西安交通大学研究生毕业论文

基于单目视觉的目标跟踪算法的研究与实现

# 绪论

## 课题的研究背景

目标跟踪是计算机视觉领域的基础研究方向之一，主要根据视频初始帧中给定目标图像信息以及上下文语义分析目标的运动模式，从而在后面的视频序列中对目标进行持续的精确定位。目标跟踪算法在自动驾驶，人机交互，视频标注，智能交通，安防监控等领域有广泛的应用前景和商业价值。在学术领域，目标跟踪技术是更深层次视频分析技术的基础（例如行为模式分析等）[[1]](目标跟踪论文/深度学习目标跟踪算法综述.pdf)。

目标跟踪问题可以看作是前景背景分类和目标状态估计两个任务的结合。第一个任务要求模型能够从复杂背景中对目标与非目标做出判别；状态估计任务要求算法能对目标的尺寸、大小、长宽比等状态做出精确估计[[2]](目标跟踪论文/SiamFC++——Towards%20Robust%20and%20Accurate%20Visual%20Tracking.pdf)。在过去几年中，以KCF[2]算法为代表的相关滤波算法凭借较快的跟踪速度和判别式的跟踪策略取得了良好的跟踪效果，并在工业界获得广泛应用。相关滤波算法的优点在于能随着跟踪的进行在线更新跟踪模型，由于将图像信息映射到了傅里叶域，变卷积为点乘，简化了计算步骤，可以保证模型在线学习的同时保持较快的跟踪速度，在CPU等一些算力成本有限的平台上可以满足实时性要求，但其缺点在于不能很好的利用语义信息，对光照、遮挡等干扰的鲁棒性较差。同时固定长宽比固定，不能很好的满足对目标状态估计的要求。近年来基于孪生网络[2]的目标跟踪算法大量出现，卷积神经网络让模型能更好的提取各种层次的图像特征，但同样存在这训练集不平衡，对类内干扰鲁棒性差，状态估计不准确等问题。同时，卷积带来的巨大计算量同样对模型的速度以及准确度之间的平衡提出了更高的要求。

## 国内外研究现状

目标跟踪算法长久以来作为高级计算机视觉应用的基础研究，受到了国内外学者的高度关注。视觉跟踪系统的基本框架一般由搜索策略，特征提取和观测模型等模块组成[1].根据观测模式和判别方式的不同，可以分为生成式和判别式模型；其次，特征的提取主要包括人工特征和深度学习特征。下面本文依次对不同的跟踪模型进行分类别阐述。

（图：目标跟踪基本流程图）

### 基于生成式模型的目标跟踪算法

生成式模型是对目标模板进行特征的构建和表征后，在后续帧中进行目标搜索和匹配，寻找最为相似的区域作为跟踪结果。基于此，生成模型的主要任务，是找到与生成的模型描述相似的目标，然后进行模板匹配以找到图像中最匹配的区域，即当前帧中的目标。因此对于生成式模型而言，对目标特征的描述和建模就显的尤为重要，对目标特征准确和详细的描述会极大影响跟踪结果的准确性。常用的特征描述方法有核技巧[4],增量学习【5】，高斯混合模型【6】，线性子空间[7]，贝叶斯网络[8]，稀疏表示[9]，隐马尔可夫模型[10]等等。最后将相似性度量函数用作置信度指标，以反映每个跟踪结果的可靠性，确定目标是否丢失。生成式模型虽然能对目标特征本身进行很好的建模，但是缺不能很好的利用上下文和图像背景信息，对于类内相似目标干扰的鲁棒性较差。

### 基于判别式模型的目标跟踪算法

判别法的基本思想是在学习过程中使用数据直接学习决策函数Y = f（X）或最大化条件概率分布P（Y | X）作为预测模型【11】。判别式模型采用的是“图像特征+机器学习”的组合方式，在对搜索图片和目标进行特征建模后，将目标物体定为正样本，背景图像当作负样本，放入分类器中进行训练。与生成式模型不同的是，判别式模型能充分利用背景图像的信息，同时对目标（前景）信息和背景信息机性特征提取，提高了模型的鲁棒性。判别式模型，将目标跟踪分为分类问题和回归问题两个问题。分类问题主要解决前景和背景的二分类；回归问题主要目的是生成回归框，对目标位置和尺度状态做出精确估计。在一些论文上，回归问题也被称为目标估计问题[2]。

在计算机视觉领域中，目标检测和目标跟踪是两个重要的研究领域，且有很多相似性。目标检测需要在图片或者视频序列中寻找并定位特定静态或动态目标；目标跟踪则致力于在利用第一帧目标图像信息的基础上，实现对后续帧动态目标的定位。跟踪算法最初用于解决检测算法的速度。它用于预测下一帧中目标的位置，然后使用检测算法标记目标的位置。后来，有人根据一定的时间段对视频序列进行分割，并在该时间段内检测出每一帧图像，从而达到类似跟踪的效果。 这种跟踪等效于检测每个帧，这是一种伪跟踪。基于判别式的跟踪算法是一种“动态检测”，是一种tracking-by-detection的方法，目前的目标检测算法可以在复杂场景下实现对目标的精确定位，这对提高跟踪器的精度和鲁棒性有很强的启示意义，但它与传统的检测方法有以下几点不同：

* 目标检测算法检测目标的类别差距相对较大（例：区别人和马），而目标跟踪算法需要实现类内分类（例：篮球比赛中的1号队员和2号队员）
* 目标检测算法在训练集中对目标物体进行了充分的训练，而目标跟踪算法在获得第一帧目标数据前是类不可知的。目标跟踪可以看作是一个单样本学习问题
* 目标跟踪算法在实时性上的要求要高于检测算法。

Tracking-by-detection类算法有两类突出带表：一种是以CSK【】和KCF【】为代表的相关滤波算（Correlation Filtering）；另一种是以SiamFC【】为代表的基于孪生网络的跟踪算法。2010年，随着David等人的MOSSE【】算法的出现，基于相关滤波的目标跟踪算法开始受到研究人员的关注。相关滤波算法是利用循环矩阵，通过傅里叶变换将图像映射到傅里叶域，变卷积为点成，大大减少了模型运算量，显著提高了运算速度。CSK和KCF算法在循环矩阵做密集采样的基础上，引入了脊回归核核函数，将低纬特征映射到高维空间，极大地提高了模型的鲁棒性。在提升精确度的同时，并没有破坏原有算法的高速性能，在CPU平台上可以达到100FPS以上的水平，满足了跟踪算法对实时性的要求。（再举例说明三个相关滤波跟踪实例）相关滤波类方法的有点在于在实现模型在线更新的同时保持了跟踪的实时性；然而，大多数相关滤波算法都是基于人工手动选择特征，特征提取的鲁棒性和泛用性不强；其次，相关滤波算法采用多尺度目标框及图像金字塔的方法对不同尺寸的目标进行状态估计，估计精度严重依赖于目标框的设计情况，跟踪结果精确度较差。

2012年，AlexNet[3]的横空出世为目标跟踪算法提供了新的思路。各种深度卷积网络的出现，其实质上是为目标以及搜索图像的特征提取提供了更具泛用性的工具。基于深度学习的跟踪算法主要利用相同的卷积网络（又称为孪生网络）同时对目标模板和搜索图进行特征提取，两个网络结构相同，参数共享。特征提取后的处理方法主要有两种思路：以SiamFC[11]将得到的目标特征和搜索图特征做关联卷积计算距离，得到搜索图中距离最近的位置即为目标在搜索图中的位置；第二种类似于单样本学习，将跟踪模板视为需要学习的单样本，通过卷积神经网络对目标模板进行“理解”；在后续帧的检测中，不再需要目标模板，只有搜索图的输入，根据搜索图提取的特征和模板特征的理解得到最后的目标位置。这种方法由于在实际应用中不再需要对目标模板进行卷积特征提取，减少了计算量，极大提高了运算效率，SiamRPN类[5]是该思路的典型代表。

随着各类计算平台的快速发展，基于孪生网络的目标跟踪算法在近年来取得长足发展。李博等人将RPN网络的思想引入目标跟踪，将“理解”后的图像特征输入RPN网络，得到正确的前景背景分类及目标估计；DaSiamRPN【】在SiamRPN算法基础上，调整训练策略，平衡训练样本中的正负样本数，提高算法对类内干扰的鲁棒性。SiamDW与SiamRPN++分析并解决了深层神经网络不能应用于目标跟踪的问题，让多层特征图的结合称为可能；GradNet实现了对目标模板的在线更新。上文所提到的跟踪算法，基本达到了对跟踪物体实时性的要求（30FPS）但由于采用了人工设置anchor，在模型中引入了先验知识，影响模型的泛化性。从实验中发现，基于anchor的目标跟踪算法对于严重偏离预设anchor或者目标物体在视频序列中严重变形（非刚体）的情况，精确度较差。ATOM【】为代表的方法，在状态回归分支中引入在线学习策略，提高了状态回归的准确性，但速度上受到一定影响。另外，目前主流的跟踪算法主要使用OTB100和VOT数据库作为测试对象，不能很好的分析目标种类或运动模式对跟踪结果的影响进行定量分析。

## 论文的主要研究内容与章节安排

### 论文的主要研究内容

针对上文描述的目标跟踪算法中的问题，本文主要对以下几个问题进行了研究：1. 去除状态估计分支中的anchor设置，减少人为先验知识对模型泛化程度的影响，提高算法对极端个例的鲁棒性。2.增加状态评估分支，提高跟踪精度。 3.探索不同运动模式及目标类别对跟踪效果的影响。4. 建立检测-跟踪-坐标定位系统，实现对目标的精确实时定位。

### 章节安排

最后再写

# 单目视觉目标跟踪算法的理论基础

## 相机模型与坐标系变换（不够再加）

## 目标检测算法

## 基于相关滤波的目标跟踪算法

## 基于深度学习的目标跟踪算法

## 目标检测与跟踪算法关系探索

## 目标分割算法（不够再加）

# 目标跟踪算法的建模与实现

算法的结构主干如图3.1所示，本文算法在经典孪生网络跟踪框架的基础上，增加回归分支和质量评估分支，以提高跟踪的精确度。在回归分支中采用anchor-free的方法生成回归框，避免先验知识对跟踪效果的影响以增强算法的鲁棒性；算法在实验中达到了90fps的跟踪速度，满足实时性要求。



#### 图3.1 算法流程图

下面，本文从主干网络，分类分支，回归分支，状态估计分支，整体架构五个方面进行详细论述。

## 基于孪生网络的特征提取

简介孪生网络

本文算法中的主干网络采用张志鹏等【siamdw】在2019年提出的CIResNet-22模型。该主干网络是在ResNet【ResNet】残差网络上的改进，解决了之前Siamse类跟踪模型在深度卷积模型上性能下降的问题。原本的深层神经网络，如VGG【】，GoogleNet【】，ResNet【】等，由于在特征提取的过程中大量使用Padding来保持特征图分辨率的不变性，在做交叉关联计算时，由于Padding的存在，扩大了感受野的范围，造成了感受野在搜索图像中容易出界的问题，尤其在对靠近边缘的目标进行跟踪时，由于感受野扩充边缘出界的问题，特征图对应的响应值平移与搜索图目标平移方向会不一致，不满足平移不变性。CIResNet-22在残差块中引入了CIR-crop对padding后得到的特征图进行边缘裁剪，如图3.2所示，裁剪后的特征图满足平移不变性。同时，感受野范围同样会对跟踪结果产生较大影响：若感受野过小，特征图提取到的语义和上下文信息过少，特征图抽象度不够；感受野过大，特征图中相邻元素间的冗余信息过多，对局部信息关注不够。CIResNet对应的感受野大小为模板帧大小的0.6-0.8倍之间。目标响应值在特征图上的平移距离与下采样总步长相关，较大的步长会降低分辨率，同时丢失部分图像信息；因为相邻帧之间目标位移一般不大，故不适合采用较大的步长，但总步长过小会造成训练中较大的计算负担，在本文中，下采样步长选为8。



#### 图3.1 CIR-crop

## 分类分支

分类分支本质上是一个前景或背景的二分类问题。如图3.1所示，分类分支是采用cross-correlation后得到的特征图作为输入，经过多层卷积计算后，得到尺寸为的特征图。特征图中的每一个像素位置预测一个分类置信度，表示该位置对应图像区域是前景（模板目标）的概率。该特征图上每个像素对应输入图片上以为中心的一片图像区域，其中代表整个特征提取网络的总步长，在本文中.在训练集中，若该图像区域中心在标签目标区域内，则被认为是前景，反之为背景。相对于基于anchor的前景背景分类算法，本文的算法直接对特征点对应的图像区域做分类，避免了同一位置多个anchor造成的分类歧义，提高了算法的鲁棒性，特征图中每个特征点只输出一个置信度，更加简单直观，方便后续操作。

## 回归分支

状态估计分支的输入和分类分支相同，以cross-correlation后的特征图为输入，经过卷积运算，最后输出一个4D偏移向量以预测目标位置。偏移向量预测了原搜索图上的对应点到训练集标定框四个边的像素距离，记为。因此，特征图位置的目标偏移向量可以表示为：

其中和分别表示特征点对应标定框的左上角和右下角。相较于基于锚点的目标跟踪算法，如SiamRPN【】，本文算法直接以位置作为训练样本，给出偏移向量和目标分类，避免了一个特征点多个锚点带来的歧义性问题。由于基于锚点的偏移向量与预定义的锚本身有很大关系，偏移向量的预测值受人为先验知识干扰，降低了算法的鲁棒性。本文算法预测的偏移向量可以直接学习到训练样本中的偏移分布，提高了目标位置预测的鲁棒性。

从以上两节可以总结得到，分类分支和偏移回归向量都是基于目标感受野直接给出，避免了人为设计anchor带来的先验知识干扰，使模型能更准确的学习到训练集的数据分布特性，鲁棒性更强。

## 质量评估分支

## 训练过程

1. 实验结果及分析
   1. 目标跟踪算法实验效果对比
      1. OTB
      2. VOT
      3. GOT-10

…………(有时间再加上其他数据库)

* 1. 消融实验

# 系统结果展示与分析

## 目标检测

## 目标跟踪

## 目标定位

## 系统框图

# 总结与展望

# 参考文献与致谢

[4] D. Comaniciu, V. Ramesh, P. Meer, Kernel-Based Object Tracking, Pattern Analysis & Machine Intelligence, 25 (2003) 564-575.

[5] D.A. Ross, J. Lim, R.S. Lin, M.H. Yang, Incremental Learning for Robust Visual Tracking, International Journal of Computer Vision, 77 (2008) 125-141.

[6] A.D. Jepson, D.J. Fleet, T.F. El-Maraghi, Robust online appearance models for visual tracking, Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on2003), pp. 1296-1311.

[7] M.J. Black, A.D. Jepson, EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation, International Journal of Computer Vision, 26 (1998) 63-84.

[8] O. Tuzel, F. Porikli, P. Meer, A Bayesian Approach to Background Modeling, Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on2005), pp. 58-58.

[9] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse Representation for Computer Vision and Pattern Recognition, Proceedings of the IEEE, 98 (2010) 1031-1044.

[10] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Readings in Speech Recognition, 77 (1990) 267–296

【11】 Shaoze You1,2, Hua Zhu1,2,\*, Menggang Li1,2, Yutan Li1,2 ：A Review of Visual Trackers and Analysis of its Application to Mobile Robot，