****

**本 科 生 毕 业 论 文（设计）**

题 目：面向视频的人体骨架提取与行为识别

院 系： 数据科学与计算机学院

专 业： 软件工程（移动信息工程）

学生姓名： 杨 邦

学 号： 14353349

指导教师： 朱俊勇

（职 称） 特聘副研究员

二〇一八 年 月

【摘要】 ————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————

【关键词】深度学习，骨架提取，行为识别

**[Abstract]** ————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————————

**[Keywords]**

目 录

第1章 前言 1

1.1行为识别的背景和意义 1

1.2行为识别的研究现状 2

1.2.1 特定场景内人数估计 2

1.2.2 经过指定线人数估计 4

1.3 本文的研究内容 6

1.4本文的组织结构 6

第2章 视频中特定区域的人数估计 8

2.1角点 8

2.2 稀疏光流法追踪 9

2.3特定区域人数估计方法 11

2.4 本章小结 14

第3章 视频中穿过指定线的人数估计 15

3.1 HOG 检测子简介 17

3.1.1计算图像梯度图 18

3.1.2空间分块投票 18

3.1.3 Block内的归一化 19

3.1.4 3-D融合检测结果 19

3.1.5 积分图 21

3.1.6 支撑向量机分类 22

3.2粒子滤波器 22

3.2.1算法框架 22

3.2.2 蒙特卡罗模拟 24

3.2.3序列重要性重采样 25

3.2.4重采样 26

3.3 在线boosting分类器 27

3.3.1 Boosting框架 28

3.3.2特征选择子 29

3.3.3特征库 30

3.4数据匹配 33

3.5 穿过特定线的人数估计方法 34

3.5.1 HOG部分 34

3.5.2 在线分类器部分 35

3.5.3 粒子滤波器部分 36

3.5.4 模型假设部分 37

第4章 实验分析 40

4.1 特定场景的人数估计 40

4.2 经过指定线的人数估计 42

第5章 结 论 48

参考文献 49

致谢 54

# 前言

## 1.1行为识别的背景和意义

动作识别是基于视觉的人体运动理解的重要任务，其广泛应用于智能监控系统，人机交互系统等应用领域。随着现代社会的不断发展，智能环境的应用已经在人们工作生活的各个方面全面展开，而为了实现更好的人机交互，更加敏锐与智能地判断人们的需求，行为识别变得愈发重要。

在行为识别中，有两个关键和互补的方面：外形和动态变化。识别系统的性能在很大程度上取决于它是否能够提取和利用相关信息。然而由于一些复杂的因素，如尺度变化、视角变化和摄像机运动，提取这些信息是不容易的。因此，设计有效的既能处理这些挑战同时能保持行为类别信息的表示方法是至关重要的。卷积神经网络近年来广泛引用于计算机视觉中，包括分割、分类、识别等任务。在行为识别领域，基于深度学习的方法同样以卷积神经网络为主流，其不需要像传统方法那样对特征提取方法进行人工设计，可以直接在视频数据上进行训练和学习，得到最有效的表征方法。

行为识别原本主要在彩色视频上进行研究，由于视频的数据十分庞大，对内存需求高，因此在视频级别上训练深度网络是不可行的。为解决该问题，大多方法采用采样或者截取小段视频区间的方法来训练网络。如[1]从视频中采样25帧，用以训练双流网络；[2]用了16帧以及60帧的视频片段用以训练3D卷积神经网络。但在如何保持关键信息的完整性这个方面，还有待考虑及进一步研究。

随着近年来Kinect、RGBDsense等感知设备的逐渐发展，以及在全三维空间捕捉人类行为可以提供更全面的信息，基于RGB-D数据和骨架序列数据的动作识别研究已经引起了许多学者的兴趣。大规模的数据集如NTU RGB+D[3]、Kinetics[4]相继涌现，让近年来逐渐以深度学习为主流的行为识别方法受益无穷，让构建及训练更深的网络成为可能。不像NTU RGB+D能够提供视频、深度图、3D骨架序列、多角度等，一些早期的数据集如HMDB51[5]、UCF101[6]仅能提供视频。这些数据集上，光流、RGB等信息从视频中被充分利用，然而很少人利用姿态估计方法提取出人体骨架信息进行进一步利用，本文则会聚焦该问题进行分析研究。

尽管他们的目标不尽相同，但是这两个任务是高度耦合的:当确定了骨架节点，相应的肢体也就能呈现出来，而肢体的运动方式无疑非常利于进行行为识别；相反，行为的正确识别也能对缺失的骨架节点的推测有着巨大帮助。

随着技术的进步，骨架提取以及行为识别的聚焦点都已从单张静态图片转移到视频上，效率方面也从十分耗时向追求实时而进行转变。目前骨架提取的数据来源主要是普通摄像头得到的RGB图像或者诸如Kinect、RGBDsense等设备捕捉的深度图。前者适用范围广，许多数据集都是由普通摄像头所拍摄的图像、所录制的视频片段组成；后者则由于对感知设备有要求，因此范围受限。对于骨架提取而言，其主要困难在于人体具有很大的姿态范围，这也导致了可能会出现遮挡等无法检测的情况；对于行为识别而言，当研究的对象是视频时，摄像头的晃动、与动作行为无关的干扰信息均会对行为的正确识别造成巨大影响。【此段大量引用】

近几年来，随着数据集的增大等诸多条件的成熟，深度学习在行为识别以及姿态估计中都取得了越来越好的性能，许多数据集上最先进的结果皆与深度学习有关。目前骨架提取大部分都用到了CNN；而在行为识别方面，深度学习取得的性能已然超越传统方法，CNN、RNN、LSTM等网络均有不错的成效。

在已有的数据集如HMDB51、UCF101及 XXX 上，利用RGB、光流信息，很少有考虑到从数据集视频中提取人体骨架信息。

## 1.2行为识别的研究现状

目前行为识别大致有以下几种方法：

**1.2.1传统行为识别方法**

传统方法基本遵循三个步骤：1）兴趣点检测；2）局部描述子提取；3）局部描述子的融合。

**1.2.1.1兴趣点检测**

为了建造一个时空兴趣点检测器,2D-Harris检测器被广泛使用，该检测器的想法是找出图像中两个正交方向有显著变化的空间位置。为了应对视频方面的表示，2D-Harris检测器被拓展成3D-Harris检测器[7]，其能识别具有大空间变化和非恒定运动的点。同样被在2D中广泛使用的Hessian检测器，同样被拓展成了3D的形式[8]。不同于3D-Harris检测器使用梯度来检测兴趣点，3D-Hessian检测器利用二阶导数进行判定。但[9]研究表明，3D检测器真正所感兴趣的时空兴趣点在复杂的环境中是十分罕见的。为了克服这个局限性，[9]提出将空间滤波与时间滤波相分离，使得检测器可以在复杂的运动中响应出空间特征。

**1.2.1.2局部描述子提取**

局部描述子主要能分成三类。1）边缘和运动描述子：方向梯度直方图（HoG）能够作为运动描述子[10]，在其基础上，为了能在时空域上同样能捕获到运动信息，HoG被拓展为HoG3D[11]。除此之外，光流场能够对视频中的像素级运动进行编码，基于此的描述子也有很多，如光流直方图（HoF）[12]、运动边界直方图（MBH）[13]等。2）像素模式描述子，局部二值模式（LBP）是一种用来描述图像局部纹理特征的算子，具有旋转不变性和灰度不变性等显著优点，其成功应用于包括人脸识别在内的多种视觉问题。3）轨迹描述子，由于兴趣点并不总是固定在同一个长方形区域中，因此从该区域提取的特征可能不一定能够描述兴趣点本身。轨迹则是随着时间的推移能够正确跟踪兴趣点的特征。轨迹的速度，轨迹之间的相对运动都可以表征某些动作类别[14][15]。

**1.2.1.3局部描述子的融合**

局部描述子的融合就是为了将一组局部特征聚集成具有区分性，同时又有固定大小的描述符，然后才能使用一些学习算法如SVM进行学习。BoV（Bag-of-Visual Words）是其中一种主流的方法。在BoV中，“视觉词”出现频次的直方图被当做描述子，其在[16]中最先被应用于行为识别中。FV（Fisher Vector）[17]近些年来也成为一个不错的选择，它同时将生成方法和判别方法的优点结合到模式分类中。除此之外，稀疏编码原则[18]、字典学习[19]、时间一致性[20]同样能用来聚合局部时空特征。

**1.2.2基于深度学习的行为识别方法**

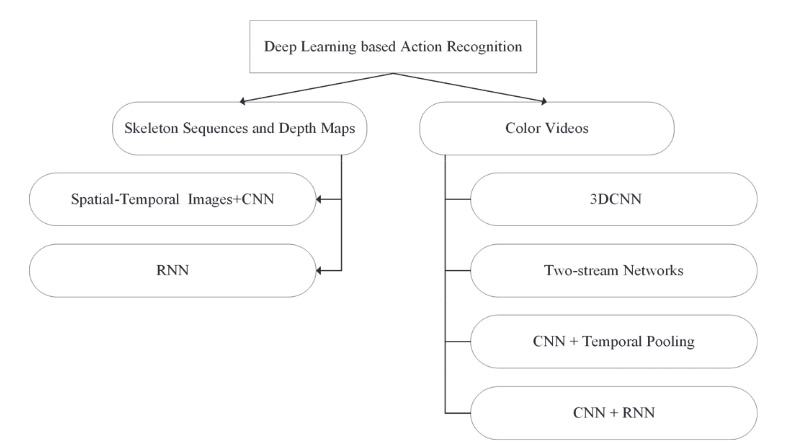


图1.2.1 基于深度学习的行为识别方法

如上图1.2.1所示，深度学习方法主要是建立在彩色视频或者骨架序列及深度图上，接下来分开介绍。

**1.2.2.1基于彩色视频的方法**

卷积神经网络（CNN）近年来广泛引用于计算机视觉中，包括分割、分类、识别等任务。基于深度学习的行为识别方法不需要像传统方法那样对特征提取方法进行人工设计，可以在视频数据上进行训练和学习，得到最有效的表征方法。基于图像分析时，CNN使用的是二维卷积（即卷积核的维度为二维），但迁移到视频分析时，2D并不能很好地捕获时序上的信息，因此3D卷积被提出[21]。Facebook提出了C3D[22]（3D convolutional networks）网络，其采用了3D卷积和3D池化构建了网络。

[23]最先提出双流网络（Two-Stream Convolutional Networks），即空间流（视频RGB图像）与时间流（光流信息）。对空间与时间信息分别训练CNN模型，接着两个分支的网络的评分进行融合，即可得到最终的分类结果。多个分支网络的评分融合有许多方式，如SVM、取平均等。之后有许多研究工作都建立在双流网络的基础上。[24]利用双流网络提取特征。[25]双流网络模型均换成了VGG16，同时利用了CNN网络进行两个分支网络的评分融合。[26]双流网络的模型则是采用VGG16或者Inception-BN，训练完毕后则从该双流网络的卷积层或全连接层提取视频的特征表示，最后利用得到的特征训练分类器预测类别。对于循环神经网络及其拓展模型，一种比较典型的做法就是将视频序列经过CNN处理提取出特征数据，然后再经过若干层LSTM进行最后的判别[27][28]。[27]除了应用视频序列外，还运用了光流信息；而[28]应用的场景更广，可以完成动作识别、图片描述以及视频描述三种任务。

**1.2.2.2基于人体骨架的方法**

谈基于人体骨架的方法前，首先了解一下人体骨架数据的来源。姿态估计，即是要准确的得到人体关节点的坐标。目前在2D多人姿态估计方面，主要有两种方法，即自上而下的方法（Top-Down）和自下而上(Bottom-up)的方法。自上而下的方法[29][30]首先使用人员检测器，框出目标帧中所有人的位置，紧接着利用现有技术[31][32]再对每个人执行单人姿态估计。但是这种方法的缺点也是明显的：1）当人十分靠近时，容易导致人员探测器检测失败，在这种情况下，整个人的姿态根本无法估计；2）显然这种方法的运行时间与人数成正比，当待检测的人数较多时，性能无法得到保障。相反，自下而上的方法有更强的稳健性，并且能够使得运行时间与目标帧人数不相关，虽然其实现较难，但在实时估计效率方面具有较大优势。自下而上的方法主要思想是先得到所有候选的关节点坐标，然后再通过聚类[33]、图表匹配[34]或其他方法决定各个节点属于哪一个人。在实践中，Cao和Simon等人[34]利用自下而上的方法能够针对多人做到实时预测。他们利用两路CNN的网络框架：一路用来获得置信图，从而得到关节点坐标；另一路则能获得PAF（Part Affinity Fields），用以展示肢体的位置和方向。最后用图论的方法将候选关节点两两连接并且不重复。该方法无论在准确率上还是在效率都有着质的飞跃，尤其是效率：其运行效率比DeeperCut[35]提高了几个数量级。

骨架提取除了上述提到从图像中提取外，还能通过感知设备直接获取。随着3D行为识别近年来吸引越来越多的注意力以及Kinect感知设备的逐渐发展，越来越多的用于3D行为识别的数据集相继出炉，如最早的MSR-Action3D[36]，还有被越来越多论文拿来当做基准的NTU RGB+D数据集[3]。LSTM（Long Short-Term Memory）已经很好地用于模拟骨架序列的时间模式,并且有许多改进。例如，[37]探讨了骨骼关节的共现特征。[38]在空间和时间领域开发了关注模型。由于CNN网络在许多图像识别任务上取得了巨大的成功，许多工作研究如何将骨架序列作为CNN网络的输入。[39]提出了骨架轨迹图的概念，利用色彩映射成功将骨架序列转变成RGB图像。[40]提出了骨架距离图的概念，通过关节点两两之间的距离以及色彩映射，同样实现了序列到三维图像的转变。[41]则是结合了LSTM以及CNN，同时结合了前人提出的骨架轨迹图以及骨架距离图，最终取得了十分可观的效果。

## 1.3 本文的研究内容

文章分为两个部分，包括1）视频中人体骨架提取，及通过骨架点信息训练模型；2）视频采样得到RGB图像，通过图像训练模型。

在第一部分，我们采用Cao等人[]提出的OpenPose模型提取人体骨架。但由于视频中存在无关背景干扰、视频抖动、人体出现部位不全等情况，导致骨架节点信息可能是稀疏的，于是我们利用Lin等人[]提出的矩阵补全方法进行改善，提高骨架信息的稠密程度。距离

在第二部分，我们将视频均匀采样，利用pHash、aHash等相似图片检测方法剔除无关背景信息，保证采样的图片是相似的。紧接着我们利用已于ImageNet上预训练过的VGG19网络及XXXX网络，提取图片的特征，利用特征训练分类器。

首先，数据集采用UCF101以及HMDB51，将视频均匀采样（这里可以改进成在感兴趣的区域重点采样，不过这个涉及到另外一个方面——定位感兴趣区域，可能时间不允许去拓展了），每个视频均采样N帧（N在这里取25）。

接下来，主要分为3部分：

1. 将每一帧RGB静态图像输入已在ImageNet上预训练好的VGG16（or VGG19）网络，在第k层提取特征（此处k根据对比试验来决定，到底采用哪个网络的哪一层输出会使得结果更精确）。接着利用该层输出的特征训练一个softmax回归模型。接着看看该模型对于行为识别的预测准确率如何
2. 利用已开源的openpose，得到RGB静态图像上人物的骨架关键点坐标信息，假设有m个关键点，n帧（这里n要不要与上面的N一致暂时未确定），则会生成[m, 2, n]的时序信息矩阵（第二维为x, y坐标）。或者在这个矩阵基础上，通过后一帧对前一帧做差生成矢量矩阵。如手腕在第n-1帧坐标为（x1,y1），在第n帧坐标为（x2, y2），则手腕关键点在n-1帧及n帧间的矢量则为（x2-x1, y2-y1）,最终矢量矩阵大小为[m, 2, n-1]。紧接着，同样训练一个softmax回归模型，看看准确率如何。

这里会有一些问题：

1. 可能很多帧检测不到全部关键点，甚至一个都没有，如何处理？
2. 受到干扰，检测到的关键点是错误的，又该如何处理？
3. 可能1与2的准确率均一般，接下来研究如何将两者耦合起来，能得到更好的结果（平均 or 带权重 or ...）

（第二部分我还是选择训练softmax回归模型，是不是会有更好的选择，如LSTM？还有第三部分的耦合，如果都是通过softmax回归得到结果，最后就很方便，只需要将结果加权起来之类的。而如果第二部分换成其他的话，两者的结合可能就又得好好考虑考虑了，老师您有什么建议呢？）

挑选骨架点提取 良好 的类别S，进行训练

记录spatial中准确率中较低的类别，看看是否存在于S中，若存在，则可以通过它来提高准确率。

Spatial的对比：

1. 是否取10部分（一张图的四个角+中心，再+其水平翻转）
2. 基于1，再进行两种网络不同层的feature的结果对比

引入骨架点之后的对比：

Spatial中最好的结果 与 +入骨架点补充的结果 对比

## 1.4本文的组织结构

第一章，前言。介绍行为识别的背景和意义。同时介绍了国内外研究现状及分析了视频中行为识别的难点。简要介绍了本文的主要方法和文章结构。

第二章，具体介绍骨架提取的方法，以及得到骨架信息矩阵后如何进一步利用以得到骨架轨迹图和骨架距离图；解决骨架信息矩阵稀疏性问题并且构建两路输入的深度网络。

第三章，具体介绍从视频中采样得到快照的方法；利用灰度直方图和图像指纹剔除不相似的采样帧；构建以CNN+LSTM为核心框架的空间信息网络。

第四章，实验结果的分析。介绍所使用的数据集，在数据集上测试所提出的方法，并评价试验结果。对试验中某些特定场景做出解释和分析。

第五章，对现有工作的总结和对未来工作的展望。如何进一步提高检测的准确率、实时性和鲁棒性。

# 骨架信息网络构建

尽管我们的研究对象是视频，但是视频是由一帧帧的RGB图像组成，因此只要实现了在RGB图像上提取人体骨架，即可得到整个视频流中人体骨架节点信息。本文骨架提取框架采用Cao等人开源的OpenPose进行骨架信息提取。

本章将会介绍：1）视频骨架信息的提取；2）骨架信息补全；3）

## 2.1 OpenPose骨架提取

在本文中，骨架节点一共18个，包括鼻子、脖子、左右眼、左右耳、左右肩、左右手肘、左右手腕、左右臀部、左右膝盖及左右脚踝。OpenPose提出了PAF（Part Affinity Fields）概念，其能展示肢体的关联方式（也即肢体的方向）。OpenPose采用了自下而上方法以及使用了二路多阶段CNN网络框架，在训练网络时，在每一个新的阶段都对上一个阶段的置信图和PAF进行修正。在训练好网络后，其骨架提取的流程可分为三步：1）得到身体部位检测的置信图，即得到关节点的位置；2）得到PAF，即得到关节点的关联方式；3）最后通过二部图匹配的方式得到每个人的骨架。

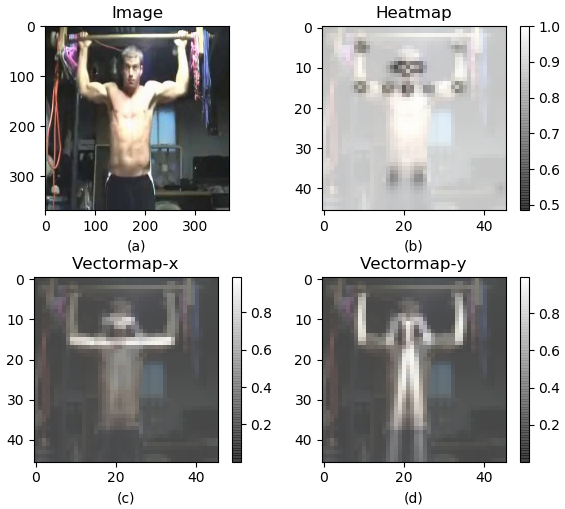
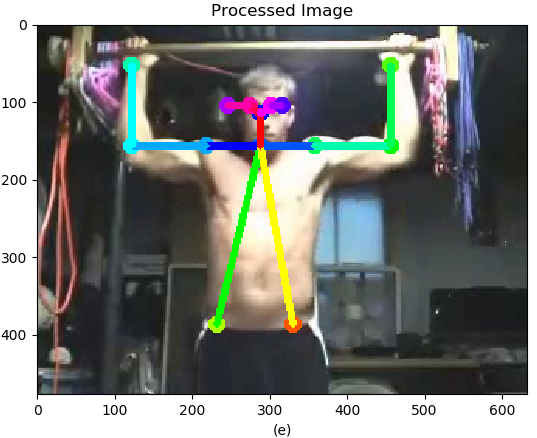
 

图2.1.1 骨架提取过程示例

如上图2.1.1所示，(a)为原图像，图中的人正在做引体向上；(b)为热图（也即上文提到的置信图结果），图中的深色圆形区域代表检测到的关节点位置，检测到了：鼻子、脖子、左右眼、左右耳、左右肩、左右手肘、左右手腕、左右臀部；(c)、(d)则是上文所提到的PAF（肢体的关联方式），其中(c)为水平方向，(d)为垂直方向。最终骨架提取的结果通过图像的方式展示出来，如(e)所示。

## 2.2 骨架轨迹图

由于卷积神经网络的输入要求是三维图像，因此我们需要将骨架矩阵信息转变成骨架轨迹图。令骨架轨迹图为*JTM*（Joint Trajectory Maps），则其可以表示为：

(2.2.1)

其中是表征在第*i*帧的时空信息的函数。假设一个视频有帧，则有帧骨架信息，设骨架节点为，则有 ,其中表示在第*i*帧的骨架点坐标信息，表示在第*i*帧第*j*个骨架节点的*2D*坐标信息。则有：

(2.2.2)

骨架轨迹*T*可以定义为：

(2.2.3)

其中,第*i*帧第*k*个骨架点轨迹为。在这种情况下：

(2.2.4)

如图2.2.2(a)所示，此时的轨迹图仅能表征出骨架信息的空间信息，即只能看到骨架节点从1至n帧的运动轨迹。为了能够在骨架轨迹图中体现出时间信息，我们采用颜色映射的方法。在这里，我们采用jet热力图，其色域是从蓝色到红色，而中间经过青色、黄色和橙色，如图2.2.1所示。设jet一共有阶，在第阶的颜色为。令表示的颜色，则。因此我们给骨架轨迹图附上了颜色，如赋为蓝色，赋为红色,在这种情况下：

(2.2.4)

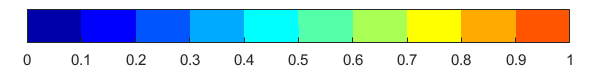


图2.2.1 10阶jet热图

如图2.2.2所示，假设这是右手腕在1至n帧进行的直线运动，(a)图仅仅携带了空间信息，(b)图同时携带了空间与时间信息。通过颜色映射，我们丰富了骨架轨迹图携带的信息。

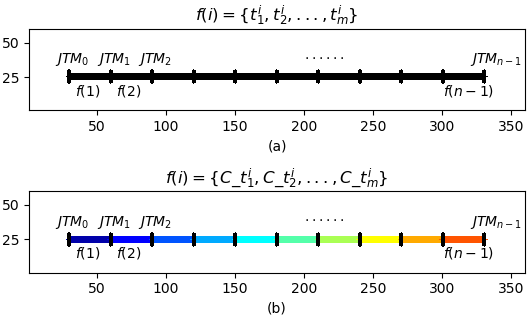


图2.2.2 与 对比

如图2.2.3所示，分别展示了a)运球、b)打高尔夫球、c)抓、d)坐起来、e)潜水、f)鼓掌这六种行为的骨架轨迹图。

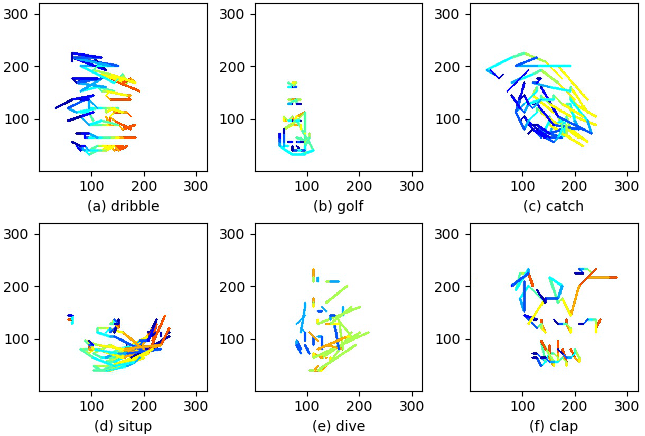


图2.2.3 几种行为的JTM图示例

## 2.3骨架距离图

在骨架轨迹图中，我们利用了骨架节点的相对位置信息。除此之外，骨架节点两两之间的距离信息同样能够被利用。由于有m个骨架节点，所以每一帧都有对距离信息。令骨架距离图为*JDM*（Joint Distance Maps）,则有：

(2.3.1)

其中表示将距离大小映射成颜色的映射，表示所有骨架节点彼此欧几里得距离的集合。将距离映射成颜色，我们同2.3一样采用jet热图，这里采用256阶。因此对于一个拥有帧的视频，最终我们能得到[n \* ]的二维矩阵。设矩阵中最大距离、最小距离分别为，对于距离，其颜色映射为：

(2.3.2)

为了满足如VGG19网络输入所需要的224\*224，我们采用双线性插值的方法将该矩阵固定成我们想要的维度。如图2.3.1所示，分别展示了a)爬楼梯、b)倒立、c)踢球、d)倾倒、e)推、f)剑术练习这六种行为的骨架距离图。

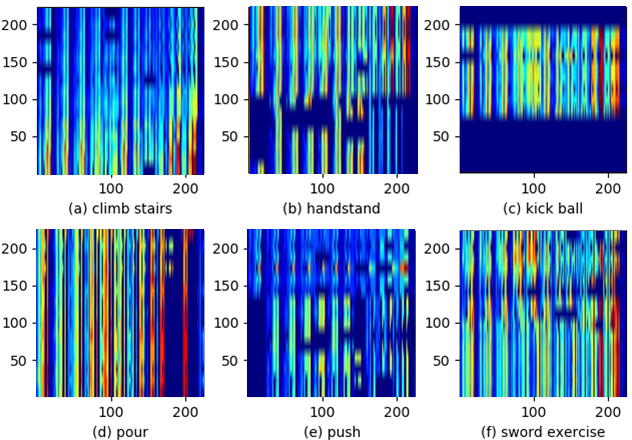


图2.3.1 几种行为的JDM图示例

## 2.4 骨架矩阵信息补全

由于遮挡、光照、噪声等原因，对于普通视频进行骨架提取最后得到的骨架信息矩阵大部分是较稀疏的。为了解决稀疏问题，本文采取线性插值的方式。令表示在[1, n]帧的视频序列中能够成功检测到的集合，则有：

(2.4.1)

所以在帧区间以及没有检测到的位置信息，这两个区间我们不执行插值。在帧区间中没有检测到的位置信息可以通过前后两个端点进行线性插值，如, 由以与为端点的线段决定。如下图2.4.1所示，我们可以看到几种行为线性插值前后的区别的例子。

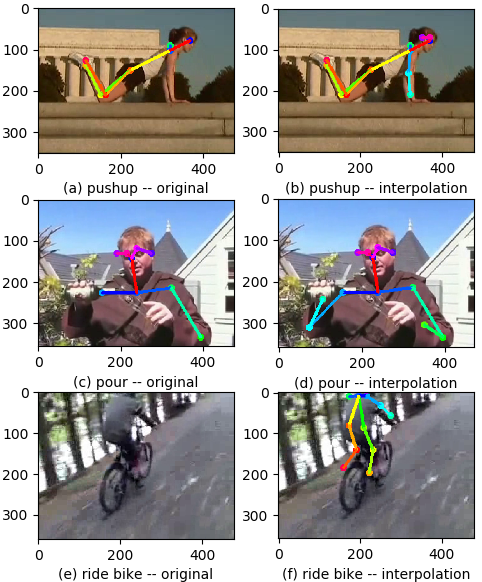


图2.4.1 几种行为的骨架信息插值前后对比

在2.4.1(a)中，人的手肘和手腕并没能识别出来，而由于前后几帧有这样的信息，(b)的插值效果上佳；同样地，(d)相较(c)成功重新获得右手肘、右手腕、左手腕的位置信息；在极端的情况下，如2.4.1(e)所示，在该帧下人体所有的关节点都没有识别出来，而线性插值后(f)还是能较好地恢复信息，尽管左膝盖以及左脚踝的位置有些偏差。然而，并不是所有的插值的效果都是理想的，如下图2.4.2所示

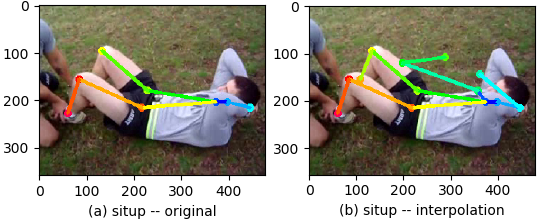


图2.4.2不理想的插值效果

见图2.4.2(a)，由于人物坐起来时能够检测到手肘的位置，之后再躺下去时就检测不到了。正因为如此，如(b)所示，手肘的位置在插值后就“残留”在了错误的位置。所以线性插值能有效降低骨架信息矩阵的稀疏度，同时它既有理想的插值效果，也有不理想的。因此在接下来的骨架信息网络构建中，我们将综合利用插值前后的效果图，实现两路输入的网络。

## 2.5 骨架信息网络

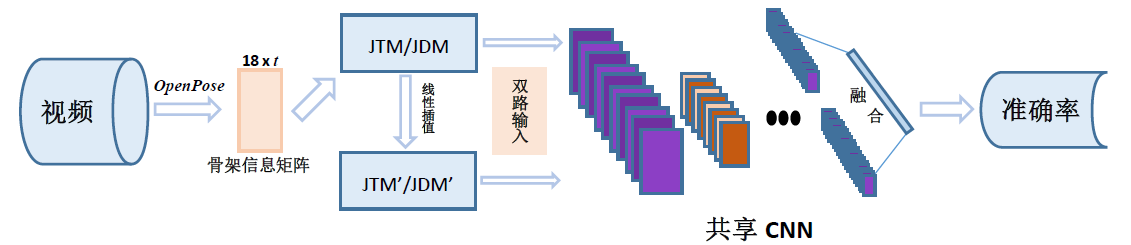


图2.5.1 骨架信息网络框架a

上文2.4中提到了使用线性插值来降低骨架信息矩阵的稀疏度，因此原本的JTM和JDM在插值后也会有所不同。为了综合利用插值前后的效果图，在构建骨架信息网络时，我们就是采用两路输入的方式，经过共享的网络（如已在ImageNet上预训练过的InceptionV3网络）得到两个特征，将这两个特征取平均融合，再通过三个全连接层得到最后得到行为分类。框架如上图2.5.1所示，其步骤可分为：a）使用OpenPose从视频中得到骨架信息矩阵。其中矩阵大小为[18 x *t*]，18表示关节点个数，*t*代表视频的帧数；b）通过上文2.2、2.3的方法，利用骨架信息矩阵生成和；c）使用线性插值得到插值后的和；d）将与，或与输入共享CNN网络，得到两个特征后进行平均融合；e）融合后的特征通过全连接层或SVM等预测行为识别类别，得到最终预测准确率。

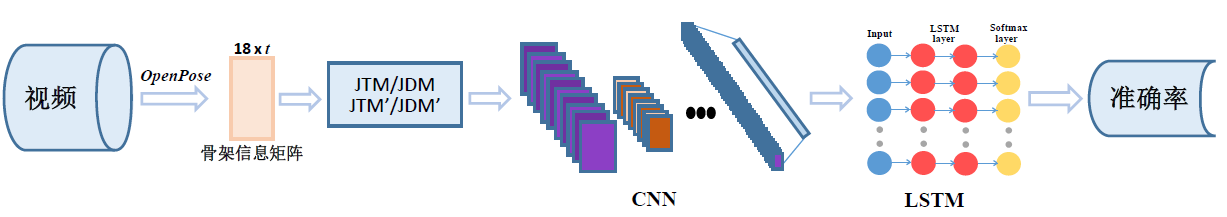


图2.5.2 骨架信息网络框架b

同时我们也考虑另外一种框架（如上图2.5.2所示），利用CNN网络提取生成的JTM/JDM图片的特征序列，再将该序列通过两层LSTM进行最后的判别。在第四章我们会研究JTM和JDM在两种框架下的表现。

## 2.6 本章小结

为了从视频序列中提取有效的人体骨架信息，我们利用OpenPose，从视频序列中得到了较稀疏的信息矩阵。为了进一步从信息矩阵中抽取出能表征时间的特征并将其转换成RGB图像，我们提出使用骨架轨迹图（JTM）和骨架距离图（JDM）：

1. 骨架轨迹图：有效展示了人体各个关节点在一系列帧中的轨迹，同时为了在轨迹中凸显出时间信息，采用jet色图赋予不同时间点不同颜色。
2. 骨架距离图：先生成时间序列中俩俩关节点的距离，再使用jet色图赋予不同距离不同颜色，成功将二维矩阵转变成RGB图像。

为了解决骨架信息矩阵的稀疏性问题，我们提出使用线性插值的方法。线性插值后，稠密程度平均上涨14.66%。由于插值后既有理想的插值效果，也有不理想的效果，我们的骨架信息网络的输入设置为两路，这样就能综合考虑，优劣互补。

在骨架信息网络中，我们提出了两种框架：a）两路输入同时经过已在ImageNet上预训练的InceptionV3网络后各得到一个1x2048维的特征，将这两个特征取平均融合，再通过三个全连接层得到最终的行为分类；b）将生成的JTM或JDM经过已在ImageNet上预训练的VGG16、VGG19网络后得到1x512维的特征，将特征经过两层LSTM层得到最终的行为分类。在第四章的实验中，我们将会研究JTM和JDM在两种框架下的表现。

# 视频序列采样及空间信息网络框架

在本章中，我们将介绍视频序列的采样方式，图像相似度的衡量以及空间信息网络框架。

## 3.1视频序列采样

由于视频相较于图像内存大得多，因此将其作为训练CNN或RNN的输入是不现实的。对于这个问题，目前的方法[A][B]是从原视频中采样少量帧或者选取很短的视频片段。本文的做法同[A],视频被均匀划分为片段：，接着从每一个视频片段中随机取得一张快照：，这样这张快照（采样帧）就能代表视频。由于视频可能来源于电影，所以视频中可能只有小部分区域才是动作发生的区域，存在无关背景的干扰。为了解决噪音问题，我们采用灰度直方图以及图像指纹两种方法来计算图像相似度，进而剔除掉不相似的采样帧。

令表示图像的灰度直方图，表示图像与的相似度，则有：

(3.1.1)

式子(3.1.1)首先是将图像转变成灰度图，然后再计算相似度。由于RGB图像是有三个通道，同样的方法，我们可以分别计算、、三个通道的相似度再取平均，则有：

(3.1.2)

除了通过灰度直方图计算两张图片的相似度，我们还能通过图像指纹的方式来计算相似度。图像指纹简单来说，就是将图像按照一定的哈希算法，经过运算后得到一组二进制数字。比较两张图像的指纹（二进制数字），计算其汉明距离，距离越大说明相似度越低。此处采用平均哈希法（aHash）以及感知哈希算法（pHash）。执行平均哈希法，首先需要将图像变成灰度图并缩小成(8 \* 8)大小得，共64个像素值，即图像的指纹长度为64。令表示图像的指纹，表示图像的平均像素值,表示图像每个位置的像素值。则有：

(3.1.3)

(3.1.4)

其中式(3.1.4)中代表汉明距离，为异或操作。执行感知哈希算法，首先需要将图像变成灰度图并缩小成(32 \* 32)大小得，对图像进行离散余弦变换，可得到(32 \* 32)大小的矩阵，取其左上角的(8 \* 8)区域得，于是有：

(3.1.5)

(3.1.6)

由于这四种评判算法各有优劣，因此我们判定图像与不相似，当且仅当满足以下条件：。采用这样的判断准则，在生成张快照时剔除不相似的图片。



图3.1.1 筛选并剔除不相似快照

筛选并剔除不相似快照的效果可见上图3.1.1，我们可以看见：(a)剔除了无效的地面背景；(b)剔除了与射箭动作无关的箭靶背景；(c)剔除了无效楼房背景；(d)剔除了无效观众背景。[]提出，从背景得到的时空特征（静态特征），特别是靠近运动区域的特征，对于动作识别是有用的。正如图3.1.1(b)的箭靶，与射箭动作是息息相关的，又如足球与踢球相关、篮球、篮球场与运球相关。我们将在第四章对比剔除前后的准确率差异。

## 3.2空间信息网络框架

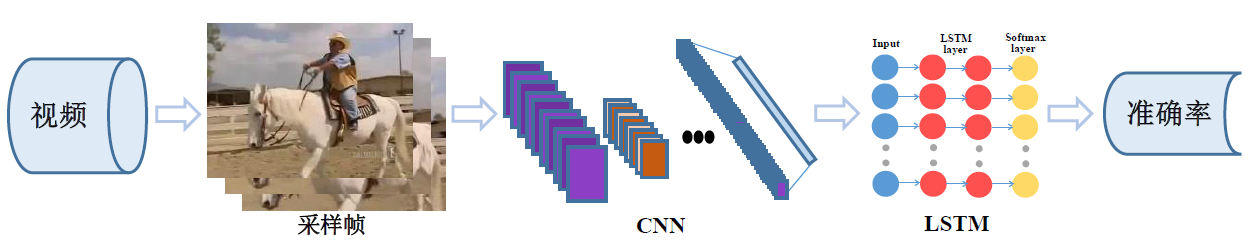


图3.2.1 空间信息网络

正如第一章中提到的：对于循环神经网络及其拓展模型，一种比较典型的做法就是将视频序列经过CNN处理提取出特征数据，然后再经过若干层LSTM进行最后的判别。借助这个思想，框架如上图3.2.1所示，其步骤可分为：a）视频采样得到张快照；b）快照经过已在ImageNet上预训练的CNN网络得到个特征；c）取时间步为，将个特征组合后通过两层层进行训练；d）最终得到预测准确率。

在第四章中，我们将测试a）不微调CNN网络提取特征；b）在选取的数据集上微调网络再提取特征；c）选取不同的，以及不同的特征组合方式。

## 3.3本章小结

在本章中，我们将视频等分段并从每段中随机采样出一帧快照，这张快照就代表了这个视频。考虑到快照中可能包含无关的背景信息（噪音）且其占的比例较少，我们采用了灰度直方图以及图像指纹的方式来衡量图像相似度，剔除不相似的快照。

在构建空间信息网络时，我们参考了前人的思想[]，将快照经过CNN处理提取出特征数据，然后再经过若干层LSTM进行最后的判别。

# 实验结果与分析

本章我们介绍选取的数据集及预处理，以及前面提及的骨架信息网络、空间信息网络的实验结果，并对实验结果做出简要的分析。

## 选取的数据集及预处理

选取的数据集为HMDB51，该数据集共有6849个视频片段，51种动作类别，每种动作至少涵盖101个视频判断。视频主要来自于电影，少部分来自YouTube、Google等。动作类别可以分成5种：1）一般面部动作（如笑、咀嚼、说话）；2）有客体互动的面部动作（如吃、喝、抽烟）；3）一般身体运动（如坐下、踢、跳）；4）有客体互动的身体运动（如骑马、梳头发）；5）人类互动的身体运动（如握手，亲吻）。

我们采用OpenPose对数据集中的视频进行骨架提取，然后对所有提取到的矩阵进行稠密度统计，得到了如下的结果。

图4.1.1 骨架信息矩阵稠密度最高的10个类别

如上图4.1.1所示，稠密度最高的10个类别为：运球、握手、高尔夫、抓、击剑、拥抱、剑术、射球、倒、射箭。如下图4.1.2所示，稠密度最低的10个类别为：落地、喝、吃、跳、抽烟、侧空翻、转、翻跟斗、咀嚼、跑步

图4.1.2 骨架信息矩阵稠密度最低的10个类别

为了能够充分利用骨架信息矩阵，我们剔除骨架信息矩阵稠密度小于0.41的类别，最终剩下30种动作类别，包括（按照稠密度从高到低排列）：运球、握手、高尔夫、抓、击剑、拥抱、剑术、射球、倒、射箭、坐起来、鼓掌、拔、俯卧撑、推、骑车、踢球、击打、踢、剑练习、攀登、爬楼梯、骑马、倒立、大笑、梳头发、谈话、笑、潜水、亲吻。紧接着对着30类的骨架信息矩阵进行线性插值，稠密度平均上涨14.66%。接下来我们将研究各种参数对最终行为识别准确率的影响。

## 运行环境与网络结构说明

运行平台为tensorflow和以tensorflow为后端的keras。电脑显卡GTX 860M。

图2.5.1所示的框架中（简称为**，Shared InceptionV3 Network**），共享网络采用已在ImageNet上预训练的InceptionV3，重新定义全连接层3层，长度分别为1024、512、51，其中前两层激活函数为relu，最后一层激活函数为softmax。在数据集上微调该网络时，使用交叉熵作为损失函数，SGD作为优化算子，droupout设置为0.5。由于GPU受限，batch size只能设置为10，共执行12个epochs（每4个epochs为一轮，也即3轮）。在第1轮，学习率设置为0.01，动量设置为0；在第2轮，学习率设置为0.01，动量设置为0.5；在第3轮，学习率设置为0.001，动量设置为0.9。

图2.5.2所示的框架中（简称为），CNN网络采用已在ImageNet上预训练的VGG16或者VGG19，特征将从网络的卷积层（具体哪一层接下来会做实验）中提取；紧接着特征通过两层LSTM层，单元数分别为128、64，dropout均为0.4，输入的time step设置为1，最后通过一层全连接层，长度为51，激活函数为softmax。训练LSTM部分时，设置batch size为256，epochs为100；采用交叉熵作为损失函数，RMSprop作为优化算子，设置学习率为0.001，rho为0.9。

图3.2.1所示的框架中（简称为），CNN网络采用已在ImageNet上预训练的VGG16或者VGG19，特征将从网络的卷积层（具体哪一层接下来会做实验）中提取；紧接着特征通过两层LSTM层，单元数分别为128、64，dropout均为0.4，输入的time step将在接下来实验中研究，最后通过一层全连接层，长度为51，激活函数为softmax。训练LSTM部分时，设置batch size为256，epochs为50；采用交叉熵作为损失函数，RMSprop作为优化算子，设置学习率为0.001，rho为0.9。

## 参数实验及最终实验结果

于VGG19提取出block5\_pool特征，即意味着我们从VGG19网络第五块的池化层提取出了特征，其维度为7x7x512，为了降低维度，我们取平均后可得到1x512的特征。我们标记VGG19的block5\_pool特征为**V19b5**，的全连接第一层为**Fc1**，第二层为**Fc2**，本身的预测输出为**Out**； JTM/JDM插值前的效果图为**Ori**，插值后的效果图为**Mc**。**Ori+Mc**在中表示两者的1x512特征进行平均融合，在**中**表示**Ori**与**Mc**两路输入；标记的输入为**Spatial**。

**4.3.1实验一：JTM/JDM不同特征、输入和框架对准确率的影响**

我们先对JTM进行实验，实验前提：生成JTM的jet阶数L为10，每个视频片段生成了10张JTM图，数据集为HMDB51 split1。实验结果如下表4.3.1.1所示，最好的准确率达到了**35.859%**。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | |  | | |
|  | **V16b4** | **V16b5** | **V19b4** | **V19b5** | **Fc1** | **Fc2** | **Out** |
| Ori | 0.302374 | 0.329330 | 0.267039 | 0.337989 |  |  | 0.268362 |
| Mc | 0.299302 | 0.316896 | 0.275838 | 0.340503 |  |  | 0.270417 |
| Ori+Mc | 0.307542 | 0.334288 | 0.275978 | **0.358590** |  |  | 0.277933 |

表4.3.1.1 JTM中，不同特征、输入和框架对准确率的影响

结论：在中，提取的特征准确率从高到低依次为**V19b5> V16b5> V16b4> V19b4**；不同输入的准确率从高到低依次为**Ori+Mc> Ori≈ Mc**。使用**Ori+Mc**准确率平均提高**3.283%**。在**中**，不同输入的准确率从高到低依次为**Ori+Mc> Ori≈ Mc**，但是总体准确率不如。因此，我们接下来对于JTM采用**V19b5**特征，**Ori+Mc**输入，框架。

接下来我们对JDM进行实验。对于一个拥有帧的视频，如果我们利用其所有帧，最终我们能得到[n \* ]的二维矩阵（式子2.3.2上方），假设此次实验采样10帧生成JDM。实验前提：生成JDM的采样帧数为15，每个视频片段生成了10张JDM图，数据集为HMDB51 split1。实验结果如下表4.3.1.2所示，最好的准确率达到了**XXX%**。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | |  | | |
|  | **V16b4** | **V16b5** | **V19b4** | **V19b5** | **Fc1** | **Fc2** | **Out** |
| Ori | 0. | 0. | 0. | 0. | 0.391088 | 0.399273 | 0.380073 |
| Mc | 0. | 0. | 0. | 0. | 0.392554 | 0.395107 | 0.376288 |
| Ori+Mc | 0. | 0. | 0. | **0.** | 0.410639 | **0.412052** | 0.397346 |

表4.3.1.2 JDM中，不同特征、输入和框架对准确率的影响

结论：在中，提取的特征准确率从高到低依次为**V19b5> V16b5> V16b4> V19b4**；不同输入的准确率从高到低依次为**Ori+Mc> Ori≈ Mc**。使用**Ori+Mc**准确率平均提高**3.283%**。在**中**，不同输入的准确率从高到低依次为**Ori+Mc> Ori≈ Mc**，但是总体准确率不如。因此，我们接下来对于JTM采用**V19b5**特征，**Ori+Mc**输入，框架。

确定二者结合准确率更高。

**4.3.2实验二：JTM中jet图阶数L对准确率的影响**

实验前提：每个视频片段生成了10张JTM图，采用**V19b5**特征，**Ori+Mc**输入，框架,数据集为HMDB51 split1。实验结果如下图4.3.2.1所示，最好的准确率达到了**35.859%**。

图4.3.2.1 JTM中jet图阶数L对准确率的影响

结论：不同阶数的准确率从高到低为：L10 > L5 > L15 > L20 > L25。因此我们接下来取L为10。

**4.3.3实验三：生成JDM所采样的帧数K对准确率的影响**

。现在我们研究采样多少帧效果更加，设采样的帧数为k

|  |  |  |
| --- | --- | --- |
|  | Vgg16b5 | Vgg19b5 |
| 15 | 0.350559 | 0.353352 |
| 30 |  |  |
| 50 |  |  |
| All |  |  |

**4.3.4实验四：Spatial中不同特征对准确率的影响**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | V16b4 | V16b5 | V19b4 | V19b5 |
| Spatial | 0. | 0. | 0. | 0. |
| Spatial10 | 0. | 0. | 0. | 0. |

**4.3.5实验五：中LSTM的输入time step对准确率的影响**

**4.3.6实验六：视频片段生成的样本数N对准确率的影响**

**4.3.7实验七：得分融合方式对准确率的影响**

|  |  |  |  |
| --- | --- | --- | --- |
|  | mean | max | Multiple-wise |
| JTM |  |  |  |
| JDM |  |  |  |
| Spatial |  |  |  |
| JTM + JDM |  |  |  |
| JTM + Spatial |  |  |  |
| JDM + Spatial |  |  |  |
| JTM + JDM + Spatial |  |  |  |

**4.3.8最终实验结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Split1 | Split2 | Split3 | 平均 |
| JTM |  |  |  |  |
| JDM |  |  |  |  |
| Spatial |  |  |  |  |
| JTM + JDM |  |  |  |  |
| JTM + Spatial |  |  |  |  |
| JDM + Spatial |  |  |  |  |
| JTM + JDM + Spatial |  |  |  |  |

Time step 直接取样本数？？

Spatial 1 & 10对比

JTM\JDM\spatial score fusion && feature fusion

If score fusion: mean, max, multiple-wise

1. 不同层提取的特征
2. 不同的网络
3. 不同的time\_step
4. 。。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | V16block4 | V16block5 | V19block4 | V19block5 | InV3 | InV3+LSTM | InV3+SVM |
| spatial | 0.51117 | 0.64664 | 0.443435 | 0.6522 | 0.62081 | 0.61159 |  |
| JDM | 0.305865 | 0.299581 | 0.250318 | 0.290502 | 0.39734 | 0.415502 |  |
| JTM | 0.28002 | 0.32192 | 0.28212 | 0.36033 | 0.27583 | 0.333798 |  |

## 模糊矩阵及特例分析

画出混淆矩阵【火狐浏览器已收藏Matlab代码】

详细分析 哪一类 容易 被误判为哪一类。

# 结 论

本文中我们介绍了视频中人数估计的两种情况，都在实际生活中有重要意义。

对第一种情况，相比其他的算法我们的的速度和准确度都比较好。通过使用角点的数目估计视频中的人数，最小生成树聚类人群。比较简单的方法取得了比较好好的结果。在比较极端的情况下仍然能取得比较好的效果。改进的方向是通过扩张我们特征的维数，即加入其他信息比如人群的坐标，覆盖面积等等，以克服一些视角和遮掩带来的误差。但是高度遮掩情况下的人数统计仍然具有一定的挑战性。

对第二种情况，我们建立了一个比较完整的体系。利用HOG寻找人群，利用粒子滤波器刻画人的位置分布，利用在线分类器区别人群彼此，加入对系统的假设使其更加鲁棒。但是由于多人在线追踪的情况仍然会出现很多问题，包括如何更好的匹配检测结果、更好的区分人与人之间和更稳定可靠地寻找行人。由于实际情况的复杂，这里能够提升的地方比较多。也包括对计算量的控制，众所周知的HOG算子是非常消耗时间的，虽然我们采用了积分图来加快计算速度，仍然不能满足实时的要求。此处可以考虑利用运动信息或者显著性图来结合HOG扫描过程，将扫描范围限制在一定范围内。需要提到的是我们应该尽量改进检测子使得误检率最够低，在追踪过程中，漏检虽然也不是很好的现象，但是由于追踪器的鲁棒性，可以很好的估计目标的分布，同时由于分类器的作用在重新发现后能够再次匹配。但是误检带来的结果很可能打乱整个追踪过程。由于我们的HOG计算过程采用了积分图的方式加快速度，从而在HOG结果中缺少插值过程，使得结果不够稳定。这部分也有很大的改进空间。由于detection的部分在整个算法中的左右尤其重要，所以这部分的改进也十分必要，无论是速度还是准确率方面。

加大粒子滤波器中粒子的数目会带来更好的效果，但是带来更大的计算消耗。所以对这部分的提升，我们应该着眼于如何更好的进行粒子状态地更新。能否结合追踪器当前的状态来更新权重，尽量结合多当前可用的信息。

在线分类器 的部分，我们相信其在线更新的能力在这里是必须的。但是在特征的选择之上我们或许可以找到更加好的选择。模型假设方面我们也相信良好的模型假设会使追踪过程更加稳定鲁棒，这在我们的实验中也得到了体现。

相信随着技术的发展，不只是人数估计，视频监控中的各项技术将会越来越成熟和受到重视。

1. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.
2. Lin Z, Chen M, Ma Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices[J]. arXiv preprint arXiv:1009.5055, 2010. 矩阵补全
3. Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//CVPR. 2017, 1(2): 7. Openpose
4. Li C, Hou Y, Wang P, et al. Joint distance maps based action recognition with convolutional neural networks[J]. IEEE Signal Processing Letters, 2017, 24(5): 624-628. JDM
5. Wang P, Li Z, Hou Y, et al. Action recognition based on joint trajectory maps using convolutional neural networks[C]//Proceedings of the 2016 ACM on Multimedia Conference. ACM, 2016: 102-106. JTM
6. Pishchulin L, Insafutdinov E, Tang S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4929-4937.
7. Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In Computer Vision and Pattern Recognition, pages 3582–3589, 2014.
8. Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. pages 627–642, 2016.
9. Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann Lecun, and Christoph Bregler. Efficient object localization using convolutional networks. pages 648–656, 2014.
10. Wanli Ouyang, Xiao Chu, and XiaogangWang. Multi-source deep learning for human pose estimation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2337–2344, 2014.
11. Insafutdinov E, Pishchulin L, Andres B, et al. Deepercut: A deeper, stronger, and faster multi-person pose estimation model[C]//European Conference on Computer Vision. Springer, Cham, 2016: 34-50.
12. Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3d points[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010: 9-14.
13. Wentao Zhu, Cuiling Lan, Junliang Xing,Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie, “Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks,” arXiv preprint arXiv:1603.07772, 2016.

# 参考文献

1. Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets[J]. arXiv preprint arXiv:1507.02159, 2015.
2. Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2017.
3. Shahroudy A, Liu J, Ng T T, et al. NTU RGB+ D: A large scale dataset for 3D human activity analysis[J]. arXiv preprint arXiv:1604.02808, 2016.
4. Kay, Will, et al. "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).
5. Jhuang H, Garrote H, Poggio E, et al. A large video database for human motion recognition[C]//Proc. of IEEE International Conference on Computer Vision. 2011, 4(5): 6.
6. Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.
7. I. Laptev, On space–time interest points, Int. J. Comput. Vis. 64 (2) (2005) 107–123
8. G. Willems, T. Tuytelaars, L. Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, Proc. European Conference on Computer Vision (ECCV), 2008. pp. 650–663.
9. P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, Proceedings of the 14th International Conference on Computer Communications and Networks, 2005. pp. 65–72.
10. A. Kläser, M. MarszaÅek, C. Schmid, A spatio-temporal descriptor based on 3D-gradients, In BMVC08, 2008. pp. 275:1-10.
11. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005. pp. 886–893. vol. 1.
12. I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2008. pp. 1–8.
13. N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, Proc. European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science (LNCS), vol. 3952, 2006. pp. 428–441.
14. P. Matikainen,M. Hebert, R. Sukthankar, Trajectons: action recognition through the motion analysis of tracked features, Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, Sept. 2009. pp. 514–521.
15. Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, C.-W. Ngo, Trajectory-based modeling of human actions with motion reference points, Proc. European Conference on Computer Vision (ECCV), 2012. pp. 425–438.
16. P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, Proceedings of the 14th International Conference on Computer Communications and Networks, 2005. pp. 65–72.
17. X. Peng, C. Zou, Y. Qiao, Q. Peng, Action recognition with stacked Fisher vectors, Proc. European Conference on Computer Vision (ECCV), 2014. pp. 581–595.
18. J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
19. J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.
20. L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (Feb. 1989) 257–286.
21. Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.
22. Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE, 2015: 4489-4497.
23. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C] //Advances in neural information processing systems. 2014: 568-576.
24. Diba A, Sharma V, Van Gool L. Deep temporal linear encoding networks[C]//Computer Vision and Pattern Recognition. 2017.
25. Feichtenhofer C, Pinz A, Zisserman A P. Convolutional two-stream network fusion for video action recognition[J]. 2016.
26. Lan Z, Zhu Y, Hauptmann A G, et al. Deep Local Video Feature for Action Recognition[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017: 1219-1225.
27. Ng J Y H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond short snippets: deep networks for video classification. arXiv: 1503.08909, 2015
28. Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. arXiv: 1411.4389, 2014.
29. Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In Computer Vision and Pattern Recognition, pages 3582–3589, 2014.
30. Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. pages 627–642, 2016.
31. Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann Lecun, and Christoph Bregler. Efficient object localization using convolutional networks. pages 648–656, 2014.
32. Wanli Ouyang, Xiao Chu, and XiaogangWang. Multi-source deep learning for human pose estimation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2337–2344, 2014.
33. Pishchulin L, Insafutdinov E, Tang S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4929-4937.
34. Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//CVPR. 2017, 1(2): 7.
35. Insafutdinov E, Pishchulin L, Andres B, et al. Deepercut: A deeper, stronger, and faster multi-person pose estimation model[C]//European Conference on Computer Vision. Springer, Cham, 2016: 34-50.
36. Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3d points[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010: 9-14.
37. Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in AAAI Conference on Artificial Intelligence, 2017.
38. Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li, “Action recognition based on joint trajectory maps using convolutional neural networks,” in Proc. ACM on Multimedia Conference, 2016, pp. 102–106.
39. Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li, “Joint distance maps based action recognition with convolutional neural networks,” IEEE Signal Processing Letters, vol. 24, pp. 624–628, 2017.
40. Li C, Hou Y, Wang P, et al. Joint distance maps based action recognition with convolutional neural networks[J]. IEEE Signal Processing Letters, 2017, 24(5): 624-628.
41. Li C, Wang P, Wang S, et al. Skeleton-based action recognition using LSTM and CNN[C]//Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on. IEEE, 2017: 585-590.

# 致谢

首先我要对指导老师赖剑煌教授表示诚挚的感谢。是赖老师一直以来的悉心指导和热情鼓励才让我在模式识别这个相对陌生的领域逐渐找到方向完成自己的工作。老师治学态度严谨，在整个论文写作过程中始终对大家严格把关，每周都能在百忙中抽出时间对大家进行指导。从论文标题、格式、设计、布局和结论都给大家提出了宝贵的意见。赖老师渊博的专业知识、严谨的治学态度、务实的敬业精神令人敬佩。很荣幸能跟随老师学习和研究。

感谢连国云师兄的无私帮助和指导，连国云师兄在自己繁忙的学业中始终愿意为我们提供热心的帮助。感谢实验室其他师兄师姐的支持和鼓励。

感谢颜咚同学，你的支持和鼓励是我不断坚持的动力，衷心地感谢你。谢谢孙世通同学提供的材料支持。

谢谢四年来教导过我的老师们，你们的指导和教诲我会铭记于心，谢谢你们传授的知识和做人的道理。

最后我要感谢我的父母和家人，感谢你们的支持和理解，使我克服困难不断进取。没有你们也就没有今天的我。

**附录**

(\* 附录是正文主体的补充。下列内容可以作为附录：  
1）攻读学位期间发表的（含已录用，并有录用通知书的）与学位论文相关的学术论文。  
2）由于篇幅过大，或取材于复制件不便编入正文的材料、数据。  
3）对本专业同行有参考价值，但对一般读者不必阅读的材料。  
4）论文中使用的符号意义、单位缩写、计算机程序全文及有关说明书。  
5）附件：光盘、与论文相关的鉴定证书、获奖奖状或专利证书的复印件等。 \*)

（论文附录依次用大写字母“附录A、附录B、附录C……”表示，附录内的分级序号可采用“附A1、附A1.1、附A1.1.1”等表示，图、表、公式均依此类推为“图A1、表A1、式A1”等。）

毕业论文成绩评定记录

|  |
| --- |
| 指导教师评语：  成绩评定：    指导教师签名： 年 月 日 |
| 答辩小组或专业负责人意见：  成绩评定：    签名（章）： 年 月 日 |
| 院系负责人意见：  成绩评定：  签名（章）： 年 月 日 |

**附表一、毕业论文开题报告**

|  |
| --- |
| 论文（设计）题目： |
| （简述选题的目的、思路、方法、相关支持条件及进度安排等）  学生签名： 年 月 日 |
| 指导教师意见：  1、同意开题（ ） 2、修改后开题（ ） 3、重新开题（ ）    指导教师签名： 年 月 日 |

**附表二、毕业论文过程检查情况记录表**

|  |  |
| --- | --- |
| 指导教师分阶段检查论文的进展情况（要求过程检查记录不少于3次）：  **第1次检查**  学生总结：  指导教师意见：  **第2次检查**  学生总结：  指导教师意见：  **第3次检查**  学生总结：  指导教师意见：  **第4次检查**  学生总结：  指导教师意见：  学生签名： 年 月 日  指导教师签名： 年 月 日 | |
| **总体**  **完**  **成**  **情**  **况** | 指导教师意见：  1、按计划完成，完成情况优（ ）  2、按计划完成，完成情况良（ ）  3、基本按计划完成，完成情况合格（ ）  4、完成情况不合格（ ）  指导教师签名： 年 月 日 |

**附表三、毕业论文答辩情况**

|  |  |  |  |
| --- | --- | --- | --- |
| 答辩人 |  | 专 业 |  |
| 论文（设计）题目 |  | | |
| 答辩小组成员 |  | | |
| 答辩记录：  记录人签名： 年 月 日 | | | |

**学术诚信声明**

本人所呈交的毕业论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。本毕业论文的知识产权归属于培养单位。本人完全意识到本声明的法律结果由本人承担。

本人签名： 日期：

毕业论文应按以下顺序装订：封面→毕业论文（设计）开题报告→毕业论文（设计）过程检查情况记录表→毕业论文（设计）答辩情况登记表→学术诚信声明→中英文摘要→目录→引言→正文→结语→参考文献→致谢→附录→成绩评定记录