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 a new kind of visual feature and we incorporate it into RCNN framework for action detection.

An example

![RGB](image1.png)  ![Flow x](image2.png)  ![Flow y](image3.png)

(a) RGB  (b) Flow x  (c) Flow y

(d) A-FCN Result  (e) M-FCN Result  (f) H-FCN Result

Figure 1: An example of actionness maps.

Actionness estimation with H-FCN

Action proposal and action detection

![Pipeline of our approach](image4.png)

Figure 2: Pipeline of our approach.

- Input
  - A-FCN takes a single frame ($W \times H \times 3$).
  - M-FCN takes two consecutive frames ($W \times H \times 4$).

- 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](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: $1/\sqrt{2}$, $1/2$, $2$).
    - These actionness maps from different scales are first up-sampled to the size of original image and then averaged.

Experimental results

Evaluation datasets: Stanford40, UCF Sports, and JHMDB.

![Evaluation of actionness estimation](image5.png)

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

![Examples of actionness maps and action proposals](image6.png)

Figure 5: Examples of actionness maps and action proposals.

![Evaluation of action detection on JHMDB dataset](image7.png)

Figure 6: Evaluation of action detection.

Table 1: Evaluation of actionness estimation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Stanford40</th>
<th>UCF Sports</th>
<th>JHMDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPM</td>
<td>85.6%</td>
<td>54.9%</td>
<td>58.2%</td>
</tr>
<tr>
<td>RankSVM</td>
<td>55.8%</td>
<td>21.9%</td>
<td></td>
</tr>
<tr>
<td>L-CNN</td>
<td>77.9%</td>
<td>76.0%</td>
<td>84.0%</td>
</tr>
<tr>
<td>M-FCN</td>
<td>77.2%</td>
<td>80.6%</td>
<td>86.5%</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of action detection on JHMDB dataset.

References