UNIVERSITY OF CALIFORNIA
UC MERCED

Efficient Large-Scale Video Understanding in The Wild

Yi Zhu and Shawn Newsam yzhu25, snewsam@ucmerced.edu

ICIP 2017
IEEE International Conference on Image Processing
September 17-20, 2017, Beijing, China

Motivation:

- Enormous explosion of user-generated videos, containing a wealth of information. However, it would take forever to manually annotate the data and make use of them.
- Need for video-based applications, like video search, video highlighting, video surveillance, etc. Recent trending topics in computer vision include video action/event recognition.

Objective:

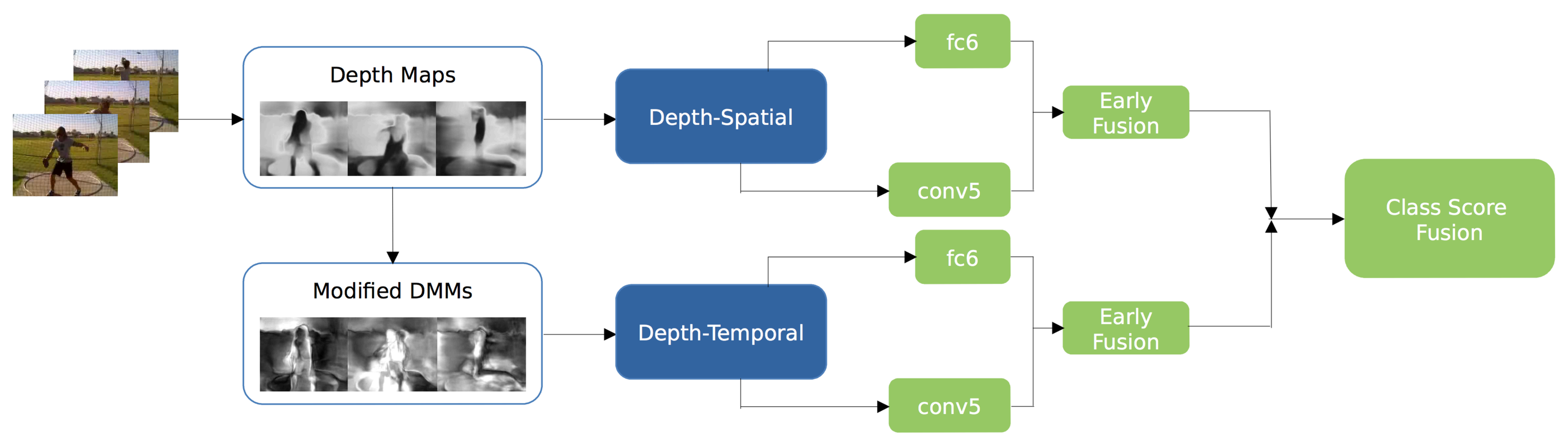
Efficient large-scale video understanding in the wild. Specifically,

- Better encoding of static frame appearance information (image classification)
- Better description of short term motion between adjacent frames (action recognition)
- Better exploration of temporal structure and extraction of video/clip-level features instead of frame-level features (event recognition)
- Better fusion of different channels of information (multi-modal learning)
- Better identification of spatial patterns geographically according to video metadata information like geo-coordinates. (smart city)

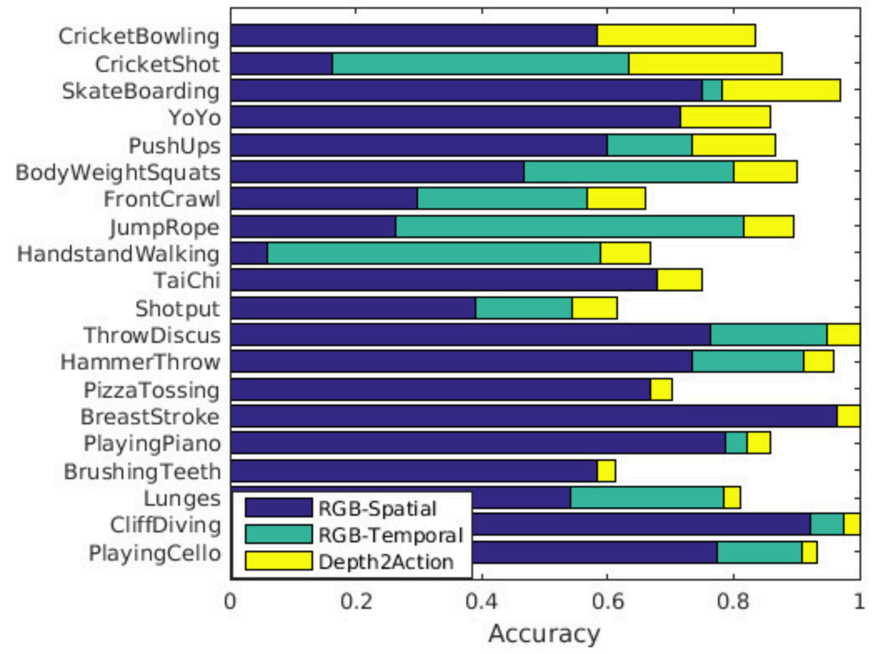
Past work:

1. Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition (ECCV 2016)

- This paper performs the first investigation into depth for large-scale human action recognition in video where the depth cues are estimated from the videos themselves



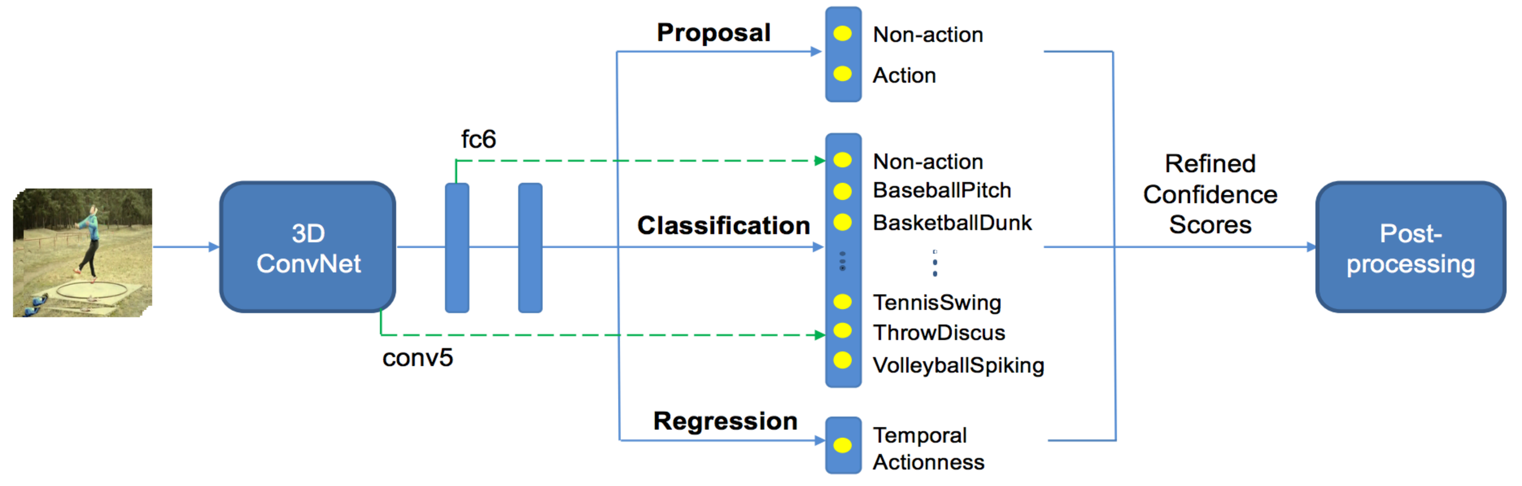
- Depth information is complementary to other channels, like static appearance and motion information.



Algorithm	UCF101	Algorithm	HMDB51	Algorithm	ActivityNet
Srivastava et al. [35]	84.3 %	Srivastava et al. [35]	44.1 %	Wang and Schmid [39]	61.3 %*
Wang and Schmid [39]	85.9 %	Oneata et al. [25]	54.8 %	Simonyan and Zisserman [31]	67.1 %*
Simonyan and Zisserman [31]	88.0 %	Wang and Schmid [39]	57.2 %	Tran et al. [37]	69.4 %*
Jain et al. [9]	88.5 %	Simonyan and Zisserman [31]	59.1 %		
Ng et al. [24]	88.6 %	Sun et al. [36]	59.1 %		
Lan et al. [17]	89.1 %	Jain et al. [9]	61.4 %		
Zha et al. [52]	89.6 %	Fernando et al. [6]	63.7 %		
Tran et al. [37]	90.4 %	Lan et al. [17]	65.1 %		
Wu et al. [47]	91.3 %	Wang et al. [43]	65.9 %		
Wang et al. [43]	91.5 %	Peng et al. [28]	66.8 %		
Depth2Action	72.5 %	Depth2Action	49.7 %	Depth2Action	52.1 %
+Two-Stream	92.0 %	+Two-Stream	67.1 %	+C3D	71.2 %
+IDT+Two-Stream	**93.0 %**	+IDT+Two-Stream	**68.2 %**	+IDT+C3D	**73.4 %**

2. Efficient Action Detection in Untrimmed Videos via Multi-Task Learning (WACV 2017)

- This paper studies a multi-task learning framework that performs the three highly related steps of action proposal, action recognition, and action localization refinement in parallel instead of the standard sequential pipeline that performs the steps in order



- Our parallel model is more robust than its sequential counterpart when limited training data is available.

Training Set	$\alpha = 0.2$	$\alpha = 0.5$
$V_T + V_U$	43.5	19.0
$\frac{3}{4}V_T + V_U$	41.7	17.9
$\frac{1}{2}V_T + V_U$	36.9	14.4

(a) Sequential Network [28]

Training Set	$\alpha = 0.2$	$\alpha = 0.5$
$V_T + V_U$	43.6	19.2
$\frac{3}{4}V_T + V_U$	42.9	18.7
$\frac{1}{2}V_T + V_U$	39.4	17.3

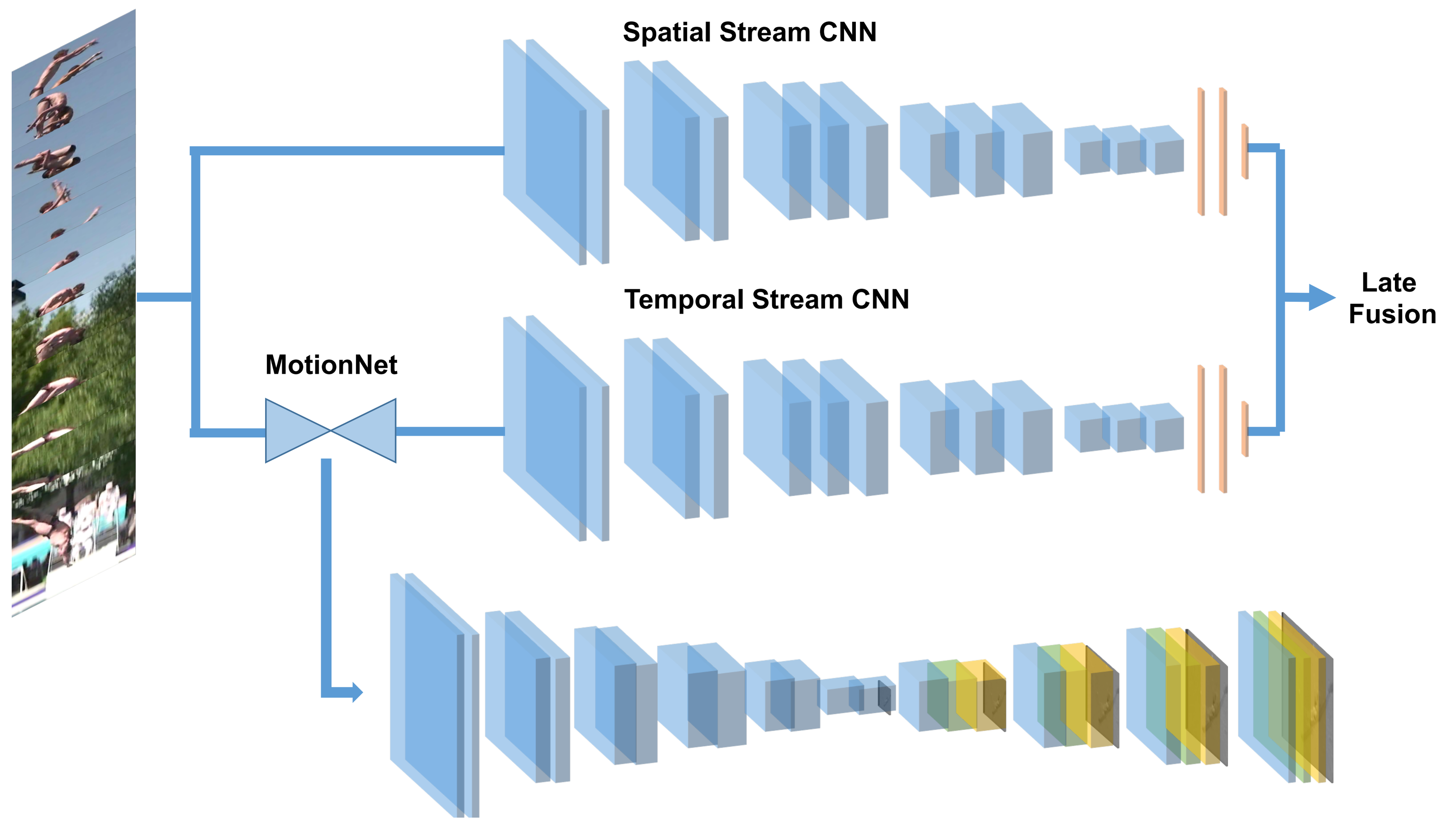
(b) Our Parallel Network

Model	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
Karaman et al. [16]	4.6	3.4	2.1	1.4	0.9
Wang et al. [36]	18.2	17.0	14.0	11.7	8.3
Oneata et al. [23]	36.6	33.6	27.0	20.8	14.4
Sun et al. [31]	12.4	11.0	8.5	5.2	4.4
Heilbron et al. [10]	36.1	32.9	25.7	18.2	13.5
Richard et al. [26]	39.7	35.7	30.0	23.2	15.2
Yeung et al. [42]	**48.9**	**44.0**	36.0	26.4	17.1
Ours fc8	45.6	41.2	34.5	26.1	17.0
Ours conv5 + fc8	46.6	42.9	35.6	28.5	18.7
Ours conv5 + fc6 + fc8	47.7	43.6	**36.2**	**28.9**	**19.0**

Work in progress:

Hidden Two-Stream Networks for Action Recognition

- We present a novel CNN architecture that implicitly captures motion information for action recognition. Our method is 10x faster that a conventional two-stage approach, does not need to cache flow estimates, and is end-to-end trainable.



Code and models available:
https://github.com/bryanyzhu/Hidden-Two-Stream

Method	Accuracy (%)	fps
TV-L1 [26]	85.65	14.75
FlowNet [22]	55.27	52.08
FlowNet2 [33]	79.64	8.05
NextFlow [48]	72.2	42.02
Enhanced Motion Vectors [32]	79.3	390.7
MotionNet (2 frames)	84.09	48.54
ActionFlowNet (2 frames)[18]	70.0	200.0
ActionFlowNet (16 frames)[18]	83.9	–
Stacked Temporal Stream CNN (a)	83.76	169.49
Stacked Temporal Stream CNN (b)	84.04	169.49
Stacked Temporal Stream CNN (c)	84.88	169.49
Two-Stream CNNs [10]	88.0	14.3
Very Deep Two-Stream CNNs[11]	**90.9**	**12.8**
Hidden Two-Stream CNNs (a)	87.50	120.48
Hidden Two-Stream CNNs (b)	87.99	120.48
Hidden Two-Stream CNNs (c)	**89.82**	**120.48**

Method	UCF101(%)	HMDB51(%)
Motion Vector + Fisher Vector Encoding [58]	78.5	46.7
ActionFlowNet (2 frames) [18]	70.0	42.6
ActionFlowNet (16 frames) [18]	83.9	56.4
C3D (1 Net) [6]	82.3	–
C3D (3 Net) [6]	85.2	–
Enhanced Motion Vector [32]	80.2	–
RGB + Enhanced Motion Vector [32]	86.4	–
RGB Diff [15]	83.0	–
RGB + RGB Diff [15]	86.8	–
Two-Stream 3DNet Initial [57]	85.2	–
Two-Stream 3DNet Mid [57]	87.0	–
Hidden Two-Stream Networks with Tiny-MotionNet	88.7	58.9
Hidden Two-Stream Networks with MotionNet	**90.3**	**60.5**