ECCV'20 ONLINE
23-28 AUGUST 2020

Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos

Anurag Arnab, Chen Sun,
Arsha Nagrani, Cordelia Schmid

Google Research

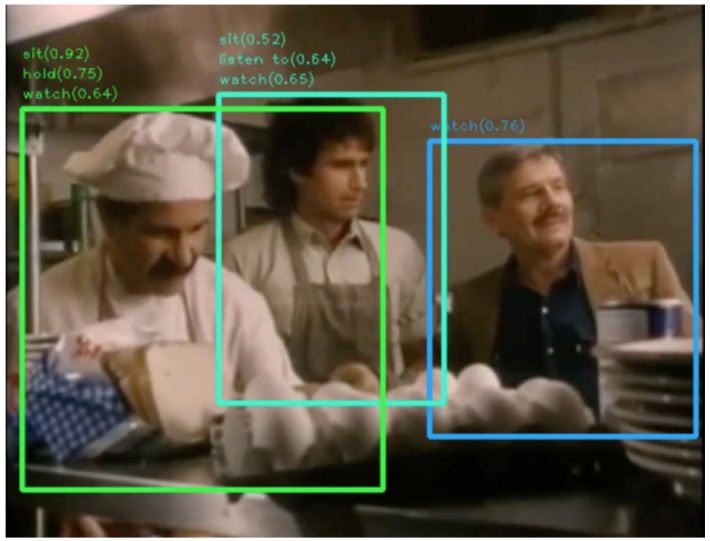
Introduction

- Spatio-temporal action detection
  - Bounding box in space and time around action of interest
- Most approaches extend detectors, such as Faster-RCNN and SSD, temporally.
- In this paper, we only use cheap, video-level labels

Annotation

Labels: *sit, hold, watch, talk to, listen to*

Prediction



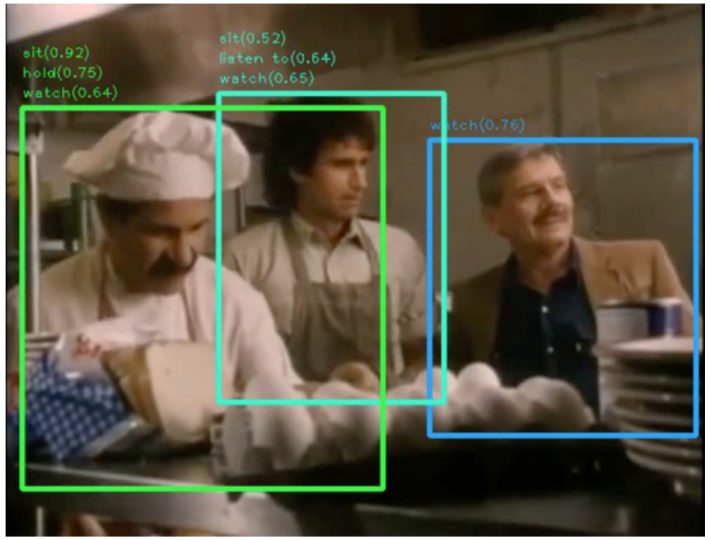
Weaker supervision

- Labelling bounding boxes per frame is too expensive
- Temporal boundaries of actions are ambiguous, annotators often do not agree with each other
- Only use cheap, video-level labels.

Annotation

Labels: *sit, hold, watch, talk to, listen to*

Prediction



Approach Overview

- Leverage off-the-shelf, per-frame person detectors to obtain person tubelets.
- Multiple Instance Learning
  - Each bag is formed from all tubelets in the video
- Due to noise, and violations of the MIL assumptions, predict the uncertainty for each bag as well.

Multiple Instance Learning

- Have a bag of examples, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.
- Only know label for the whole bag, y.
- Key assumption is that one or more instances in the bag have label y.
- Want to train an instance-level classifier.
    - Classify each instance in the bag.

Multiple Instance Learning

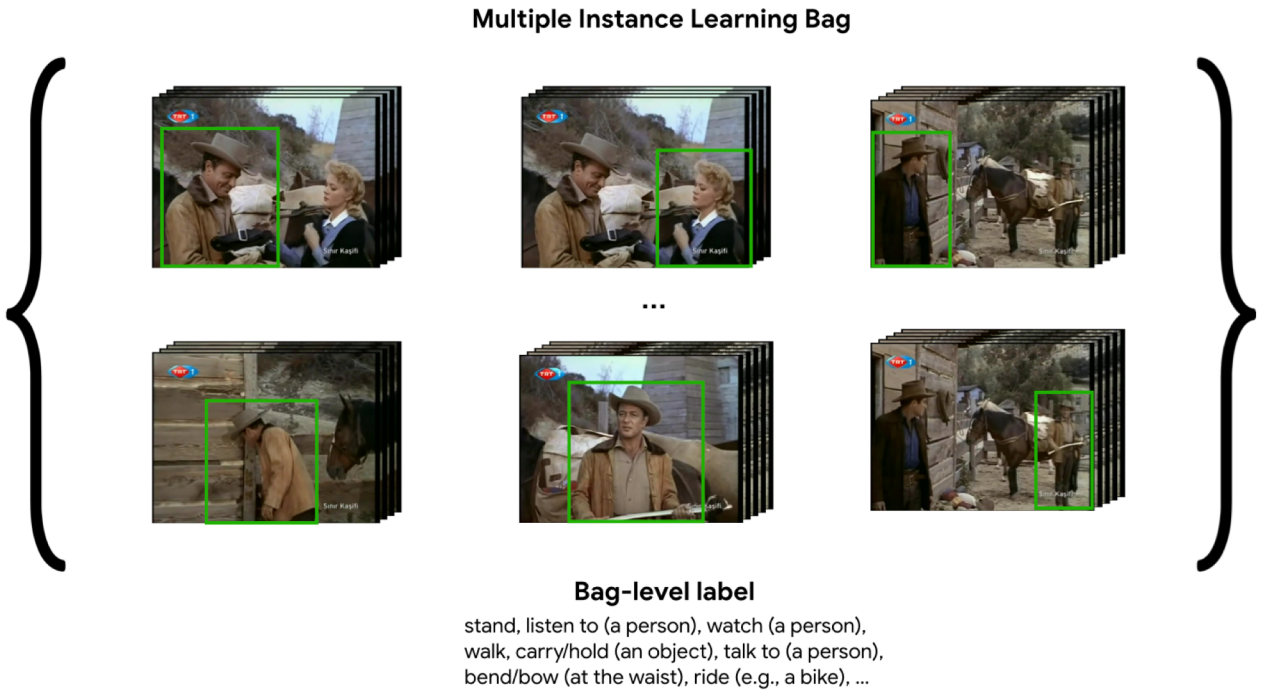
- Have a bag of examples, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.
- Only know label for the whole bag, y.
- Want to train an instance-level classifier.
- Aggregate instance-level predictions into a bag-level prediction.

$$p(y_l = 1 | x_1, x_2, \ldots, x_n) = g(p_i, p_2, \ldots, p_n)$$

- Use standard loss function
- Common aggregation functions: max, log-sum-exp, average, attention

Multiple Instance Learning (MIL)

- All the person tubelets within a video form a “bag”
  - Person tubelets are detections linked over at most K frames.
- The standard MIL assumption is that at least one tubelet in the bag has the video-level label.



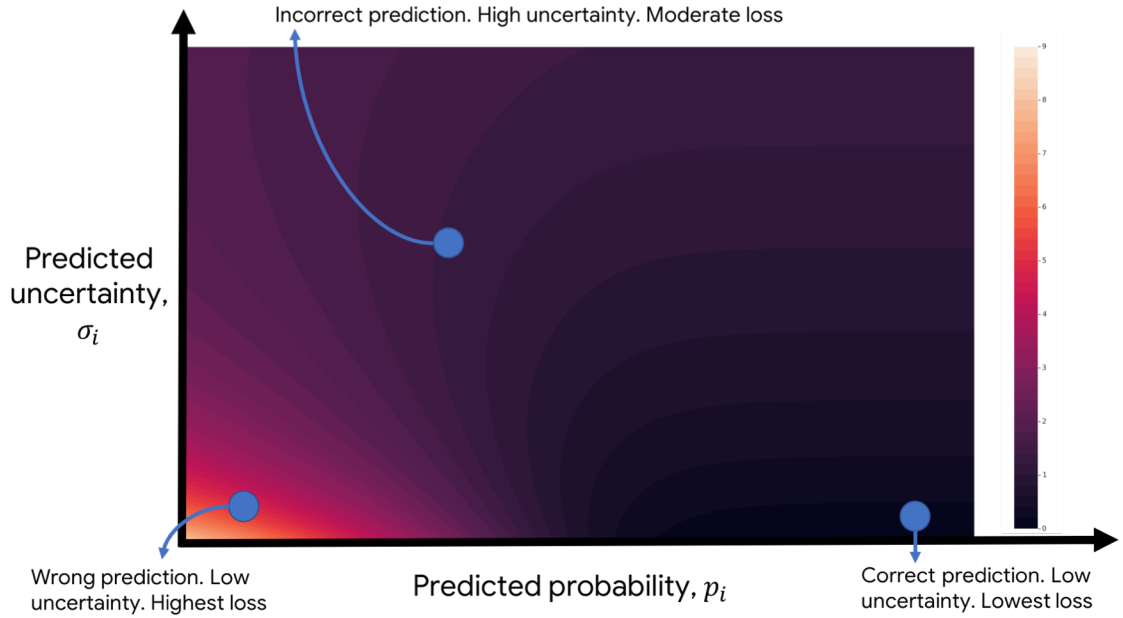
Label Noise and Violations of MIL Assumption

- MIL assumption is often violated
- Sampling bags
  - Cannot fit a whole bag in memory
  - Particularly as videos get longer
  - Uniformly sample tubes
- Person detector errors
  - Due to domain gap
  - False positives as some datasets don’t label actors exhaustively

All person detections besides the pole-vaulter are considered false-positives in this dataset.

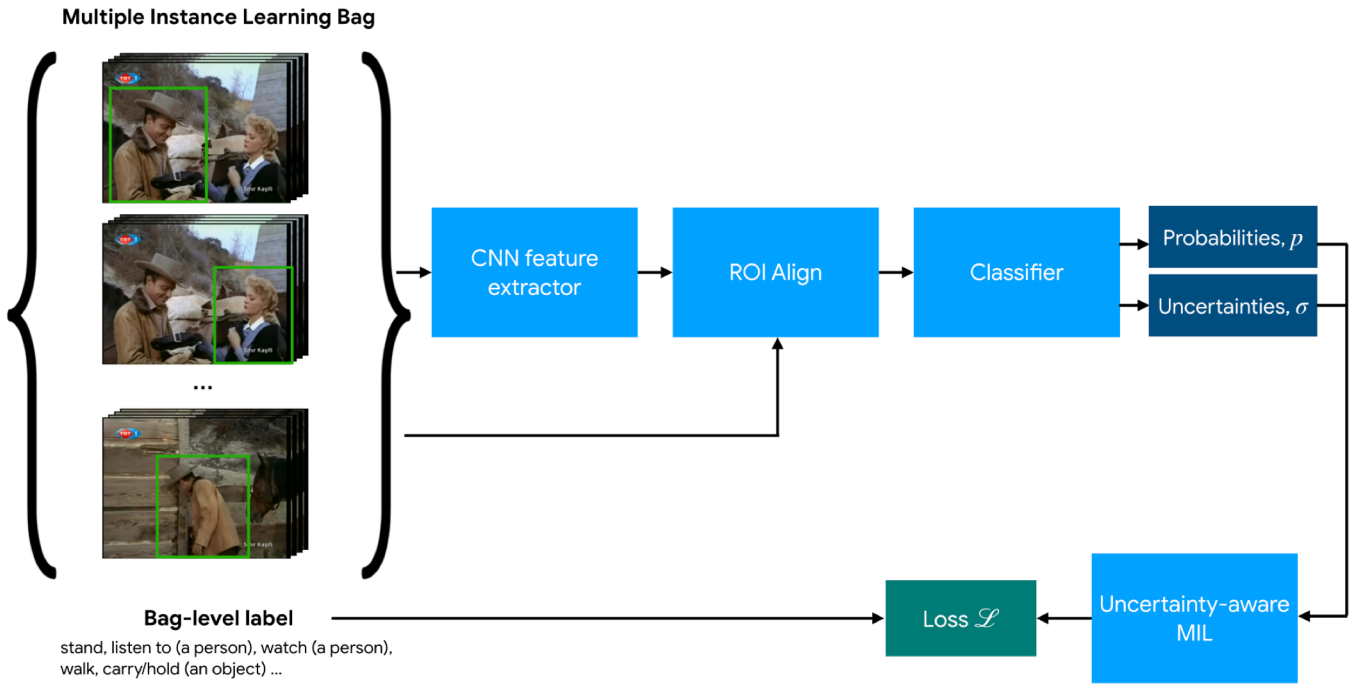
Uncertainty Estimation

- Predict uncertainty for each instance in the bag
- Intuition:
  - When possible, predict correct label with low uncertainty
  - Otherwise, predict incorrect label with high uncertainty.
- $L(x, y, \sigma) = \frac{1}{\sigma^2}\mathcal{L}_{ce}(x, y) + \lambda \log(\sigma^2)$



Network Architecture

- Fast-RCNN style detector
- Use person tubelets as proposals



Evaluation Datasets

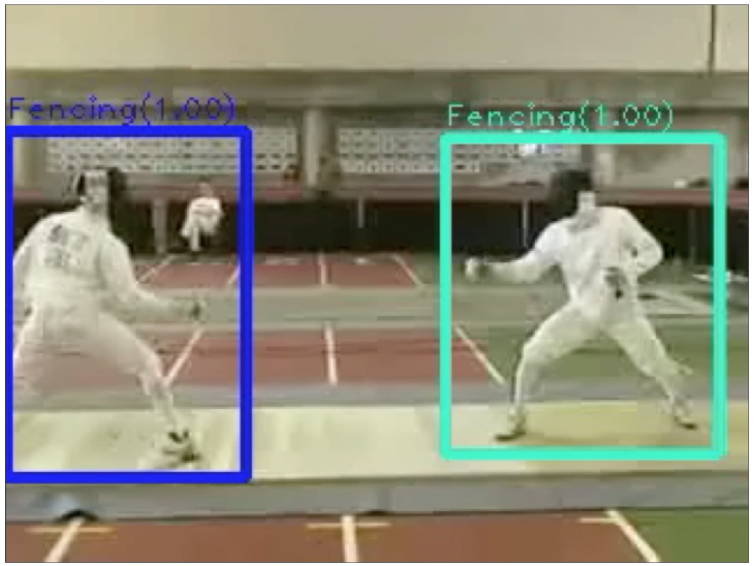
- UCF101-24
  - Most common dataset
  - Sports videos from YouTube, 24 classes
  - Many “background people” not doing the labelled action
  - Evaluate Video AP

Evaluation Datasets

- UCF101-24
  - Most common dataset
  - Sports videos from YouTube, 24 classes
  - Many “background people” not doing the labelled action
  - Evaluate Video AP
- AVA
  - 60 atomic actions, from 15 minute movie clips
  - Keyframes at 1Hz, are labelled. Predict actions at keyframe given temporal context
  - Evaluate Frame AP

UCF101-24 Ablation

	Video AP	
	0.2	0.5
Weakly supervised baseline	54.3	29.7
MIL - LSE pooling	60.1	33.1
MIL - mean pooling	60.3	33.0
MIL - max pooling	60.7	33.5
MIL - max pooling, uncertainty	61.7	35.0
Fully supervised	69.3	43.6



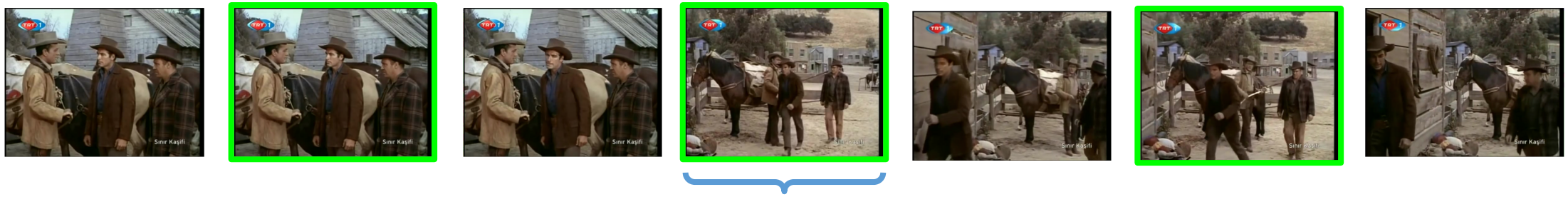
- Big domain gap between COCO and UCF
- Detector, trained only on COCO, has 47% recall and 21% precision on UCF training set.
- Sampling tubelets is necessary: Average of 33.1 tubelets per video, V100 GPU can hold 16.

UCF101-24 Comparison

	Video AP at 0.2	Video AP at 0.5
Fully supervised		
Peng *et al.* [35]	42.3	35.9
Hou *et al.* [17]	47.1	–
Weinzaepfel *et al.* [50]	58.9	–
Saha *et al.* [38]	63.1	33.1
Singh *et al.* [41]	73.5	46.3
Zhao *et al.* [52]	78.5	50.3
Singh *et al.* [40]	79.0	50.9
Kalogeiton *et al.* [19]	77.2	51.4
Ours	69.3	43.6
Weakly supervised		
Escorcia *et al.* [8]	45.5	–
Chéron *et al.* [6]	43.9	17.7
Ours	61.7	35.0

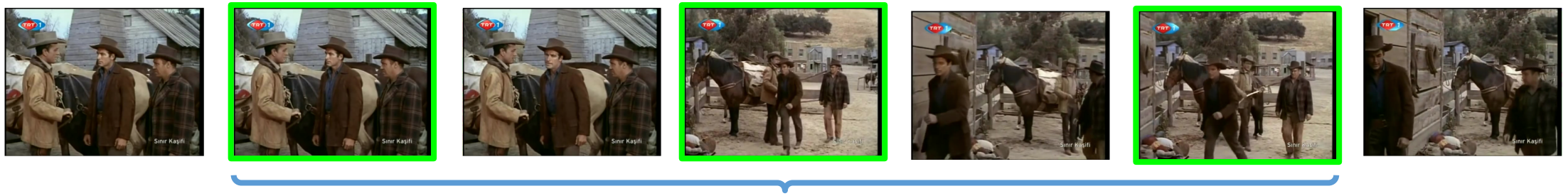
AVA

- AVA labels keyframes at 1Hz, videos are 15 minutes long.
- Vary the subclip of the video from which we take clip-level annotation
- Problem gets harder as the subclip duration is increased.



AVA

- AVA labels keyframes at 1Hz, videos are 15 minutes long.
- Vary the subclip of the video from which we take clip-level annotation
- Problem gets harder as the subclip duration is increased.



AVA Results

	Sub-clip duration (seconds)						
	FS	1	5	10	30	60	900
Frame AP	24.9	22.4	18.0	15.8	11.4	9.1	4.2

Conclusion

- Weakly-supervised spatio-temporal action detection with Multiple Instance Learning
- Predict uncertainty to better handle noise and violations of standard MIL assumption.

- Visit our poster session, or download our paper, for more details
  - A Arnab *et al*. Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos. ECCV 2020 [PDF]