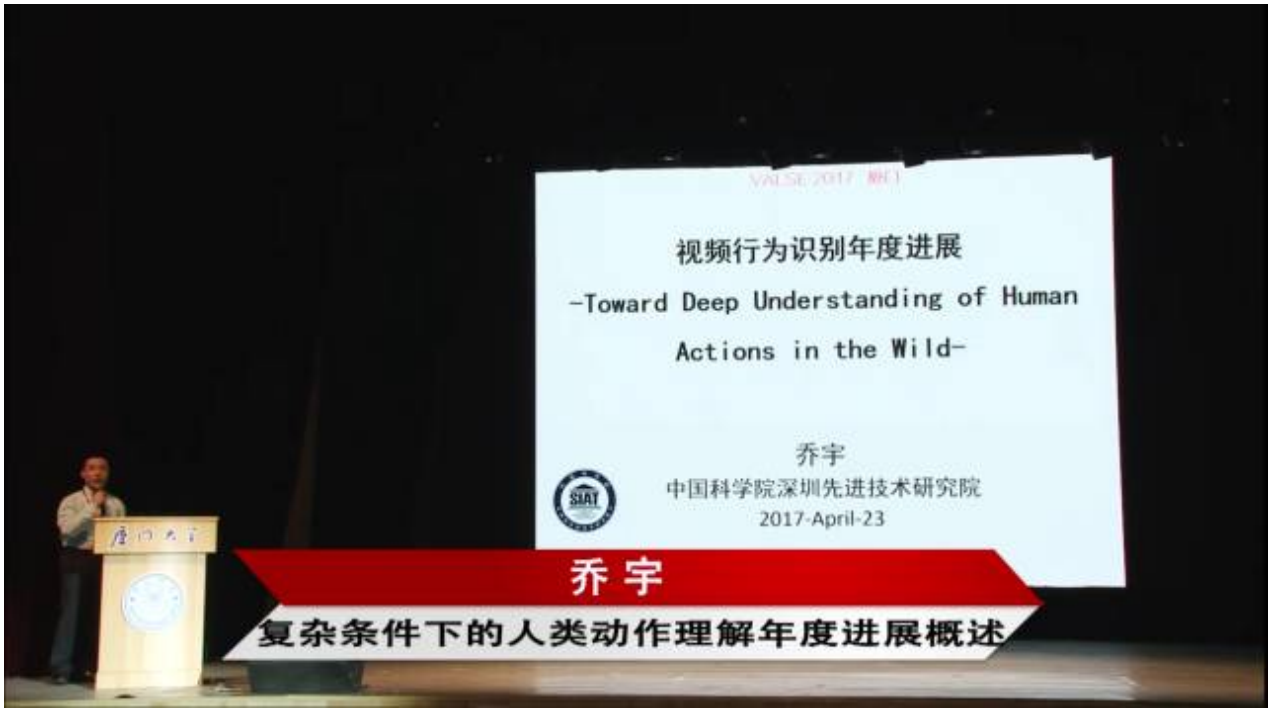
视频行为识别年度进展

2017-06-12 19:02 乔宇 0 3 阅读 567

点击上方“深度学习大讲堂”可订阅哦！

深度学习大讲堂是由中科视拓运营的高质量原创内容平台，邀请学术界、工业界一线专家撰稿，致力于推送人工智能与深度学习最新技术、产品和活动信息！

编者按：行为识别技术在智能监控、人机交互、视频序列理解、医疗健康等众多领域扮演着越来越重要的角色，而视频行为识别技术受到遮挡，动态背景，移动摄像头，视角和光照变化等因素的影响而具有很大的挑战性。来自中科院深圳先进技术研究院的乔宇研究员，将带着大家回顾过去一年中视频行为识别领域的研究进展。文末提供开源代码下载链接及文中提到论文的下载链接。



VALSE 2017 厦门

视频行为识别年度进展

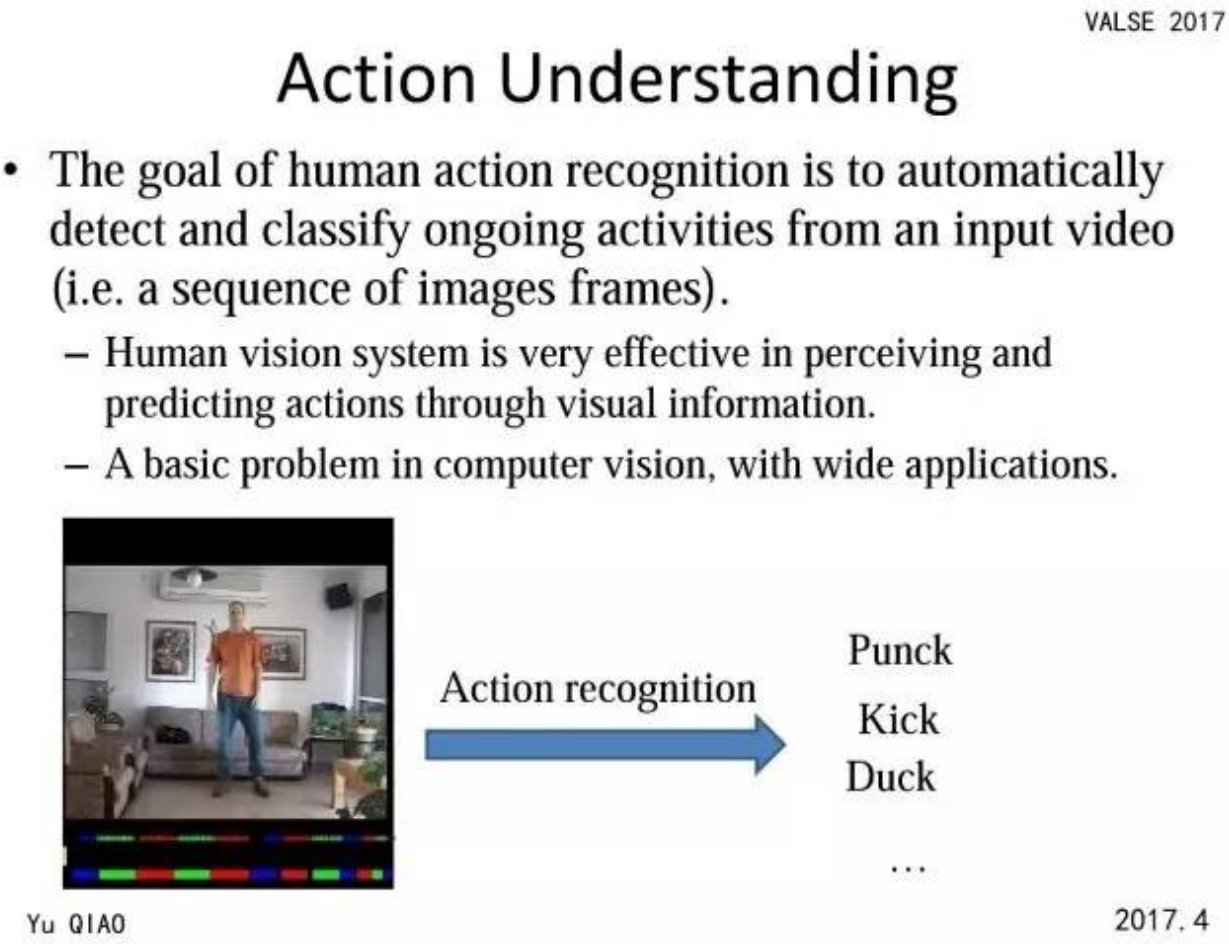
-Toward Deep Understanding of Human Actions in the Wild-

乔宇

中国科学院深圳先进技术研究院

2017-April-23

视频行为识别，通俗来讲就是给出一段视频，来判断人或者感兴趣的物体在进行什么行为。



行为识别是计算机视觉的一个基本问题。我们对这个问题感兴趣可能有两方面的原因：一方面，人类非常擅长解决这个任务，人在日程生活中需要识别和预测周边人的行为如走路、跑、体育活动等等；另一方面，这个任务有非常多的应用，比如监控视频、互联网的视频检索处理、人机交互等非常多领域有非常多的应用。

视频行为识别数据集简介

很多视觉的问题都和数据库有密切的关系，这里列出了跟行为相关的一些数据库。

其实，我们可以把他们分为三种类型：一种类型的数据库是早期实验室采集的，背景相对简单和固定，只有一个人在视频中央进行某个特定行为。后来，人们逐渐开始从体育转播的视频或者从一些电影中录取一些行为视频，因为这些视频中的人是一些专业的演员或者体育运动员，而且是使用专业拍摄设备拍摄，质量比较高；最后人们关心的行为是activity in the wild，也就是指一些用户拍摄的各种各样的视频数据。

VALSE 2017

Action Video Dataset list

Dataset	Year	Actions	Clips per Action	Settings	SoTA
KTH [82]	2004	6	400	controlled settings	≥ 0.95
Weizmann [34]	2005	9	9	controlled settings	≥ 0.95
IXMAS [114]	2006	11	33	multiview	≥ 0.95
Hollywood [54]	2008	8	30 - 140	movie	≥ 0.60
UCF Sports [76]	2009	9	14 - 36	sports broadcast	≥ 0.95
Hollywood2 [61]	2009	12	61 - 278	movie	$\simeq 0.70$
UCF YouTube [59]	2009	11	100	web video	≥ 0.90
MSR [120]	2009	3	14-25	indoor and outdoor	≥ 0.95
High Five [68]	2010	4	50	TV shows	$\simeq 0.65$
UT-interaction [78]	2010	6	20	controlled settings	≥ 0.95
Olympic Sports [63]	2010	16	21 - 67	web video	$\simeq 0.90$
UCF50 [75]	2010	50	min. 100	web video	$\simeq 0.95$
HMDB51 [51]	2011	51	min. 101	movie & web video	$\simeq 0.65$
Cooking Dataset [77]	2012	65	-	controlled settings	$\simeq 0.50$
UCF101 [89]	2012	101	min. 100	web video	$\simeq 0.90$
Sports-1M [47]	2014	487	average 2327	web video	$\simeq 0.64$

4/27/2017
Yu QIAO

7

2017. 4

这里列出了从2004年到2014年开发的比较主要的数据库。可以看出，这些数据库随着时间发展，其种类从传统的几个类别到几十个类别，再到几百个类别，一直在扩充。当然这实际上也只是行为的一个子类。就是说行为的类别成千上万，非常多，字典中很多动词或者动名词都可以对应于一个行为。另外，也可以看到数据库中视频的数据量也变得越来越大。

VALSE 2017

ActivityNet 2016

ACTIVITYNET

200 categories， 648 hrs video， 10k for training , 5k for testing

http://activity-net.org/challenges/2016/

Achieve NO 1 in classification task in ActivityNet 2016 among 24 teams.

Validation Set	mAp	Top-3 Acc.
Visual	90.4%	95.2%
Audio	15.2%	29.1%
Visual + Audio	90.9%	95.6%
Testing Set	**mAP**	**Top-3 Acc.**
Visual CNN (Single)	91.2%	95.6%
Final Ensemble	93.2%	96.4%

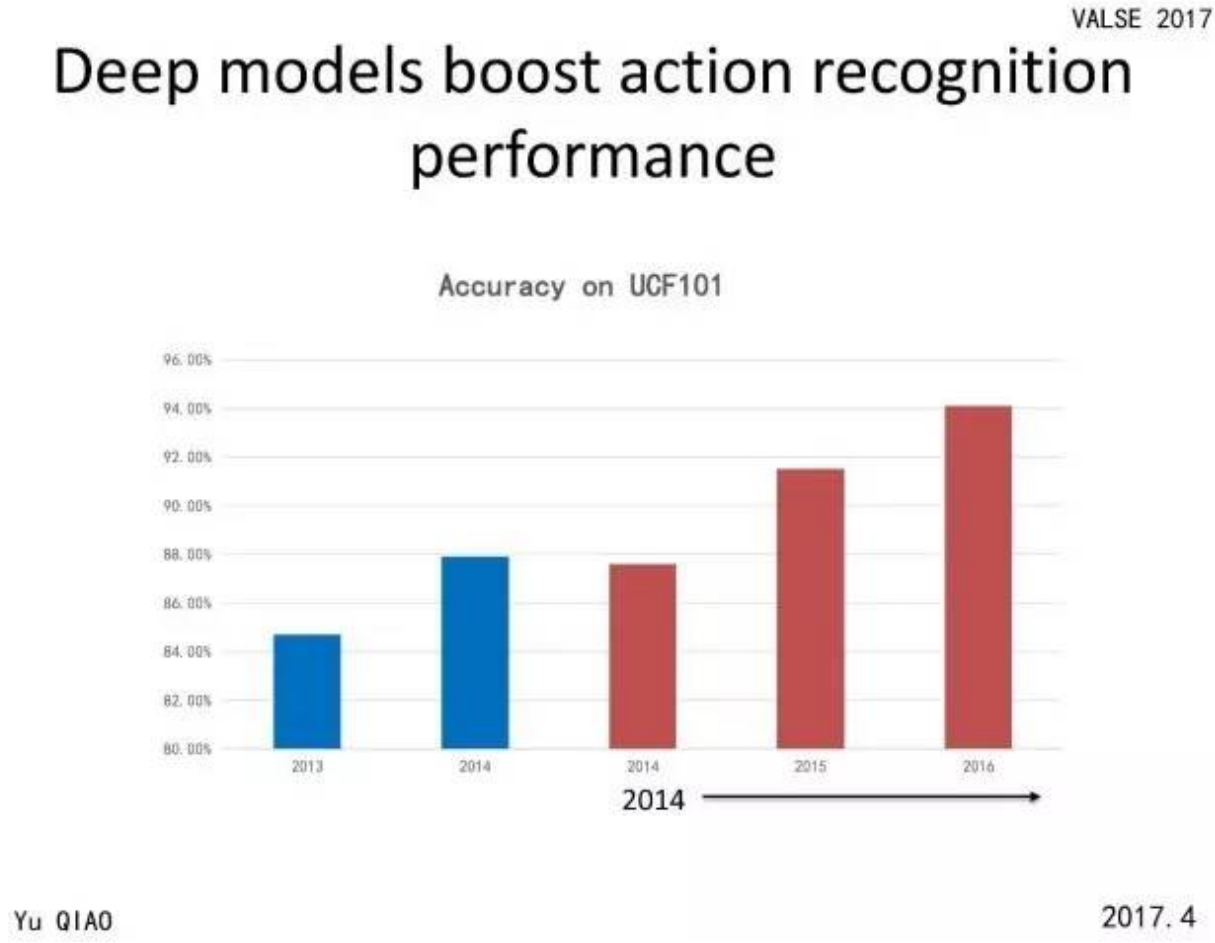
ETH zürich

Yu QIAO

2017. 4

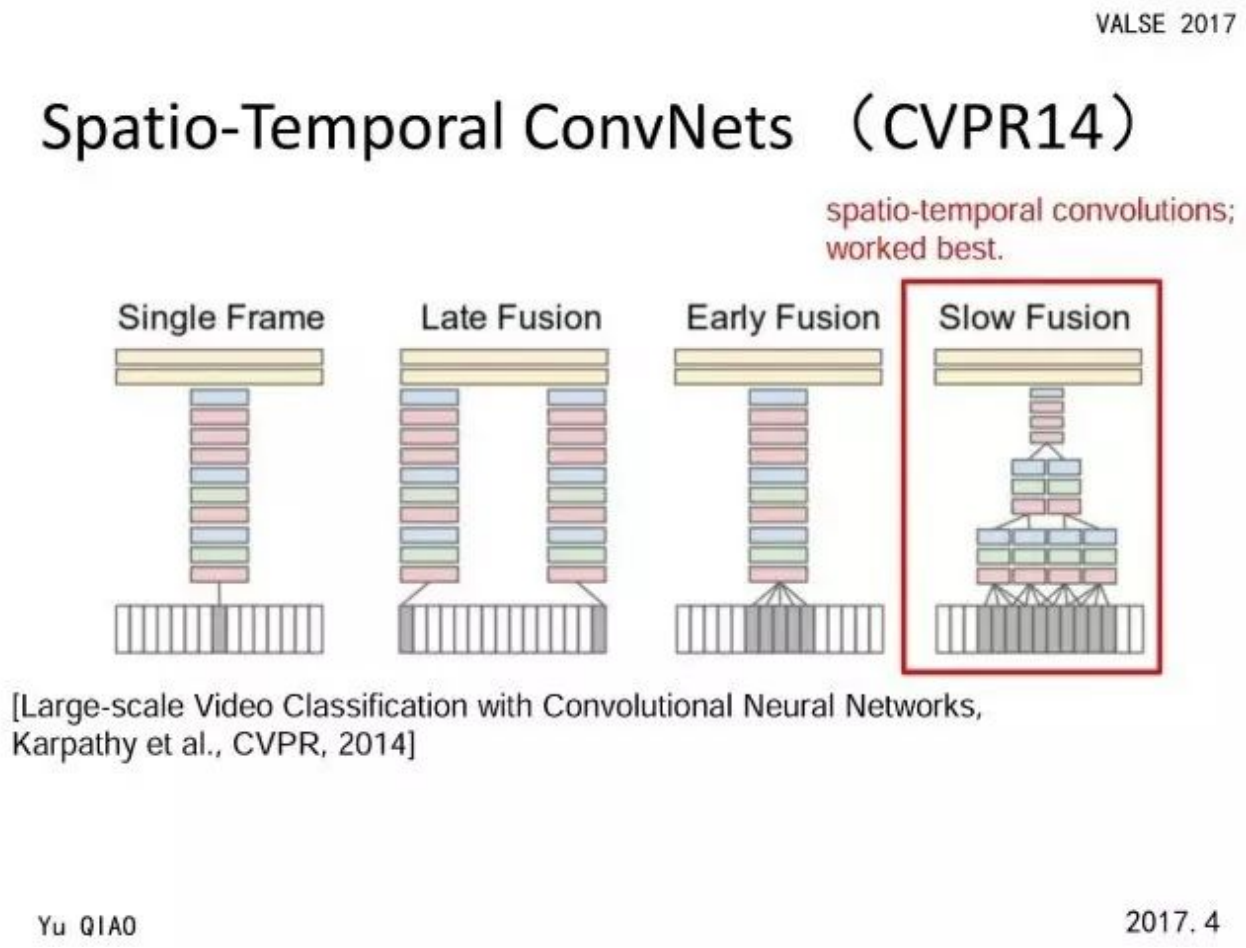
在过去的一年中，可以看到有两个比较大的数据库，一个叫做ActivityNet。这是一个行为的数据库，有200个类别，600多个小时的视频。在CVPR2016上，该数据库建设方围绕该库组织了一个竞赛。我们和香港中文大学、苏黎世联邦理工学院（ETH Zurich）一起获得了这个比赛的第一名。通过融合音频的方法，在这个数据库上可以得到93%的识别率。

另一个更大规模的数据库叫YouTube-8M。这个数据库由谷歌建立，其中的数据都来自YouTube，总共有700多万，包括45万个小时的视频，4700多个类别。谷歌今年用这个数据库在Kaggle上也组织了一个竞赛。



如果用一句话概括过去几年视频行为识别的进展，可能跟很多领域一样：深度模型大大推动了行为识别率的提高。事实上，在视频领域取得成功比在图像上要慢一些，甚至在开始的一段时间，深度学习的方法并不是特别成功。这里列出了一个从2013年到2016年大家用的非常广泛的UCF101数据库上的一些结果。在2014年，当图像领域深度学习方法已经远远甩出非深度学习方法的时候，在行为识别领域，都是非深度学习做的方法要更好一些。当然，从2015年深度学习方法开始取得了一些进展，然后到2016年、2017年，现在深度学习方法已经非常有效了。所以，下面主要围绕深度学习的方法，来回顾一下视频行为识别领域的发展历史。

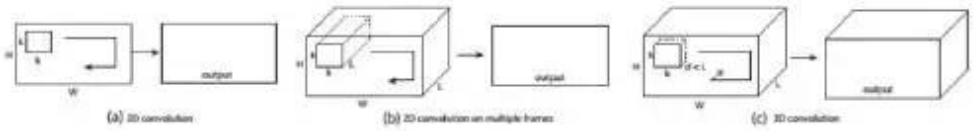
深度学习在视频行为识别上的进展



在2014年，受卷积神经网络在图像分类领域取得的成功的启发，大家开始考虑把卷积神经网络用于视频分类，但是，早期通过微调在ImageNet数据集上训练的网络获得的结果都不太好，大概有5到6个点的识别率的提高。当时大家的一个共识就是数据量不够，这时谷歌就建立了一个100万的数据库，叫做Sports-1M。在这个数据库上，他们通过融合卷积神经网络特征的方法构造了几种融合（fusion）策略，但是这个方法的结果并不是很好，比如说，当时这个方法在UCF101的识别率只有百分之六十几。

VALSE 2017

C3D （CVPR 15）



2D Convolution → 3D Convolution

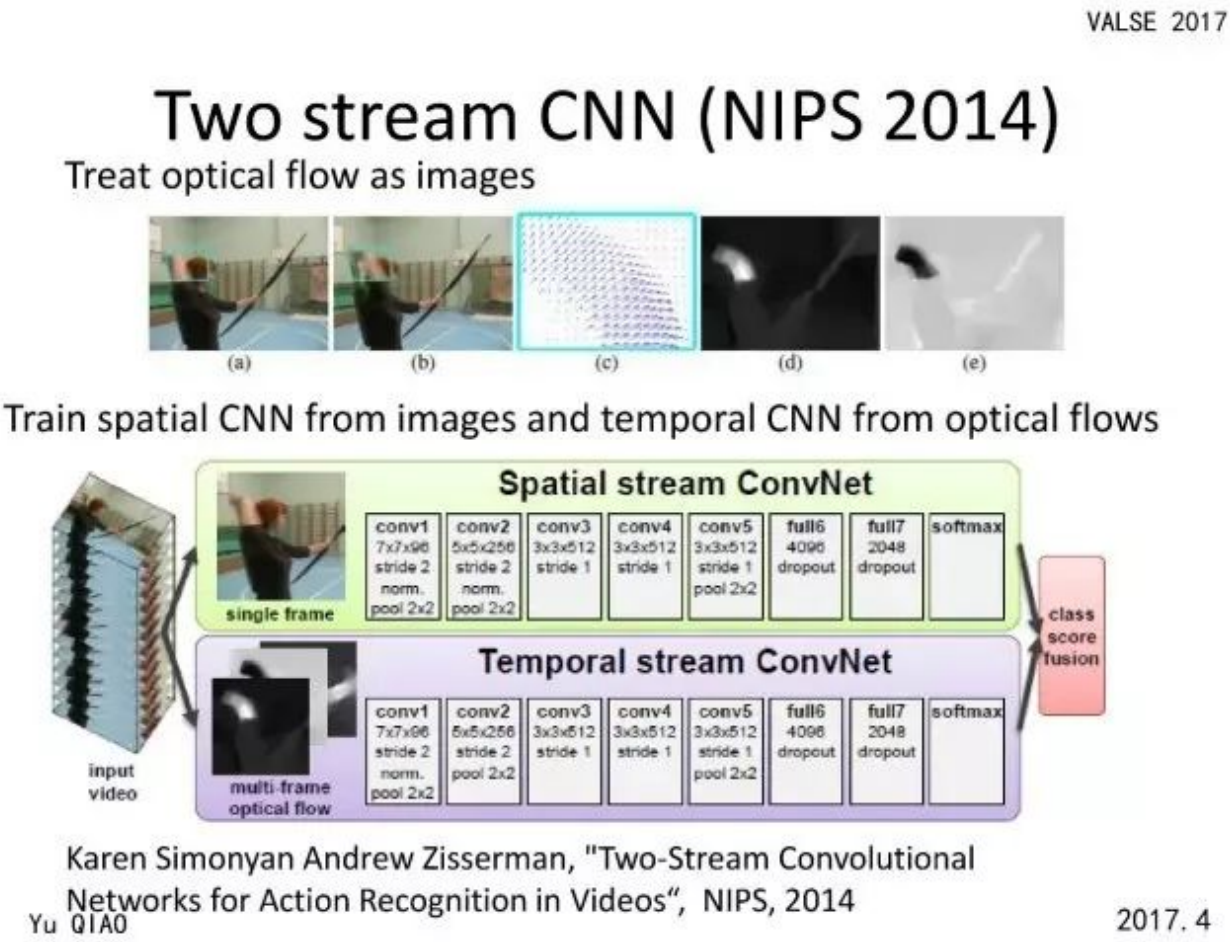
Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from `pool1` to `pool5`. All pooling kernels are $2 \times 2 \times 2$, except for `pool1` is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

C3D: 3D VGGNet

[Learning Spatiotemporal Features with 3D Convolutional Networks, Tran et al. 2015]

Yu QIAO 2017.4

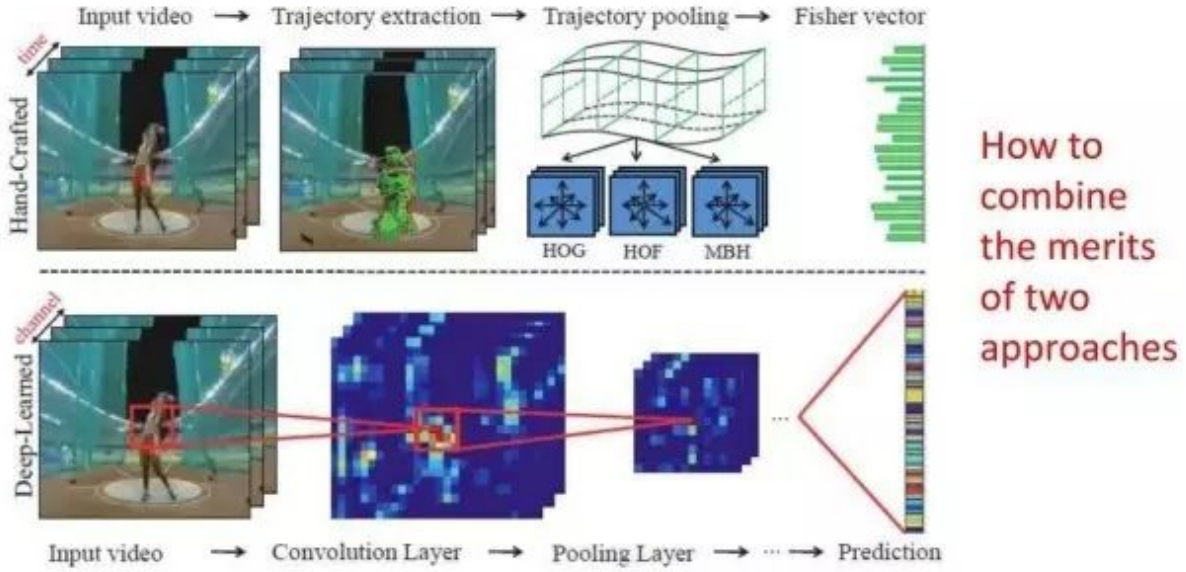
此后研究人员希望提升深度学习方法在视频分类的应用，一个比较成功的工作是Facebook的C3D。简单来说，这个工作就是把二维卷积推广到三维。把VGG的网络中3*3的卷积核变成了3*3*3。虽然说起来简单，真正想把这个网络训练起来是需要相当的功力的。



效果更好的是基于VGG的一个工作，Two stream CNN。以前做视频识别的时候，对于运动或行为信息，都会使用光流的信息，这个网络把光流当成一个图像，光流本身是一个向量，可以把x方向y方向当成两张图像，然后再对光流图像训练一个卷积神经网络。这个方法与比传统非深度学习方法相比只有不到一个百分点的差距。

VALSE 2017

Trajectory-Pooled Deep-Convolutional Descriptors (CVPR15)



Limin Wang, Yu Qiao, Xiaoou Tang "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015

Yu QIAO

2017. 4

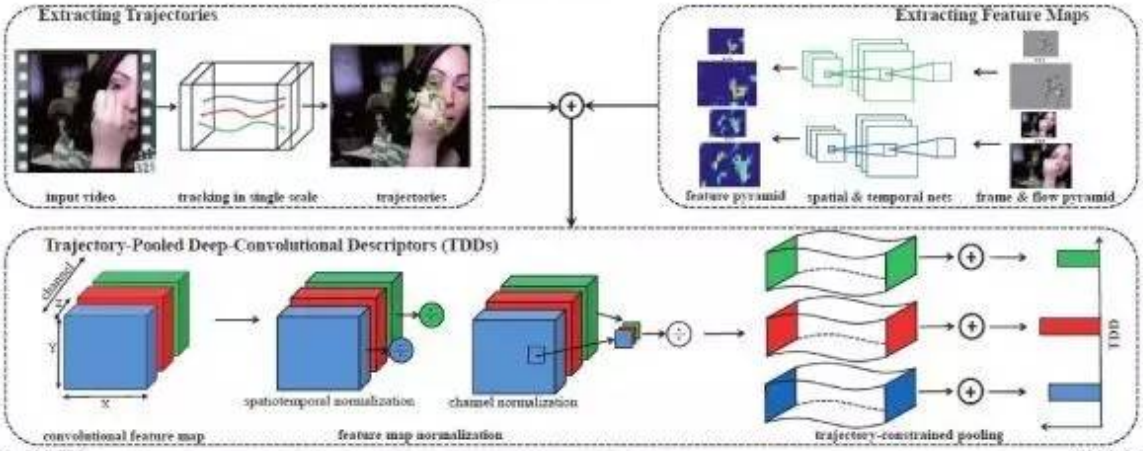
后面，我们把在传统方法中积累的一些经验和深度学习的方法做一个结合。在传统方法中我们会使用一些运动轨迹——通过光流跟踪的运动轨迹，使用运动轨迹的好处是，可以比较好的在运动比较显著的区域进行特征的集中提取，然后对传统方法沿着运动轨迹提取的一些卷积特征做下采样操作。这个是CVPR 2015的工作。

VALSE 2017

Framework of the proposed methods

Propose *trajectory-pooled deep convolutional descriptor* (TDD) to integrate the key factors from handcrafted and deep approaches.

- Utlize two-stream ConvNets to obtain multi-scale deep convolutional features.
- Pool the local ConvNet responses over the spatiotemporal tubes centered at the trajectories.



Yu QIAO 2017. 4

这个工作第一次在UCF101数据库上将识别率提升到了90%以上。

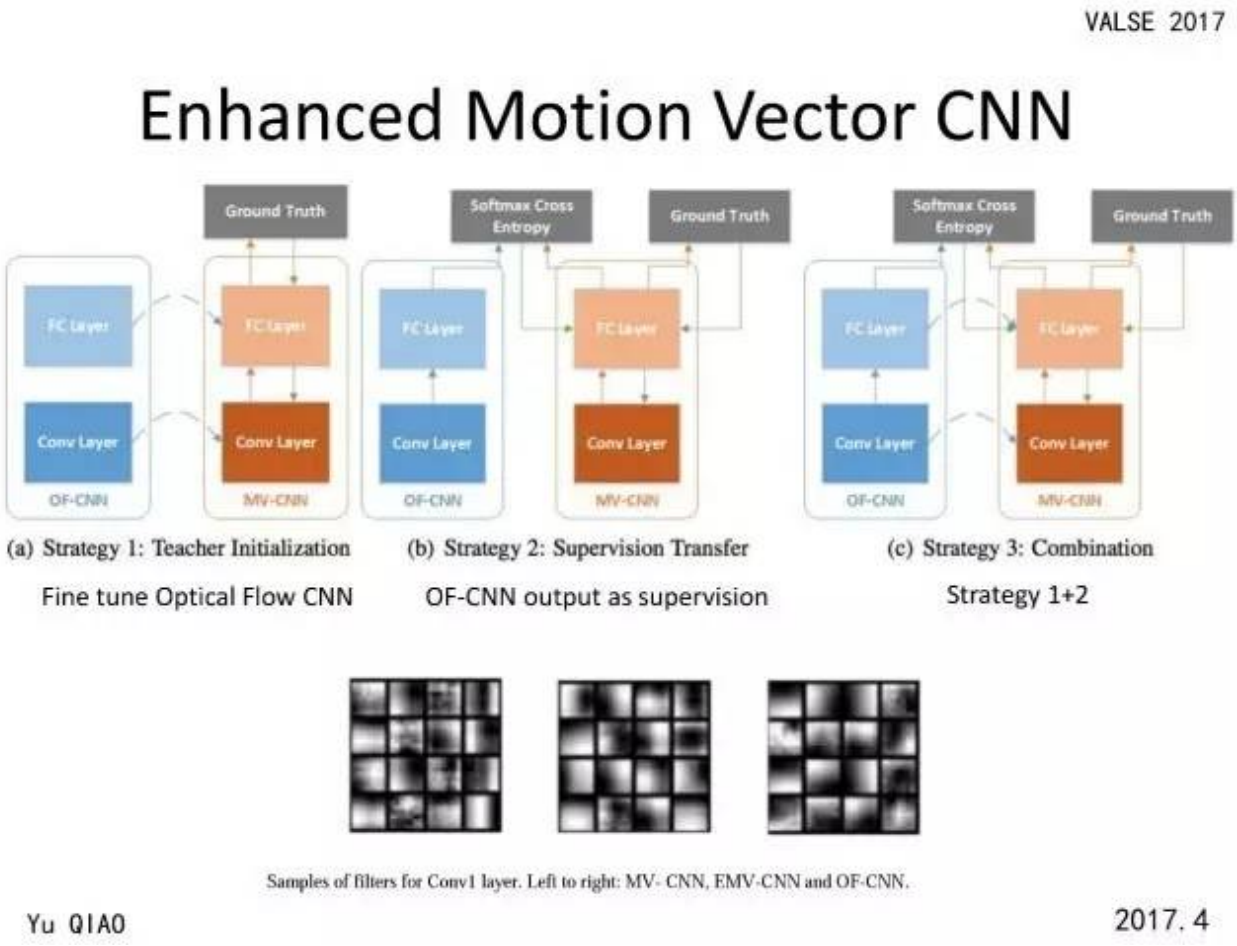
VALSE 2017

Motion Vector CNN (CVPR15)

- Many deep learning approaches for video based action recognition are computationally expensive, due to the calculation of optical flows
- Motion vector also includes motion information of local regions

B. Zhang, L. Wang, Z. Wang, Yu Qiao, and H. Wang " Real-time Action Recognition with Enhanced Motion Vector CNNs ", Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2016 2017. 4

处理视频的时候数据量比较大，但很多视频应用对实时性有要求。未来加速，我们就提出了一个用运动向量代替光流的方法，运动向量只存在于压缩视频MPEG或者h.264中，无需计算就可获得；当然直接用运动向量会造成识别率很大的降低。

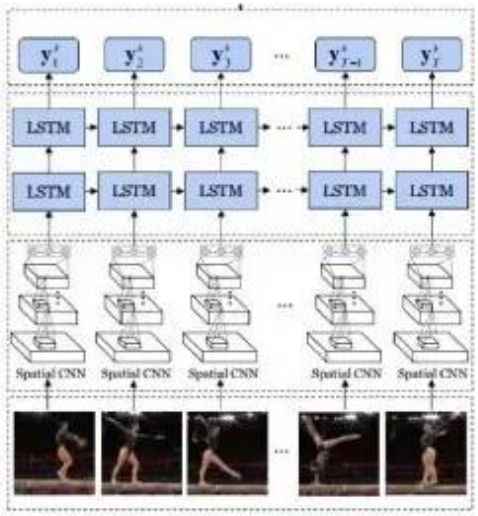


事实上，我们是把光流训练出来的网络作为一个老师来教运动向量的网络。通过这样的方法，在识别率没有下降太大的情况下，可以每秒钟用GPU做到400帧，也就是使用一个GPU可以支撑大概10路视频。

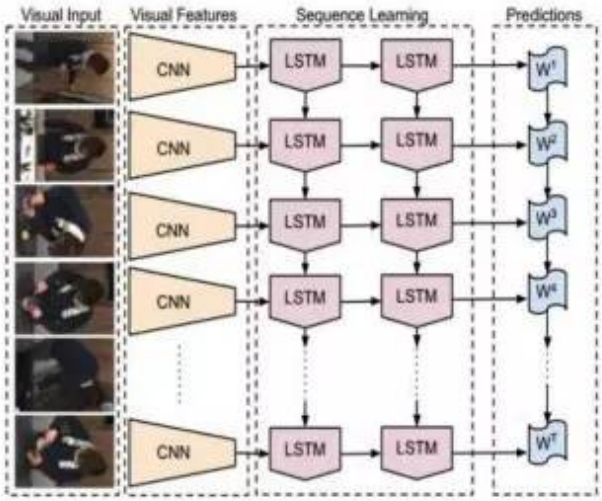
长时间序列的视频行为识别方法

VALSE 2017

Recurrent Neural Network/LSTM for Action Recognition



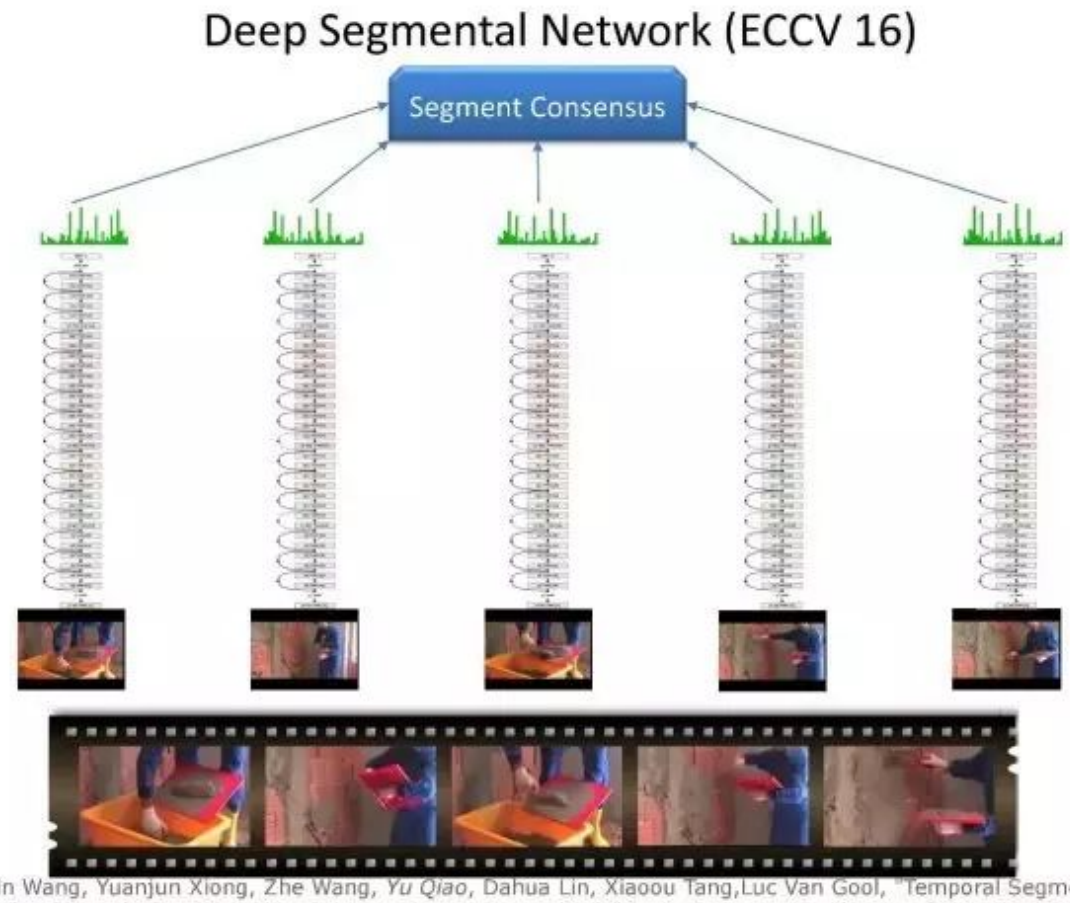
Wu Z, Wang X, Jiang Y, et al. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification[C]. acm multimedia, 2015: 461-470.



[Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al., 2015]

Yu QIAO 2017. 4

在解决这个短时的视频行为识别问题后，人们越来越关注长时的序列，自然就考虑把递归神经网络（RNN、LSTM）这些模型用于时序建模。常见的方式是，使用卷积神经网络（CNN）提取图像帧的特征，把CNN抽取的特征送到长短期记忆网络（LSTM）中去，然后做分类，这个框架也是后来很多做video caption工作的一个基础。



1.Limin Wang, Yuanjun Xiong, Zhe Wang, *Yu Qiao*, Dahua Lin, Xiaoou Tang,Luc Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," Proc. European Conference Computer Vision, 2016

另一方面，在序列建模的时候，视频是分段的，不同段有各自不同的语义，我们与香港中大合作做了一个工作叫Deep Segmental Model，就是把视频分成很多段，针对每一段抽取特征，当然我们也注意到很重要的一点：不同段的特征的重要性是不一样的，然后需要把重要性考虑到识别模型。

VALSE 2017

Performance of TSN

HMDB51		UCF101		THUMOS14		ActivityNet	
iDT+FV [2]	57.2%	iDT+FV [70]	85.9%	iDT+FV [70]	63.1%	iDT+FV [70]	66.5%
DT+MVSV [46]	55.9%	DT+MVSV [46]	83.5%	object+motion [71]	71.6%	Depth2Action [72]	78.1%
iDT+HSV [73]	61.1%	iDT+HSV [73]	87.9%				
MoFAP [51]	61.7%	MoFAP [51]	88.3%				
Two Stream [1]	59.4%	Two Stream [1]	88.0%	Two Stream [1]	66.1%	Two Stream [1]	71.9%
VideoDarwin [23]	63.7%	C3D (3 nets) [17]	85.2%	EMV+RGB [18]	61.5%	C3D [17]	74.1%
MPR [74]	65.5%	Two stream +LSTM [4]	88.6%				
$F_{ST}CN$ [55]	59.1%	$F_{ST}CN$ [55]	88.1%				
TDD+FV [5]	63.2%	TDD+FV [5]	90.3%				
LTC [24]	64.8%	LTC [24]	91.7%				
KVMF [75]	63.3%	KVMF [75]	93.1%				
TSN (3 seg)	70.7%	TSN (3 seg)	94.2%	TSN (3 seg)	78.8%	TSN (3 seg)	89.0%
TSN (7 seg)	**71.0%**	TSN (7 seg)	**94.9%**	TSN (7 seg)	**80.1%**	TSN (7 seg)	**89.6%**

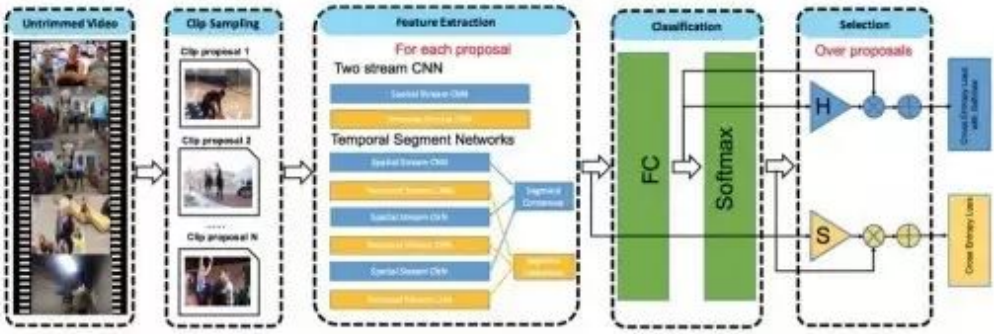
L. Wang et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, in ECCV 2016.
L. Wang et al. Temporal Segment Networks for Action Recognition in Videos, to appear.

Yu QIAO 2017. 4

这里列出了我们最新的一些结果，包括几个比较大的数据库以及ActivityNet。这个方法也是一个比较高的基准（baseline）了。

VALSE 2017

UntrimmedNet: Directly Learning from Untrimmed Video (CVPR17)



Leverage **attention modeling** in TSN for **weakly supervised action recognition and detection.**

L. Wang et al. UntrimmedNet for Weakly Supervised Action Recognition and Detection, in CVPR 2017.

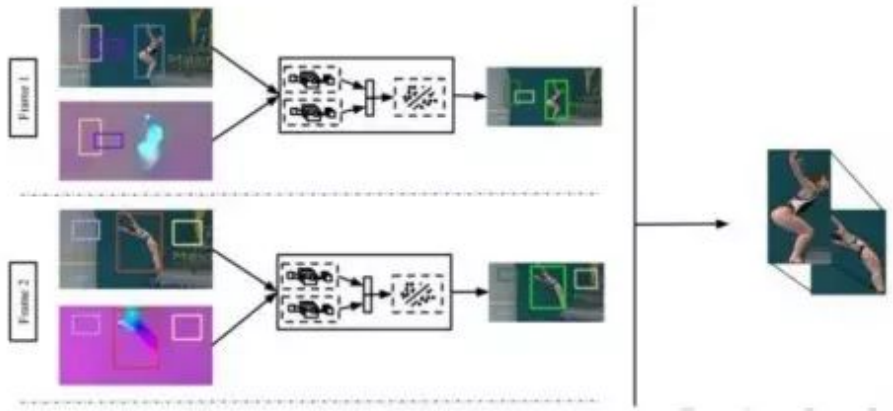
Yu QIAO 2017.4

这个工作把TSN网络推广到弱监督的识别和检测，视频中许多时间段并不包括我们感兴趣的行为，这个方法把注意机制用于非截断视频行为的识别与检测。

视频行为检测

VALSE 2017

R-CNNs for Pose Estimation and Action Detection (CVPR'14)



R-CNN based frame-level detection + linking with dynamic programming

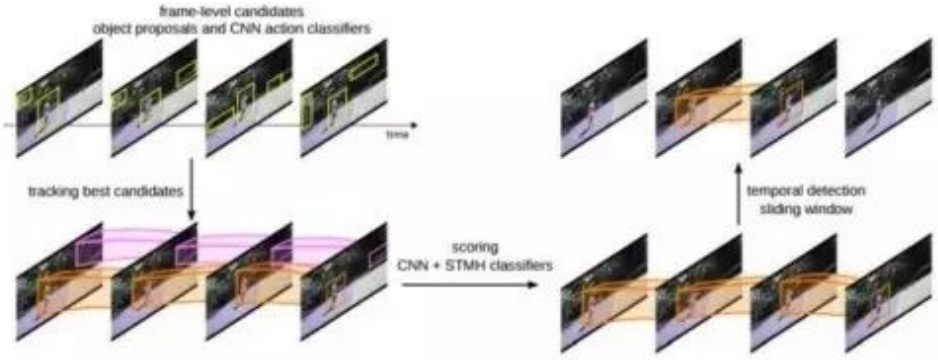
Gkioxari, G , Hariharan, B , Girshick, R, J. Malik, R-CNNs for Pose Estimation and Action Detection, IEEE Conference on Computer Vision & Pattern Recognition, 2014

Yu QIAO 2017. 4

除了视频的识别之外，视频中的行为检测也是一个非常重要的问题，这个问题很大程度上是跟随物体检测方法的进步。较早的时候，就是伯克利的一个组把RCNN也运用到视频检测中，通过动态规划的方法把RCNN检测的框连接起来。

VALSE 2017

Learning to Track for Spatio-Temporal Action Localization (CVPR15)



R-CNN based frame-level detection (replacing SS by a better one{EdgeBoxes) + best candidates tracking

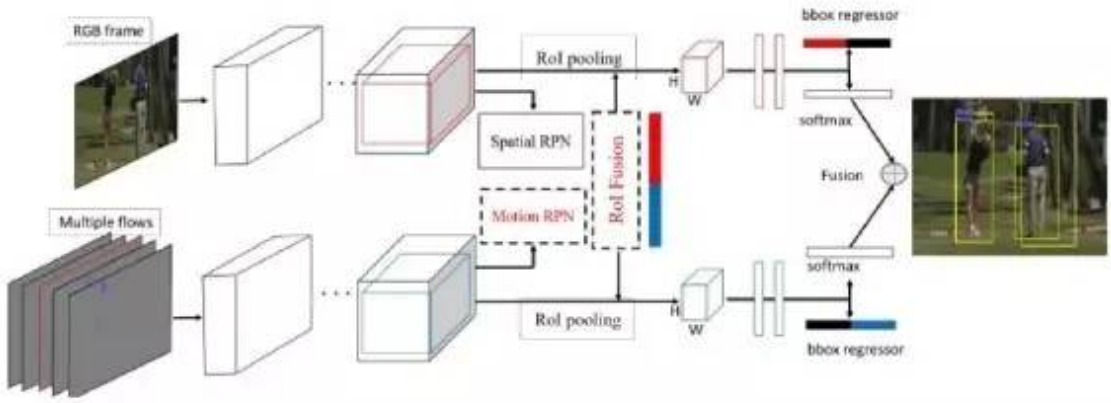
P Weinzaepfel，Z Harchaoui，C Schmid, "Learning to Track for Spatio-Temporal Action Localization", CVPR, 2015

Yu QIAO 2017. 4

后面，法国国家信息与自动化研究所（INRIA）改进了这个工作，一个是改进了提proposal的方法，另外一个是加了跟踪的环节。

VALSE 2017

Multi-region two-stream R-CNN for action detection (ECCV 16)

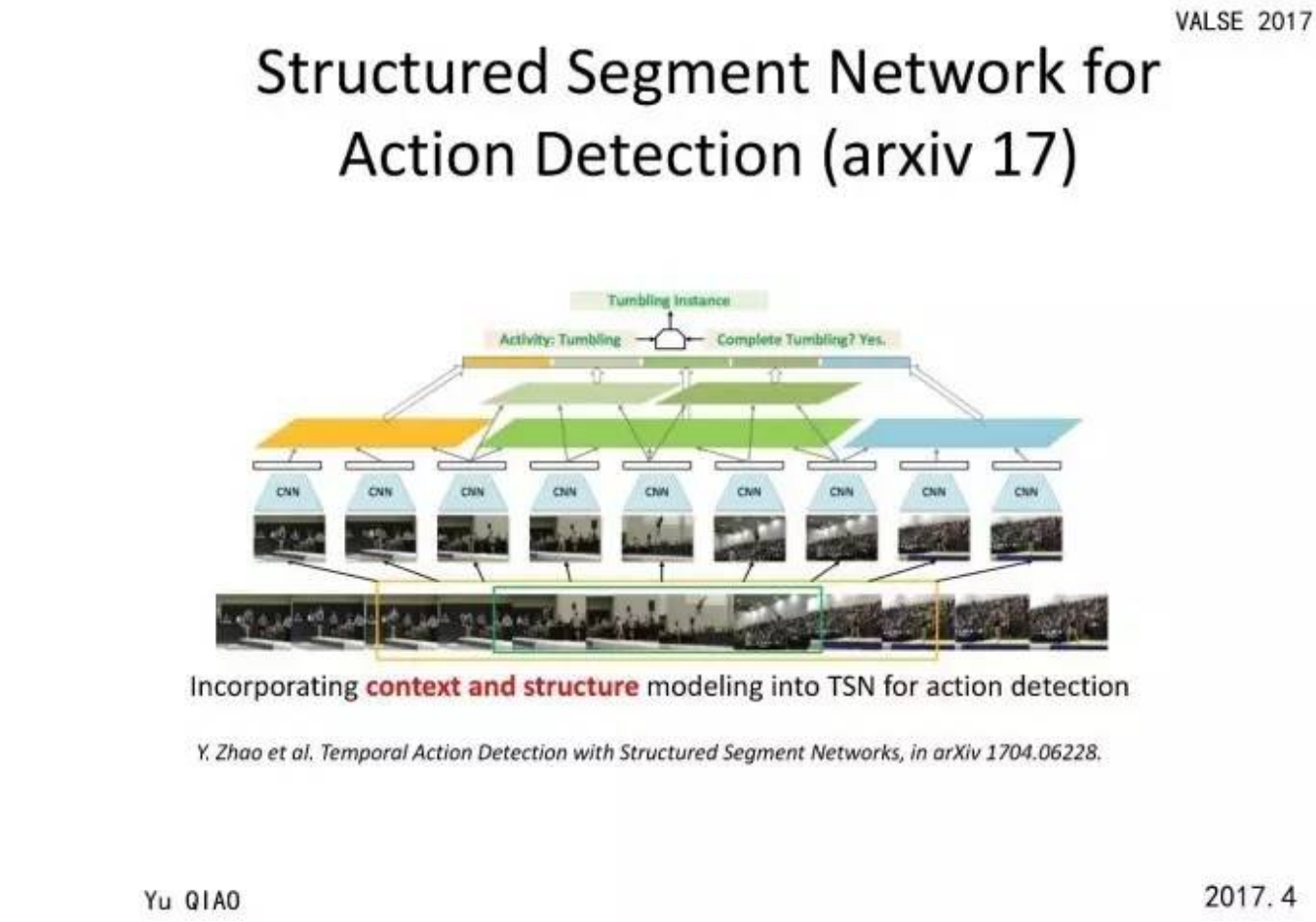


RPN proposals with multiple RGB frames & optical flows + multi-region scheme for action detection

Xiaojiang Peng, Cordelia Schmid, Multi-region two-stream R-CNN for action detection European Conference on Computer Vision (ECCV), 2016.

Yu QIAO 2017. 4

后面又有研究人员把时空的特征联合起来形成proposal，在跟踪的时候还加入了一些框的合并机制来进一步提高精度。



最近的一个工作是，香港中文大学研究组将视频的结构信息，以及上下文信息用到行为的检测中去，取得了很好的效果。

总结：

行为识别现在是一个正在进行的领域。随着更大的数据库和更复杂的挑战的出现，我想这个问题远远还没有到解决的时候，从短时特征的提取到长时时间序列的建模，还有很多工作需要去做，包括后面提到的检测、跟踪、姿态估计。以及相关问题。另外行为分析识别还和video caption有很大的相关性，都属于视频理解。这里列出了一些关键词，大家选研究方向的话，可以进行参考，包括注意力机制、记忆、强化学习等。

这里是我们之前工作的一些代码，欢迎大家下载和使用。

文中提到所有论文的下载链接为：
http://pan.baidu.com/s/1pLx2Sxd

致谢：
本文主编袁基睿，诚挚感谢志愿者范琦、王超 、朱婷对本文进行了细致的整理工作。

作者信息：

乔宇，中科院深圳先进技术研究院研究员，集成所所长副所长，博士生导师。入选中国科学院“百人计划”，深圳市“孔雀计划”海外高层次人才，广东省引进创新团队的核心成员。研究兴趣包括计算机视觉、深度学习、机器人等。已在包括IEEE T-PAMI，IJCV， IEEE Trans. on Image Processing， IEEE Trans. on Signal Processing， CVPR，ICCV， ECCV，AAAI等会议和期刊上发表学术论文110余篇。获卢嘉锡人才奖。带领团队多次在ChaLearn，LSun，THUMOUS，ACTIVITYNet等国际评测中取得第一，获ImageNet 2016场景分类任务第二名。主持国家重大研究计划子课题，国家自然科学基金重点、中国科学院国际合作重点，粤港合作，深圳市基金研究“杰青”、日本学术振兴会等资助的多个项目。

VALSE是视觉与学习青年学者研讨会的缩写，该研讨会致力于为计算机视觉、图像处理、模式识别与机器学习研究领域内的中国青年学者提供一个深层次学术交流的舞台。2017年4月底，VALSE2017在厦门圆满落幕，近期大讲堂将连续推出VALSE2017特刊。VALSE公众号为：VALSE，欢迎关注。

往期精彩回顾

VALSE2017系列之四：目标跟踪领域进展报告 (/html/580/201705/2650326508/1.html)

深度学习大讲堂改版纪念：一个陌生女人的来信 (/html/580/201705/2650326459/1.html)

VALSE2017系列之三：人体姿态识别领域年度进展报告 (/html/580/201705/2650326421/1.html)

VALSE2017系列之二: 边缘检测年度进展概述 (/html/580/201705/2650326339/1.html)

行人检测、跟踪、与检索领域年度进展报告 (/html/580/201705/2650326280/1.html)

欢迎关注我们！

深度学习大讲堂是由中科视拓运营的高质量原创内容平台，邀请学术界、工业界一线专家撰稿，致力于推送人工智能与深度学习最新技术、产品和活动信息！

中科视拓（SeetaTech）将秉持“开源开放共发展”的合作思路，为企业客户提供人脸识别、计算机视觉与机器学习领域“企业研究院式”的技术、人才和知识服务，帮助企业在人工智能时代获得可自主迭代和自我学习的人工智能研发和创新能力。

中科视拓目前正在招聘： 人脸识别算法研究员，深度学习算法工程师，GPU研发工程师， C++研发工程师，Python研发工程师，嵌入式视觉研发工程师，运营经理。有兴趣可以发邮件至：hr@seetatech.com，想了解更多可以访问，www.seetatech.com

中科视拓

深度学习大讲堂

点击阅读原文打开中科视拓官方网站

分享:

阅读 567 3

登录

来说两句吧...

还没有评论，快来抢沙发吧！

深度学习大讲堂 更多文章

- VALSE2017系列之四：目标跟踪领域进展报告 (/html/580/201705/2650326508/1.html)
- 深度学习大讲堂改版纪念：一个陌生女人的来信 (/html/580/201705/2650326459/1.html)
- VALSE2017系列之三：人体姿态识别领域年度进展报告 (/html/580/201705/2650326421/1.html)
- VALSE2017系列之二: 边缘检测领域年度进展报告 (/html/580/201705/2650326339/1.html)
- 行人检测、跟踪与检索领域年度进展报告 (/html/580/201705/2650326280/1.html)

猜您喜欢