[行为检测|行为识别]调研综述

流浪者
计算机视觉爱好者，微信公众号“faiculty”

行为检测

标签（空格分隔）： 计算机视觉 行为检测 视频理解

[toc]

---

1. 背景

视频理解是目前计算机视觉领域非常热，也是极具挑战力的一个方向。视频理解方向包含众多的子研究方向，以CVPR组织的ACTIVITYNET为例，2017年总共有5个Task被提出。

- Task1：未修剪视频分类(Untrimmed Video Classification)。这个有点类似于图像的分类，未修剪的视频中通常含有多个动作，而且视频很长。有许多动作或许都不是我们所关注的。所以这里提出的Task就是希望通过对输入的长视频进行全局分析，然后软分类到多个类别。
- Task2：修剪视频识别(Trimmed Action Recognition)。这个在计算机视觉领域已经研究多年，给出一段只包含一个动作的修剪视频，要求给视频分类。
- Task3：时序行为提名(Temporal Action Proposal)。这个同样类似于图像目标检测任务中的候选框提取。在一段长视频中通常含有很多动作，这个任务就是从视频中找出可能含有动作的视频段。
- Task4：时序行为定位(Temporal Action Localization)。相比于上面的时序行为提名而言，时序行为定位于我们常说的目标检测一致。要求从视频中找到可能存在行为的视频段，并且给视频段分类。
- Task5：密集行为描述(Dense-Captioning Events)。之所以称为密集行为描述，主要是因为该任务要求在时序行为定位(检测)的基础上进行视频行为描述。也就是说，该任务需要将一段未修剪的视频进行时序行为定位得到许多包含行为的视频
  man playing a piano

2. 国内外研究现状

在该方向上，国内有许多机构和学校也是主要的研究者，所以这里不再区分国内外，直接描述当前的研究现状。目前为止ActivityNet已经举办两届，下面是2017年的State-of-art。

表 2.1 2017 年 State-of-art

	No1			No2		
	UserName	Orgnization	Accuracy/Avg.Error	UserName	Orgnization	Accuracy
Untrimmed	Jiankang Deng	IBUG(伦敦帝国学院)	0.0883	Yuanjun Xiong	CUHK	0.09788
Trimmed	Chuang Gan	Tsinghua	0.1239(Error)	Yuanjun Xiong	CUHK	0.13917
Proposal	Tianwei Lin	上交	0.662688	Vendetta V	中科大	0.661473
Localization	Tianwei Lin	上交	0.35015	Yuanjun Xiong	CUHK	0.31863
Captioning	Peter Li	-	0.12956	Ting Yao	MSRA	0.1284

3. 行为分类

行为分类(Trimmed Action Recognition)是视频理解方向很重要的一个问题，至今为止已经研究多年。深度学习出来后，该问题被逐步解决，现在在数据集上已经达到了比较满意的效果。如第2章所述 。行为分类问题简单的来说就是：对于给定的分割好的视频片段，按照其中的人类行为进行分类。比如女孩化妆、男生打球、跑步等等。该任务不需要确定视频中行为的开始时间和结束时间。

在深度学习出现之前，表现最好的算法是iDT [1][2]，之后的工作基本上都是在iDT方法上进行改进。IDT的思路是利用光流场来获得视频序列中的一些轨迹，再沿着轨迹提取HOF，HOG，MBH，trajectory4中特征，其中HOF基于灰度图计算，另外几个均基于dense optical flow(密集光流计算)。最后利用FV(Fisher Vector)方法对特征进行编码，再基于编码训练结果训练SVM分类器。深度学习出来后，陆续出来多种方式来尝试解决这
C3D(Convolution 3 Dimension) [6]，还有RNN [7] 方

行为识别虽然研究多年，但是至今还是处于实验室数据集测试阶段，没有真正的实用化和产业化。由此可见该任务目前还是没有非常鲁棒的解决方案。下面简单阐述一下本人对于该问题的看法。

任务特点：行为识别和图像分类其实很相似，图像分类是按照图像中的目标进行软分类，行为识别也类似。最开始的时候类似于UCF数据集，都是采用的单标签，也就是一段视频只对应一个标签。现在CPVR举办的Activitynet(Kinetics 数据集)每段视频中包含多个标签。相比于图像分类，视频多了一个时序维度，而这个问题恰恰是目前计算机领域令人头疼的问题。

任务难点

- 如上所说，行为识别处理的是视频，所以相对于图像分类来说多了一个需要处理的时序维度。
- 行为识别还有一个痛点是视频段长度不一，而且开放环境下视频中存在多尺度、多目标、摄像机移动等众多的问题。这些问题都是导致行为识别还未能实用化的重要原因。

3.2 数据集介绍

目前还比较常用的数据库主要有3个，UCF101、HMDB51和Kinetics.

表 3.1 数据集比较

	视频来源	视频数	动作数
UCF-101	YouTube	13320	101
HMDB51	YouTube	7000	51
Kinetics	YouTube	300k	400

3.3 传统方法

在深度学习之前，iDT(improved Dense Trajectories)方法是最经典的一种方法。虽然目前基于深度学习的方法已经超过iDT，但是iDT的思路依然值得学习，而且与iDT的结果做ensemble后总能获得一些提升。iDT的思路主要是在《Dense Trajectories and Motion Boundary Descriptors for Action Recognition》和《Action Recognition with Improved Trajectories》两篇文章中体现。

下面本文简单的介绍DT(Dense Trajectories)方法。

3.3.1 密集采样特征点

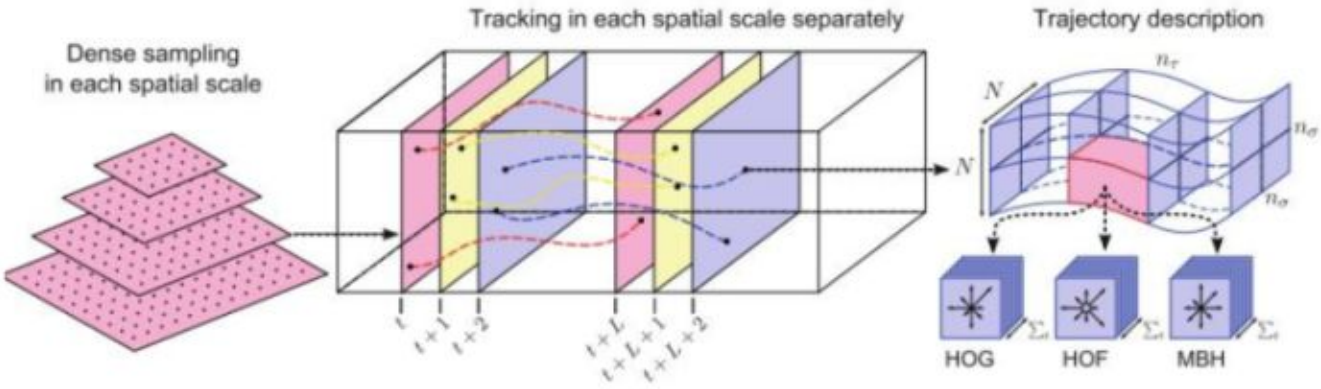


图3.1 iDT算法框架图

DT方法通过网格划分的方式在多尺度图像中分别密集采样特征点。

3.3.2 轨迹与轨迹描述子

假设上一步骤中密集采样到的某个特征点的坐标为 $P_t = (x_t, y_t)$，再用下面的公式计算该特征点在下一帧图像中的位置。

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * w_t)|x_t, y_t$$

上式中 w_t 为密集光流场，是 I_t 和 $I_(t+1)$ 计算得到的。M代表的是中值滤波器，尺寸为3x3,，所以这个式子是通过计算特征点领域内的光流中值来得到特征点的运动方向。

3.3.3 运动描述子

除了轨迹形状特征，还需要更有力的特征来描述光流，DT/iDT中使用了HOF，HOG和MBH三种特征。下面简单的阐述一下这几种特征。

HOG特征：HOG特征计算的是灰度图像梯度的直方图。直方图的bin数目为8。所以HOG特征的长度为2*2*3*8=96.

HOF特征：HOF计算的是光流的直方图。直方图的bin数目取为8+1，前8个bin与HOG都相同。额外的一个用于统计光流幅度小于某个阈值的像素。故HOF的特征长度为2*2*3*9=108.

MBH特征：MBH计算的是光流图像梯度的直方图，也可以理解为在光流图像上计算的HOG特征。由于光流图像包括X方向和Y方向，故分别计算MBHx和MBHy。MBH总的特征长度为2*96=192.
最后进行特征的归一化，DT算法中对HOG，HOF和MBH均使用L2范数进行归一化。

Two-Stream方法是深度学习在该方向的一大主流方向。最早是VGG团队在NIPS上提出来的[3]。其实在这之前也有人尝试用深度学习来处理行为识别，例如李飞飞团队 [8] ，通过叠加视频多帧输入到网络中进行学习，但是不幸的是这种方法比手动提取特征更加糟糕。当Two-Stream CNN出来后才意味着深度学习在行为识别中迈出了重大的一步。

3.4.1 TWO-STREAM CNN

Two-Stream CNN网络顾名思义分为两个部分，一部分处理RGB图像，一部分处理光流图像。最终联合训练，并分类。这篇文章主要有以下三个贡献点。

- 首先，论文提出了two-stream结构的CNN网络，由空间(RGB)和时间(光流)两个维度的网络组成
- 其次，作者提出了利用网络训练多帧密度光流，以此作为输入能在有限训练数据的情况下取得不错的结果。
- 最后，采用多任务训练的方法将两个行为分类的数据集联合起来，增加训练数据，最终在两个数据集上都取得了更好的效果(作者提到，联合训练也可以去除过拟合的可能)。

网络结构:

因为视频可以分为空间和时间两个部分。空间部分，每一帧代表的是空间信息，比如目标、场景等等。而时间部分是指帧间的运动，包括摄像机的运动或者目标物体的运动信息。所以网络相应的由两个部分组成，分别处理时间和空间两个维度。

每个网络都是由CNN和最后的softmax组成，最后的softmax的fusion主要考虑了两种方法：平均，在堆叠的softmax上训练一个SVM。网络的结构图如下所示。

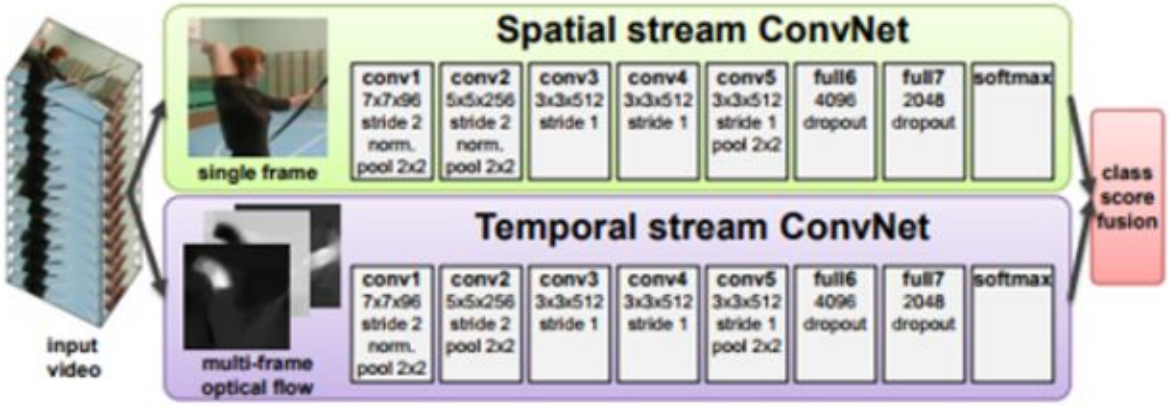


图 3.2 two-stream 网络结构图

光流栈(Optical flow Stacking):

假设考虑做动作的分类(行为识别主要包含两个方向，一个是动作分类，给出一个视频截断，判断视频的动作类别，或者称为offline。 另一个就是动作识别，给出一个自然视频，没有进行任何的裁剪，这个时候需要先知道动作的开始时间和结束时间，然后还要知道动作的类别)。 考虑对一小段视频进行编码，假设起始帧为T，连续L帧(不包含T帧)。计算两帧之间的光流，最终可以得到L张光流场，每张光流场是2通道的(因为每个像素点有x和y方向的移动)。

最后，我们将这些光流场输入，得到相应的特征图。

实验结果:

最终该方法在UCF-101和HMDB-51上取得了与iDT系列最好的一致效果。在UCF-101上准确度为88.0%，在HMDB上准确度为59.4%。

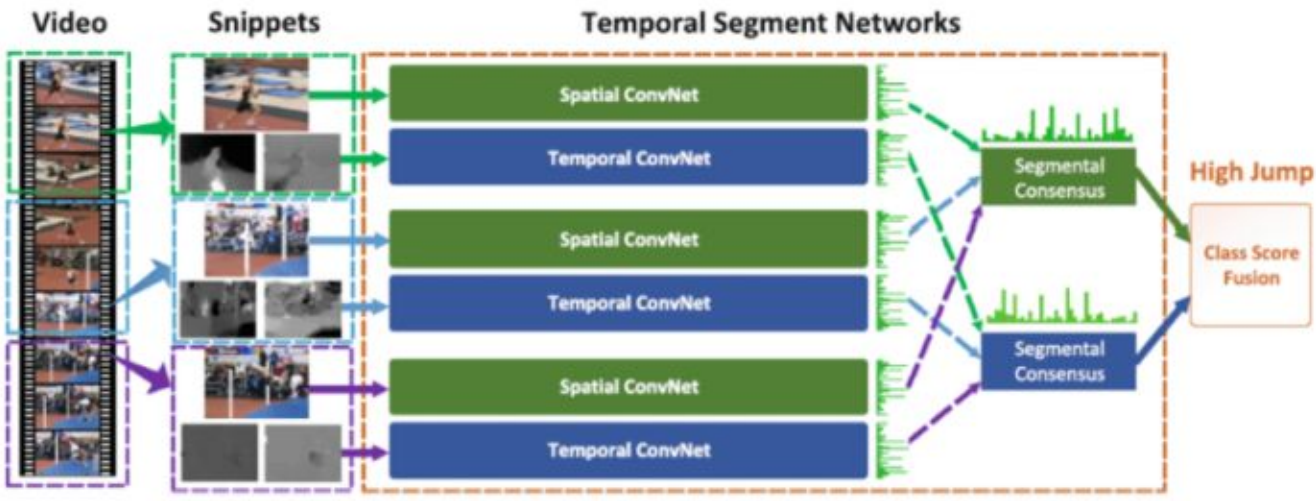
表 3.2 Two-Stream 实验结果表

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**

3.4.2 TSN

TSN(Temporal Segments Networks)[5]是在上述基础的two-Stream CNN上改进的网络。目前基于two-stream的方法基本上是由TSN作为骨干网络，所以这里进行简单的阐述。

3.4.1小节所述的two-stream的方法很大的一个弊端就是不能对长时间的视频进行建模，只能对连续的几帧视频提取temporal context。为了解决这个问题TSN网络提出了一个很有用的方法，先将视频分成K个部分，然后从每个部分中随机的选出一个短的片段，然后对这个片段应用上述的two-stream方法，最后对于多个片段上提取到的特征做一个融合。下图是网络的结构图。



3.5 C3D方法

C3D(3-Dimensional Convolution) [6] 是除了Two-Stream后的另外一大主流方法，但是目前来看C3D的方法得到的效果普遍比Two-Stream方法低好几个百分点。但是C3D任然是目前研究的热点，主要原因是该方法比Two-Stream方法快很多，而且基本上都是端到端的训练，网络结构更加简洁。该方法思想非常简单，图像是二维，所以使用二维的卷积核。视频是三维信息，那么可以使用三维的卷积核。所以C3D的意思是：用三维的卷积核处理视频。

网络结构

C3D共有8次卷积操作，5次池化操作。其中卷积核的大小均为3*3*3，步长为1*1*1。池化核为2*2*2，但是为了不过早的缩减在时序上的长度，第一层的池化大小和步长为1*2*2。最后网络在经过两次全连接层和softmax层后得到的最终的输出结果。网络的输入为3*16*112*112，其中3为RGB三通道，16为输入图像的帧数，112*112是图像的输入尺寸。

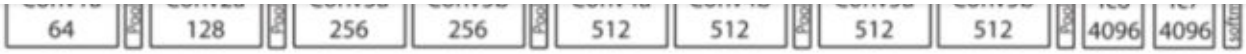


图 3.4 C3D 网络结构图

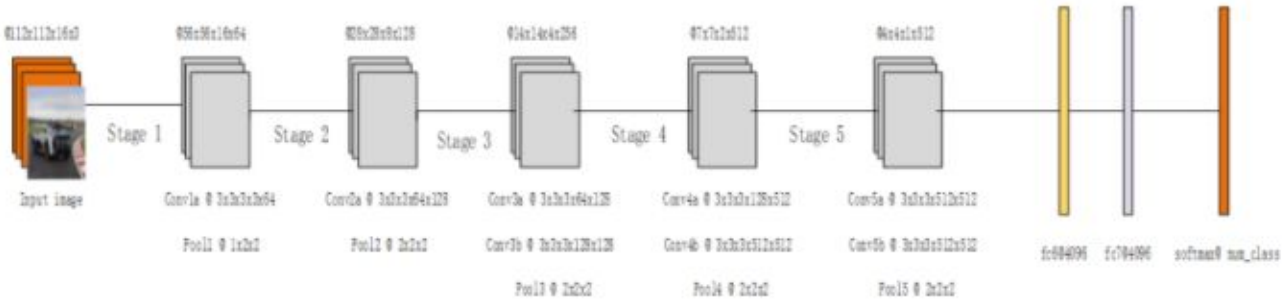


图 3.5 网络结构详解图

3.6 RNN方法

因为视频除了空间维度外，最大的痛点是时间序列问题。如果能很好的处理这个维度，那么效果是不是会显著提升呢？而众所周知，RNN网络在NLP方向取得了傲人的成绩，非常适合处理序列。所以除了上述两大类方法以外，另外还有一大批的研究学者希望使用RNN网络思想来解决这个问题。目前最新的进展是中科院深圳先进院乔宇老师的工作：《RPAN：An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos》[7]。这篇文章是ICCV2017年的oral文章。但是与传统的Video-level category训练RNN不同，这篇文章还提出了Pose-attention的机制。

这篇文章主要有以下几个贡献点。

- 不同于之前的pose-related action recognition，这篇文章是端到端的RNN，而且是spatial-temporal evolutionos of human pose
- 不同于独立的学习关节点特征(human-joint features)，这篇文章引入的pose-attention机制通过不同语义相关的关节点(semantically-related human joints)分享attention参数，然后将这些通过human-part pooling层联合起来
- 视频姿态估计，通过文章的方法可以给视频进行粗糙的姿态标记。(这个方法还挺不错)。

3.6.1 网络结构

RPAN网络框架可以分为三个大的部分。

- 特征生成部分：用Two-Stream的方法生成

下图是RPAN网络的结构图。

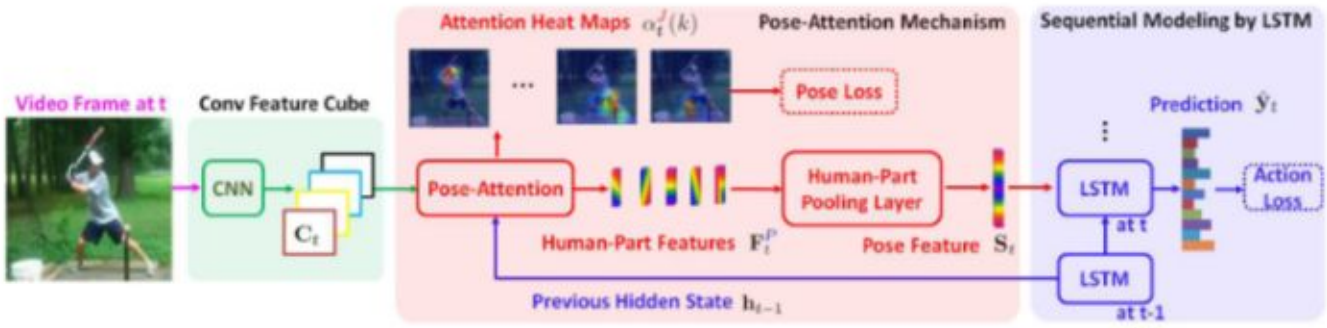


图 3.6 RPAN 网络结构图

3.6.2 特征生成

RPAN网络中采用TSN(Temporal Segments Network)的网络框架生成Convolution Cubes。包含空间和时间上两个维度。具体内容可以查看3.4.2小节。

3.6.3 姿态注意机制

经过上述Two-Stream网络后生成了K1 *K2*dc的特征图。之后作者经过一系列的操作将姿态和上述的特征图结合起来得到姿态特征图，最后输入LSTM中。具体的，文章进行了如下几步操作。

Step1：空间特征向量文章中定义一个 C_t ，表示第t个视频帧在不同空间位置上的特征向量。空间图是K1xK2的大小，共dc个通道。所以 C_t 是K1xk2个dc维的向量。Ct的定义如下。

$$C_t = \{C_t(1), \ldots\ldots, C_t(K_1 * K_2)\}$$

Step2：人体部位定义
因为要涉及到姿态检测，所以文章中先定义了一个关节点，总共13个。然后由这些13个关节点，定义了5个身体的部位。定义如下图所示。

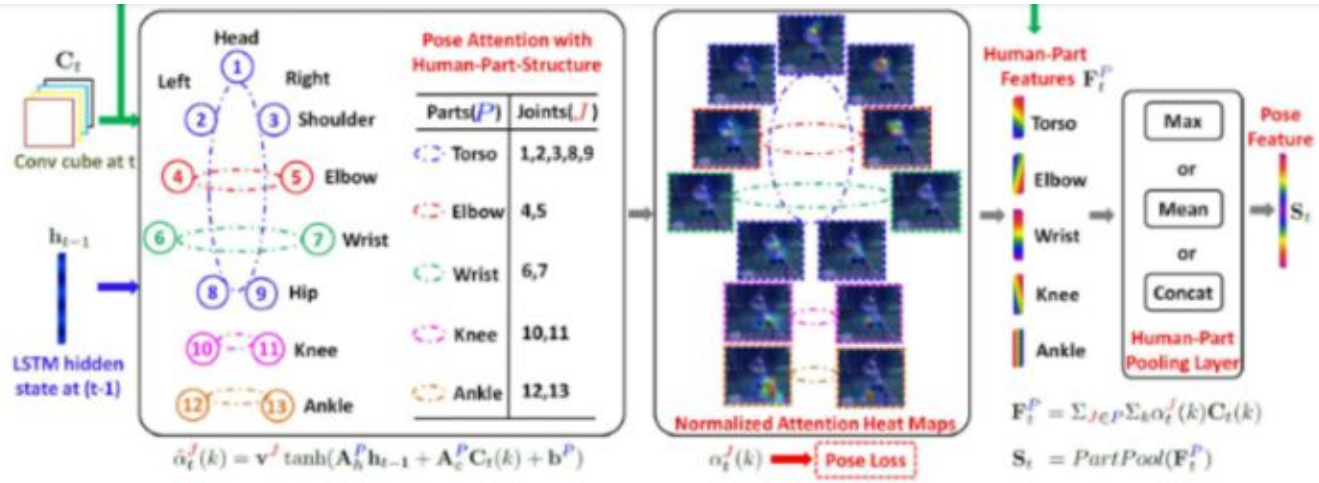


图 3.7 人体关键点和部位定义

3.6.4 LOSS FUNCTION

文章中定义了一个联合训练的Loss Function，将行为损失和姿态损失联合起来。

$$\mathcal{L}_{total} = \lambda_{action}\mathcal{L}_{action} + \lambda_{pose}\mathcal{L}_{pose} + \lambda_{\Theta} \parallel \Theta \parallel_2$$

$$\mathcal{L}_{action} = -\Sigma_{t=1}^{T}\Sigma_{c=1}^{C}\mathbf{y}_{t,c}\log\hat{\mathbf{y}}_{t,c}$$

$$\mathcal{L}_{pose} = \Sigma_J\Sigma_{t=1}^{T}\Sigma_{k=1}^{K_1\times K_2}(\mathbf{M}_t^J(k) - \alpha_t^J(k))^2$$

3.7 总结

行为识别目前还是视频理解方向的热点，而且至今为止也没有得到很好的解决。由于视频中目标复杂，场景复杂，所以单纯的Two-Stream和C3D方法表现得都不太如意。RPAN中引入了姿态监督的机制，或许能提高视频分类的效果。

4 行为检测

行为检测也是目前视频理解方向的研究主要热点，因为该任务更加贴近生活，在监控安防中有潜在的巨大价值。但是相比于行为分类，行为检测难度更高，不仅需要定位视频中可能存在行为动作的视频段，还需要将其分类。而定位存在行为动作的视频段是一个更加艰巨的任务。

回归操作。这类方法包含，利用Faster R-CNN框架[9][10]思路，利用SSD框架思路[11]，还有基于TAG网络[12]等等。还有一类方法是基于C3D做帧分类(Frame Label)，然后预测存在行为的视频段并分类，例如2017年ICCV的CDC网络[13]。

4.1 研究难点

上面简单阐述了行为检测的难点，这里总结一下主要有以下三点。

- 时序信息。与行为识别/分类一样，视频理解的通用难点就是时序信息的处理。所以针对这一点目前的主要方法基本上都是使用RNN读入CNN提取的特征，或者直接使用C3D一样的时序卷积。
- 边界不明确。不同于行为识别的是，行为检测要求做精确的动作区间检测，而生活中一个动作的产生往往边界不是十分确定的，所以这也是导致目前行为检测mAP偏低的原因。
- 时间跨度大。在生活中，一个行为动作往往跨度非常大，短的动作几秒左右，比如挥手。长的动作有的持续数十分钟，比如攀岩、骑行等等。这使得我们在提取Proposal的时候变得异常的艰难。

4.2 数据集介绍

行为检测方向常用的数据集主要是THUMOS 2014和ActivityNet。THUMOS 2014来自于THUMOS Challenge 2014,。它的训练集为UCF101数据集，验证集和测试集分别包括1010和1574个未分割的视频片段。在行为检测任务中只有20类动作的未分割视频是有时序行为片段标注的，包括200个验证集(3007个行为片段)和213个测试集视频(包含3358个行为片段)。

MEXaction2：MEXaction2数据集中包含两类动作：骑马和斗牛。该数据集由三个部分组成：YouTube视频，UCF101中的骑马视频以及INA视频。其中YouTube视频片段和UCF101中的骑马视频是分割好的短视频片段，被用于训练集。而INA视频为多段长的未分割的视频，时长共计77小时，且被分为训练，验证和测试集三部分。训练集中共有1336个行为片段，验证集中有310个行为片段，测试集中有329个行为片断。且MEXaction2数据集的特点是其中的未分割视频长度都非常长，被标注的行为片段仅占视频总长的很低比例。

ActivityNet: 目前最大的数据库，同样包含分类和检测两个任务。这个数据集仅提供视频的youtube链接，而不能直接下载视频，所以还需要用python中的youtube下载工具来自动下载。该数据集包含200个动作类别，20000（训练+验证+测试集）左右的视频，视频时长共计约700小时。由于这个数据集实在太大了，我的实验条件下很难完成对其的实验，所以我之前主要还是在THUMOS14和MEXaction2上进行实验。

4.3 CDC网络

- 第一次将卷积、反卷积操作应用到行为检测领域，CDC同时在空间下采样，在时间域上上采样。
- 利用CDC网络结构可以做到端到端的学习。
- 通过反卷积操作可以做到帧预测(Per-frame action labeling)。

4.3.1 网络结构

CDC网络在C3D的基础上用反卷积，将时序升维。做到了帧预测。以下是CDC网络的结构图。

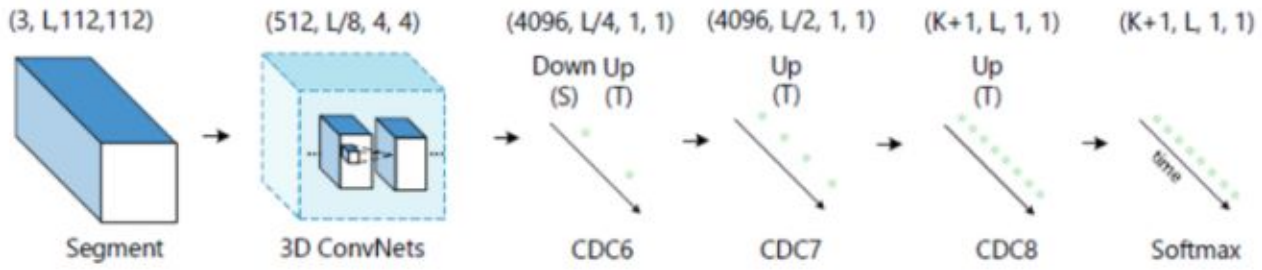


图 4.1 CDC 网络结构图

网络步骤如下所示。

- 输入的视频段是112x112xL，连续L帧112x112的图像
- 经过C3D网络后，时间域上L下采样到 L/8, 空间上图像的大小由 112x112下采样到了4x4
- CDC6: 时间域上上采样到 L/4, 空间上继续下采样到 1x1
- CDC7: 时间域上上采样到 L/2
- CDC8：时间域上上采样到 L，而且全连接层用的是 4096xK+1, K是类别数
- softmax层

4.3.2 CDC FILTER

文章的还有一大贡献点是反卷积的设计，因为经过C3D网络输出后，存在时间和空间两个维度，文章中的CDC6完成了时序上采样，空间下采样的同时操作。

如下图所示，一般的都是先进行空间的下采样，然后进行时序上采样。但是CDC中设计了两个独立的卷积核(下图中的红色和绿色)。同时作用于112x112xL/8的特征图上。每个卷积核作用都会生成2个1x1的点，如上conv6，那么两个卷积核就生成了4个。相当于在时间域上进行了上采样过程。

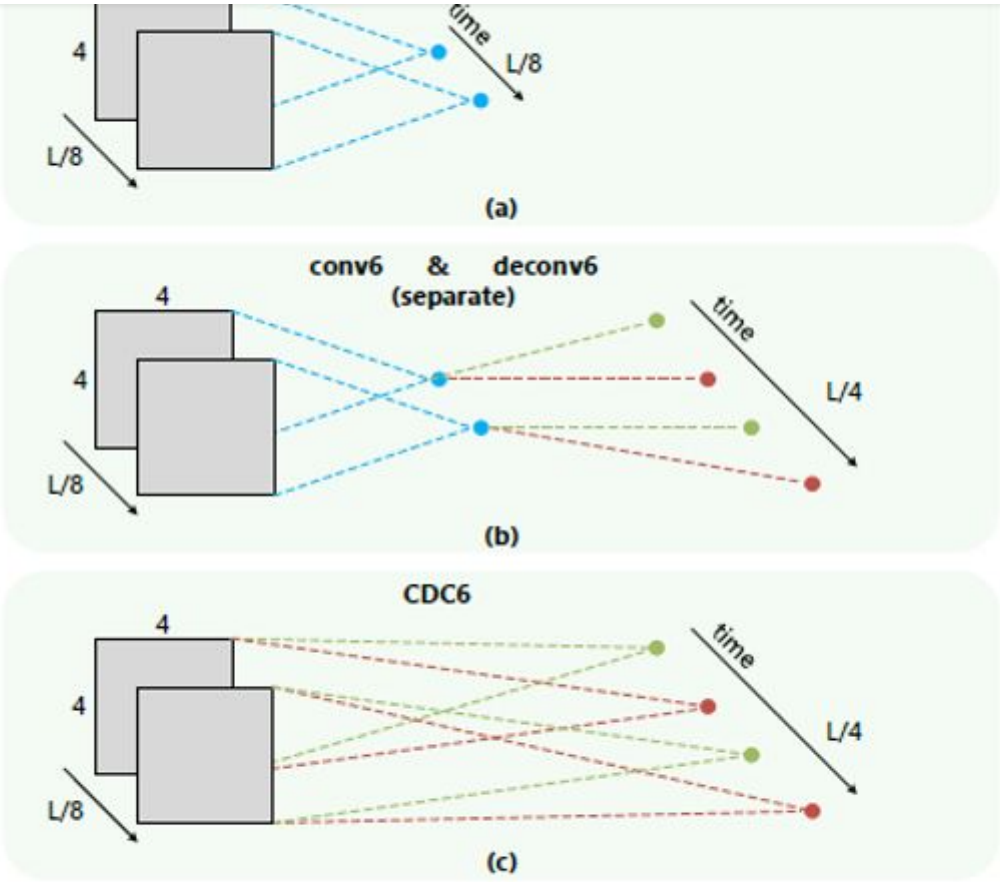


图 4.2 CDC6 filter 设计

4.3.3 LOSS FUNCTION

根据上述的网络结构图可以知道，经过softmax后会输出 (K+1, 1, 1)，也就是说针对每一帧，都会有一个类别的打分输出。所以作者说做到了每帧标签。

假设总共有N个training segments，我们取出第n个training sample，那么经过网络后会得到(K+1, 1, 1),经过CDC8后的输出为On[t], 然后经过softmax层，针对这个样本的第t帧，我们能得到它对应的第i个类别的打分如下。

$$P_N^{(i)}[t] = \frac{e^{O_n^{(i)}[t]}}{\sum_{j=1}^{K+1} e^{O_n^{(j)}[t]}}$$

最终总的Loss Function如下。

$$L = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{L}(-log(P_n^{(z_n)}[t]))$$

R-C3D(Region 3-Dimensional Convolution)网络[10]是基于Faster R-CNN和C3D网络思想。对于任意的输入视频L，先进行Proposal，然后用3D-pooling，最后进行分类和回归操作。文章主要贡献点有以下3个。

- 可以针对任意长度视频、任意长度行为进行端到端的检测
- 速度很快(是目前网络的5倍)，通过共享Progposal generation 和Classification网络的C3D参数
- 作者测试了3个不同的数据集，效果都很好，显示了通用性。

4.4.1 网络结构

R-C3D网络可以分为4个部分。

- 特征提取网络：对于输入任意长度的视频进行特征提取
- Temporal Proposal Subnet: 用来提取可能存在行为的时序片段（Proposal Segments）
- Activity Classification Subnet: 行为分类子网络
- Loss Function

下图是整个网络结构图。

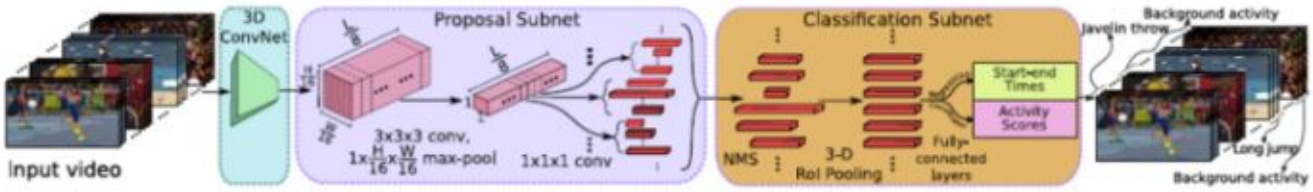


图 4.3 R-C3D 网络结构图

4.4.2 特征提取网络

骨干网络作者选择了C3D网络，经过C3D网络的5层卷积后，可以得到512 x L/8 x H/16 x W/16大小的特征图。这里不同于C3D网络的是，R-C3D允许任意长度的视频L作为输入。

Temporal Proposal Subnet

这一部分是时序候选框提取网络，类似于Faster R-CNN中的RPN，用来提取一系列可能存在目标的候选框。

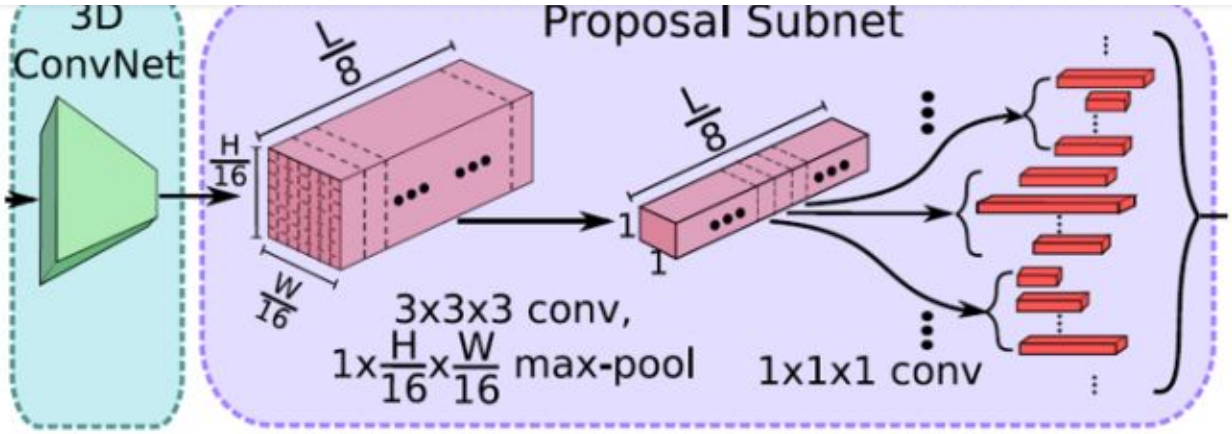


图 4.4 Temporal Proposal Subnet

Step1：候选时序生成

输入视频经过上述C3D网络后得到了512 x L/8 x H/16 x W/16大小的特征图。然后作者假设anchor均匀分布在L/8的时间域上，也就是有L/8个anchors，每个anchors生成K个不同scale的候选时序。

Step2: 3D Pooling

得到的 512xL/8xH/16xW/16的特征图后，为了获得每个时序点（anchor）上每段候选时序的中心位置偏移和时序的长度，作者将空间上H/16 x W/16的特征图经过一个3x3x3的卷积核和一个3D pooling层下采样到 1x1。最后输出 512xL/8x1x1.

Step3： Training类似于Faster R-CNN，这里也需要判定得到的候选时序是正样本还是负样本。文章中的判定如下。 正样本：IoU > 0.7，候选时序帧和ground truth的重叠数 负样本： IOU < 0.3 为了平衡正负样本，正/负样本比例为1:1.

4.4.3 ACTIVITY CLASSIFICATION SUBNET

行为分类子网络有如下几个功能：

- 从TPS（Temporal Proposal subnet）中选择出Proposal segment
- 对于上述的proposal，用3D RoI 提取固定大小特征
- 以上述特征为基础，将选择的Proposal做类别判断和时序边框回归

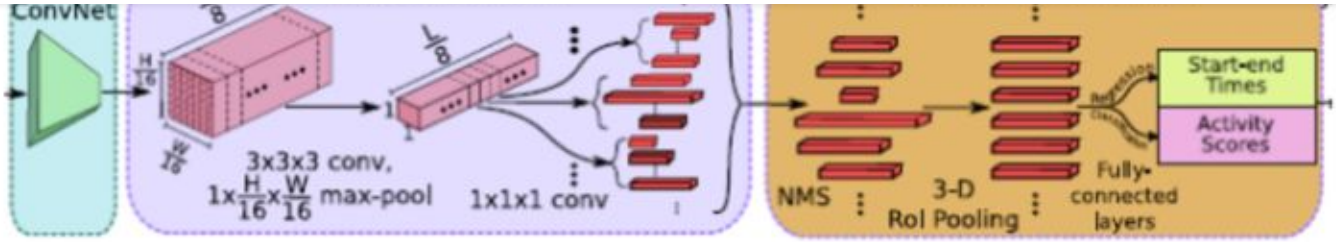


图 4.5 行为分类子网络

Step1： NMS

针对上述Temporal Proposal Subnet提取出的segment，采用NMS(Non-maximum Suppression)非极大值抑制生成优质的proposal。NMS 阈值为0.7.

Step2：3D RoI

RoI (Region of interest,兴趣区域).这里，个人感觉作者的图有点问题，提取兴趣区域的特征图的输入应该是C3D的输出，也就是512xL/8xH/16xW/16，可能作者遗忘了一个输入的箭头。 假设C3D输出的是 512xL/8x7x7大小的特征图，假设其中有一个proposal的长度（时序长度）为lp，那么这个proposal的大小为512xlpx7x7，这里借鉴SPPnet中的池化层，利用一个动态大小的池化核，ls x hs x ws。最终得到 512x1x4x4大小的特征图

Step3: 全连接层

经过池化后，再输出到全连接层。最后接一个边框回归(start-end time)和类别分类(Activity Scores)。

Step4： Traning

在训练的时候同样需要定义行为的类别，如何给一个proposal定label？同样采用IoU。

- IoU > 0.5，那么定义这个proposal与ground truth相同
- IoU 与所有的ground truth都小于0.5，那么定义为background

这里，训练的时候正/负样本比例为1:3。

4.4.4 LOSS FUNCTION

文章将分类和回归联合，而且联合两个子网络。分类采用softmax，回归采用smooth L1。

- 其中的N都代表batch size
- lamda 为1

[1] Wang H, Schmid C. Action recognition with improved trajectories[C]//Computer Vision (ICCV), 2013 IEEE International Conference on. IEEE, 2013: 3551-3558.
[2] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.
[3] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576. [4] Feichtenhofer C, Pinz A, Zisserman A P. Convolutional two-stream network fusion for video action recognition[J]. 2016.
[5] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European Conference on Computer Vision. Springer, Cham, 2016: 20-36.
[6] 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.
[7] Du W, Wang Y, Qiao Y. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3725-3734.
[8] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.
[9] Dai X, Singh B, Zhang G, et al. Temporal Context Network for Activity Localization in Videos[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 5727-5736.
[10] Xu H, Das A, Saenko K. R-c3d: Region convolutional 3d network for temporal activity detection[C]//The IEEE International Conference on Computer Vision (ICCV). 2017, 6: 8.
[11] Lin T, Zhao X, Shou Z. Single shot temporal action detection[C]//Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017: 988-996.
[12] Zhao Y, Xiong Y, Wang L, et al. Temporal action detection with structured segment networks[C]//The IEEE International Conference on Computer Vision (ICCV). 2017, 8.
[13] Shou Z, Chan J, Zareian A, et al. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 1417-1426.

> **<个人网页blog已经上线，一大波干货即将来袭：faiculty.com/>**
>

发布于 2018-05-11

机器学习

文章被以下专栏收录

推荐阅读

3 条评论

切换为时间排序

写下你的评论...

爱吃兔子的猪 7 个月前

发现有一处笔误哦：特征描述子 HOG的特征长度应该是2 * 2 * 3

1

流浪者 (作者) 回复 爱吃兔子的猪 7 个月前

谢谢提醒～

知乎 首发于 **计算机视觉**

夕舞雪薇 7 个月前

最近开始做这方面，感谢大佬分享

赞