Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs

Zheng Shou, Dongang Wang, Shih-Fu Chang
Columbia University

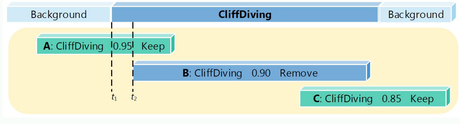
IEEE 2016 Conference on Computer Vision and Pattern Recognition
CVPR 2016

Overview:

- **Motivation:** Localizing actions in untrimmed long videos, such as those in surveillance, can save tremendous time and costs.
- **Problem definition:** given an untrimmed video, "when does each specific action instance start and end, and which action class does it belong to?"
- **Framework of Segment-CNN:**
  - Generate candidate segments via multi-scale sliding window (16, 32, 64, 128, 256, 512 frames with 75% overlap in time)
  - Segment-based CNNs:
    - Network architecture: C3D [Tran et al.]
    - Training: Proposal → Classification → Localization (init by the trained cls model)
    - Prediction: Proposal → Localization
  - Conduct post-processing steps such as non-maximum suppression

Training details of Segment-CNN:

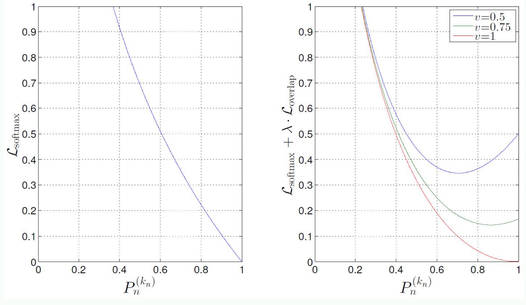
- **Proposal network:**
  - Two categories (background and being action)
  - Identify candidate segments that may contain actions
- **Classification network:**
  - Train N+1 multi-class classification model using Softmax Loss: $\mathcal{L}_{\text{softmax}} = \frac{1}{N}\sum_n \left(-\log\left(P_n^{(k_n)}\right)\right)$
  - Serve as initialization for the localization network
- **Localization network:**
  - Motivation: NMS might remove segment of small score but of large overlap with ground truth.

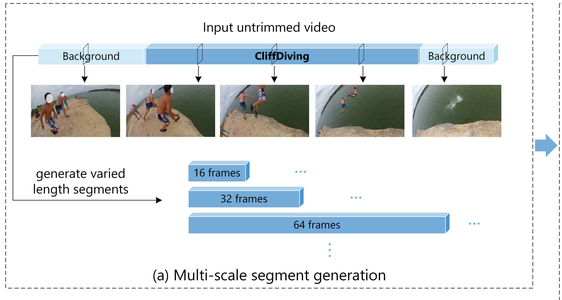


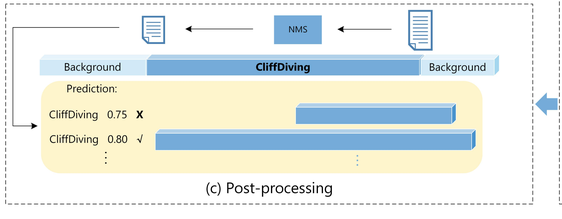
- Solution: explicitly consider temporal overlap in the loss function $\mathcal{L} = \mathcal{L}_{\text{softmax}} + \lambda \cdot \mathcal{L}_{\text{overlap}}$

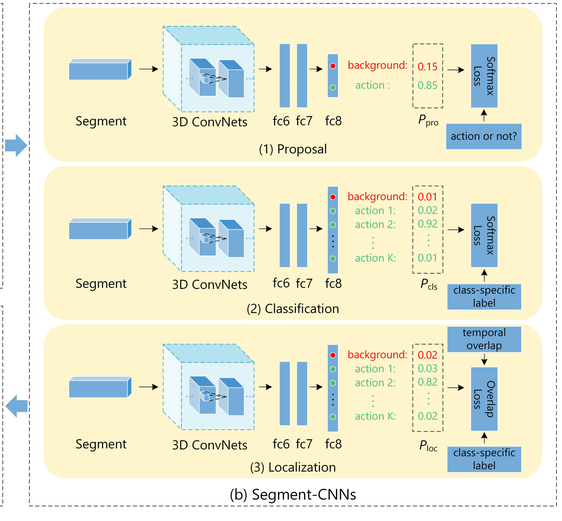
$$\mathcal{L}_{\text{overlap}} = \frac{1}{N}\sum_n \left(\frac{1}{2}\cdot\left(\frac{\left(P_n^{(k_n)}\right)^2}{\left(v_n\right)^\alpha} - 1\right)\cdot[k_n > 0]\right)$$

- The overlap loss term rewards segment with higher temporal overlap with the ground truth









Experimental setup:

- **Evaluation metrics:**
  - Regard as retrieval problem and evaluate mAP
  - What is a correct action instance prediction: correct category prediction + IoU with ground truth instance larger than the evaluation threshold (set to 0.5 unless specified)
  - Redundant detections are not allowed
- **THUMOS14 temporal action detection task:**
  - 20 sports categories
  - Training data: 2,755 trimmed videos from UCF101 + 1,010 untrimmed YouTube videos of 3,007 instances
  - Test data: 213 untrimmed videos of 3,358 action instances

Contact: zheng.shou@columbia.edu

Code available: https://github.com/zhengshou/scnn

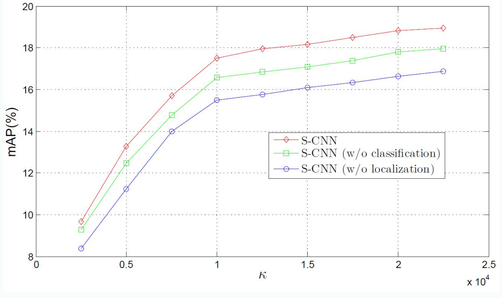
Experimental results on THUMOS14:

- **Comparison with state-of-the-art systems:**
  - Mainly use FV encoding of iDTF and frame-level CNN features
  - Some work leveraged video-level classifiers trained on multiple features

IoU threshold in evaluation	0.1	0.2	0.3	0.4	0.5
Karaman et al.	1.5	0.9	0.5	0.3	0.2
Wang et al.	19.2	17.8	14.6	12.1	8.5
Oneata et al.	39.8	36.2	28.8	21.8	15.0
Segment-CNN	**47.7**	**43.5**	**36.3**	**28.7**	**19.0**

- **Efficiency analysis:**
  - Speed: GTX980. ~1s per batch. ~0.5s to process 1s video.
  - Storage: less than 1 GB. do not need to cache intermediate features.
- **Impact of individual networks:**
  - Proposal: S-CNN (19.0%) **VS** S-CNN w/o proposal (17.1%)
  - Classification:
  - Localization:

Here we perform top-K selection on the final prediction results to select K segments with maximum confidence scores



- **Examples:**