清 华 大 学

综 合 论 文 训 练

题目：多目标对象追踪关联问题研究

系 别：软件学院

专 业：计算机软件

姓 名：陈政宇

指导教师：丁贵广 教授

2016年6月8日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文

**（涉密的学位论文在解密后应遵守此规定）**

签 名：\_\_\_\_\_\_\_\_\_\_\_\_导师签名：\_\_\_\_\_\_\_\_\_\_\_\_\_\_日 期：\_\_\_\_\_\_\_\_\_\_\_

中文摘要

随着近年来人工智能、深度学习和神经网络的发展，其在实际场景中的应用效果有了显著的提高，尤其是在计算机视觉领域。随着2012年AlexNet的提出，人工智能系统进行图像的识别，分类，检索已经达到接近甚至超过人类的水平。目标追踪是计算机视觉中一个十分重要且有用的分支，其中又分为单目标追踪和多目标追踪两个分支。目标追踪在安防，数据分析，监控等多个领域都有着广泛而重要的应用。在多目标追踪之中，目前主流的有基于滤波器和前向帧信息的追踪方法和基于卷积神经网络检测出不同帧的目标再进行匹配的两种方法。在实际应用之中，第二种的效果较好，但一般会需要更高的计算能力。在基于检测结果的多目标追踪之中，其效果暂时还难以与人眼观察的结果媲美。最主要的问题在于检测结果的误差，以及物体之间的互相遮挡。本文通过探究多目标追踪的各种方法，实现了一种基于匈牙利算法的多目标追踪系统，并通过研究其重识别策略，优化了算法并解决了在部分场景下遮挡之后的追踪目标丢失的问题，并且还让系统的运行速度保持在可即时运行范围之内。

**关键词：神经网络， 多目标追踪， 行人重识别， 匈牙利算法**

ABSTRACT

In recent years ,the development of Artificial Intelligence, Deep Learning and Neural Networks has effectively improved their performance in real scene applications, especially in Computer Vision field. After Alex and Hinton have come up with AlexNet in 2012, the ability of AI systems to do image recognition, classification and retrieval has come close to or even exceed human level. Object Tracking is a very important and useful branch of Computer Vision, and itself has divided into two branches, Single Object Tracking and Multi Object Tracking. Object Tracking has shown wide and important usage in security, data analysis and monitoring. Among all the methods in Multi Object Tracking, two methods are widely used, one is based on inferior frame information and filters and the other is based on associating the bounding box of objects given by CNN in different frames. The later one, in general, gives a better and more accurate result but requires more computing power. The performance of those multi object tracking methods are still hard to compare with human result, because of the errors of detections and the occulusions of bounding boxes. In this article, I surveyed multiple multi objects tracking methods, and implemented a method based on Hungarian Algorithm. Meanwhile, I researched the tactics of object re-id, optimized the algorithm and solved the problem of losing tracking targets after occulusion in some scenes while still make the system a real-time tracking system.

**Keywords:** Neural Network; Multi Object Tracking; Person Re-id; Hungarian Algorithm

目 录

[第1章 引言 2](#_Toc516188702)

[1.1 课题背景、目的与意义 2](#_Toc516188703)

[1.2 主要工作 3](#_Toc516188704)

[1.2.1 技术调研 3](#_Toc516188705)

[1.2.2 算法优化 3](#_Toc516188706)

[1.3 论文的组织与安排 3](#_Toc516188707)

[第2章 相关工作与预备知识 5](#_Toc516188708)

[2.1 神经网络 5](#_Toc516188709)

[2.2 匹配算法 6](#_Toc516188710)

[第3章 基于深度特征的匈牙利匹配追踪算法 8](#_Toc516188711)

[3.1 目标检测 8](#_Toc516188712)

[3.2 状态预测 8](#_Toc516188713)

[3.3 特征提取 9](#_Toc516188714)

[3.4 代价矩阵 9](#_Toc516188715)

[3.5 匹配流程 13](#_Toc516188716)

[第4章 实验探究与分析 14](#_Toc516188717)

[4.1 MOT数据集 14](#_Toc516188718)

[4.2 实验结果 15](#_Toc516188719)

[4.2.1 实验度量标准 15](#_Toc516188720)

[4.2.2 实验结果分析 16](#_Toc516188721)

[4.3 本章小结 18](#_Toc516188722)

[第5章 总结与展望 20](#_Toc516188723)

[5.1 总结 20](#_Toc516188724)

[5.2 展望 20](#_Toc516188725)

[插图索引 21](#_Toc516188726)

[表格索引 22](#_Toc516188727)

[参考文献 23](#_Toc516188728)

[致 谢 25](#_Toc516188729)

[声 明 26](#_Toc516188730)

[附录A 外文资料的书面翻译 27](#_Toc516188731)

[A.1 介绍 27](#_Toc516188732)

[A.2 含有深度关联度量模型的SORT 28](#_Toc516188733)

[A.2.1. 轨迹处理和状态估计 28](#_Toc516188734)

[A.2.2. 关联问题 29](#_Toc516188735)

[A.2.3. 级联匹配 30](#_Toc516188736)

[A.2.4. 深度外貌特征描述子 31](#_Toc516188737)

[A.3 实验 32](#_Toc516188738)

[A.4 结论 33](#_Toc516188739)

主要符号对照表

CNN 卷积神经网络（Convolutional Neural Networks）

MOT 多目标追踪（Multi Object Tracking）

DPM 可变性部件模型（Deformable Part Model）

FRCNN 快速区域卷积神经网络（Fast Region of CNN features）

SDP 尺度依赖池化（Scale-dependent Pooling）

Tracklet 指一个目标的一段轨迹

Track 指轨迹上在其中某一帧的状态

# 引言

## 课题背景、目的与意义

多目标追踪是计算机视觉的一个重要分支，其在实际生活中有重要的应用，包括安防，数据分析，监控等应用场景。随着深度学习近些年来的快速发展，计算机视觉领域已经有了显著的进步，并且在许多领域，类似于图像的识别、检索、分类等方面拥有了接近乃至超越人类水平的准确率。但是不同于在静态图像上的问题，在动态视频中计算机视觉系统一直难以达到人眼的识别正确率。多目标追踪系统是追踪系统中的一类，其主要目的是在一段视频之中检测出动态的目标物体（如行人，车辆，动物，轮船等），并且追踪其运动的轨迹，给每个物体赋予一个正确的ID，并在理想情况下，在整个视频周期之内，对每个物体只赋予唯一的一个ID。目前主流的方法是基于神经网络首先进行目标检测，随后对检测到的目标进行帧间的相互匹配的Tracking By Detection 方法。这种方法有如下两个关键步骤：

1. 使用神经网络模型（如CNN等）侦测出各个帧上的目标物体，并给出目标物体的物体边界框。
2. 对1中给出的物体边界框，采用匹配算法进行帧间匹配，给匹配上的物体赋予相同的ID。

在这两步中都存在着限制性能瓶颈的因素。我们采用Tracking By Detection的方法，是因为当前神经网络对图像的识别定位已经达到了较高的精度。但是这一般针对的是简单场景下的情况。在复杂的场景下，如目标之间相互遮挡，或者是目标挤在一个狭小的区域之内，或是由于光线等因素目标的全貌无法清晰识别之时，神经网络给出的检测结果与定位结果都会有一定的误差，并且这种误差也难以在2中消除。

在第二步中，由于1给出的结果可能是不准确不充分的，算法就需要在考虑可能误差的情况下将不同帧之间的物体边界框进行匹配。当前的多目标追踪系统一般对相邻帧的匹配结果较好，因此当物体被遮挡而1中给出的连续的检测框中断时，系统往往会出现误差。

为了让系统达到可用的性能，我们不能够允许其处理速度过于缓慢。对于一个可用的追踪系统来说，我们追求的目标最好是达到实时处理的速度。这就要求我们使用一种快速有效的算法完整侦测和匹配。匈牙利算法是一种基于代价函数计算两组目标之间相互匹配最优结果的算法，对于一个代价矩阵，经过匈牙利算法计算之后，每行每列最多有一个非零项。将该非零项所处的行和列代表的对象互相匹配起来，就会得到一个在当前代价矩阵下的最优匹配解。匈牙利算法拥有较快的速度，因此对于我们实现实时的目标追踪系统是一个很好的方法。

在当前的研究进展之中，多目标追踪系统无法达到理想效果的一个主要原因就是来自于物体之间的遮挡。遮挡导致的物体侦测框丢失，或者是定位不准确，都会显著影响性能。因此，若能够对物体遮挡进行有效的处理或改进，可以显著增强目标追踪系统在实际中的可用性。因此，本文调查研究了多种多目标追踪系统，并基于一种实时的多目标追踪系统SORT(Simple Online Realtime Tracking)进行改进，力争在保证系统处理实时性的情况下解决在物体遮挡时遇到的部分问题。

## 主要工作

本文的主要工作是调查研究当前的多目标追踪技术，找到一种合适的实时处理方法，并研究其中在目标遮挡之间出现的问题，解决遮挡中出现的一些re-id错误以及丢失跟踪目标的问题

### 技术调研

目前前言的多目标追踪比赛是MOTChallenge[1] 竞赛，其中提交的竞赛结果包含多个当前最前言的多目标追踪处理技术。其中，在目前有开放论文或开源代码的项目之中，如下几个在综合表现上有较好的性能，如德国马克思普朗克研究所的Siyu Tang提出的基于图分割算法的匹配方法[2]，还有科布伦茨-兰道大学的Nicolai Wojke提出的一种基于匈牙利算法的实时追踪系统[3]，以及斯坦福大学的Silvio Savarese等人提出的一种基于RNN的匹配追踪系统[4]。最终发现，SORT系统在实时场景下效果较好，方法易于实现，鲁棒性也较高，因此决定以SORT为基础研究改进实时多目标追踪系统在复杂场景下的效果。

### 算法优化

对于一个多目标追踪系统，在帧间匹配之中，代价函数是其核心问题。如何找到一个更好的描述场景的代价函数很大程度上会影响最终系统的性能。对于多目标追踪问题，我们可以自然的想到，位置信息和目标的特征信息应该是主要决定我们进行判断的标的。因此，对于该目标追踪系统，我们将位置信息与目标特征信息进行整合，得到匹配的代价函数。同时，我们也将对特征信息与特征向量进行改进，争取在最终的系统中得到更好的实验效果。

## 论文的组织与安排

本文的其它章节安排如下：

第二章介绍了多目标对象追踪问题需要使用的技术、方法的基本情况，和部分当前研究进展

第三章本文中使用的算法的主要流程架构，与针对关联问题的改进。

第四章介绍了实验整体方法，实验数据与实验结果分析。

第五章对所做的工作进行了总结与展望。

# 相关工作与预备知识

与本文研究内容相关的工作主要包括两个方面：神经网络检测子和匹配算法。我们将对这两个领域中的相关工作和一些本文中用到的预备知识进行介绍。

## 神经网络

在多目标追踪问题之中，特别是Tracking By Detection问题中，神经网络一般是用于在检测部分给出一个合适的Bounding Box。一般情况下，都会使用卷积神经网络(Convolutional Neural Networks)。

卷积神经网络是神经网络的一种，区别于全连接网络，卷积神经网络层之间采用卷积的方式进行连接，这样下一层的每一个神经元只与前一层的一个特定区域的神经元相关联，这样可以识别出更多的局部特征。



图 2.1 CNN与全连接网络结构图

如图2.1.1所示，一个典型的神经网络一般至少由三层组成，输入层，输出层，和隐藏层。左边是一个典型的全连接网络，而右边则是一个卷积神经网络。卷积神经网络的每一层的深度，由想要抽取的特征数量决定，而其高度和宽度则由前一层的大小和卷积核大小共同决定。理论上，一个足够深的卷积神经网络可以模拟任意形式的函数。对于一个卷积层，其中每一个神经元的值等于 ， 其中为权重，为对应的前一层的神经元，b为bias。

## 匹配算法

在多目标追踪的研究之中，区别于神经网络的检测子是近些年来得益于深度学习的发展才快速进步的，匹配问题则是一个更加古老的问题。匹配问题有许多不同的解决思路，但其主要都是通过定义一个代价矩阵，并通过求解出对给定代价矩阵的最优分配解来实现的。

一个典型的代价矩阵得到的方式是IOU Matching。IOU指的是Intersection Over Union，是一种度量两个Bounding Box之间重合程度的标准。其定义如下：

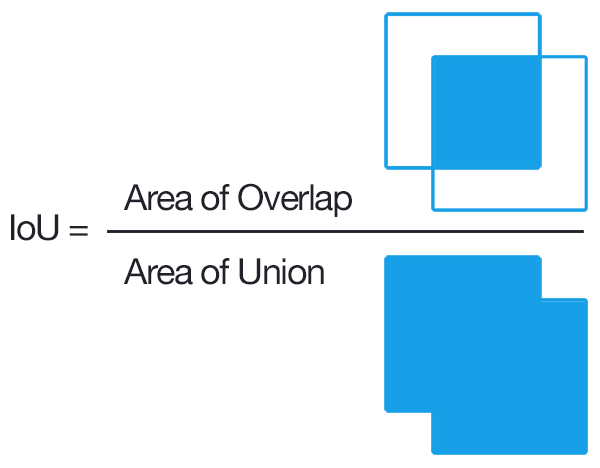


图 2.2 IOU示意图

这种方式的一个直观思想是认为对于当前的摄像设备，速度一般都是30fps以上，在一帧的时间之内，目标物体的移动距离不可能太远，因此，若两帧间有目标的Bounding Box拥有高重合度，则可以认为这很有可能是一个物体。

另一种常见的方式是对Bounding Box之内的目标进行特征抽取进行匹配。这也是一种符合直觉的方法，因为对于人眼而言，我们在跟踪物体是其实也是潜在的先将物体识别出来在于之前的记忆进行联系的。

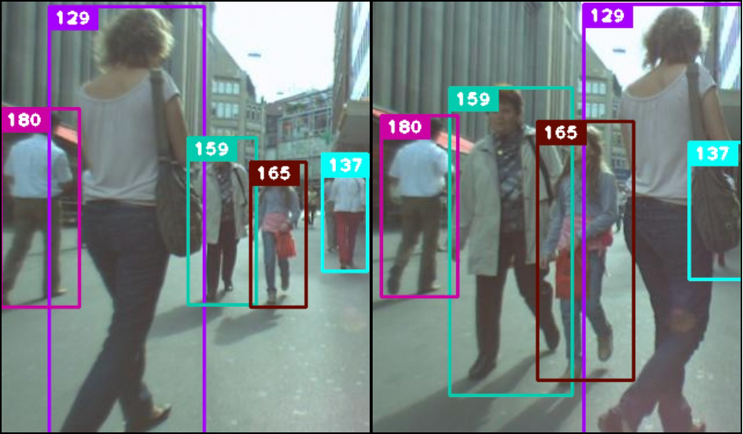


图 2.3 常见场景示意

如对于上图所示的实例，在两帧不同的画面中，有不少行人被框选出，通过训练一个小型的神经网络，提取出Bounding Box中的特征，在特征上进行匹配，就将问题转化成了一个近似于Classification的问题。利用对于Bounding Box之内的目标提取出的Features 向量进行处理，计算帧间不同Bounding Box之间的距离作为Cost，就可以得到这种方式定义的代价矩阵，再计算出最优解即可。

对于解除给定代价矩阵的最优解问题，常见的方法是使用匈牙利算法[5]，这是一种能够在O(n3)时间之内得到指派问题最优解的快速算法，也是多目标追踪里常见的算法之一。

# 基于深度特征的追踪算法与关联问题处理

算法的整体流程是这样的。首先我们通过检测算法得到每一帧上目标的Bounding Box，包括其位置，长宽信息，置信度等。随后，针对每一帧给出的目标结果，对Bounding Box内的目标用一个深度模型进行特征抽取。对每一个Bounding Box我们都会得到一个表示其特征信息的128维向量。同时，我们可以用卡尔曼滤波算法计算出每个物体在下一帧之中的可能位置，通过马哈拉诺比斯距离描述物体实际的位置和我们估计位置之间的距离。同时用余弦距离描绘不同帧之间的Bounding Box内目标的特征差别。将两个距离整合，用于描绘不同帧之间每两个Bounding Box之间的Cost，最后通过匈牙利算法解出给定代价矩阵的最优解，得到结果。下面，本文将更详细的阐述算法思路。

## 目标检测

本文使用的检测数据来自于MOT Challenge之中，在该benchmark中，作者使用SDP，FRCNN，DPM三种检测方法给出了三个精度各异的检测结果。

## 状态预测

对于Tracklet上的每一个track，我们采用一个多维数组，其中，代表目标框所在中心点的位置，表示长宽之比，h表示高度，而表示的则是目标Bounding Box在各个参数方向上的导数。我们可以用一个标准的卡尔曼滤波器作为一个预测状态的模型，并将Bounding Box的各个坐标作为目标状态的直接观测。

对于每一个轨迹k，我们会记录其最后一次出现的时间（以帧数为单位），并计算从那时起到现在经过的时间。卡尔曼滤波器会根据来计算那个轨迹在当前帧中可能出现的范围以及概率。当轨迹k再次成功匹配之后，计数器将会归零。

状态预测的结果将反应目标可能的运动位置，而一般来说，大部分的物体运动都是有一定的目标的而不是杂乱无章的。事实上，当我们考虑人眼跟踪快速移动的目标的时候我们也会发现，我们会倾向于提前预判物体的运动状态并将头部和眼球随之转动，当物体突然发生剧烈的运动状态变化时，我们很容易跟丢物体。因此，通过物体的运动状态来预测未来的位置是符合直觉的。

除了对状态的预测，我们还将设定一个参数max\_age， 当时，我们倾向于认为该目标已经脱离了我们的跟踪范围，因此我们将将其轨迹删除。而对于刚刚出现的新的Bounding Box（即无法与任何已存在的tracklet匹配上），我们会设置一个init\_num，在其出现的前init\_num帧我们会认为该目标处在待定状态。

## 特征提取

由于计算代价矩阵需要一个能够描述各个Bounding Box内目标特征的模型，因此，我们采用了一个预训练好的残差神经网络[6]，其结构为两个卷基层后接上六个残差block。通过这个模型，我们可以对每一个Bounding Box生成一个128维的向量作为Bounding Box内目标的描述向量。在一块GeForce GTX 1050的GPU上，该模型可以做到在30ms内处理32个Bounding Box，因此，模型的运算速度是符合实时需求的。

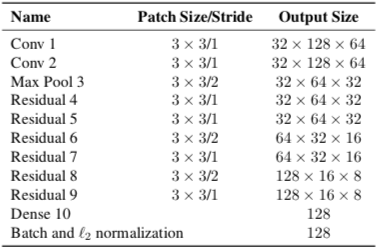


图 3.1 特征提取网络的结构

## 代价矩阵

当我们提取出Bounding Box内目标的特征，并知晓其位置等状态信息之后，我们可以据此定义Cost，以便我们随后应用匈牙利算法解决问题。

如3.2所说，我们采用马氏距离来衡量检测到的Bounding Box与我们用卡尔曼滤波器预测结果之间的差距，其定义为：

(3-1)

这里，表示的是实际检测到的Bounding Box的状态信息，而表示的是用卡尔曼滤波预测的Bounding Box的分布在度量空间中的投影。同时，我们可以在设定合适的阈值的情况下，过滤掉一些非常不可能的匹配对。我们不妨将置信度设为95%，通过逆卡方分布计算出合适的的阈值，当,我们认为这一组匹配是及不可能的。对于四维度量空间来说，我们可以计算出阈值

对于由3.3中得到的特征向量，我们可以使用余弦距离计算不同的Bounding Box之间的距离，对于一个轨迹上的一帧i和检测框j，我们用

和分别表示i和j的特征向量那么我们可以将余弦距离表示为：

(3-2)

同样的，我们可以设置一个阈值，当时，我们便将这一组匹配(i,j)视作不可行的。

这两个度量距离分别从不同的角度考虑了匹配中的因素，简而言之，马氏距离从位置信息上提供了参考，这种参考在短时期内是较为有效的，但是特征度量从另一个角度提供参考，而特征是不会随着时间而显著变化的，这可以为我们在物体被遮挡较长之后重新出现的Re-id提供一个更好的帮助。我们可以用参数将两个度量距离结合，得到最终的Cost如下式表达：

(3-3)

但在实际场景之中，我们可以想见，行人在行走时其形态是不断变化的。虽然人眼能够识别出来是同一个人，但是神经网络给出的特征向量必然是有所不同的。并且，在复杂场景之下，还有可能出现相互遮挡的情况，这都导致只用每一个轨迹最后一帧的特征向量进行匹配的话会增大误差。



图 3.2 辅助说明图1

如图3.1所示，ID为5的黑衣男子和绿上衣的女性产生了重叠，他们的Bounding Box互相之间也有交叠， 因此此时无论是ID为1还是ID为5的框内提取出的特征都可能是不准确的。基于此，我们对于一个轨迹序列，可以存储其中之前帧的特征向量，在计算特征向量之间的距离时，对于新的检测框，我们不仅要计算其与轨迹最后一帧的特征向量之间的距离，还要计算其与之前帧之间的距离，并从中得到一个最小的距离作为新的余弦距离。

因此我们可以重新计算得到一个新的，我们令表示一个轨迹上所有被存储的帧的特征向量，其中，我们有

(3-4)

通过这种方式，我们应该可以更好的提高检测的准确程度。

但是在实验中发现，通过这种方法依然会遇到一些问题，其中一个最主要的问题是在遭遇这档之后，物体重新出现重识别时难以与原先的轨迹匹配的问题。



图 3.3 辅助说明图2

如图3.4.2所示我们可以看出，黑衣男子在从蓝衣男子背后重新出现之后，系统给他赋予了一个新的id，而不是与之前的id进行匹配。这是因为尽管我们已经对之前的ID为5的tracklet存储了多个特征向量但是新出现的detection的Bounding Box依然有大部分被遮挡，这严重影响了特征向量抽取的效果。因此，尽管我们已经为ID为5的tracklet存储了相当数量的特征向量，但是新的detection与ID为5的tracklet上的任意一帧的特征向量匹配时，其距离依然过大。因此我们采用另一种方式来辅助判定。

对于抽取出的128维特征向量，其每一维度都代表着一种高层次的特征，虽然我们无法准确的判断每一维度的含义，但大体上，他们会对不同的特征产生响应。例如，有的维度可能对头部较为敏感，有的可能对手更为敏感[7]。对于一个轨迹上个被存储的特征向量，我们将其堆叠起来会得到一个维度的矩阵。用detection的特征向量和每一个128维向量做元素间乘积，我们可以得到一个大小不变的矩阵。对于这个矩阵的128维度中的每一维，我们可以找到一个最大值，然后再将所有的最大值相加得到一个新的度量距离，进行公式化表示，我们有：

(3-5)

其中，表示编号为id的tracklet所存储的特征向量列表。

我们将这个新的度量距离考虑进去，可以得到一个新的Cost函数，

(3-6)

其中和都是0~1之间的实数。并可以据此得到当前帧的所有detection 和之前任意一个tarcklet之间的代价矩阵。对于一组tracklet和一组detection，我们可以用匈牙利算法求得其最优解。当一个轨迹和一个检测框匹配时，我们可以将该检测框添加到该轨迹上去。

## 匹配流程

在3.1~3.4的基础之上，我们来考虑具体的算法流程。一般来说，临近的两帧之前进行匹配，其准确性应该是较好的，因为在短时间内，目标的位置，形态学参数都没有出现较大的改变，可以得到一个更准确的描述。因此我们可以采用如下的算法流程。

当我们得到新一帧（第n帧）上所有的侦测框时，我们可以首先将他和n-1帧上的track进行匹配。当匹配完成之后，对于被匹配上的检测框，我们可以将其标记为Confirmed状态，而未成功匹配的仍未Unconfirmed状态。对于轨迹我们同样处理。随后，我们再将其与第n-2帧的轨迹中仍为unconfirmed状态的进行匹配，重复这个操作，直到没有更多的帧或者达到预先设定的上限。

对于循环结束后仍然未能成功匹配的检测框，我们将其视作新的目标，让其处于初始化状态，而对于仍未能匹配的轨迹，我们将维护其参数，直至大于我们设定的max\_age之后将这个轨迹标记为删除状态。

对于处于初始化状态的检测框，我们将将他们和之前检测到的仍处在初始化状态的检测框进行匹配，匹配成功的将视为一个轨迹上的。如果一个新出现的处于初始化状态的轨迹连续存在了init\_num帧，我们将赋予其一个新的ID，视作一条新的轨迹。这就是整个算法的大体流程。

# 实验探究与分析

我们的多目标追踪算法是一个轻量级的实时追踪算法，因此我们使用个人电脑就可以完成运算。实验中我使用的是苹果电脑MacBook Pro 2017版本。实验中，我们使用MOT16数据集，并验证了不同算法模型在实验中的具体效果，并从不同方面分析了不同算法的优劣以及可能的应用场景。

## MOT数据集

近年来计算机视觉领域的许多分支都举行了许多在线的竞赛，并产生了不少高质量的数据集，例如ImageNet、COCO、Pascal VOC、KITTI等。在多目标追踪领域，MOT竞赛和MOT数据集无疑是现在质量较好，较为完备和流行的。

MOT16数据集包含14段视频，其中七段为测试视频，七段为训练视频，总大小约2.1G。MOT数据集的训练数据都包含了人工标注的准确良好的检测框，并对属于同一个目标的检测框赋予了同一的ID。同时，MOT16数据集还提供了基于DPM[8]数据集的检测结果，用于模拟在Tracking问题中Detection的检测结果，这样可以让算法更注重于Tracking的过程，并且统一度量，利于比较。

七段视频的信息概况如下：

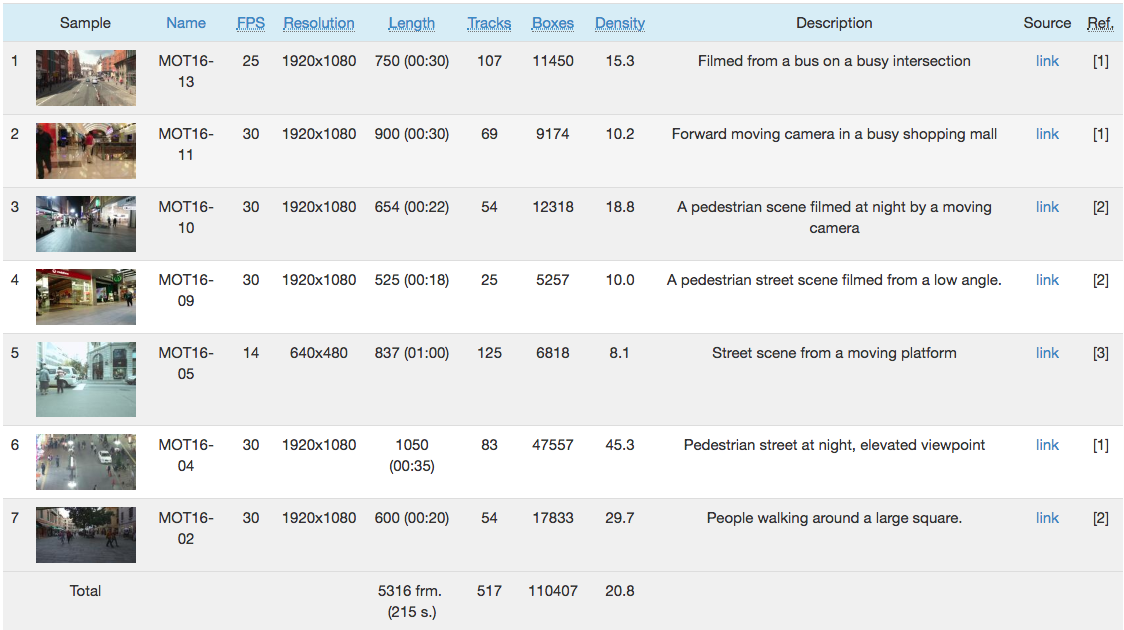


图 4.1 MOT数据集注解

视频的清晰度各异，场景各异，甚至帧数都各不相同，条件较为复杂，并且分为静态镜头和动态镜头两种情况，极大程度上模拟了各种场景下的多目标追踪需求。

## 实验结果

### 实验度量标准

MOT提供了自动化的测评工具，可以对测试结果从多个维度进行评价。

重要评价指标如下：

MOTA(Multiple Object Tracking Accuracy):是主要的总体衡量标准，包括false positive, missed targets 和 id switches.

MOTP(Multiple Object Tracking Precision):表示预测的Bounding Box和标注之间的不重合程度

IDs(Identity switches):表示标注的ground truth上的轨迹变换id的次数

同时，我们也会以人眼观察的方式判断检测结果中是否有显眼的错误。因为直觉上来说，人的注意力一般都是集中在更靠近前景的位置，同时也更关注在画面中占中心地位和较多区域的物体，而上述的评价指标对于所有的轨迹都给予相同的权重，其实并不能很显著的反映人观察的感受。

### 实验结果分析

我们先比较仅仅使用特征模型进行匹配的方法和其他的几种MOT上的Online方法：

表 4.1 与其余几种Online MOT算法的比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | FP | FN | IDs | FM | MOTA | MOTP | Runtime |
| Ours | 8063 | 56668 | **781** | 1344 | 59.5 | 79.1 | **40Hz** |
| POI[9] | 5061 | **55914** | 805 | 3093 | **66.1** | **79.6** | 10Hz |
| EAMTT[10] | **4407** | 81223 | 910 | **1321** | 52.5 | 78.8 | 12Hz |

通过数据分析可以看出，使用特征模型进行匹配的方法在给出较好精度的情况下有极快的处理速度，同时给出了较低的ID switch数量。

但是，观察检测结果我们也会发现，方法存在一些较为显眼的问题。如图3.2所示的ID switch现象。

我们接下来使用公式3-6中定义的Cost来进行试验。

观察试验结果可以发现：



图 4.2 算法辅助说明图3

可以看出我们可以解决一些显式的ID匹配错误问题，如图4.2所示，ID为5的人成功的被匹配到了一起。

统计其度量评测标准可得：

表 4.2 算法改进前后比较

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | FP | FN | IDs | FM | MOTA | MOTP | Runtime |
| Origin | 8063 | 56668 | 781 | 1344 | 59.5 | 79.1 | 40Hz |
| New | 8065 | 55695 | 905 | 1365 | 59.0 | 80.1 | 25Hz |

通过数据分析我们看出，这个改动并没能给性能带来很大的提升，在某些指标上有所上升，但是同时也有些指标有所下降。为什么在解决了一些显式的错误之后并不能提高整体的性能呢？

通过观察视频后我们发现，在行人目标较为拥挤的时候，由于检测网络的不准确和漏报的产生，对象的Bounding Box内很有可能涵盖了别人的信息（如图4.3），这就使得其提取出的特征向量极其不准确。而在这种情况下，对象数量大，因此这里的追踪实际上会较明显的影响最后的算法性能评估结果。但是我们的改进暂时对这种情况没有明显效果，所以暂时在评价指标上看不出显著差异。

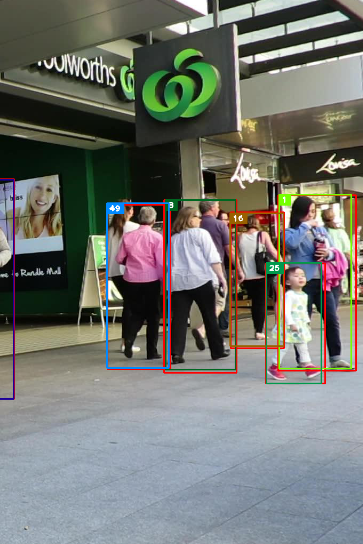


图 4.3 典型的拥挤行人场景示意

## 本章小结

在第三章介绍了基于匈牙利算法的关联算法的主要流程之后，本章依照第三章的方法进行了实现与拓展，并简要的介绍了MOT数据集并在其上进行测试。随后我们通过实验对比了本算法与其他方法的效果和本算法修改前后的差别，从中可以看出，我们的算法与其他在线算法相比还是具有竞争力的。我们随后进行的一个改进也可以解决一些显式的问题，但是由于在4.2中分析的原因，并没有在评估指标上产生显著影响。总而言之，本模型还是一个在在线场景下具有应用价值的易实现的快速算法。

# 总结与展望

## 总结

在本文中，我们通过引入距离度量标准和改进特征度量方法，实现了一个可以在普通机器上实时运行的在线算法，并且本算法与其他算法相比，在保持一定跟踪效果上有较快的速度，具有一定的竞争力。

通过本次毕设，我更详尽的了解了计算机视觉中多目标追踪领域的研究进展，热点问题，与领域研究的困难。并通过实现，改进算法加深了自己对于对目标追踪系统的理解，对我以后的学习研究大有裨益。

## 展望

在4.2节中，我们分析了系统当前的问题，可以看出，系统在目标行人较为拥挤时暂时还没有很好的效果。。这一方面是因为，检测子返回的Bounding Box本身存在误差，另一方面也是因为，在拥挤的时候我们的特征提取网络不能准确的抽取对象的特征向量。

对于跟踪问题，我们很难再检测子部分进行改进（那被归类为Detection问题），因此我们只能再第二方面下手。目前我考虑的一个可能的改进是，为每一个Bounding Box计算几个Confidence，这个置信度用于表示我们当前提取的特征在多大程度上是可信的。可以考虑利用IOU或别的方法来求得这个置信度的值。得到置信度之后，我们首先考虑在置信度高的Frame之间进行匹配，这样可以尽量减少Bounding Box内无关对象的干扰，并尝试获得更好的效果。

插图索引

[图 2.1 CNN与全连接网络结构图 5](#_Toc516187823)

[图 2.2 IOU示意图 6](#_Toc516187824)

[图 2.3 常见场景示意 7](#_Toc516187825)

[图 3.1 特征提取网络的结构 9](#_Toc516187826)

[图 3.2 辅助说明图1 11](#_Toc516187827)

[图 3.3 辅助说明图2 12](#_Toc516187828)

[图 4.1 MOT数据集注解 15](#_Toc516187829)

[图 4.2 算法辅助说明图3 17](#_Toc516187830)

[图 4.3 典型的拥挤行人场景示意 18](#_Toc516187831)

表格索引

[表 4.1 与其余几种Online MOT算法的比较 16](#_Toc516098498)

[表 4.2 算法改进前后比较 17](#_Toc516098499)

参考文献

1. Milan A, Lealtaixe L, Reid I, et al. MOT16: A Benchmark for Multi-Object Tracking[J]. 2016.
2. Tang S, Andres B, Andriluka M, et al. Multi-Person Tracking by Multicut and Deep Matching[M]// Computer Vision – ECCV 2016 Workshops. Springer International Publishing, 2016:100-111.Zhou, Zimu, et al. "Towards omnidirectional passive human detection." INFOCOM, 2013 Proceedings IEEE. IEEE, 2013.
3. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]// IEEE International Conference on Image Processing. IEEE, 2017:3645-3649.Sauter, Martin. From GSM to LTE: an introduction to mobile networks and mobile broadband. John Wiley & Sons, 2010.
4. Sadeghian A, Alahi A, Savarese S. Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies[J]. 2017:300-311.Seidel, Scott Y., and Theodore S. Rappaport. "914 MHz path loss prediction models for indoor wireless communications in multifloored buildings." Antennas and Propagation, IEEE Transactions on 40.2 (1992): 207-217.
5. H. W. Kuhn, Bryn Yaw. The Hungarian method for the assignment problem[C]// Naval Res. Logist. Quart. 1955:83--97.Bahl, Paramvir, and Venkata N. Padmanabhan. "RADAR: An in-building RF-based user location and tracking system." INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. Vol. 2. Ieee, 2000.
6. S. Zagoruyko and N. Komodakis, “Wide residual net- works,” in *BMVC*, 2016, pp. 1–12. Proakis, John G. Digital Communications Fourth Edition. New York, McGraw-Hill, 2000.
7. Halperin, Daniel, et al. "Tool release: gathering 802.11 n traces with channel state information." ACM SIGCOMM Computer Communication Review 41.1 (2011): 53-53. Matthew D. Zeiler, Rob Fergus. Visualizing and Understanding Convolutional Networks[J]. 2013, 8689:818-833.
8. Pedro Felzenszwalb, Ross Girshick, David McAllester, et al. Visual object detection with deformable part models[J]. Communications of the ACM, 2013, 56(9):97-105.Mitola, Joe. "The software radio architecture." Communications Magazine, IEEE 33.5 (1995): 26-38.
9. F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, and J. Yan, “Poi: Multiple object tracking with high performance detec- tion and appearance feature,” in *ECCV*. Springer, 2016, pp. 36–42.
10. Sanchez-Matilla R, Poiesi F, Cavallaro A. Online Multi-target Tracking with Strong and Weak Detections[C]// European Conference on Computer Vision. Springer, Cham, 2016:84-99.

致 谢

衷心感谢丁贵广老师对于本人的悉心教导和大力支持

非常感谢实验室的李晨、陈高勋、温发琥等学长学姐的帮助与指导

非常感谢室友们在撰写论文时为我提供的帮助

感谢清华大学，感谢软件学院对我的培养

感谢父母对我一直的照顾与支持

感谢所有关心，帮助过我的人

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名： 日 期：

附录A 外文资料的书面翻译

**利用深度度量模型的简单在线实时追踪**

简单在线实时跟踪（SORT）是一种实用的，简单有效的多目标跟踪方法。 在本文中，我们整合了外观特征信息来提高SORT的性能。 由于这种改进，我们能够在更长的遮挡时间后来跟踪对象，从而有效减少ID Switch的次数。 遵照最初框架的想法，我们将大部分计算置于离线预训练阶段，在这个阶段中我们学习了大规模人员重新识别(Person Reidentification)数据集上的深度关联度量模型。 在算法在线运行期间，我们在外观特征空间中使用最近邻查询建立测量到轨迹的关联。 实验评估表明，我们的改进将ID switch的数量减少了45％，在高处理帧速率下维持了全面的竞争性能。

1. 介绍

由于目标检测方面的最新进展，基于检测的追踪已经成为多目标跟踪的领先范例。在这种范例中，对象轨迹通常存在于一次处理整个视频批次的全局优化问题中。例如，流网络公式和概率图模型已经成为这种类型的流行框架。但是，由于批处理，这些方法不适用于在每帧必须给出检测对象ID的在线追踪场景。更传统的方法是多假设跟踪（MHT）和联合概率数据关联滤波器（JPDAF）。这些方法在逐帧的基础上执行数据关联。在JPDAF中，单个状态假设是通过对各个测量值的关联可能性进行加权而生成的。在MHT中，跟踪所有可能的假设，但为了降低计算复杂度必须进行剪枝。这两种方法最近都在逐个检测场景中被重新审视，并显示出可喜的结果。但是，这些方法的性能会增加计算和实现的复杂性。

简单在线实时追踪（SORT）是一个简单得多的框架，它在图像空间中执行卡尔曼滤波，并使用利用Bounding Box的重叠度的度量方法的匈牙利算法来进行帧间匹配。 这种简单的方法在高处理帧率下实现了良好的性能。 在MOT挑战数据集[13]中，SORT在标准检测上平均高于MHT。 这不仅强调了对象检测模型性能对整体跟踪结果的影响，而且也是从实践者角度看的一个重要见解。

在跟踪精度和准确度方面实现总体良好性能的同时，SORT会出现相对较多的身ID switch。这是因为，只有当状态估计不确定性较低时，所使用的关联度量值才是准确的。因此，SORT在经过遮挡后进行跟踪时存在缺陷，而且这种情况经常出现在正视图相机的场景中。我们通过将关联度量标准替换为结合了运动和外观信息的更明智的度量标准来解决此问题。特别是，我们应用了一个卷积神经网络（CNN），该网络经过训练可以在大规模人员重新识别数据集上区分行人。通过整合这个网络，我们可以提高针对漏报和遮挡时系统的鲁棒性，同时保持系统易于实现，高效并适用于在线场景。我们的代码和预先训练的CNN模型已公开发布，以促进研究实验和实际应用开发。

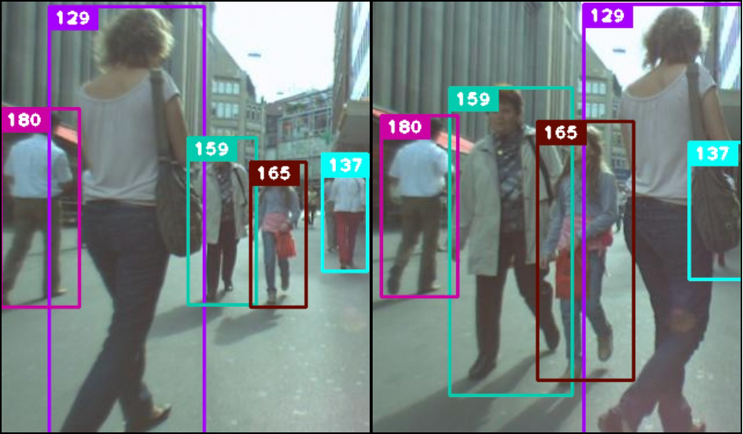


图 A.3.1 一个MOT中含有频繁遮挡的场景的示例

1. 含有深度关联度量模型的SORT

我们采用递归卡尔曼滤波和逐帧数据关联的传统单一假设跟踪方法。 在下一节中，我们将更详细地描述此系统的核心内容

1. 轨迹处理和状态估计

对于轨迹的处理和卡尔曼滤波的框架基本上是与文献同一的。我们假定这是一个非常平凡的跟踪场景，也就是说相机是未矫正过的，并且也没有自身的运动信息。虽然这些条件为追踪带来了困难，但它确是MOT竞赛中最常见的情况。因此，我们的跟踪方案被定义在了一个八维的状态空间(u,v,γ,h,u’,v’,γ’,h’)上，其中包含Bounding Box的位置信息(u,v)，宽高比γ，以及高度h，还有他们各自在图片坐标系上的相对速度。我们采用一个带有恒定运动速度和线性观察模型，也就是说把Bounding Box的坐标(u,v,γ,h)当做直接的观察状态的标准卡尔曼滤波。

对于每个轨迹k，我们计数自上次成功关联以来的帧数为ak。 该计数器在卡尔曼滤波器预测期间递增，并且当轨迹已经重新与新的观测相关联时重置为0。 超过预定最大年龄Amax的对象被认为已离开场景并从轨迹集中删除。 对于不能与现有轨迹相关联的每个检测，我们初始化新的轨迹假设。 这些新轨迹在其出现的前三帧中被分类为暂定。在这个时候，我们期望在每一帧都能成功地对其进行匹配。 在暂定状态下前三帧内没有成功关联的轨迹将被删除。

1. 关联问题

一个传统的解决预测的卡尔曼状态和新到来的测量结果之间的关联问题的方法是建立一个可以用匈牙利算法解决的关联问题。在这个问题里我们通过融合外观和运动信息来生成一个合适的度量方式。

为了纳入运动信息，我们采用平方马氏距离来衡量预测的卡尔曼滤波状态和实际侦测到的结果之间的距离:

在上式中我们用(yi, Si)来表示第i个轨迹在测量空间中的投影，并用dj表示第j个Detection的Bounding Box。马氏距离通过计算估计状态距离测量状态偏离了多少个标准差来估计状态的不确定性。更进一步，通过将马氏距离的阈值设置为逆卡方分布在95%置信度时的值，我们可以过滤掉大部分不太可能的关联配对。我们用如下的指示符来表示这个方法

如果第i道和第j道检测之间的关联是可接受的，则评估为1。 对于我们的四维测量空间，相应的Mahalanobis阈值是t（1）= 9.4877。

当运动不确定性较低时，Mahalanobis距离是一个合适的关联度量，但在我们的图像空间问题表达式中，从卡尔曼滤波框架获得的预测状态分布仅提供对象位置的粗略估计。 尤其是，未考虑的摄像机运动可能会在图像平面中引入快速位移，从而使得马哈拉诺比斯距离成为跟踪遮挡情况下的一个非常不明确的指标。 因此，我们将第二个度量集成到分配问题中。对于每一个检测的Bounding Box dj我们计算几个外貌特征描述子rj并且||rj|| = 1. 同时，我们也维护一个集合，其中包含每一个轨迹k的最后Lk=100个外貌特征描述子。随后，我们的第二个度量标准计算第i个轨迹到第j个侦测在外貌空间上的最小余弦距离。

同样的，我们引入一个阈值t(2)来对外貌特征的结果进行筛选。

我们在一个单独的训练数据集上找到了适合这个指标的阈值。 在实践中，我们应用预先训练的CNN来计算边界框外观描述符。 2.4节描述了该网络的架构。

两种指标相结合，通过服务于指派问题的不同方面而相互补充。 一方面，Mahalanobis距离提供了基于运动的可能物体位置的信息，这对于短期预测特别有用。 另一方面，余弦距离考虑外观信息，这些信息对于在长时间遮挡后恢复身份特别有用，当运动信息较难以区分时。为了建立一个指派问题的模型，我们将两种度量结合得到

当距离同时满足两个阈值条件时，得到这个匹配是可能被接受的。

对于组合中的每一个度量的影响，可以通过超参数来控制。在我们的实验中我们发现超参数是一个可行的选择，特别是当持续的相机位移存在时。在这种情况下，实际上只有外貌信息被用于在匹配代价项中。但是，马氏距离依然被用于消除一些通过卡尔曼滤波预测的不太可行的对象位置的检测框。

1. 级联匹配

我们没有在全局分配问题中解决测量到轨迹关联的问题，而是引入了一个解决一系列子问题的级联。为了启发这种方法，请考虑以下情况：当物体被遮挡较长时间时，随后的卡尔曼滤波预测会增加与物体位置相关的不确定性。因此，概率质量在状态空间中扩展并且观测可能性变得较小。有意思的是，关联度量应该通过增加测量到轨迹的距离来解释概率质量的这种扩展。直觉上，当两条轨道竞争相同的检测时，马氏距离有利于较大的不确定性，因为它有效地减少了任何检测的标准偏差距预计轨迹平均值的距离。这是一种不希望的行为，因为它会导致轨迹切割的增加和轨道不稳定。因此，我们引入一个匹配的级联，该级联优先考虑更频繁出现的对象，以便对关联概率中概率扩展的概念进行编码。

Listing 1概述了我们的匹配算法。 作为输入，我们提供轨道T和检测D指标以及最大存留时间Amax。 在第1行和第2行中，我们计算关联成本矩阵和允许关联矩阵。 然后，我们迭代轨迹年龄n来解决年龄增长轨迹的线性分配问题。 在第6行中，我们选择最后n帧中未与检测关联的轨道Tn的子集。 在第7行中，我们解决了Tn和为匹配的检测U之间的线性匹配问题。在第8行和第9行，我们更新了匹配和不匹配检测的集合，我们在第11行完成后返回。请注意，该匹配级联优先考虑较小年龄（指最后一次出现到现在的时间）的轨迹，即最近已经看到的轨迹。

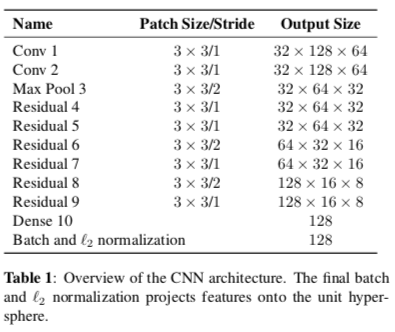
在最后的匹配阶段，我们运行原始SORT算法中提出的IOU算法，该算法包含age n = 1的未确认和未匹配轨道。这有助于解决突然出现的变化，例如：部分遮挡，并增加系统对错误初始化的鲁棒性。



1. 深度外貌特征描述子

我们的方法需要一个离线预训练好的特征提取网络才能够使得系统成功的在使用最近邻查询的情况下进行在线应用。为此，我们采用了一个CNN网络，该网络已经在大规模人员重新识别数据集[21]上进行了训练，该数据库包含超过1,100,000个图像的1,261位行人，使其非常适合在行人跟踪应用中进行深度学习。

我们的CNN网络的架构如表1所示。总之，我们采用了一个宽的残差网络，其中有两个卷积层，后面是六个残差块。 在dense layer 10中计算维度128的全局特征图。，本网络共有2,800,864个参数，在Nvidia GeForce GTX 1050 GPU上，一个包含32个Bounding Box的前向传播大约需要30 ms时间。 因此，只要有一个现代的GPU可用，该网络非常适合用于在线追踪。但是网络的细节不是我们本篇文章的重点。



1. 实验

我们评估了在MOT16基准测中跟踪器的性能。 该基准评估七个具有挑战性的测试序列的跟踪性能，包括带移动摄像头的正面视图场景以及自顶向下的监视设置。 作为我们跟踪器的输入，我们依赖Yu等人提供的检测。 他们在公共和私人数据集进行了更Fast RCNN训练，以提供出色的性能。 为了公平比较，我们重新运行基于相同检测子的SORT算法。

我们将实验参数设置为和Amax=30 帧。检测的阈值被设置为0.3 。我们剩余的参数是从训练数据集上独立获取的。评估主要关注以下几个度量标准：

* MOTA(Multiple Object Tracking Accuracy):是主要的总体衡量标准，包括false positive, missed targets 和 id switches.
* MOTP(Multiple Object Tracking Precision):表示预测的Bounding Box和标注之间的不重合程度
* MT(Mostly Tracked):至少在其生命周期内80%被跟踪到的轨迹的比例。
* ML(Mostly Lost)：在其生命周期之内至多跟踪到20%的轨迹的比例。
* IDs(Identity switches):表示标注的ground truth上的轨迹变换id的次数。
* FM(Fragementation)：追踪轨迹被丢失的检测打断的次数

我们的评估结果如表2所示。我们的适当的成功地减少了IDs的数量。 与SORT相比，IDs从1423减少到781，这减少了大约45％。 与此同时，由于在经过遮挡和检测遗漏后仍要保持对象的ID，轨迹碎片略有增加。 我们还看到MT的数量显着增加，和ML的减少。总的来说，由于整合了外观信息，我们在经过较长的遮挡后成功地保持了原有ID。 这也可以通过我们在补充材料中提供的跟踪输出的质量分析来看出。 我们的跟踪器的示例性输出如图1所示。

我们的方法也是其他在线追踪框架的强有力竞争者。具体而言，我们的方法返回了所有在线方法中最少的身份切换(IDs)次数，同时保持了MOTA的较高得分。追踪器的跟踪精度主要受到大量假正例(False Positive)的影响。鉴于它们对MOTA分数的总体影响，对检测结果应用较大的置信度阈值可能会大大增加算法在评估报告显示的性能。然而，跟踪输出的视觉检查表明，这些误报主要是由静态几何场景中的偶发检测器响应产生的。由于我们的最大允许跟踪时间相对较长，因此这些偶发错误响应更频繁的被加入对象轨迹。同时，我们没有观察到经常在误报之间跳转的轨迹。相反，跟踪器通常会在被报告的对象位置产生相对稳定的静止轨道。

我们的系统以大约20Hz的频率运行，其中大约一半时间用于特征生成。因此，给定一个现代化的高效GPU，该系统能够保持高效且实时运行。

1. 结论

我们已经展示了一个基于SORT的融合了从预训练外貌特征提取网络的改进。由于这些改进，我们使得SORT能够在更长时间的遮挡之后保持原追踪轨迹，这使得SORT成为现今的在线追踪算法的一个有力竞争者。同时，算法依然较容易实现并可以实时运行。

**原文索引**

1. Wojke N, Bewley A, Paulus D. Simple Online and Realtime Tracking with a Deep Association Metric[J]. 2017:3645-3649.
2. Alex Bewley, Zongyuan Ge, Lionel Ott, et al. Simple online and realtime tracking[J]. 2016:3464-3468.

综合论文训练记录表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学生姓名** | 陈政宇 | **学号** | 2014013461 | **班级** | 软件42 |
| **论文题目** | 多目标追踪关联问题研究 | | | | |
| **主要内容以及进度安排** | 主要内容：  探究了多目标追踪研究中的关联问题，通过研究调研文献，实现了一个在线的实时多目标追踪系统，并对其关联方法进行了改进  进度安排：  第2-4周：前期调研与方案研究  第5-8周：文献研究、确定初步的方法和实验框架  第9-12周：进行初步的实验，总结问题  第12-16周：改进实验与论文撰写  **指导教师签字：**  **考核组组长签字：**  **年 月 日** | | | | |
| **中期考核意见** | **考核组组长签字：**  **年 月 日** | | | | |
| **指导教师评语** | **指导教师签字：**  **年 月 日** | | | | |
| **评阅教师评语** | **评阅教师签字：**  **年 月 日** | | | | |
| **答辩小组评语** | **答辩小组组长签字：**  **年 月 日** | | | | |

**总成绩：**

**教学负责人签字：**

**年 月 日**