Actionness Estimation Using Hybrid Fully Convolutional Networks

Limin Wang[1,3] Yu Qiao[1] Xiaoou Tang[1,2] Luc Van Gool[3]

[1]MMLAB, SIAT, CAS, China [2]MMLAB, CUHK, Hong Kong [3]CVL, ETH Zurich, Switzerland

Introduction

- **Background**
  - Action classification can simply answer the question of "whether there is an action instance present in the video".
  - Action detection is able to provide the information of "where it is if there is an action instance in the video".
  - Sliding window is computationally prohibitive due to the huge number of spatio-temporal tube candidates.
  - Action proposal is a promising direction to reduce computational costs of action detection.
- **Action and actionness**
  - Action is defined as **intentional** bodily **movement** of biological **agents** (e.g. people, animals).
  - Actionness describes the confidence score of containing an action instance at this location.
  - Two important visual cues for actionness estimation: appearance and motion.
- **Method**
  - We formulate actionness estimation as a dense estimation problem by using fully convolutional network (FCN).
  - We propose a hybrid convolutional architecture (H-FCN), including appearance FCN (A-FCN) and motion appearance (M-FCN).
  - Additionally, actionness map is new kind of visual feature and we incorporate it into RCNN framework for action detection.

An example

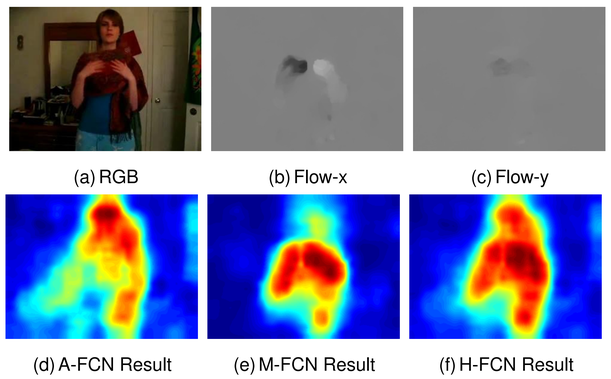


Figure 1: **An example of actionness maps**.

References

1. W. Chen, C. Xiong, R. Xu, and J. J. Corso. Actionness ranking with lattice conditional ordinal random fields. In CVPR, 2014.
2. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.
3. K. Simonyan, and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.
4. G. Gkioxari, and J. Malik. Finding action tubes. In CVPR, 2015.
5. L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream ConvNets. In arXiv, 2015.

Actionness estimation with H-FCN

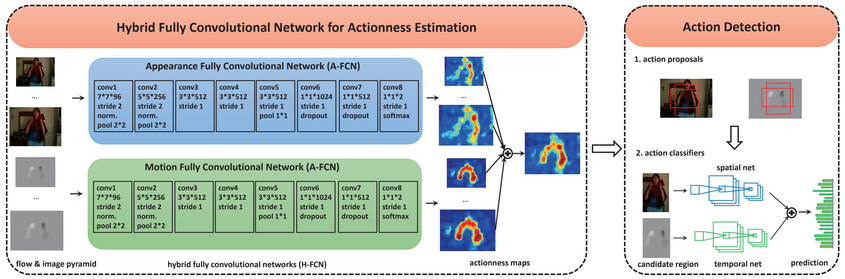


Figure 2: **Pipeline of our approach**.

- **Input**
  - A-FCN takes a single frame ($\boldsymbol{W} \times \boldsymbol{H} \times \mathbf{3}$).
  - M-FCN takes two consecutive frames ($\boldsymbol{W} \times \boldsymbol{H} \times \mathbf{4}$).
- **Architecture**
  - ClarifaiNet is adapted for H-FCN design by replacing FC layers and Conv layers.
- **H-FCN training**
  - Bounding boxes act as weak supervision signal for actionness estimation.
  - Pretraining A-FCN with ImageNet models (for object recognition).
  - Pretraining M-FCN with UCF101 models (for action recognition).
  - Training code is available at https://github.com/yjxiong/caffe.git [5].
- **H-FCN testing**
  - In realistic videos, action instances may vary in scales and we propose a multi-scale testing scheme.
  - We construct pyramid representations of RGB frames and stacking optical flow fields (four scales: $\mathbf{1}/\sqrt{\mathbf{2}}, \mathbf{1}, \sqrt{\mathbf{2}}, \mathbf{2}$).
  - These actionness maps from different scales are first up-sampled to the size of original image and then averaged.

Action proposal and action detection

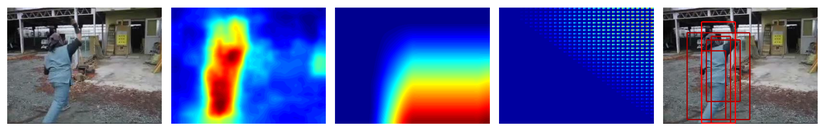


Figure 3: **Procedure of generating action proposals**.

- **Action proposal**
  - We first resize actionness maps into size of $\mathbf{32} \times \mathbf{32}$.
  - We then use integral image representation to speed up the calculation of average actionness score in any boxes.
  - We sample boxes according to their scores and spatial overlaps.
- **Action detection**
  - Following R-CNN, we train an action classifier with two-stream CNNs by cropping positive examples and mining negative examples.
  - At test time, we directly use the output of two-stream CNNs as the detection score for each action proposal.

Experimental results

Evaluation datasets: Stanford40, UCF Sports, and JHMDB.

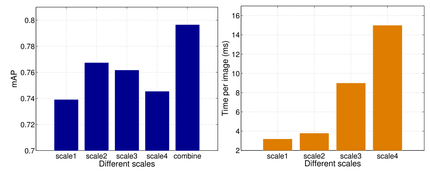


Figure 4: Exploration of multi-scale image representation for actionness estimation.

Method	Stanford 40	UCF Sports	JHMDB
L-CORF [1]	72.5%	60.8%	69.1%
DPM	**85.6%**	54.9%	58.2%
RankSVM	55.8%	21.9%	-
MBS	-	22.8%	-
A-FCN	79.7%	75.0%	80.7%
M-FCN	-	77.2%	80.6%
H-FCN	-	**82.7%**	**86.5%**

Table 1: Evaluation of actionness estimation.

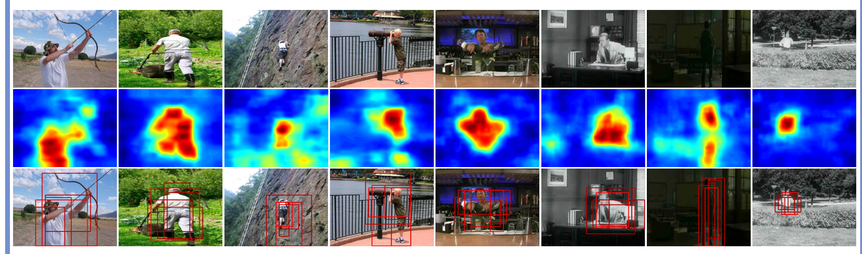
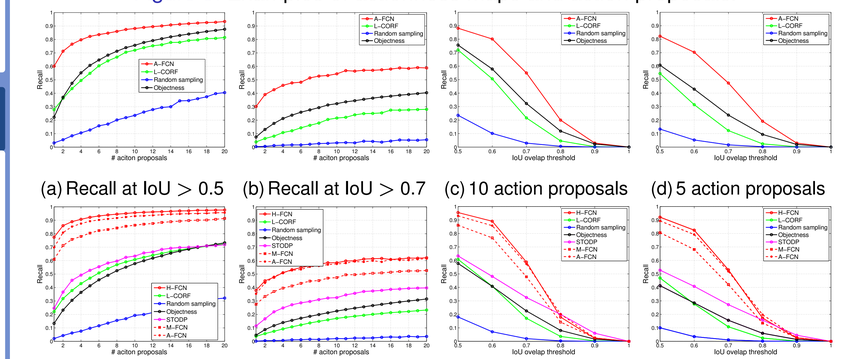


Figure 5: Examples of actionness maps and action proposals.



(a) Recall at IoU $>$ 0.5 (b) Recall at IoU $>$ 0.7 (c) 10 action proposals (d) 5 action proposals

(e) Recall at IoU $>$ 0.5 (f) Recall at IoU $>$ 0.7 (g) 10 action proposals (h) 5 action proposals

Figure 6: Evaluation of action proposals.

frame-AP(%)	brush-hair	catch	clap	climb	golf	jump	kick-ball	pick	pour	pullup	push	run	shoot-ball	shoot-bow	shoot-gun	sit	stand	swing-baseball	throw	walk	wave	mAP
spatial-CNN [4]	55.8	25.5	25.1	24.0	77.5	1.9	5.3	21.4	68.6	71.0	15.4	6.3	4.6	41.1	28.0	9.4	8.2	19.9	17.8	29.2	11.5	27.0
motion-CNN [4]	32.3	5.0	35.6	30.1	58.0	7.8	2.6	16.4	55.0	72.3	8.5	6.1	3.9	47.8	7.3	24.9	26.3	36.3	4.5	22.1	7.6	24.3
full [4]	**65.2**	18.3	38.1	**39.0**	79.4	7.3	9.4	25.2	**80.2**	**82.8**	**33.6**	11.6	5.6	66.8	27.0	**32.1**	34.2	33.6	15.5	**34.0**	**21.9**	36.2
our s-net	56.5	34.7	40.1	43.1	76.9	2.7	17.7	15.6	71.2	51.5	17.9	12.4	12.9	65.4	53.3	5.3	16.4	22.6	27.6	13.2	15.3	32.5
our t-net	42.9	19.0	49.6	28.9	71.8	14.0	20.4	36.6	60.1	66.0	18.0	17.3	8.3	73.5	26.0	11.6	44.1	53.7	17.6	22.4	11.5	34.0
our full net	60.1	**34.2**	**56.4**	38.9	**83.1**	**10.8**	**24.5**	**38.5**	71.5	67.5	21.3	**19.8**	**11.6**	**78.0**	**50.6**	10.9	**43.0**	**48.9**	**26.5**	25.2	15.8	**39.9**
video-AP(%)																						
spatial-CNN [4]	67.1	34.4	37.2	36.3	93.8	7.3	14.4	29.6	80.2	93.9	17.4	10.0	8.8	71.2	45.8	17.7	11.6	38.5	20.4	40.5	19.4	37.9
motion-CNN [4]	66.3	16.0	60.0	51.6	88.6	18.9	10.8	23.9	83.4	96.7	18.2	17.2	14.0	84.4	19.3	72.6	61.8	76.8	17.3	46.7	14.3	45.7
full [4]	**79.1**	33.4	53.9	**60.3**	**99.3**	18.4	26.2	42.0	**92.8**	**98.1**	29.6	24.6	13.7	**92.9**	42.3	**67.2**	57.6	66.5	27.9	58.9	**35.8**	53.3
our s-net	66.2	45.7	54.6	42.2	83.9	4.2	33.5	31.7	75.0	76.6	24.8	18.5	28.3	82.3	70.8	18.2	32.6	31.7	31.7	23.9	18.8	42.6
our t-net	64.2	38.1	80.1	39.0	91.8	34.7	57.4	74.6	74.5	77.6	31.3	40.9	18.5	89.4	59.0	32.3	69.3	82.9	25.8	46.1	22.2	54.8
our full net	76.4	**49.7**	**80.3**	43.0	92.5	**24.2**	**57.7**	**70.5**	78.7	77.2	**31.7**	**35.7**	**27.0**	88.8	**76.9**	29.8	**68.6**	**72.8**	**31.5**	**44.4**	26.2	**56.4**

Table 2: Evaluation of action detection on JHMDB dataset.