1st VALSE Workshop on Methods and Technologies for Looking At People (MATLAP)

面向人体姿态行为理解的深度学习方法

Deeply Understanding Human Poses and Actions in the Wild

乔宇

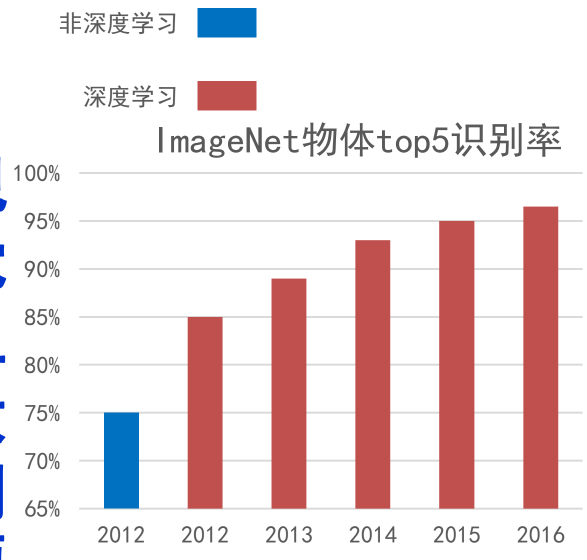
中国科学院深圳先进技术研究院

2018年4月22日

深度学习方法推动计算机视觉技术快速发展

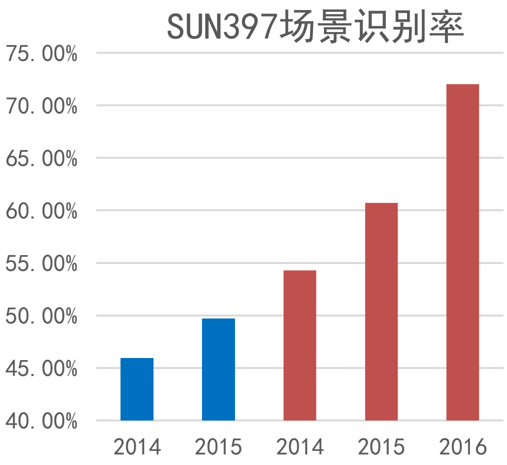
深度神经网络已被成功用于物体识别、场景分类、行为识别等视觉核心任务，极大地推动了计算机视觉的发展

视觉三大问题



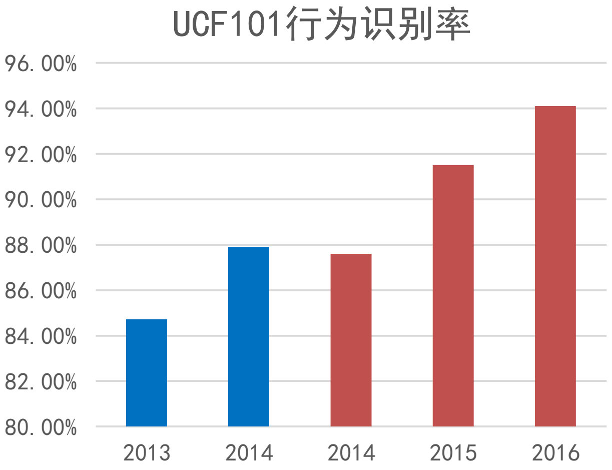
物体识别

回答“有什么”



场景分类

回答了“在哪里”



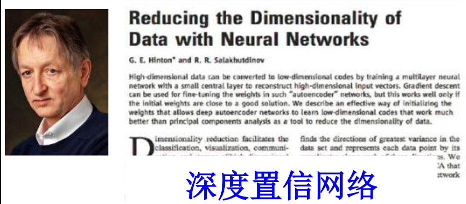
行为姿态识别

回答“干什么”

深度学习方法的兴起

深度神经网络已经在语音、视觉、自然语言处理等领域取得了巨大的成功，在学术界和工业界都引起了极大关注。

深度学习理论突破

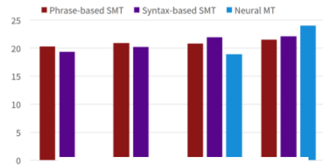


深度置信网络

ImageNet竞赛: 74% vs. 85%

1000类，1百万数据

机器翻译和文本生成



Machine Translation

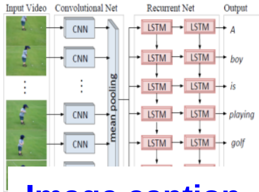
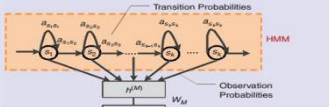


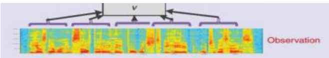
Image caption

2006 2011 2012 2013 2015 2016

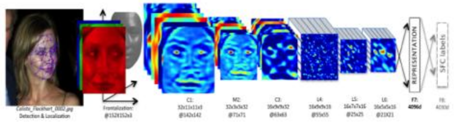
大规模语音识别



Switchboard：错误率降低8.9%



人脸识别



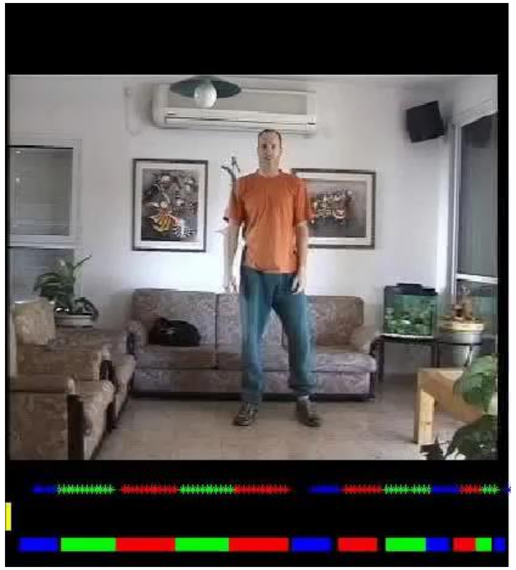
在LFW上识别率99%,超过人类

围棋

Alphago 4:1 李世乭

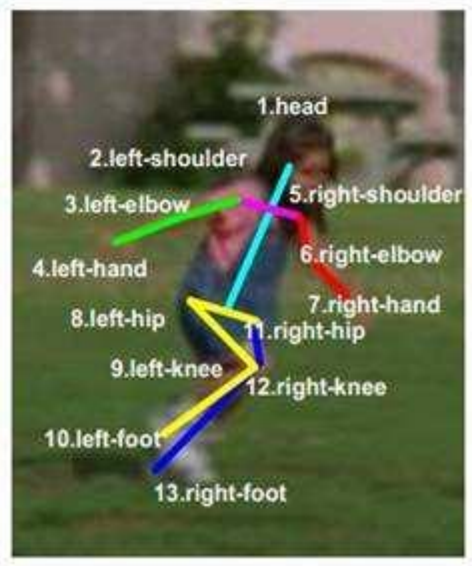
行为和姿态识别

- The goal of human action recognition is to automatically detect and classify ongoing activities from an input video (i.e. a sequence of images frames).
  - Human vision system is very effective in perceiving and predicting actions through visual information.
  - A basic problem in computer vision, with wide applications.



Punch
Kick
Duck
…

Action recognition



Pose estimation

应用：互联网和监控视频理解

互联网
图像
视频

facebook

WeChat 微信

2500亿张照片，视频日播放**80亿**次

日上传图片10亿张，视频播放**20亿**次

监控
视频

深圳、广州等大型城市，市内监控探头总数超40万。
以720p计算，**每秒产生数据>4T**。

谭铁牛院士

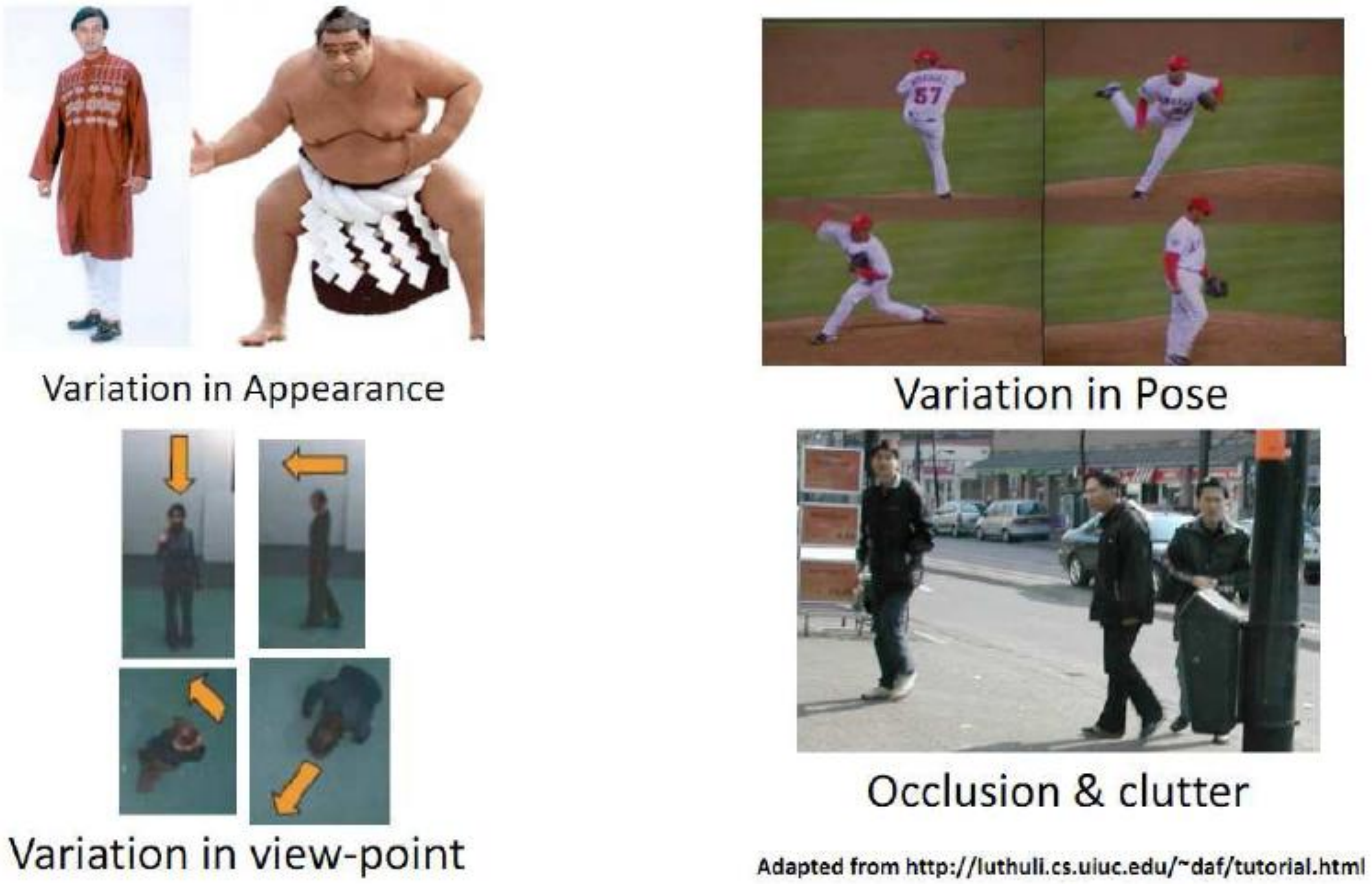
“图像视频大数据是人工智能的突破口，
是信息产业新的增长点。”

更多应用


YouTube
YAHOO!
msn
search
surveillance
sport analysis
Human
Motion
Analysis
synthesis
biomechanics
motion capture
game interfaces
sign
language

挑战–数据

- High dimension
- Large variation and complexity

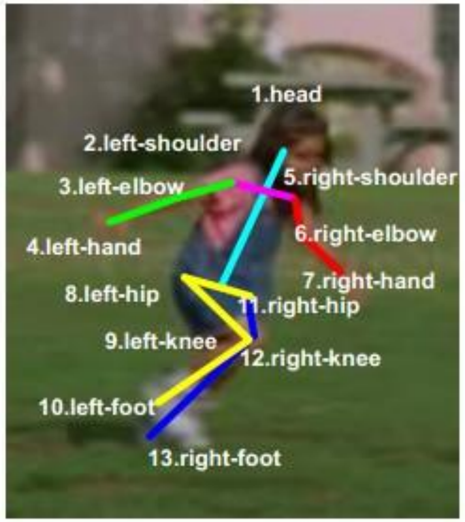


挑战–标定

- Huge number of categories

From UCF101

Bounding box



Pose

行为数据库

KTH

UCF Sports

Hollywood

Olympic

Action recognition "in the lab": KTH, Weizmann etc.
Action recognition "in TV, Moive": UCF Sports, Hollywood etc.
Action recognition "in the wild": Olympic, HMDB51, UCF101 etc.

Tal Hassner A Critical Review of Action Recognition Benchmarks CVPR'12

视频样例

KTH
In the lab

KTH-Boxing

KTH-Jogging

Movie

Hollywood-Talking

Hollywood-Driving

Sports

Basketball

Bowling

视频样例

HMDB51

chewing

Brushing hair

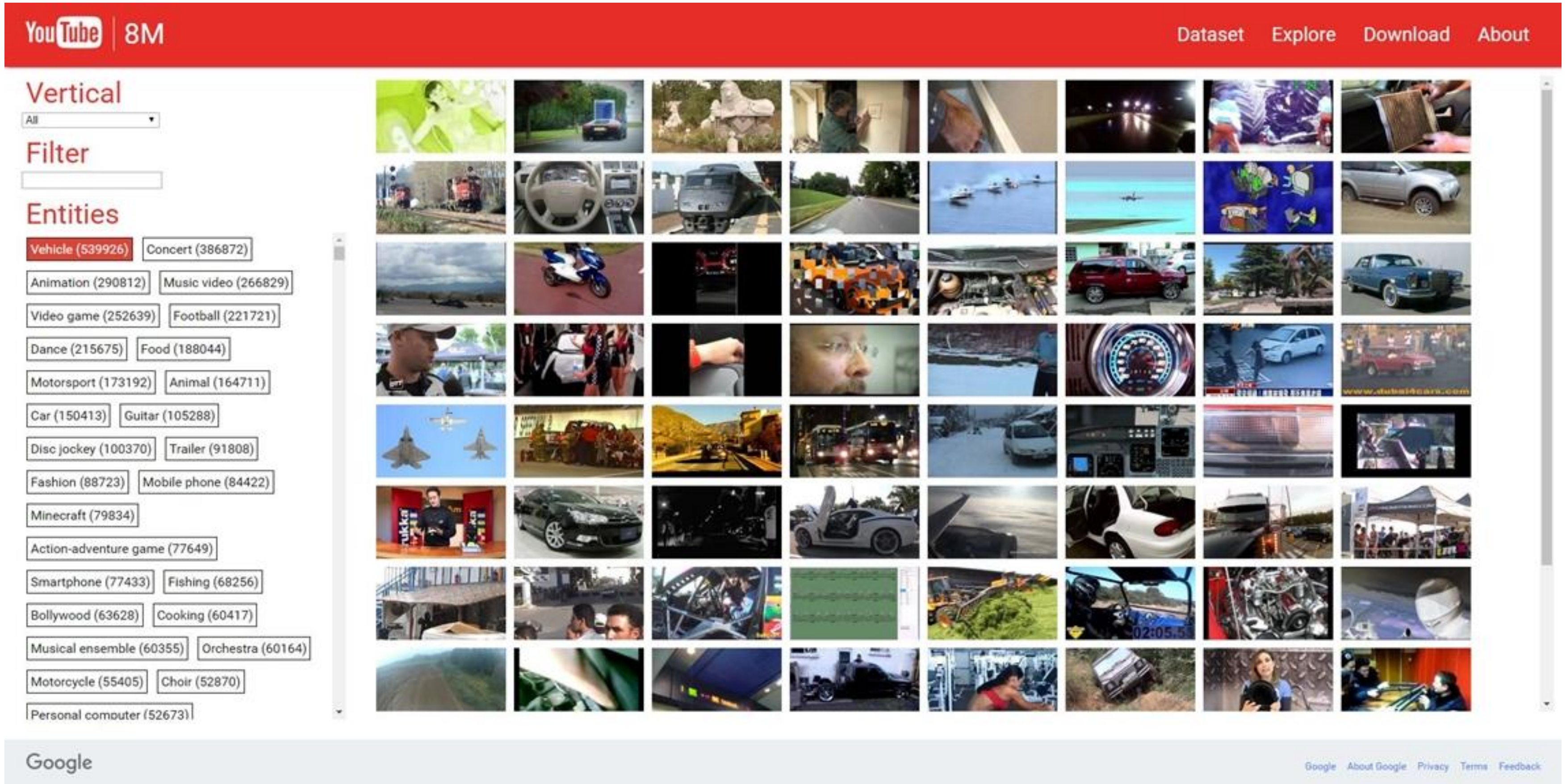
UCF101

Eye makeup

Baby crawling

Youtube-8M 2017

https://research.google.com/youtube8m/



7 Million	450,000	3.2 Billion	4716	3.4
Video URLs	Hours of Video	Audio/Visual Features	Classes	Avg. Labels / Video

更多大规模视频数据库

Kinetics

- 306,245 videos in total
- 400 action classes
- Each clip lasts around 10s

Moments in Time Dataset

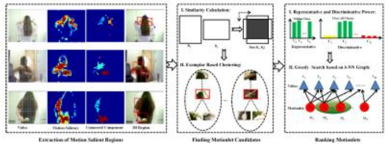
- over 1,000,000 videos
- 339 Moment classes
- 3-second video

AVA Dataset

- 80 atomic actions
- 192 clips (15 mins per clip)
- 740k annotations

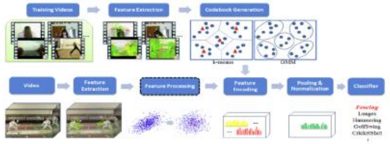
课题组开展的工作

从视频行为理解和识别是计算机视觉的基本和热点问题，在监控、互联网等有着广泛的应用。在CVPR，ICCV，IJCV,TIP等重要视觉会议和期刊发表了20多篇论文，其中2篇论文分别被ICCV和CVPR录用为Oral。



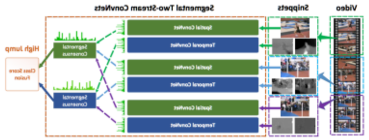
视频中层表示和结构模型

CVPR 2013，139引用
ICCV 2013，59引用
ECCV 2014，72引用
IJCV 2016



视频特征编码学习

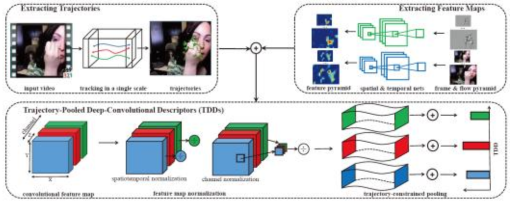
CVPR 2014 Oral， 92引用
ECCV 2014，211引用
CVIU 2016，292引用
非深度学习领域UCF和HMDB性能最好方法



时序分割模型TSN

ECCV 2016，260引用
解决显存限制下，视频端到端的训练
ActivityNet 2016竞赛第一

2013 2014 2015 2016 2017

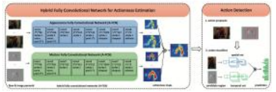


轨迹卷积特征TPD

CVPR 15， 430引用
第一个在UCF和HMDB数据库上全面超越传统方法的的深度模型

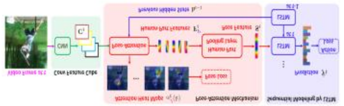
运动向量网络EMV-CNN

CVPR 2016，80引用
深度实时行为识别模型



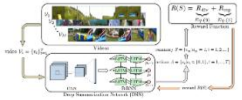
行为显著性估计

CVPR 2016，67引用
行为检测与显著性估计



姿态递归注意网络RPAN

ICCV2017 Oral
利用姿态估计引导动态行为序列建模



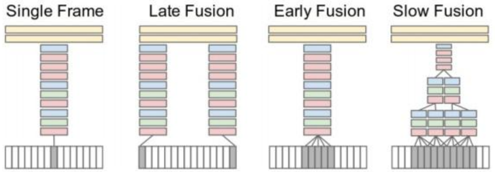
非监督强化视频概要

AAAI 2018 Oral

早期视频行为识别DL方法

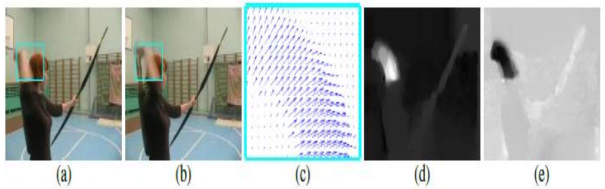
Spatial Temporal CNN

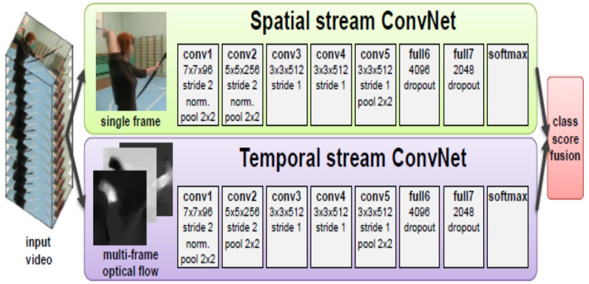
Sports-1M Dataset



[Karpathy et al., CVPR, 2014]

Two Stream-CNN





[Karen NIPS, 2014]

C3D: 3D VGGNet

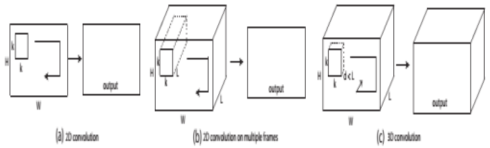
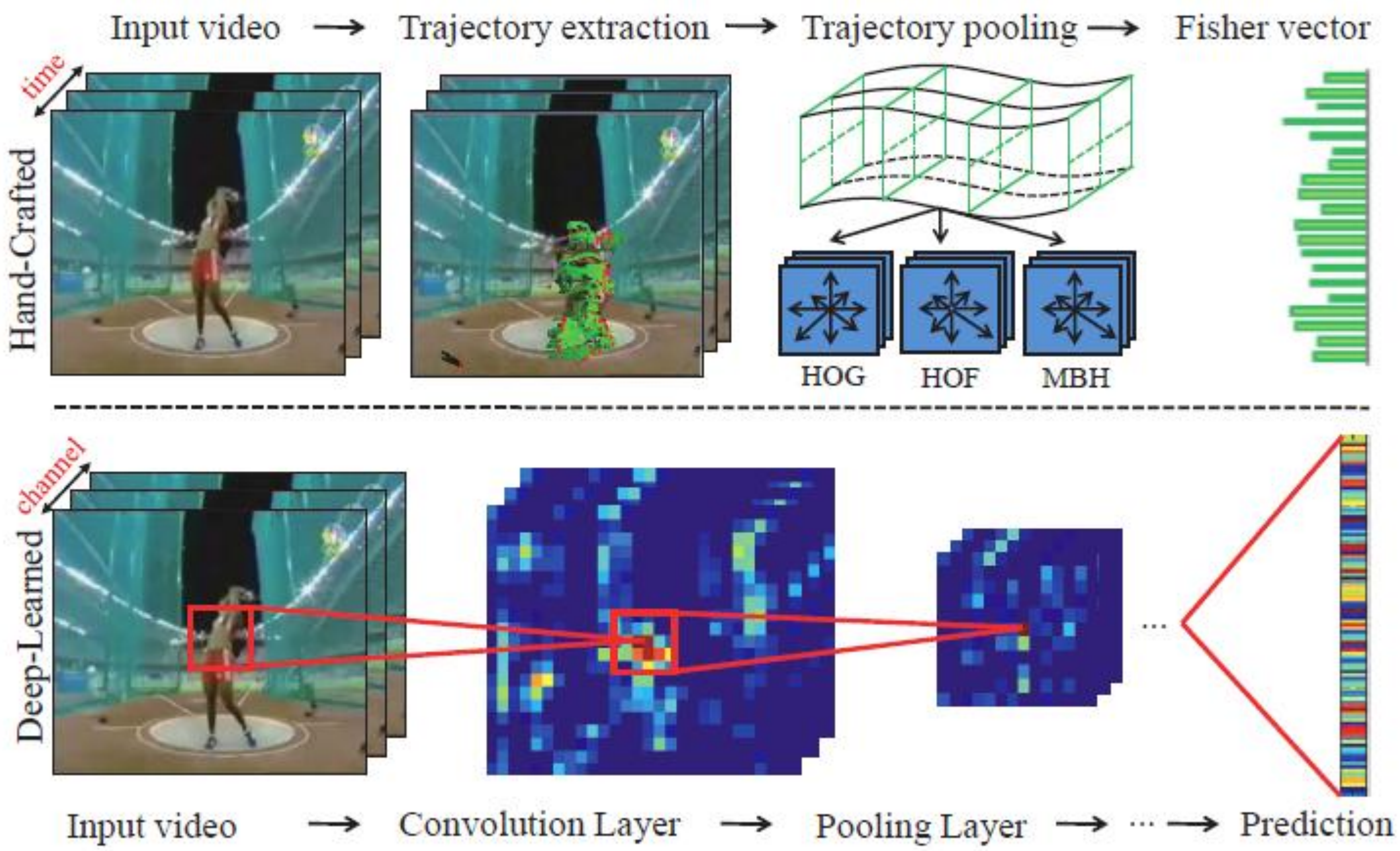


Figure 3. **C3D architecture**. C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling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

[Tran et al. CVPR 2015]

在UCF101的表现并没有明显好于非传统方法

工作1：轨迹池化卷积特征TDD(CVPR15)



如何利用传统方法与深度学习方法的优势。

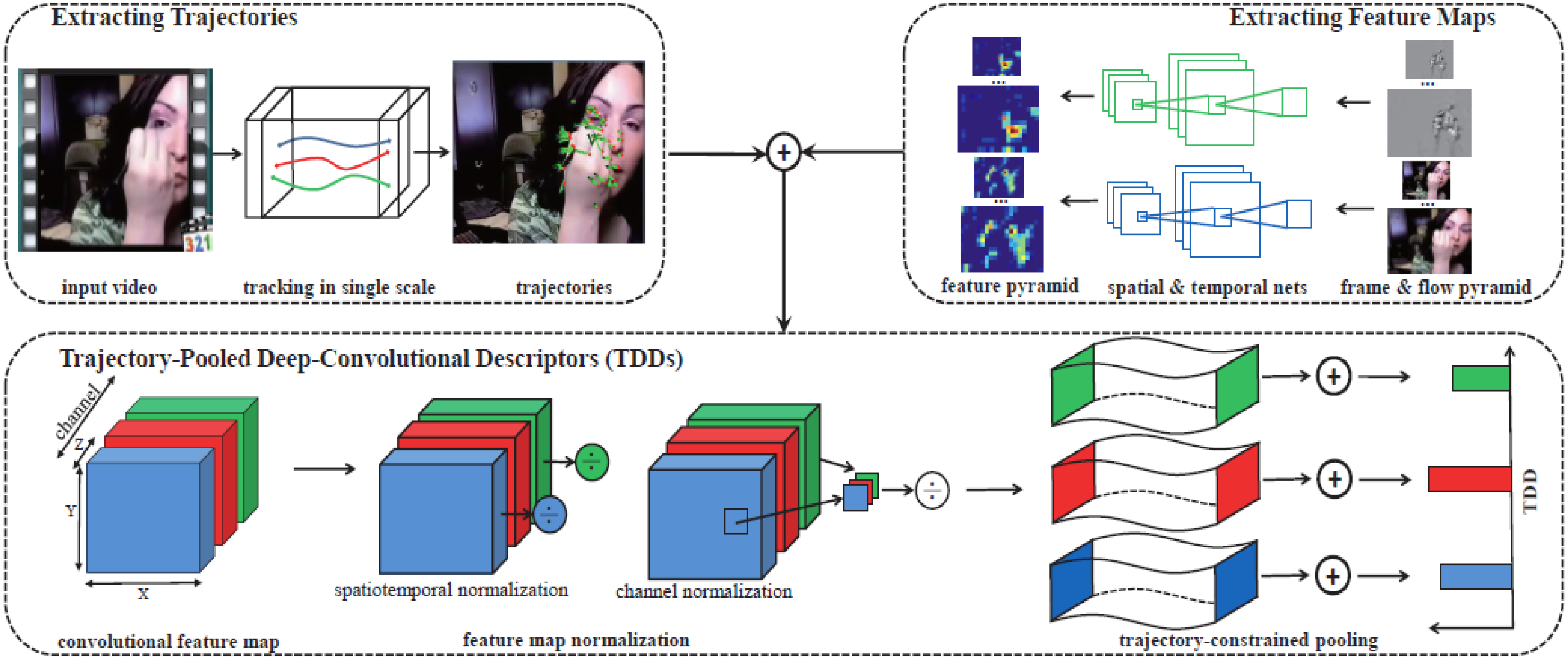
Limin Wang, Yu Qiao, Xiaoou Tang “Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors ”, Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR), 2015 (430引用)

TDD的框架

Trajectory-pooled deep convolutional descriptor (TDD) 特征结合了传统方法的轨迹跟踪和深度学习方法的卷积特征提取。

提取运动轨迹（传统方向）

提取卷积特征（深度学习方法）



沿着运动轨迹对特征进行编码

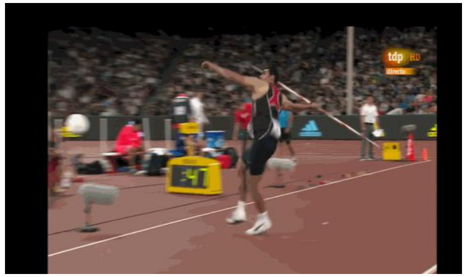
TDD的性能

第一个在UCF和HMDB上全面超越传统浅层模型的深度学习方法。

Algorithm	HMDB51	UCF101
HOG [31, 32]	40.2%	72.4%
HOF [31, 32]	48.9%	76.0%
MBH [31, 32]	52.1%	80.8%
HOF+MBH [31, 32]	54.7%	82.2%
iDT [31, 32]	**57.2%**	**84.7%**
Spatial net [24]	40.5%	73.0%
Temporal net [24]	54.6%	83.7%
Two-stream ConvNets [24]	**59.4%**	**88.0%**
Spatial conv4	48.5%	81.9%
Spatial conv5	47.2%	80.9%
Spatial conv4 and conv5	**50.0%**	**82.8%**
Temporal conv3	54.5%	81.7%
Temporal conv4	51.2%	80.1%
Temporal conv3 and conv4	**54.9%**	**82.2%**
TDD	63.2%	90.3%
TDD and iDT	**65.9%**	**91.5%**

工作2：深度时序分割模型TSN (ECCV 16)

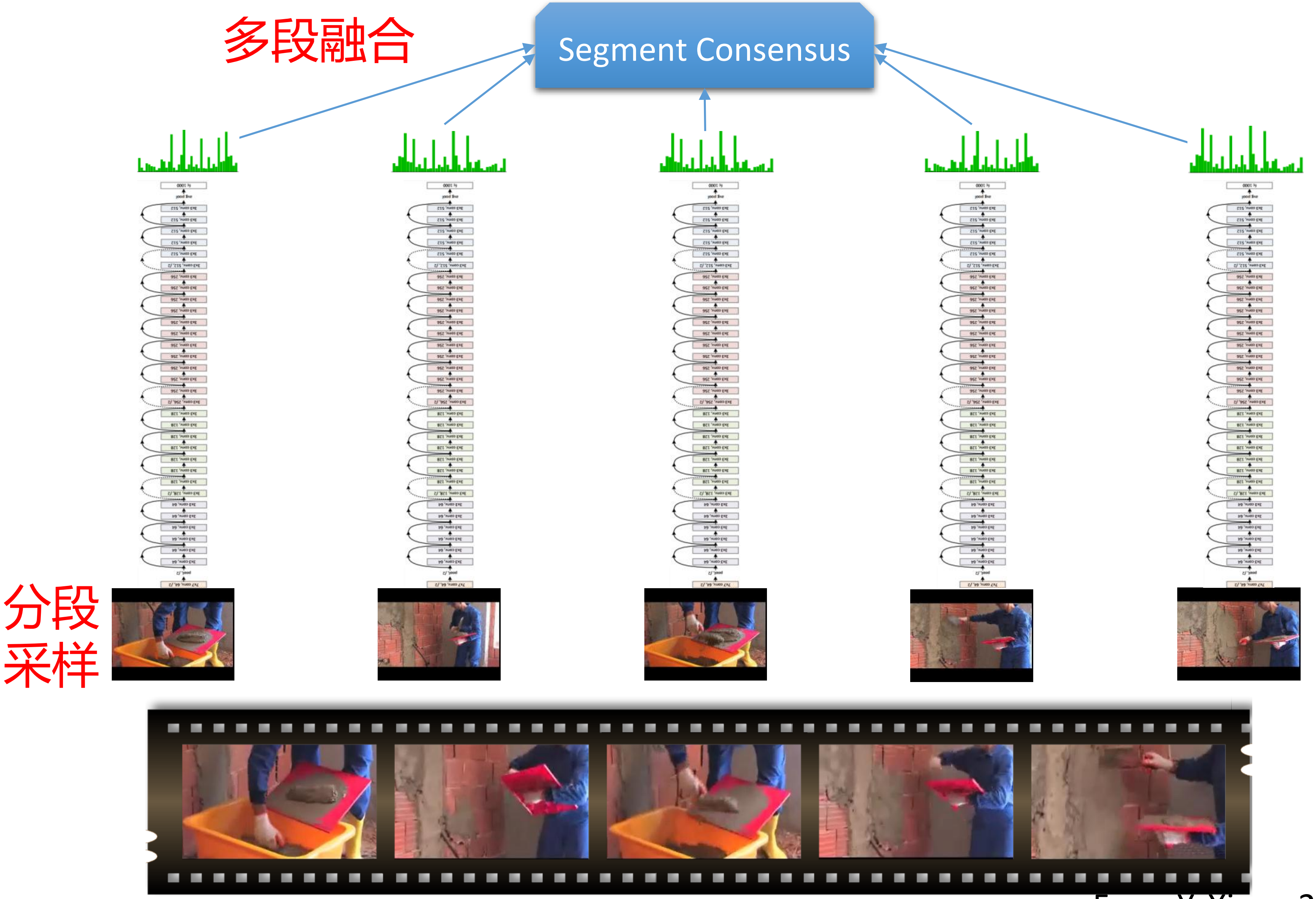
如何对视频序列进行建模和深度学习？



核心问题：视频的数据量大，特征维度很高，但深度学习的训练受制于显存和SGD算法。

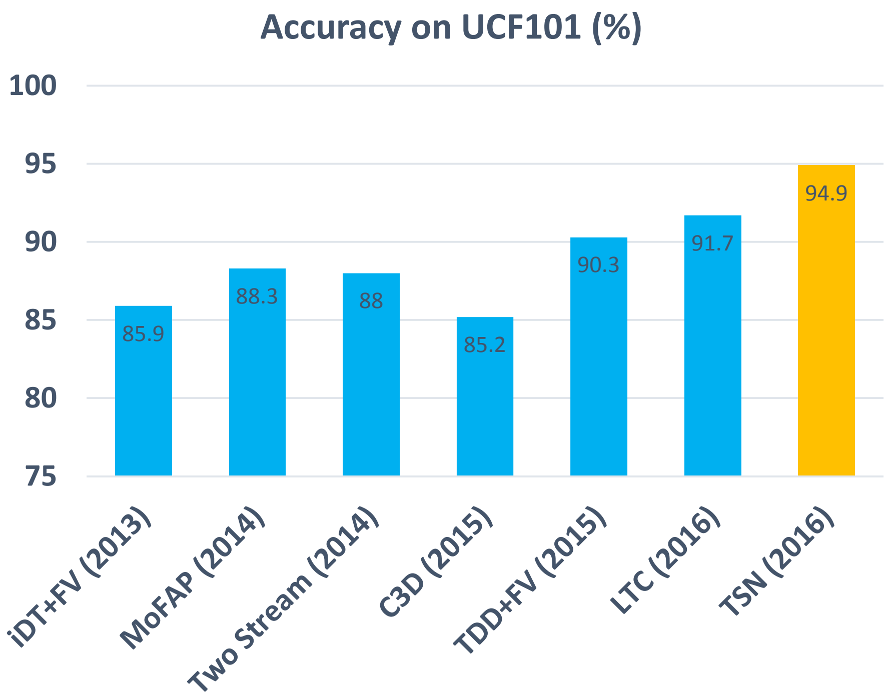
Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao et al, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," Proc. European Conference Computer Vision (ECCV), 2016 (260引用)

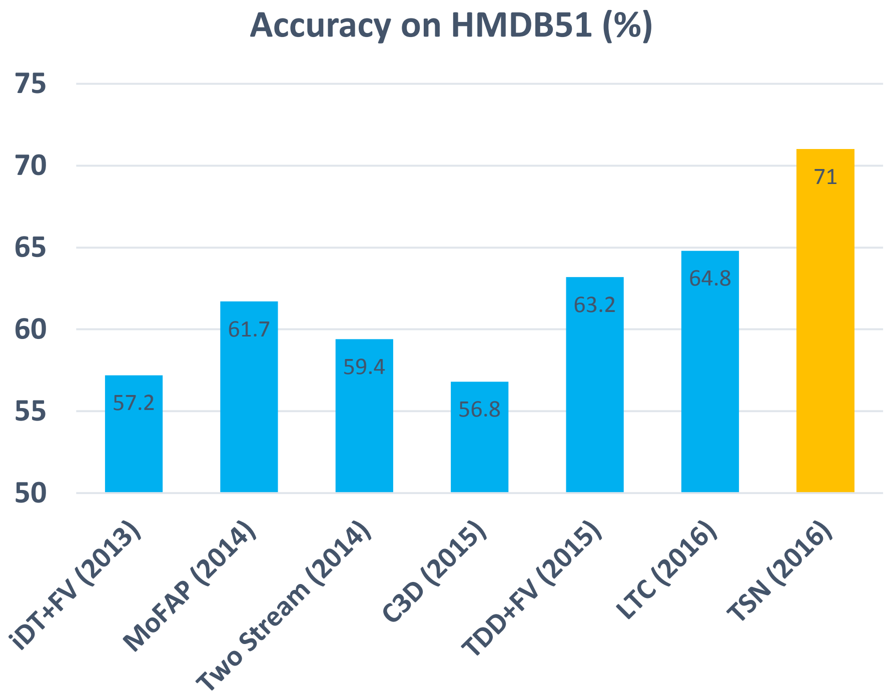
TSN框架



From Y. Xiong 2016

TSN的性能评价


Accuracy on UCF101 (%)
100
95
90
85
80
75
85.9
88.3
88
85.2
90.3
91.7
94.9
iDT+FV (2013)
MoFAP (2014)
Two Stream (2014)
C3D (2015)
TDD+FV (2015)
LTC (2016)
TSN (2016)


Accuracy on HMDB51 (%)
75
70
65
60
55
50
57.2
61.7
59.4
56.8
63.2
64.8
71
iDT+FV (2013)
MoFAP (2014)
Two Stream (2014)
C3D (2015)
TDD+FV (2015)
LTC (2016)
TSN (2016)

ActivityNet 2016

200个类别，648小时视频，10k训练 , 5k测试

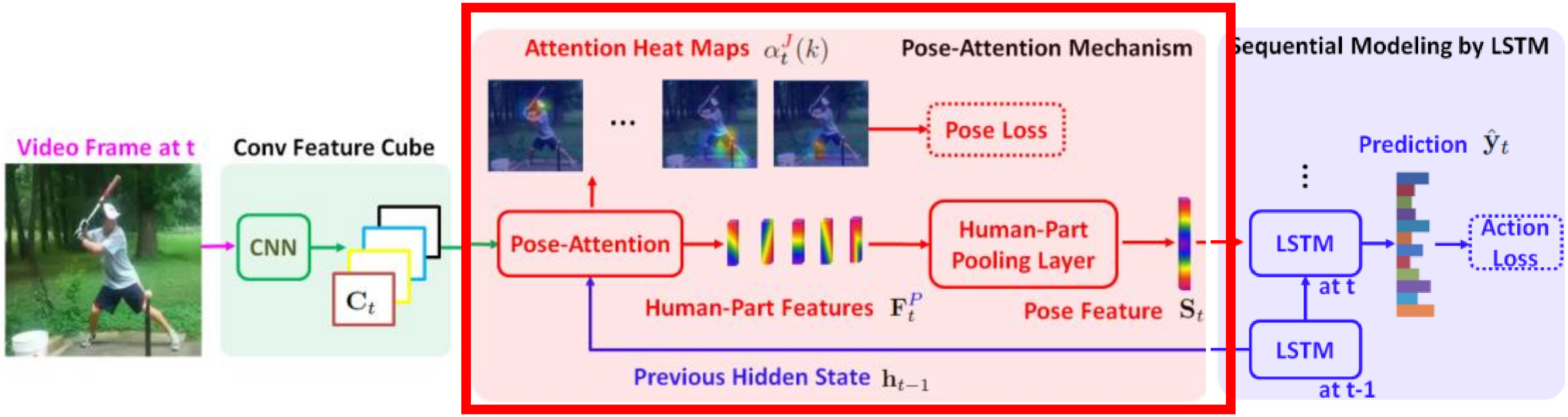
http://activity-net.org/challenges/2016/

在24个队中排名第一。

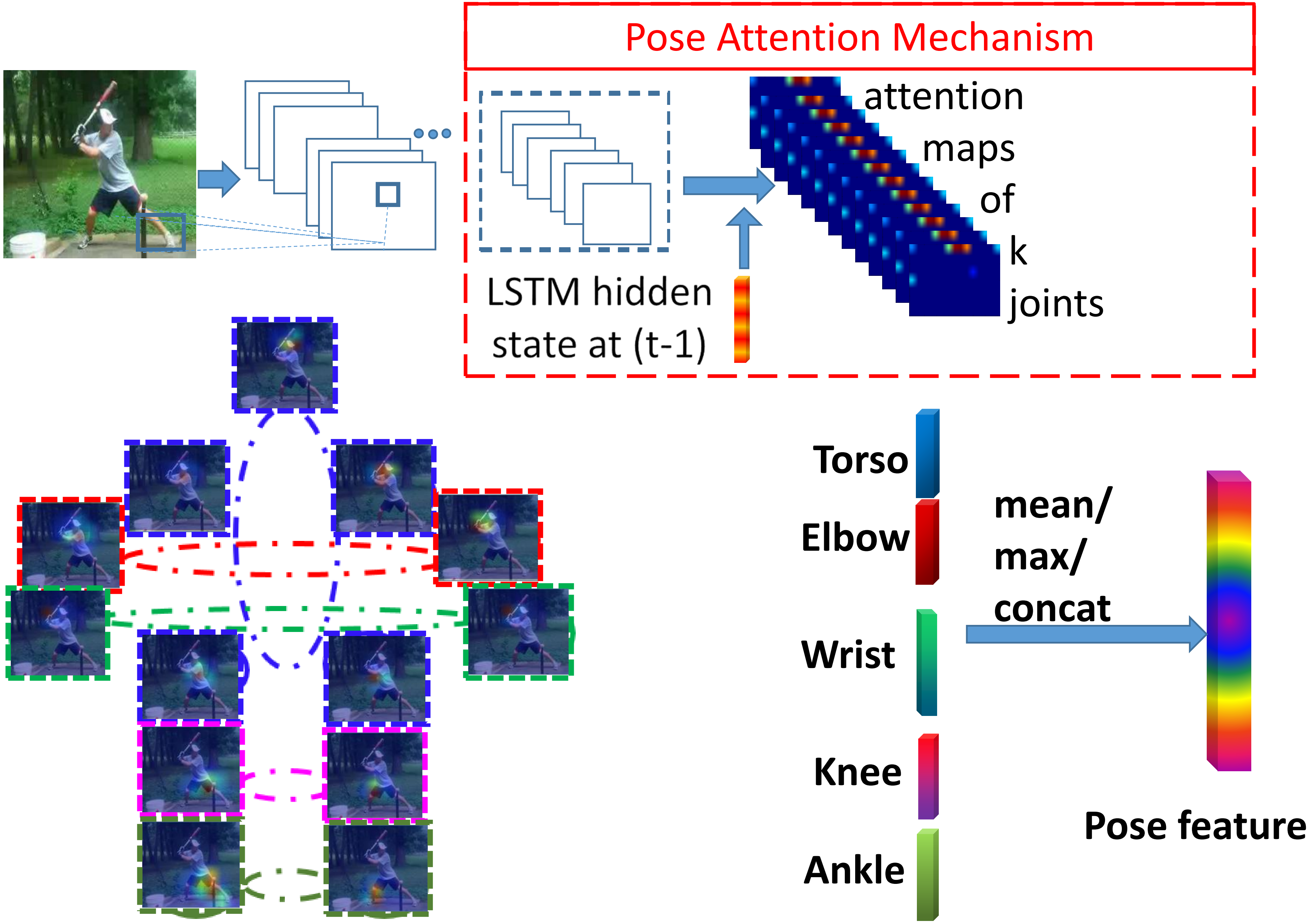
Validation Set	mAp	Top-3 Acc.
Visual	90.4%	95.2%
Audio	15.2%	29.1%
Visual + Audio	90.9%	95.6%
Testing Set	mAP	Top-3 Acc.
Visual CNN (Single)	91.2%	95.6%
Final Ensemble	93.2%	96.4%

工作3：递归姿态注意网络RPAN（ICCV17 Oral）

提出姿态注意机制RPAN对行为的动态过程进行建模。



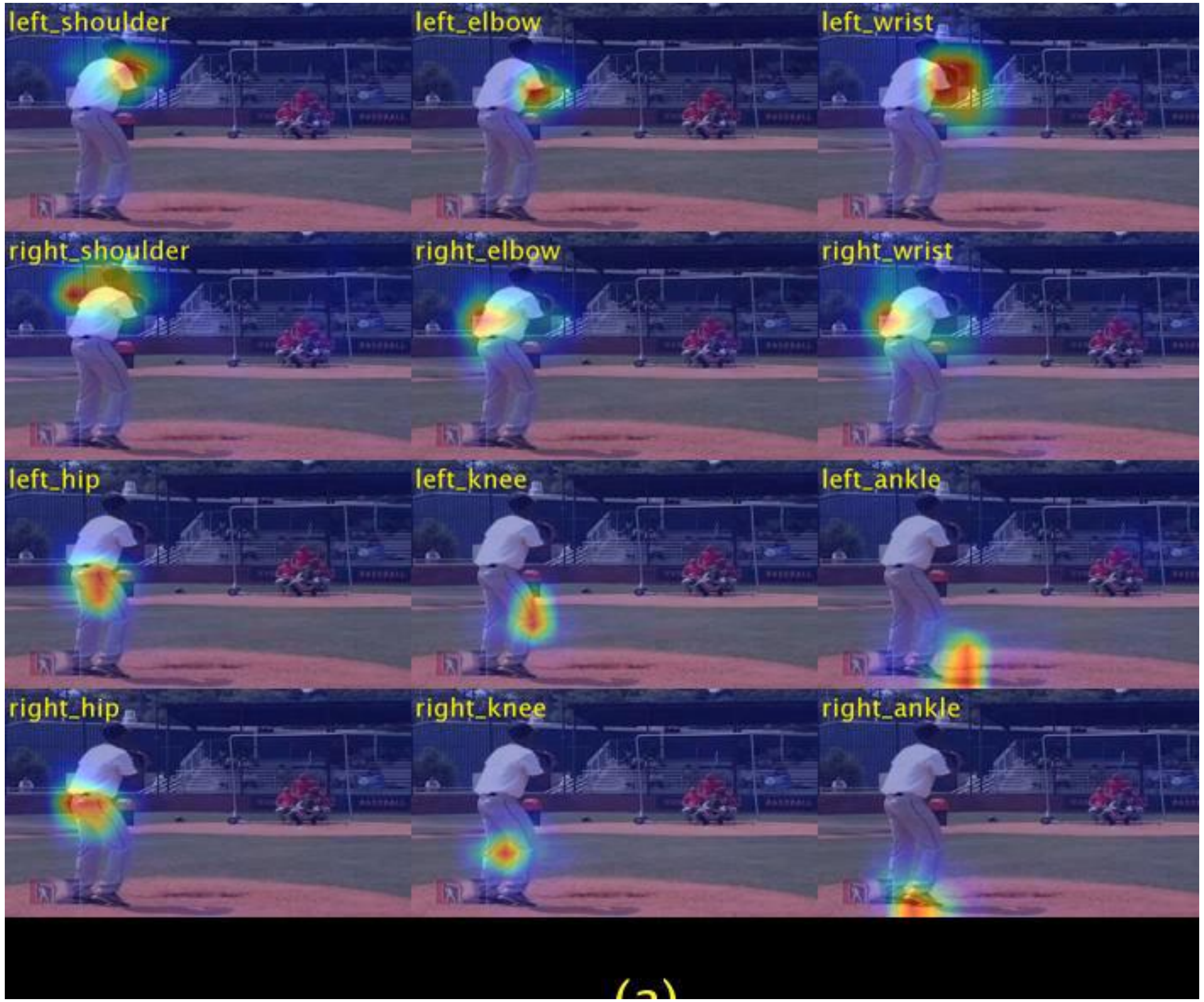
- 把行为识别和姿态估计两个任务进行结合。
- 利用姿态的变化，引导递归神经网络对行为的动态过程进行建模。


Pose Attention Mechanism
attention maps of k joints
LSTM hidden state at (t-1)
Torso
Elbow
Wrist
Knee
Ankle
mean/ max/ concat
Pose feature

RPAN的性能

State-of-the-art	**Authors**	Year	Sub-JHMDB	PennAction
Dense+Pose	H. Jhuang, et al	2013	52.9	-
STIP	W. Zhang, et al	2013	-	82.9
Action Bank	W. Zhang, et al	2013	-	83.9
MST	J. Wang, et al	2014	45.3	74.0
AOG	B. X. Nie, et al	2015	61.2	85.5
P-CNN	G. Cheron et al	2015	66.8	-
Hierarchical	I. Lillo et al	2016	77.5	-
C3D	C. Cao, et al	2016	-	86.0
JDD	C. Cao, et al	2016	77.7	87.4
idt-fv	U. Iqbal et al	2017	60.9	92.0
Pose+ idt-fv	U. Iqbal et al	2017	74.6	92.9
Our RPAN			**78.6**	**97.4**

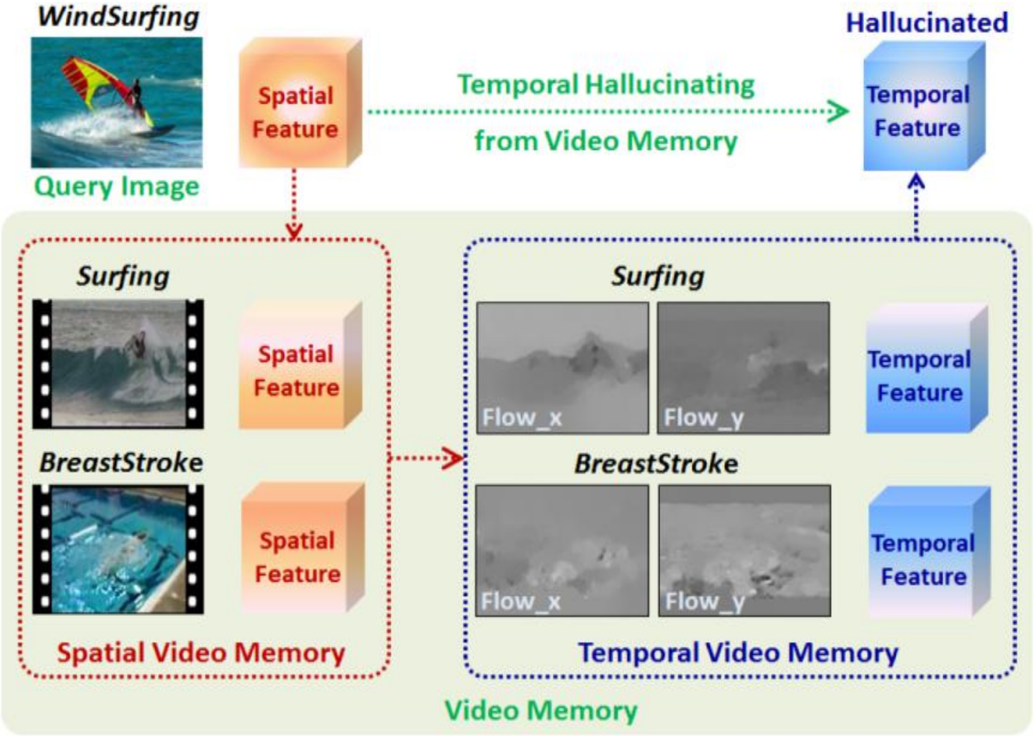
RPAN用于姿态估计


left_shoulder
left_elbow
left_wrist
right_shoulder
right_elbow
right_wrist
left_hip
left_knee
left_ankle
right_hip
right_knee
right_ankle

工作4：从静态图像中估计运动（CVPR18）

Humans can classify new action categories after **seeing few images**:

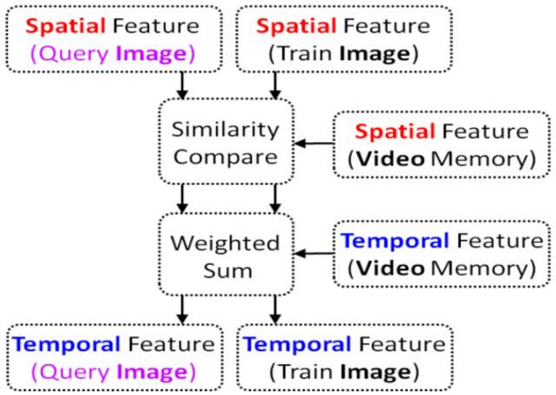
1. Comparing appearance similarities between images on hand
2. Recalling importance motion cues from relevant action videos in memory



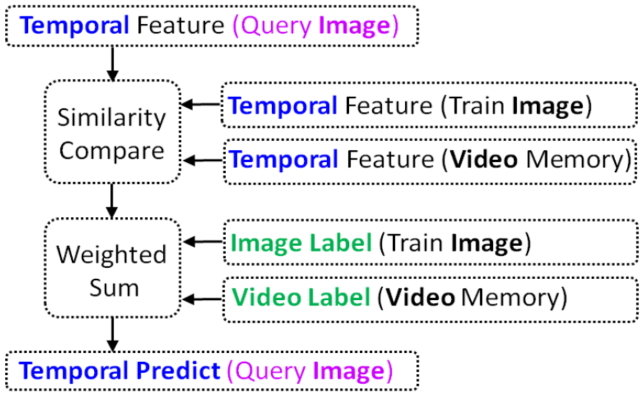
Lei Zhou Y Wang, Yu Qiao., Temporal Hallucinating for Action Recognition with Few Still Images, CVPR 2018

Hybrid Video Memory (HVM) Machine

- **Temporal Memory Module**
  1. Hallucinating temporal features for still images
  2. Predicting action labels by hallucinated features
- **Spatial Memory Module**
  1. Predicting action labels by spatial features
  2. Spatial and temporal prediction are complementary
- **Memory Selection Module**: most-relevant videos as memory



Temporal Hallucinating

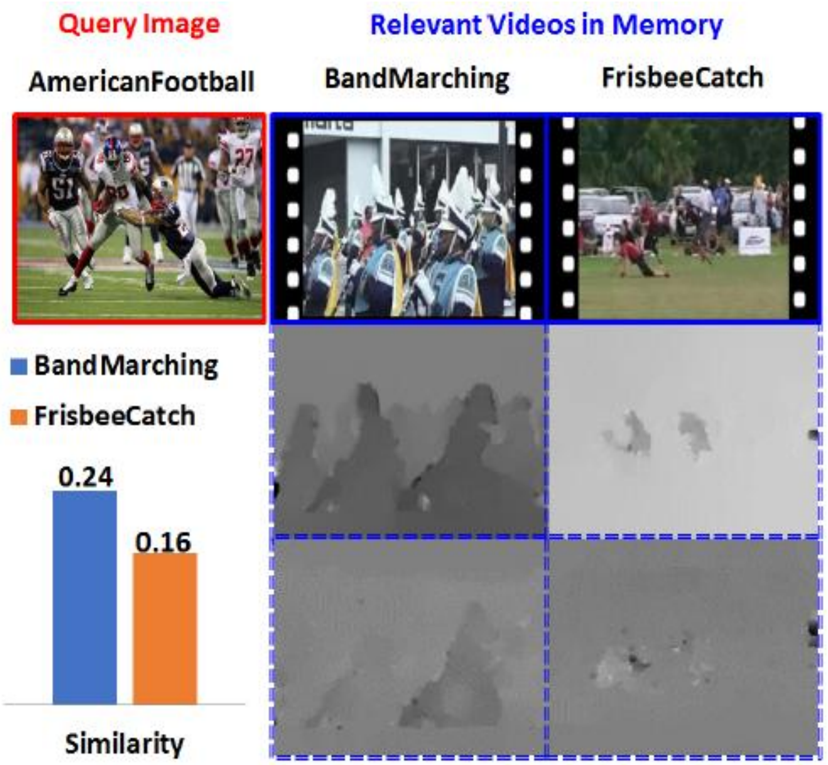


Temporal Predicting

实验结果

1-training-image per category

Approaches	WEB101	VOC	DIFF20
KNN	26.1	38.3	55.7
SVM	22.3	32.0	54.2
TGPN [36]	15.5	30.5	35.2
TSN [37]	26.1	40.3	56.3
R*CNN [8]	n/a	28.3	n/a
KV-MemNNs [21]	24.4	39.5	52.1
Matching Network [34]	26.6	39.9	56.7
Our HVM	**35.4**	**42.2**	**60.2**



模型和代码公开

场景理解与分类

•MR-CNNs (2nd in scene classification task ImageNet 2016, 1st in LSUN 2016)

•Weakly Supervised PatchNets (Top performance in MIT Indoor67 and SUN397)

行为识别和检测

•Temporal Segment Networks (NO1 in ActivityNet 2016)

•MV-CNNS(Speed:300帧/s)

•Trajectory-Pooled Deep-Convolutional Descriptors （Top performance in UCF101 and HMDB51)

人脸检测与识别

•MJ-CNN face detection (top performance in FDDB & WIDE)

•HFA-CNN face recognition (single model 99% in LFW)

场景文字检测与识别

•Connectionist Text Proposal Network for Scene Text Detection (Top performance in ICDAR)

下载地址

http://mmlab.siat.ac.cn/yuqiao/Codes.html

谢谢！

Q&A

敬请批评指正