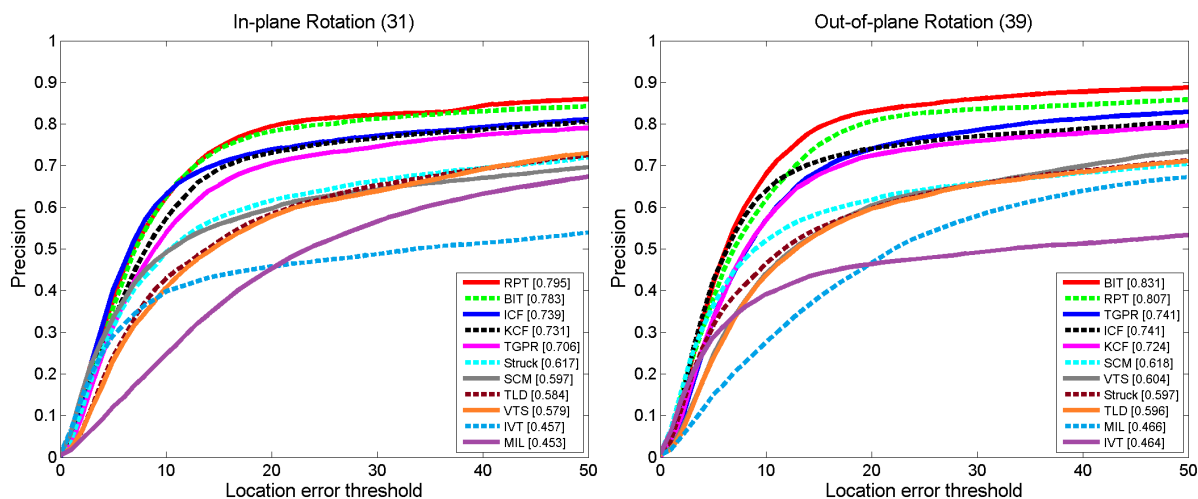


图 4-10 TB50 中旋转 (IPR/OPR) 的预测曲线

法。快速运动和运动模糊共同存在的挑战是跟踪目标的运动速度过快，基于稀疏采样方法（如 TGPR, SCM, VTS, MIL, IVT）因为撒粒子空间不够大而导致采样框无法落到目标位置上，从而导致跟踪失败。RPT 方法在这两类测评中均取得最好的结果，得益于其同样密集采样的方法，并且引进多个更大尺度碎片跟踪辅助整体跟踪，保证了对于大空间的搜索能力，但同时也损失了实时性能（详见节 4.2.4 的实时性分析）。

图4-10分别为二维旋转和三维旋转下算法对比结果，BIT 在这两类评测中均取得优异的结果，其中对于二维旋转取得 78.3% 的预测精度，仅略低于 RPT 的 79.5%；对于三维旋转取得 83.1% 远高于第二位的 80.7%。BIT 对于旋转的鲁棒性来自于 S1 单元多方向 Gabor 滤波器和 S2 单元对于多方向滤波器的 AVG 池化。S2 单元中的 AVG 池化实现了不同类型 Gabor 特征的融合，实现方向激励的竞争选择。

图4-11为超越视野类别的测试对比，BIT 取得 65.4% 最高的预测准确率。对于超出

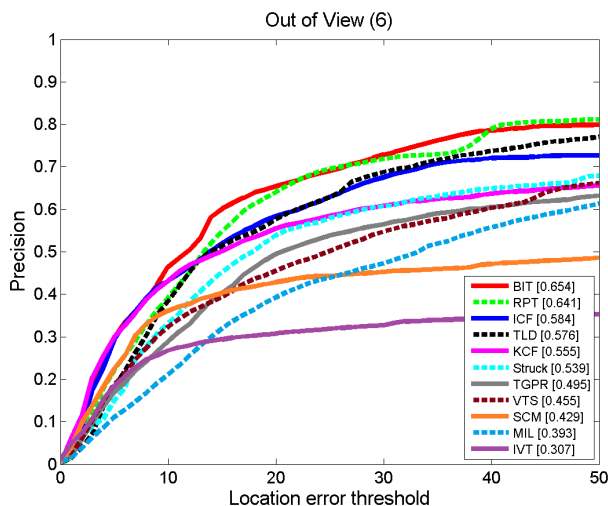


图 4-11 TB50 中超越视野 (OV) 的预测曲线

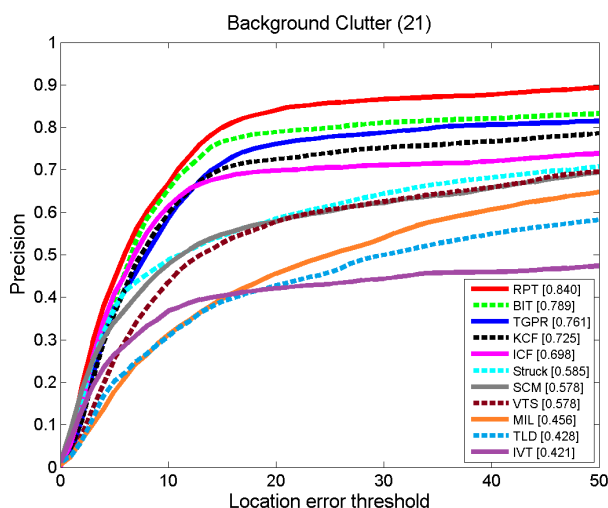


图 4-12 TB50 中背景干扰 (BC) 的预测曲线

视野的物体跟踪，BIT 中融合了生成模型 (S2 单元) 与判别模型 (C2 单元) 的 BITM 起了关键作用。当跟踪目标超出视野或被背景完全遮挡，单一的跟踪模型容易将背景信息当作目标信息来学习，而对于 BIT 的混合跟踪模型在做目标匹配的同时也判别背景与目标，在保证稳定跟踪的同时够避免单一判别模型对于背景信息错误学习。

图4-12为背景干扰类别视频的测试对比结果，BIT 取得 78.9% 的预测准确率，仅低于第一位 RPT 方法 84.0%。BIT 的生成模型对于背景干扰的挑战起到主要作用，生成模型能够更好地发现目标细节，避免背景中相似目标的干扰。RPT 方法通过扩大感受碎片，采用远大于跟踪目标尺度的碎片可以更好地学习到背景的干扰信息。

图4-13为低分辨率类别视频的测试对比结果，这是唯一一种 BIT 并不能取得较好跟踪结果的挑战类别。因为 BIT 中的表观模型中的 C2 单元采用了池化来保证特征鲁棒性，同时也损失了跟踪目标的分辨率。因此并不能对于低分辨率目标很好跟踪，只取得

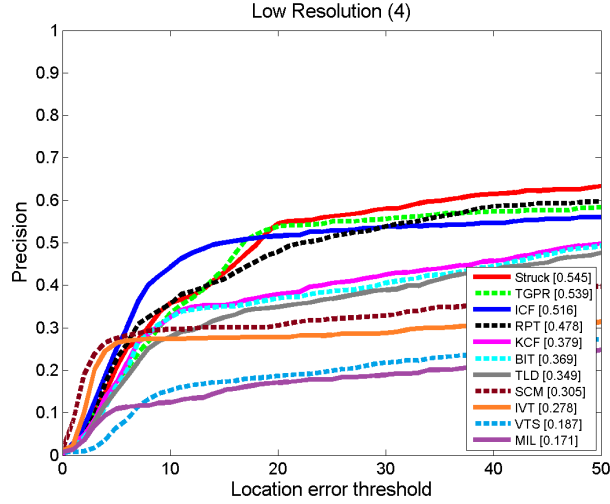


图 4-13 TB50 中低分辨率 (LR) 的预测曲线

36.9% 的预测精度。

### 4.2.3 BIT 模型分析

根据第一章介绍,跟踪模型可以分为生成模型和判别模型。对于生成模型,跟踪器在上一帧跟踪目标领域内搜索与目标最接近作为跟踪的结果。对于判别模型,跟踪器学习一个目标与背景的分类器,判别性地定位到与目标最接近且与背景差异最大的跟踪结果。然而,生成模型忽略了背景中负样本对于跟踪目标定位的辅助作用,而判别模型关注区分性而忽略目标细节。在本文中提出的 BIT 跟踪算法,在跟踪模型中融合了生成模型和判别模型,其中 S2 单元对应于生成模型, C2 单元对应于判别模型。在视觉调谐学习 (S2 单元) 中,卷积匹配寻找与跟踪模板最相近的目标;在独立任务学习 (C2 单元) 中,在跟踪目标和周围背景间学习一个卷积神经网络分类器。

为了证明本文提出的混合模型能够更好地解决跟踪任务,本小节在 TB50 测试数据库上分别对比剔除掉 C2 单元和 S2 单元后的跟踪性能。对于单生成模型跟踪器 (不包含 C2 单元),目标定位可通过最大化 S2 的响应图得到,表示如下:

$$(\hat{x}, \hat{y}) = \arg \min_{(x,y)} S2(x, y) \quad (4-2)$$

对于单判别模型跟踪器 (不包含 S2 单元),直接学习从 C1 到 C2 的卷积神经网络,因此 C2 的响应图可表示如下:

$$C2(x, y) = 1/K \sum_k \mathcal{F}^{-1} [\mathcal{F} [W(x, y, k)] \odot \mathcal{F} [C1(x, y, k)]] \quad (4-3)$$

图4-14展示了混合模型和单一跟踪模型在 TB50 测试数据下 50 段视频的对比结果。可以明显地看出,混合模型取得 81.7% 的预测精度,远远高于单一模型的 74.9% 和

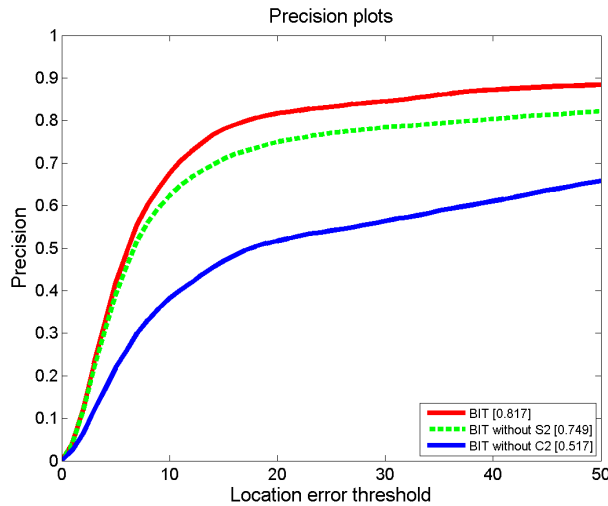


图 4-14 BIT 中 S2 单元和 C2 单元对比分析

51.7%。此外，生成模型与判别模型的预测精度差距为 23.2%，这说明对于单一模型，判别方法可以取得更好的跟踪结果。因为背景信息的引入可以增加样本空间，更好地辅助模型学习。

#### 4.2.4 实时性分析

本节的实时性分析在一台配置 Intel i7 3770 (3.4GHz) 处理器的个人电脑上评测，对比了 BIT 和其他 10 个跟踪算法<sup>注 1</sup>。其中定义平均帧率为在 TB50 中所有视频跟踪的平均帧率，定义实时性的标准为 24 帧/秒（标准电影帧率）。表 4-4 和图 4-15 展示了各个方法的跟踪速度和跟踪精度的对比。其中达到实时标准的跟踪算法有 BIT、ICF、KCF、TLD、MIL、IVT。本文的 BIT 取得 44.9 帧/秒的跟踪速度，位列所有跟踪算法第 3 位，但取得第一的跟踪精度；跟踪速度最高为 KCF，达到 284.4 帧/秒，但跟踪精度位列第五位只达到 73.9%。对比与跟踪精度第二的 RPT 方法只能达到 4.1 帧/秒，BIT 具有明显的性能优势。综上所述，BIT 是一个兼顾跟踪精度与处理速度的优秀跟踪算法，此外，其跟踪速度是实时标准的 2 倍，这为后续的跟踪精度改进留下巨大的空间。

### 4.3 基于深度学习的 BIT

本文提出一种基于深度特征对于 BIT 中生物启发表观模型的改进。通过 VGG-19 在 ILSVRC 2012 预训练的网络模型取代 BIT 框架中 S1 和 C1 单元，实现对于跟踪目标

<sup>注 1</sup>实时性分析基于 TB50<sup>[104]</sup> 提供的工具包 ([http://cvlab.hanyang.ac.kr/tracker\\_benchmark/](http://cvlab.hanyang.ac.kr/tracker_benchmark/))，其中包括 Struck<sup>[108]</sup>、SCM<sup>[32]</sup>、VTS<sup>[109]</sup>、TLD<sup>[39]</sup>、MIL<sup>[38]</sup>、IVT<sup>[22]</sup>。此外 ICF<sup>[96]</sup>、KCF<sup>[7]</sup>、RPT<sup>[106]</sup>、TGPR<sup>[107]</sup> 均采用论文作者公布的代码。

表 4-4 TB50 数据库下各算法的跟踪速度与精度

	跟踪算法	帧率 (fps)	预测精度 (%)	代码类型
实时	BIT	44.9	<b>81.7</b>	MC
	ICF <sup>[96]</sup>	68.8	76.4	MC
	KCF <sup>[7]</sup>	<b>284.4</b>	73.9	MC
	TLD <sup>[39]</sup>	28.1	60.8	MC
	MIL <sup>[38]</sup>	38.1	49.9	C
	IVT <sup>[22]</sup>	33.4	47.5	MC
非实时	RPT	4.1	81.1	MC
	TGPR <sup>[107]</sup>	0.7	76.6	C
	Struck <sup>[108]</sup>	20.2	65.6	C
	SCM <sup>[32]</sup>	0.5	64.9	MC
	VTS <sup>[109]</sup>	5.7	57.5	MC-E

(M: Matlab, C:C/C++, MC: Mixture of Matlab and C, E: Binary code)

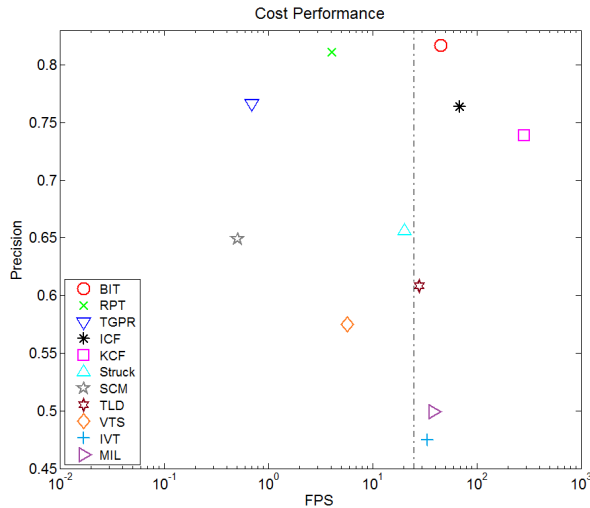


图 4-15 TB50 跟踪数据库下各算法性价比（虚线为实时/非实时的分界线）

的特征提取。本小节在 TB50 数据库上测试改进后 BIT 跟踪算法在 50 段测试视频中的预测精度，其中参与对比算法包括节4.2中的所有算法。此外，采用 VGG-19 网络不同层次的特征输出，分析 CNN 的深度对于跟踪精度的影响。

### 4.3.1 TB50 对比实验

在图4-16为基于深度学习的 BIT 算法（BIT+VGG）与其他跟踪算法在 50 段视频下的测试对比结果。其中，VGG 网络采用 Conv5 的输出，即 VGG-19 最后一个卷积层的输出。BIT+VGG 在中心位置误差 20 阈值下的预测精度为 84.9%，相较于原始的 BIT

算法有了明显的提高（提高了 3.2 个百分点）。以上结果说明了，深度学习的特征比手工的生物启发模型具有更好的鲁棒性。VGG-19 卷积神经网络是生物启发表观模型的一种，因此能很好地与生物启发跟踪模型配合工作。其次，基于学习方法的滤波器相较于 Gabor 滤波器具有更丰富的多样性，因此能够更鲁棒地实现跟踪目标表示。此外，本文从 TB50 数据库的 11 种类别中挑选两个提升最为明显的类别进行分析。

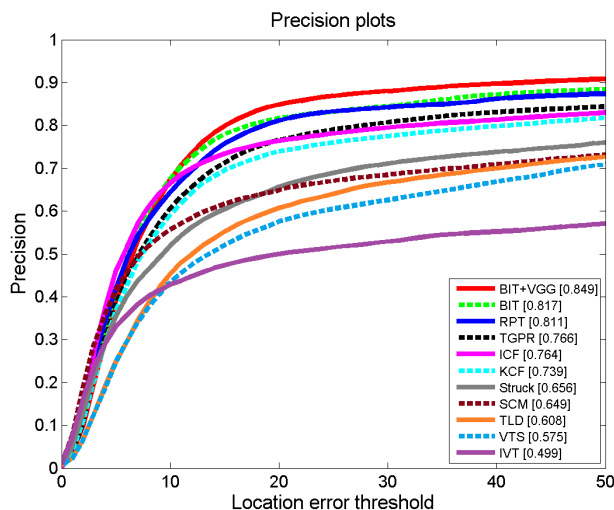


图 4-16 基于深度学习的 BIT 在 TB50 的预测精度对比

其中深度学习的引进对于低分辨率的跟踪视频提升最为明显，图4-17为低分辨率类别的对比结果。基于深度学习的 BIT 改进方法相较于原始 BIT 跟踪算法提高了近 30 个百分点，远高于现有的其他算法。首先这得益于 VGG-19 采用了小尺度的卷积层 ( $3 \times 3$ )，小的卷积核可以更好地刻画地分辨率目标的细节。其次在进行特征抽取前对不同尺度的跟踪目标都采取了尺度归一化 ( $256 \times 256$ ) 以适应 VGG 网络结构，这对于低分辨率的目标近似于超分辨率 (Super-solution) 操作以拟补分辨率的不足。

图4-18为二维旋转类别的测试对比。基于深度学习的 BIT 改进方法取得 82.4% 的预测精度，相较于原始 BIT 跟踪算法提高了近 4 个百分点，同时也高于原处于第一位的 RPT (79.5%)。生物启发表观模型中仅采用了 16 个方向的 Gabor 滤波器对目标进行建模，16 个方向的滤波器并不能很好刻画连续变化的方向特征。深度学习特征在卷积单元中最多采用多达 512 个卷积核，因此能够更好地保证对于旋转的鲁棒性。

### 4.3.2 深度特征分析

节3.1.3中指出不同层次的深度学习特征对于目标的刻画能力不同，本小节对比分析各层次深度特征对于跟踪结果的影响。图4-19为不同层次的 VGG-19 特征在 TB50 下的测试对比结果，其中 Conv3、Conv4、Conv5 分别取得 79.4%、81.3%、84.9% 的预测



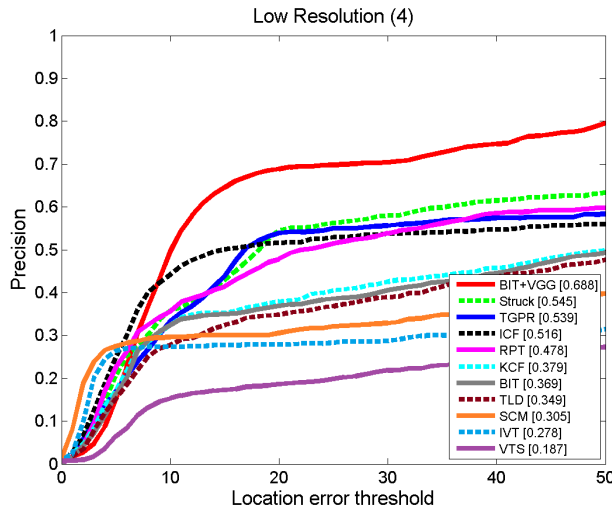


图 4-17 基于深度学习的 BIT 在低分辨率 (LR) 的预测曲线

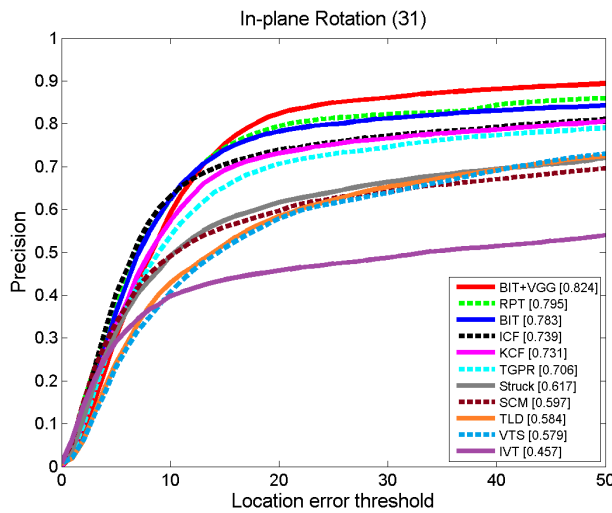


图 4-18 基于深度学习的 BIT 在二维旋转 (IPR) 的预测曲线

精度。由此可见抽象能力越强的表观模型对于目标跟踪越鲁棒，具有高层语义表示的 Conv5 特征可以对跟踪目标的外观更好地建模。此外，跨层 (Cross-layer) 的特征融合将进一步提升跟踪的性能。

此外，Conv3、Conv4 也在跟踪中取得不错的结果，但均略低于采用生物启发特征的 BIT。可见第 2 章提出的生物启发表观模型具有很好的目标刻画能力，可以通过极少的特征参数取得很好的跟踪结果。BIT 在单个 CPU 的运算环境下取得 45 帧/秒的平均跟踪速度，基于深度特征的改进在配备 Nvidia K40 GPU 的环境下仅能取得 11 帧/秒的平均跟踪速度。相比于深度特征，采用生物启发表观模型的 BIT 表现出极好的实时性。

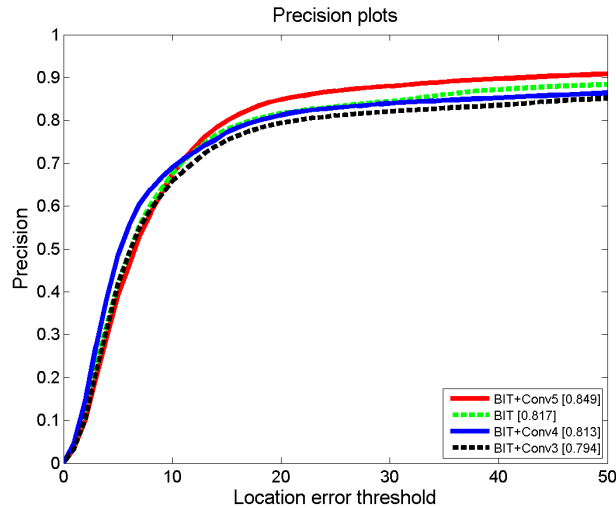


图 4-19 不同 VGG 特征在 BIT 框架下的预测精度对比

## 4.4 自适应尺度的 BIT

本文第 3 章中提出一种基于 DSST 和 SPP 的快速尺度方法应用于 BIT 的尺度自适应。本小节将在 TB50 和 ALOV300++ 上通过对于尺度估计敏感的指标评测对于改进后的 BIT 进行性能评价。其中对比的算法包括 RPT, DSST, ICF 等具有尺度自适应的算法, 同时也包括其他一些优秀的跟踪算法。其中包括: RPT<sup>[106]</sup>、ICF<sup>[96]</sup>, TGPR<sup>[107]</sup>, KCF<sup>[7]</sup>, SCM<sup>[32]</sup>, Struck<sup>[108]</sup>, TLD<sup>[39]</sup>, VTS<sup>[109]</sup>, MIL<sup>[38]</sup>, IVT<sup>[22]</sup>, FBT<sup>[110]</sup>。

### 4.4.1 TB50 对比实验

在图4-20中所示为加入尺度自适应后的 BIT 方法与其他 10 个跟踪算法在 TB50 数据库下 50 段视频的测试对比结果。本文算法 BIT 在 50% 重叠率下的成功率为 77%, 高于第二位的 RPT 算法 73% 四个百分点。对比与预测精度评测下 BIT 与 RPT 的差距并不明显, 可以看出本文提出的尺度自适应方法对结果有显著的提高。图4-20中, 中括号内的数字表示各个算法成功率曲线下面积 (Area Under Curve, AUC), 描述在各个阈值下平均的跟踪成功率。本文算法 BIT 的 AUC 值为 0.593, 同样也高于第二位的 RPT 算法 0.575。此外, BIT 与 DSST 的改进算法 ICF 做对比, ICF 在 TB50 测试数据库上的 AUC 值只取得 0.563, 远低于本文提出的自适应尺度的 BIT 算法。综上所述, 本文第 3 章所提出的自适应尺度方法为生物启发跟踪器引入了尺度估计功能, 并且能有效且高速地进行尺度空间的估计, 相比于其他尺度估计算法表现出更好的性能。

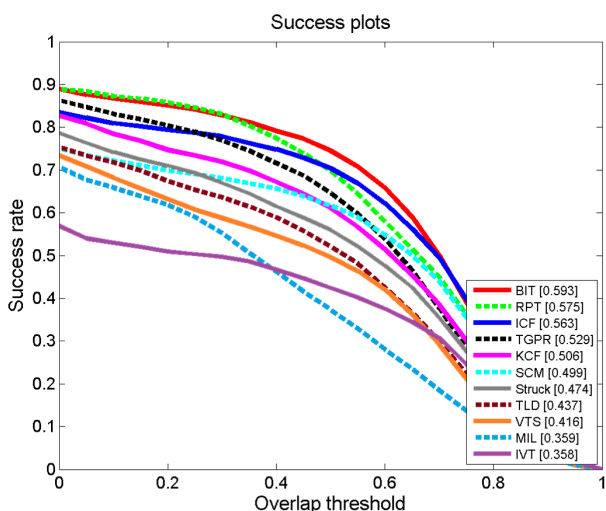


图 4-20 BIT 与其他算法在 TB50 跟踪数据库上的成功率图

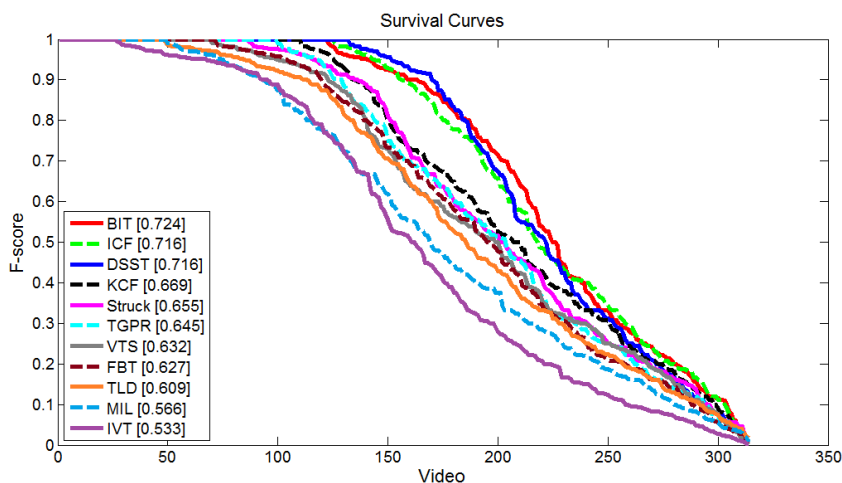


图 4-21 ALOV300++ 数据库下各个算法生存曲线

#### 4.4.2 ALOV300++ 对比实验

图4-21为加入尺度自适应后的 BIT 方法与其他 10 个跟踪算法在 ALOV300++ 数据库下 314 段视频的测试对比结果。当跟踪算法的预测结果与标定结果重叠率大于 50% 时认为跟踪正确，并计算各个算法的 F-Score。此外，采用生存曲线（Survival curve）可视化对比各个算法的跟踪成功率。生存曲线通过按各个视频样本的预测成功率排列，表示达到某一成功率以上的视频样本数。本文提出的自适应尺度改进的 BIT 在 50% 重叠率的下取得最高的 F-Score 为 0.724，均高于 ICF 和 DSST 的 0.716。实验结果表明，基于自适应尺度改进的 BIT 在大数据集测试上也能高效地进行尺度估计，并取得与其他算法更高的准确率。

## 4.5 本章小结

本章基于 TB50 和 ALOV300++ 两个大型的跟踪测试数据库，详细地分析了本文提出的生物启发跟踪器（BIT）的跟踪性能，其中包括综合性能分析、模块化分析、实时性分析等。此外，基于深度学习和尺度自适应的改进上做了对比分析。综合以上实验结果表明，本文提出的 BIT 方法在综合性能上相比于现有的其他跟踪算法取得了更优的效果。

## 第五章 总结与展望

### 5.1 本文总结

随着计算机科学和人工智能的高速发展，机器视觉逐渐从研究阶段正式走向产业化实施。目标跟踪作为机器视觉领域三大核心问题之一，在视频监控、人机交互、智能驾驶等领域有着重要的应用前景。然而，对于非约束环境下的重重挑战，相比于特定目标的检测、识别任务，通用化的跟踪算法仍有待深入研究。近年来，国内外研究者在目标跟踪领域，提出了一系列具有创新性和实用性的目标跟踪算法。然而，受限于传统方法的局限性，没有单个算法能完美地解决非约束环境中各种各样的挑战，如光线变化、尺度变化、外观形变等。受到灵长动物视皮层感知机制的启发，本文引入生物学研究理论设计了一个新的视觉跟踪算法——生物启发跟踪器（BIT）。

本文主要围绕视觉跟踪算法中的表观模型和跟踪模型，基于生物启发模型提出一种新的视觉跟踪算法 BIT。本文主要的工作包括：

- (1) 本文重点提出一种新的目标跟踪算法。生物启发跟踪器（Biologically inspired tracker, BIT）是基于 HMAX 模型，模仿灵长动物视皮层中腹侧流通路的感知机制设计的视觉跟踪器。借鉴于 HMAX 模型在机器视觉领域中的成功应用，BIT 以 HMAX 模型为基础进行针对于视觉跟踪任务的设计和改造，包括生物启发表观模型（BIAM）与生物启发跟踪模型（BITM）。其中，BIAM 主要模仿视皮层初级简单细胞（S1 单元）和视皮层复杂细胞（C1 单元），实现对物体外观特征的鲁棒提取；BITM 主要模仿高级视皮层中的视觉调谐学习（S2 单元）和独立任务学习（C2 单元）此外，为了保证跟踪算法的实时性，本文提出了一种快速 Gabor 近似（FGA）方法优化 BIAM 模块的特征提取效率，并应用快速傅里叶变化（FFT）实现 BITM 模块的算法加速，使 BIT 成为一个可以真正实用的实时视觉跟踪器。在 TB50 综合性实验中表明，相比于现有的其他跟踪算法，BIT 是一个兼顾性能和效率的优秀跟踪器。
- (2) 针对 BIT 的表观模型采用手工特征的不足，本文引进深度学习特征取代 S1 和 C1 单元对其进行改进。本文采用在 ILSVC2012 上预训练的神经网络取 VGG-19 代生物启发表观模型，其中卷积层对应于 S1 单元，池化层对应于 C1 单元。在 TB50 数据库上评测，基于深度学习特征改进的 BIT 有了性能上的提升，取得了

最优的跟踪结果，并高于目前的其它跟踪算法。此外分层特征的实验对比说明，高层的语义特征对于跟踪具有更好的鲁棒性。

- (3) 针对 BIT 的跟踪模型在尺度估计上的问题，引进 DSST 提供独立的尺度估计。此外，本文在 DSST 的基础在尺度特征上进行改进，针对原尺度特征（HOG）计算复杂，重复裁剪和缩放的问题，提出了一种快速尺度特征。快速尺度特征引入 SPP 的思想，采用空间金字塔池化获得对于尺度敏感的特征，在保证尺度估计精度上避免了原来特征的计算冗余。本文在实验部分针对这一问题提出了改进的方案，使得独立双尺度空间尺度自适应算法能够更好地应对不同的跟踪情况。在 TB50 和 ALOV300++ 两个数据库上评测，基于尺度自适应改进的 BIT 其它算法的尺度估计更为精确，取得了鲁棒的跟踪结果。

## 5.2 未来展望

随着计算机与人工智能的发展，机器视觉将成为推动社会进步，解放和发展生产力的动力，目标跟踪技术的市场需求将推动其研究热度的持续。本文提出的 BIT 虽然在多个测试任务中有着优异的表现，但是依然存在以下瓶颈及可能的改进方向：

- (1) 生物启发模型中的 C2 特征是通过选择成千上万个认知碎片实现局部到整体的判别。本文中提出的 BIT 仅仅采用一个整体的认知碎片，以保证跟踪的实时性。但在目标跟踪中，局部建模被证明是有效<sup>[106]</sup>：一方面，可以应对目标较大幅度的旋转等形变；另一方面，可以应对局部变化（光照不均，局部遮挡等）。在未来的工作中拟引入图论、拓扑学等知识，研究融合 C2 单元的认知碎片改进，并解决局部到整体的匹配难题。
- (2) 神经科学和解剖学研究指出，视觉信息在大脑中按照腹侧（形状）与背侧（运动）两条通路进行传递，并在高层感知中融合。双通路的生物启发模型<sup>[111]</sup>在行为识别中得到了成功的应用。视觉跟踪本身是一个形状和运动的信息融合判别，本文提出的 BIT 框架忽略了背侧流对于运动信息的敏感。在今后的研究中，将通过模拟腹侧通路视皮层的 MT 区、MST 区等视皮层的反应，融合双通路模型优化视觉跟踪。
- (3) 长短时记忆池机制是神经系统有别现有传统模型的优点，长短时模型是解决长时（Long-term）跟踪问题的重点。近年来的众多跟踪算法开始引入长短时的记

忆模型，其中包括 MUSTer<sup>[96]</sup>，LTCT<sup>[112]</sup>，TGPR<sup>[107]</sup> 等。PFC 区的记忆池效应将为生物启发跟踪器引入记忆模型提供依据，将成为今后研究的方向。

## 参考文献

- [1] Yilmaz A, Javed O, Shah M. Object tracking: A survey[J]. *Acm computing surveys (CSUR)*, 2006, 38(4):13.
- [2] Wang N, Shi J, Yeung D Y, et al. Understanding and diagnosing visual tracking systems[A]. In: *Proceedings of the IEEE International Conference on Computer Vision[C]*, 2015. 3101–3109.
- [3] Bradski G R. Computer vision face tracking for use in a perceptual user interface[J]. 1998.
- [4] Sevilla-Lara L, Learned-Miller E. Distribution fields for tracking[A]. In: *Computer Vision and Pattern Recognition[C]*, 2012. 1910–1917.
- [5] Leichter I. Mean shift trackers with cross-bin metrics[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012, 34(4):695–706.
- [6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[A]. In: *Computer Vision and Pattern Recognition[C]*, 2005. 1:886–893.
- [7] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. *Pattern Analysis and Machine Intelligence*, 2015, 37(3):583–596.
- [8] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. *Pattern Analysis and Machine Intelligence*, 2002, 24(7):971–987.
- [9] Dinh T B, Vo N, Medioni G. Context tracker: Exploring supporters and distracters in unconstrained environments[A]. In: *Computer Vision and Pattern Recognition[C]*, 2011. 1177–1184.
- [10] Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection[A]. In: *ICIP[C]*, 2002. 1:I–900.
- [11] Zhang K, Zhang L, Yang M H. Real-time compressive tracking[M]//. In: *Computer Vision–ECCV[C]*. [S.l.]: Springer, 2012. 864–877.
- [12] Ng P C, Henikoff S. SIFT: Predicting amino acid changes that affect protein function[J]. *Nucleic acids research*, 2003, 31(13):3812–3814.
- [13] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[M]//. In: *Computer vision–ECCV 2006[C]*. [S.l.]: Springer, 2006. 404–417.
- [14] Cao X, Tao Z, Zhang B, et al. Self-Adaptively Weighted Co-Saliency Detection via Rank Constraint[J]. *Image Processing, IEEE Transactions on*, 2014, 23(9):4175–4186.



- 
- [15] Zhou H, Yuan Y, Shi C. Object tracking using SIFT features and mean shift[J]. *Computer vision and image understanding*, 2009, 113(3):345–352.
- [16] Tang F, Tao H. Probabilistic object tracking with dynamic attributed relational feature graph[J]. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2008, 18(8):1064–1074.
- [17] He W, Yamashita T, Lu H, et al. Surf tracking[A]. In: *Computer Vision, 2009 IEEE 12th International Conference on[C]*, 2009. 1586–1592.
- [18] Grabner H, Roth P M, Bischof H. Eigenboosting: Combining discriminative and generative information[A]. In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on[C]*, 2007. 1–8.
- [19] Wang S, Lu H, Yang F, et al. Superpixel tracking[A]. In: *Computer Vision (ICCV), 2011 IEEE International Conference on[C]*, 2011. 1323–1330.
- [20] Yang F, Lu H, Yang M H. Robust superpixel tracking[J]. *Image Processing, IEEE Transactions on*, 2014, 23(4):1639–1651.
- [21] Yang M, Yuan J, Wu Y. Spatial selection for attentional visual tracking[A]. In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on[C]*, 2007. 1–8.
- [22] Ross D A, Lim J, Lin R S, et al. Incremental learning for robust visual tracking[J]. *International Journal of Computer Vision*, 2008, 77(1-3):125–141.
- [23] Skocaj D, Leonardis A. Weighted and robust incremental method for subspace learning[A]. In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on[C]*, 2003. 1494–1501.
- [24] Brand M. Incremental singular value decomposition of uncertain data with missing values[M]// In: *Computer Vision—ECCV 2002[C]*. [S.l.]: Springer, 2002. 707–720.
- [25] Levey A, Lindenbaum M. Sequential Karhunen-Loeve basis extraction and its application to images[J]. *Image Processing, IEEE Transactions on*, 2000, 9(8):1371–1374.
- [26] Ross D A, Lim J, Lin R S, et al. Incremental learning for robust visual tracking[J]. *International Journal of Computer Vision*, 2008, 77(1-3):125–141.
- [27] Li X, Hu W, Zhang Z, et al. Robust visual tracking based on incremental tensor subspace learning[A]. In: *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on[C]*, 2007. 1–8.
- [28] Wen J, Li X, Gao X, et al. Incremental learning of weighted tensor subspace for visual tracking[A]. In: *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on[C]*, 2009. 3688–

3693.

- [29] Donoho D L. Compressed sensing[J]. Information Theory, IEEE Transactions on, 2006, 52(4):1289–1306.
- [30] Mei X, Ling H, Wu Y, et al. Minimum error bounded efficient l1 tracker with occlusion detection[A]. In: Computer Vision and Pattern Recognition[C], 2011. 1257–1264.
- [31] Zhang T, Ghanem B, Liu S, et al. Robust visual tracking via multi-task sparse learning[A]. In: Computer Vision and Pattern Recognition[C], 2012. 2042–2049.
- [32] Zhong W, Lu H, Yang M H. Robust object tracking via sparsity-based collaborative model[A]. In: Computer Vision and Pattern Recognition[C], 2012. 1838–1845.
- [33] Li H, Shen C, Shi Q. Real-time visual tracking using compressive sensing[A]. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on[C], 2011. 1305–1312.
- [34] Jin J, Dundar A, Bates J, et al. Tracking with deep neural networks[A]. In: Information Sciences and Systems (CISS), 2013 47th Annual Conference on[C], 2013. 1–5.
- [35] Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[A]. In: Advances in Neural Information Processing Systems[C], 2013. 809–817.
- [36] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[A]. In: Computer Vision and Pattern Recognition[C], 2009. 248–255.
- [37] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking[J]. Pattern Analysis and Machine Intelligence, 2003, 25(5):564–577.
- [38] Babenko B, Yang M H, Belongie S. Visual tracking with online multiple instance learning[A]. In: Computer Vision and Pattern Recognition[C], 2009. 983–990.
- [39] Kalal Z, Matas J, Mikolajczyk K. P-N learning: Bootstrapping binary classifiers by structural constraints[A]. In: Computer Vision and Pattern Recognition[C], 2010. 49–56.
- [40] Wang D, Lu H, Yang M H. Online object tracking with sparse prototypes[J]. Image Processing, IEEE Transactions on, 2013, 22(1):314–325.
- [41] Suykens J A, Vandewalle J. Least squares support vector machine classifiers[J]. Neural processing letters, 1999, 9(3):293–300.
- [42] Zhang S, Yu X, Sui Y, et al. Object Tracking With Multi-View Support Vector Machines[J]. Multimedia, IEEE Transactions on, 2015, 17(3):265–278.

- 
- [43] Grabner H, Grabner M, Bischof H. Real-Time Tracking via On-line Boosting.[A]. In: BMVC[C], 2006. 1:6.
- [44] Grabner H, Leistner C, Bischof H. Semi-supervised on-line boosting for robust tracking[M]//. In: Computer Vision–ECCV[C].[S.l.]: Springer, 2008. 234–247.
- [45] Avidan S. Support vector tracking[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004, 26(8):1064–1072.
- [46] Adam A, Rivlin E, Shimshoni I. Robust fragments-based tracking using the integral histogram[A]. In: Computer Vision and Pattern Recognition[C], 2006. 1:798–805.
- [47] Kwon J, Lee K M. Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling[A]. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on[C], 2009. 1208–1215.
- [48] Li Y, Ai H, Yamashita T, et al. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2008, 30(10):1728–1740.
- [49] Panizza B. Osservazioni sul nervo ottico[M].[S.l.]: Bernardoni, 1855.
- [50] Barlow H B. Summation and inhibition in the frog's retina[J]. The Journal of physiology, 1953, 119(1): 69.
- [51] McIlwain J, Buser P. Receptive fields of single cells in the cat's superior colliculus[J]. Experimental brain research, 1968, 5(4):314–325.
- [52] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of physiology, 1962, 160(1):106–154.
- [53] Mishkin M, Ungerleider L G, Macko K A. Object vision and spatial vision: two cortical pathways[J]. Trends in neurosciences, 1983, 6:414–417.
- [54] Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex[J]. Nature neuroscience, 1999, 2(11):1019–1025.
- [55] Li C Y. Integration fields beyond the classical receptive field: organization and functional properties[J]. Physiology, 1996, 11(4):181–186.
- [56] Serre T, Kouh M, Cadieu C, et al. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex[R].[S.l.]: DTIC Document, 2005.

- [57] Grill-Spector K, Malach R. The human visual cortex[J]. *Annu. Rev. Neurosci.*, 2004, 27:649–677.
- [58] McAlonan K, Cavanaugh J, Wurtz R H. Guarding the gateway to cortex with attention in visual thalamus[J]. *Nature*, 2008, 456(7220):391–394.
- [59] Serre T, Riesenhuber M. Realistic modeling of simple and complex cell tuning in the HMAX model, and implications for invariant object recognition in cortex[R].[S.l.]: DTIC Document, 2004.
- [60] Culham J C, Danckert S L, De Souza J F, et al. Visually guided grasping produces fMRI activation in dorsal but not ventral stream brain areas[J]. *Experimental Brain Research*, 2003, 153(2):180–189.
- [61] Granlund G H. In search of a general picture processing operator[J]. *Computer Graphics and Image Processing*, 1978, 8(2):155–173.
- [62] Knutsson H E, Wilson R, Granlund G H. Anisotropic Nonstationary Image Estimation and Its Applications: Part I—Restoration of Noisy Images[J]. *Communications, IEEE Transactions on*, 1983, 31(3):388–397.
- [63] Daugman J G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters[J]. *JOSA A*, 1985, 2(7):1160–1169.
- [64] Gabor D. Theory of communication. Part 1: The analysis of information[J]. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 1946, 93(26):429–441.
- [65] Serre T, Wolf L, Bileschi S, et al. Robust object recognition with cortex-like mechanisms[J]. *Pattern Analysis and Machine Intelligence*, 2007, 29(3):411–426.
- [66] Siagian C, Itti L. Rapid biologically-inspired scene classification using features shared with visual attention[J]. *Pattern Analysis and Machine Intelligence*, 2007, 29(2):300–312.
- [67] Xu G, Xu X, Xing X, et al. Multi-invariance appearance model for object tracking[A]. In: *Digital Signal Processing (DSP), 2015 IEEE International Conference on[C]*, 2015. 347–351.
- [68] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1998(11):1254–1259.
- [69] Van De Weijer J, Schmid C, Verbeek J, et al. Learning color names for real-world applications[J]. *Image Processing, IEEE Transactions on*, 2009, 18(7):1512–1523.
- [70] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. *The Journal of physiology*, 1962, 160(1):106.
- [71] Lampl I, Ferster D, Poggio T, et al. Intracellular measurements of spatial integration and the MAX

- operation in complex cells of the cat primary visual cortex[J]. *Journal of neurophysiology*, 2004, 92(5): 2704–2713.
- [72] Guo G, Mu G, Fu Y, et al. Human age estimation using bio-inspired features[A]. In: *Computer Vision and Pattern Recognition*[C], 2009. 112–119.
- [73] Schwartz E L. Anatomical and physiological correlates of visual computation from striate to infero-temporal cortex[J]. *Systems, Man and Cybernetics, IEEE Transactions on*, 1984(2):257–271.
- [74] Fuster J M. *Prefrontal cortex*[M].[S.l.]: Springer, 1988.
- [75] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[A]. In: *Advances in neural information processing systems*[C], 2012. 1097–1105.
- [76] Ruck D W, Rogers S K, Kabrisky M, et al. The multilayer perceptron as an approximation to a Bayes optimal discriminant function[J]. *Neural Networks, IEEE Transactions on*, 1990, 1(4):296–298.
- [77] Li M, Zhang Z, Huang K, et al. Robust visual tracking based on simplified biologically inspired features[A]. In: *ICIP*[C], 2009. 4113–4116.
- [78] Mahadevan V, Vasconcelos N. Biologically inspired object tracking using center-surround saliency mechanisms[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2013, 35(3):541–554.
- [79] Jia X, Lu H, Yang M H. Visual tracking via adaptive structural local sparse appearance model[A]. In: *Computer Vision and Pattern Recognition*[C], 2012. 1822–1829.
- [80] Haykin S, Van Veen B. *Signals and systems*[M].[S.l.]: John Wiley & Sons, 2007.
- [81] Nussbaumer H J. *Fast Fourier transform and convolution algorithms*[M]. Vol. 2.[S.l.]: Springer Science & Business Media, 2012.
- [82] Cooley J W, Tukey J W. An algorithm for the machine calculation of complex Fourier series[J]. *Mathematics of computation*, 1965, 19(90):297–301.
- [83] Wang S C. *Artificial neural network*[M]//. In: *Interdisciplinary Computing in Java Programming*[C]. [S.l.]: Springer, 2003. 81–100.
- [84] Hubel D H, Wiesel T N. Ferrier lecture: Functional architecture of macaque monkey visual cortex[J]. *Proceedings of the Royal Society of London B: Biological Sciences*, 1977, 198(1130):1–59.
- [85] Olshausen B A, et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. *Nature*, 1996, 381(6583):607–609.

- [86] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. *Biological cybernetics*, 1980, 36(4):193–202.
- [87] Hosmer Jr D W, Lemeshow S. *Applied logistic regression*[M].[S.l.]: John Wiley & Sons, 2004.
- [88] Chauvin Y, Rumelhart D E. *Backpropagation: theory, architectures, and applications*[M].[S.l.]: Psychology Press, 1995.
- [89] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11):2278–2324.
- [90] Le Cun B B, Denker J S, Henderson D, et al. Handwritten digit recognition with a back-propagation network[A]. In: *Advances in neural information processing systems*[C], 1990.
- [91] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv preprint arXiv:1409.1556*, 2014.
- [92] Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models[A]. In: *Proc. ICML*[C], 2013. 30:1.
- [93] Zhang K, Zhang L, Liu Q, et al. Fast visual tracking via dense spatio-temporal context learning[M]//. In: *Computer Vision–ECCV 2014*[C].[S.l.]: Springer, 2014. 127–141.
- [94] Boddeti V, Kanade T, Kumar B. Correlation filters for object alignment[A]. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*[C], 2013. 2291–2298.
- [95] Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking[A]. In: *British Machine Vision Conference, Nottingham, September 1-5, 2014*[C], 2014.
- [96] Hong Z, Chen Z, Wang C, et al. MUlti-Store Tracker (MUSTer): A Cognitive Psychology Inspired Approach to Object Tracking[A]. In: *Computer Vision and Pattern Recognition*[C], 2015.
- [97] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015, 37(9):1904–1916.
- [98] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[A]. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*[C], 2006. 2:2169–2178.
- [99] Wallach H M. Topic modeling: beyond bag-of-words[A]. In: *Proceedings of the 23rd international conference on Machine learning*[C], 2006. 977–984.

- 
- [100] Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel[A]. In: Proceedings of the 6th ACM international conference on Image and video retrieval[C], 2007. 401–408.
- [101] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2016, 38(1):142–158.
- [102] Girshick R. Fast r-cnn[A]. In: Proceedings of the IEEE International Conference on Computer Vision[C], 2015. 1440–1448.
- [103] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[A]. In: *Advances in Neural Information Processing Systems*[C], 2015. 91–99.
- [104] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[A]. In: *Computer Vision and Pattern Recognition*[C], 2013. 2411–2418.
- [105] Smeulders A W, Chu D M, Cucchiara R, et al. Visual tracking: an experimental survey[J]. *Pattern Analysis and Machine Intelligence*, 2014, 36(7):1442–1468.
- [106] Li Y, Zhu J, Hoi S C. Reliable Patch Trackers: Robust Visual Tracking by Exploiting Reliable Patches[A]. In: *Computer Vision and Pattern Recognition*[C], 2015.
- [107] Gao J, Ling H, Hu W, et al. Transfer learning based visual tracking with gaussian processes regression[M]//. In: *Computer Vision–ECCV*[C],[S.l.]: Springer, 2014. 188–203.
- [108] Hare S, Saffari A, Torr P H. Struck: Structured output tracking with kernels[A]. In: *International Conference on Computer Vision*[C], 2011. 263–270.
- [109] Kwon J, Lee K M. Tracking by sampling trackers[A]. In: *ICCV*[C], 2011. 1195–1202.
- [110] Nguyen H T, Smeulders A W. Robust tracking using foreground-background texture discrimination[J]. *International Journal of Computer Vision*, 2006, 69(3):277–293.
- [111] Cai B, Xu X, Qing C. Bio-inspired model with dual visual pathways for human action recognition[A]. In: *Communication Systems, Networks & Digital Signal Processing (CSNDSP), 2014 9th International Symposium on*[C], 2014. 271–276.
- [112] Ma C, Yang X, Zhang C, et al. Long-term correlation tracking[A]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C], 2015. 5388–5396.

## 攻读硕士学位期间取得的研究成果

已发表（包括已接受待发表）的论文，以及已投稿、或已成文打算投稿、或拟成文投稿的论文情况：

序号	作者	题目	发表或投稿刊物名称、级别	发表的卷期、年月、页码	相当于学位论文的哪一部分	被索引收录情况
1	<b>B. Cai, X. Xu, X. Xing, K. Jia, J. Miao, D. Tao</b>	BIT: Biologically Inspired Tracker	IEEE Transactions on Image Processing (IF=3.63)	vol. 25(3) 206.03 1327-1339	第 2 章	SCI 二区
2	<b>B. Cai, X. Xu, X. Xing, C. Qing</b>	BIT: Bio-inspired tracker	IEEE International Conference on Image Processing	2015.09 2850-2854	2.2 节	EI
3	G. Xu, X. Xu, X. Xing, <b>B. Cai, C. Qing</b>	Multi-invariance appearance model for object tracking	IEEE International Conference on Digital Signal Processing	2015.07 347-351	2.1 节	EI
4	<b>B. Cai, X. Xu, C. Qing</b>	Bio-inspired model with dual visual pathways for human action recognition	International Symposium on CSNDSP	2014.09 271-276	2.1 节	EI



## 致谢

光阴荏苒，三年即逝，时间划过了指尖，岁月徒留发际线。值此拙文完成之际，回想起在华南理工大学度过的青春年华，心中颇为感慨，衷心地向各位亲友、老师和同学致以诚挚的敬意和谢意。

首先，感谢华南理工大学电子与信息学院诸位老师的悉心培养和指导。衷心感谢指导老师徐向民教授，在过去三年中对我的关心和帮助，给了我不少指导和支持，使我可以更顺利地开展研究，如期完成毕业设计。此外，感谢导师毕淑娥教授的支持，还有实验室的邢晓芬老师、青春美老师对我的工作上的指导。同时，特别地感谢贾奎老师、陶大程老师在我科研路上的传业授道解惑。

其次，感谢家人对我无微不至的关怀。感谢父母的养育之恩，父母的期望指引着我二十六年的人生路；感谢爷爷奶奶从小对我的悉心照顾，我的成长离不开他们的关爱。此外，特别感谢兄长蔡博昆二十六年的陪伴与对我学业的支持，一起学习，一起生活，一起玩耍，一同长大，不一样的路也有不一样的精彩。

最后，感谢 13 级实验室的小伙伴们，你们的陪伴让我度过了快乐和成长的三年。虽已不记得为什么是 TIF09，但永记什么是 TIF09，还有那些一起学习，一起工作，一起吃饭，一起笑过傻过的日子。此外，特别感谢 614 的舍友们，感谢猴哥、骚哥和不是腾讯的 QQ。缘份让你我相聚，时间让一切逝去，各奔东西后，留下的只剩回忆。

盛开的花已凋零，修过的路还在修；博学楼已拔地而起，曾经的图书馆还没开。从西伯利亚到西湖的风，渐渐怀念起北二的粥；从三号楼到宏生楼，回想起曾经的测试室。东区的体育馆放起 3D 电影，北区的菜园变成一湾涟漪，南区的牌坊不再为人民服务，西八的空调还没机会打开。七年前的华园在记忆里不再清晰，明天的华园还在烟雨中朦胧。那一年木棉树下，那一季桂花绽开，凝望着此刻浪漫的校园，我穿过了正装剪起了胡须，曾经的青涩已随风而去。走过的路已草木深，远方的海还等乌篷，感谢所有关心我的人，路漫漫其修远兮，吾将上下而求索。

蔡博仑

2016 年 4 月 15 日

#### IV - 2 答辩委员会对论文的评定意见

在计算机视觉中，视觉跟踪算法一直是一个前沿的研究方向。论文研究了基于生物启发模型，具有重要现实意义和实用价值。研究内容丰富，研究方法科学，结果于实际应用性较强，有较强创新性，主要贡献如下：

- 1、提出一种快速 Gabor 近似方法优化 BIAM 模块的特征提取效率，引入 FFT 优化 BITM 模块的卷积操作，保证了 BIT 的效率和精度。
- 2、引入深度学习对 BIT 的 BIAM 进行了改进，改善了跟踪精度。
- 3、在 BITM 中加入尺度估计，并提出了一种快速尺度特征，减少了原来特征的计算冗余。

论文结构合理，分析正确，论述清楚，文献材料收集详实。另外，论文格式正确，书写规范，语言流畅，是一篇优秀的工学硕士学位论文。答辩过程中，展示表达清楚，回答问题准确。

答辩委员会认为，论文达到了学术硕士学位论文水平。经表决，一致通过蔡博仑同学学位论文答辩，建议授予专业硕士学位。

论文答辩日期： 2016 年 6 月 8 日

答辩委员会委员共 5 人，到会委员 5 人

表决票数：优秀（4）票；良好（1）票；及格（  ）票；不及格（  ）票

表决结果（打“√”）：优秀（√）；良好（  ）；及格（  ）；不及格（  ）

决议：同意授予硕士学位（√） 不同意授予硕士学位（  ）

答辩  
委员  
会成  
员签  
名

殷瑞峰 (主席)

徐永

钟海波

姜士波

晋建秀