# Action Recognition from Single Timestamp Supervision in Untrimmed Videos


Davide Moltisanti[1] Sanja Fidler[2] Dima Damen[1]

[1]University of Bristol, UK [2] University of Toronto, Canada | NVIDIA | Vector Institute


CVPR LONG BEACH CALIFORNIA June 16-20, 2019

## Abstract


Typical approaches for action recognition in videos rely on labelled start and end times for training.

This supervision is not only expensive to acquire but importantly highly subjective.

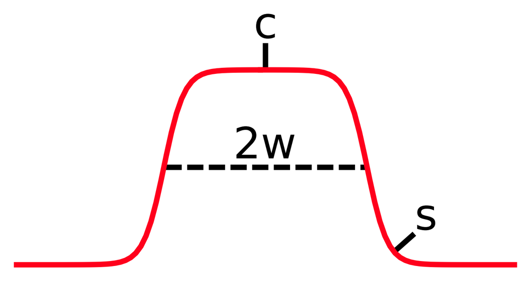
In this paper, we:

- Use **single timestamps** located around each action instance in **untrimmed** videos as weak supervision;
- Temporally refine the supervision used to train a classifier, starting from the single timestamps;
- Testing the classifier on *trimmed* video segments, we show that our method converges to the discriminative action segments, for 3 different datasets (THUMOS 14, BEOID and EPIC Kitchens).


## Approach

- We start from single timestamps, roughly located close to the action instances;
- We replace unavailable action boundaries with sampling distributions modelled by a plateau function:

$$g(x \mid c, w, s) = \frac{1}{(e^{s(x-c-w)}+1)(e^{s(-x+c-w)}+1)}$$




- We initialise one sampling distribution per action, centring the plateau on the single timestamp;
- Initial plateaus might enclose irrelevant frames. We thus update the sampling distributions, fitting multiple update proposals per distribution, using the softmax scores;
- We rank the proposals to select the most confident updates, using a Curriculum Learning approach. We reward proposals whose plateaus contain frames that on average score higher than the frames enclosed by the current plateau;
- We iteratively update the sampling distributions until convergence, which is measured using the proposals' scores.

## References


[1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS Dataset. In *ECCV*, 2018.

[2] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. UntrimmedNets for weakly supervised action recognition and detection. In *CVPR*, 2017.


## Initialising the Model

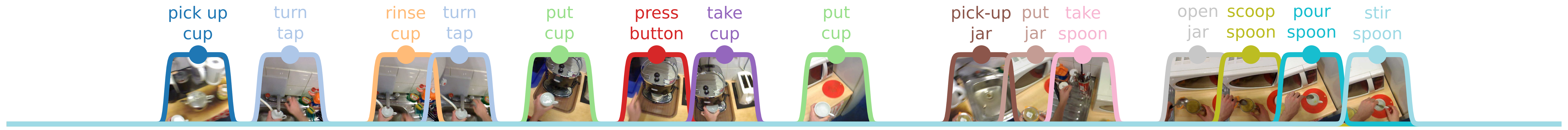
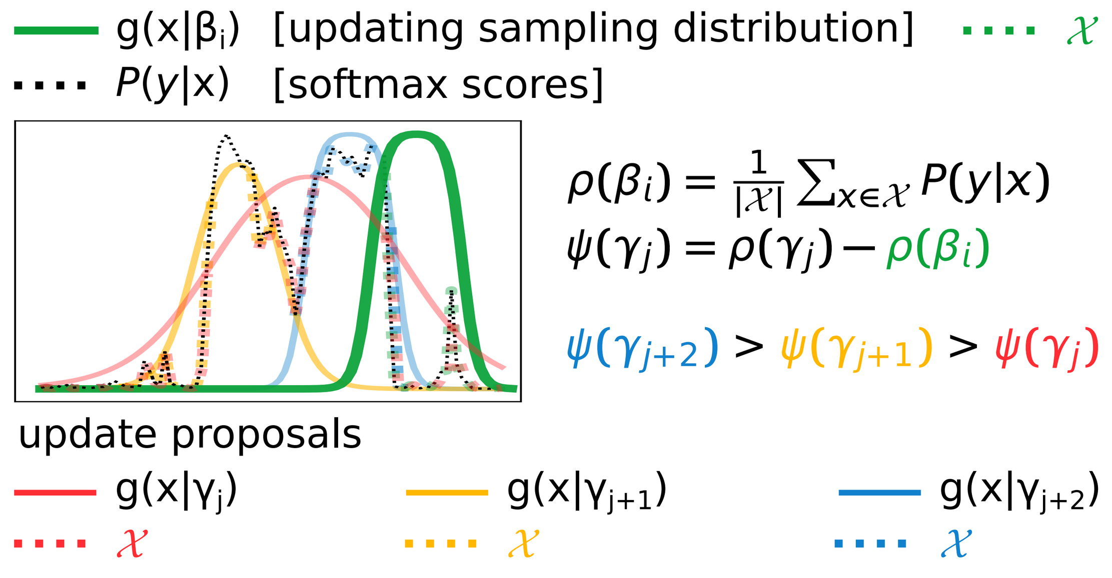



**Figure 1:** Replacing action boundaries with sampling distributions in an untrimmed video, given single timestamps (coloured dots at the centre of each plateau). The initial distributions may overlap (e.g. 'put jar', 'take spoon') and contain background frames. We iteratively refine the distributions using the classifier response during training.

## Updating the Sampling Distributions




$$\rho(\beta_i) = \frac{1}{|\mathcal{X}|}\sum_{x \in \mathcal{X}} P(y|x)$$

$$\psi(\gamma_j) = \rho(\gamma_j) - \rho(\beta_i)$$

$$\psi(\gamma_{j+2}) > \psi(\gamma_{j+1}) > \psi(\gamma_j)$$

**Figure 4:** Fitting and ranking update proposals.


**Figure 5:** Updating the sampling distribution using the classifier response - example from action 'open fridge' in EPIC Kitchens [1]. Different colours indicate different training iterations.

## Results

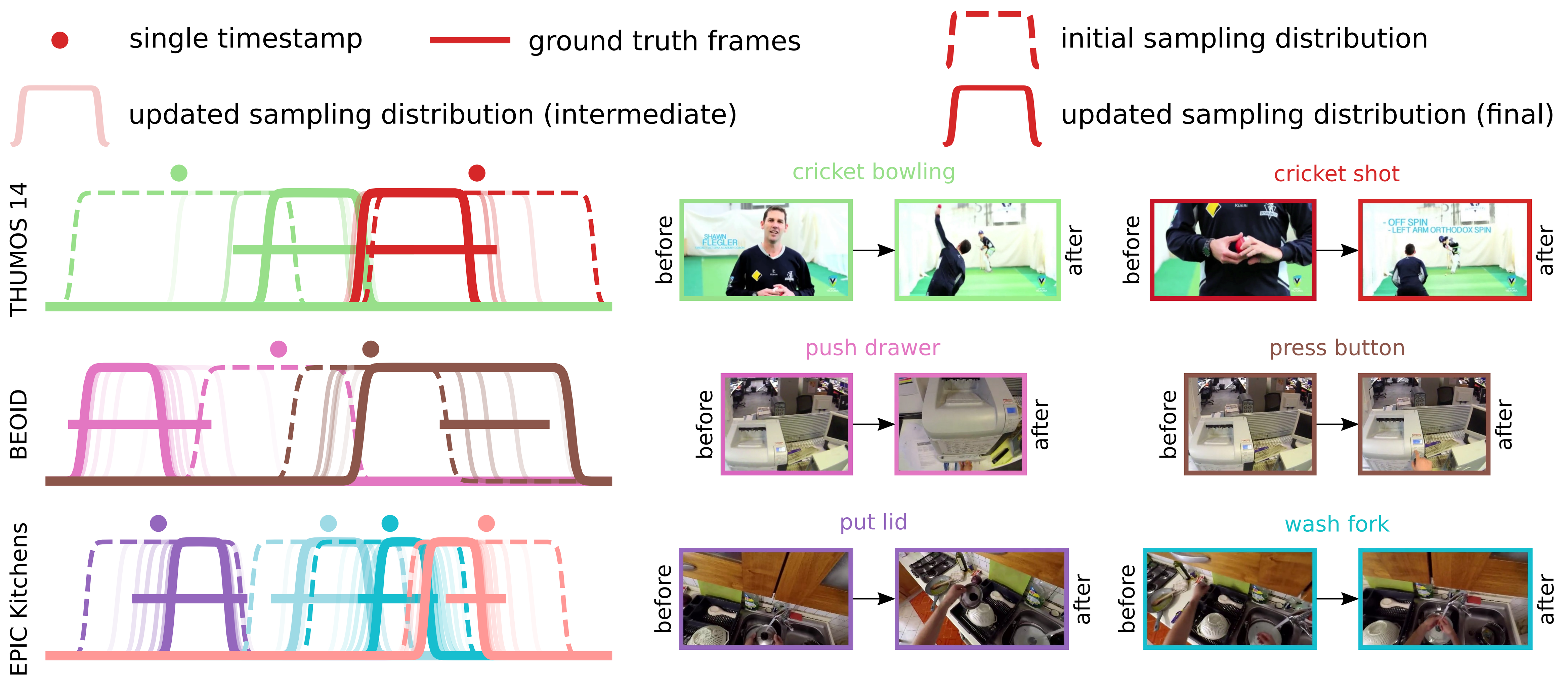



**Figure 6:** Qualitative results on the three datasets. Ground truth frames used only for plotting.

| Dataset | CL $h$ | Before update | After update |
|---|---|---|---|
| THUMOS 14 | 0.25 | 26.10 | 28.88 |
| | 0.50 | 32.69 | 55.15 |
| | 0.75 | 33.59 | 56.42 |
| | 1.00 | 63.41 | 63.53 |
| BEOID | 0.25 | 47.97 | 52.70 |
| | 0.50 | 71.62 | 83.11 |
| | 0.75 | 74.32 | 83.11 |
| | 1.00 | 64.86 | 70.27 |
| EPIC Kitchens | 0.25 | 20.47 | 22.83 |
| | 0.50 | 21.39 | 25.35 |
| | 0.75 | 20.73 | 23.86 |
| | 1.00 | 23.55 | 24.17 |

**Table 3:** Top-1 accuracy obtained with single timestamp supervision before and after update.

## Links

Code available at: bitbucket.org/dmoltisanti/action_recognition_single_timestamps

Project webpage: people.cs.bris.ac.uk/damen/single_timestamps

## Comparing Levels of Temporal Supervision

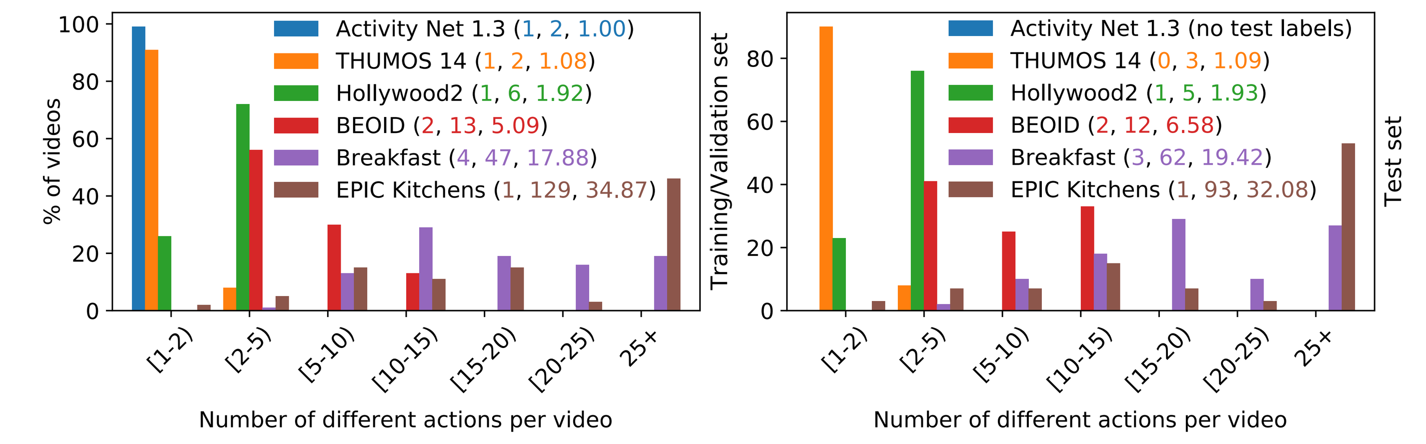



**Figure 2:** Different actions per video for various datasets.

| Set | Dataset | N. of classes | N. of videos | N. of actions | Avg video length | Avg classes per video | Avg actions per video |
|---|---|---|---|---|---|---|---|
| Train | THUMOS 14 | 20 | 200 | 3003 | 208.90 | 1.08 | 15.01 |
| | BEOID | 34 | 46 | 594 | 61.31 | 5.09 | 12.91 |
| | EPIC Kitchens | 274 | 79 | 7060 | 477.37 | 34.87 | 89.36 |
| Test | THUMOS 14 | 20 | 210 | 3307 | 217.16 | 1.09 | 15.74 |
| | BEOID | 34 | 12 | 148 | 57.78 | 6.58 | 12.33 |
| | EPIC Kitchens | 274 | 26 | 1949 | 399.62 | 32.08 | 74.96 |

**Table 1:** Datasets information. Average video length is in seconds.

| Baseline | | U. Net[2] | Ours | | |
|---|---|---|---|---|---|
| Supervision | APV | Video-level | TS | TS in GT | Full |
| THUMOS 14 | 1.08 | 64.92 | 66.68 | 64.53 | 67.10 |
| BEOID | 5.09 | 28.37 | 85.14 | 88.51 | 87.83 |
| EPIC Kitchens | 34.87 | 2.20 | 26.22 | 32.53 | 35.97 |

**Table 2:** Comparison between different levels of temporal supervision. APV indicates the average number of unique actions per training video.

## Convergence

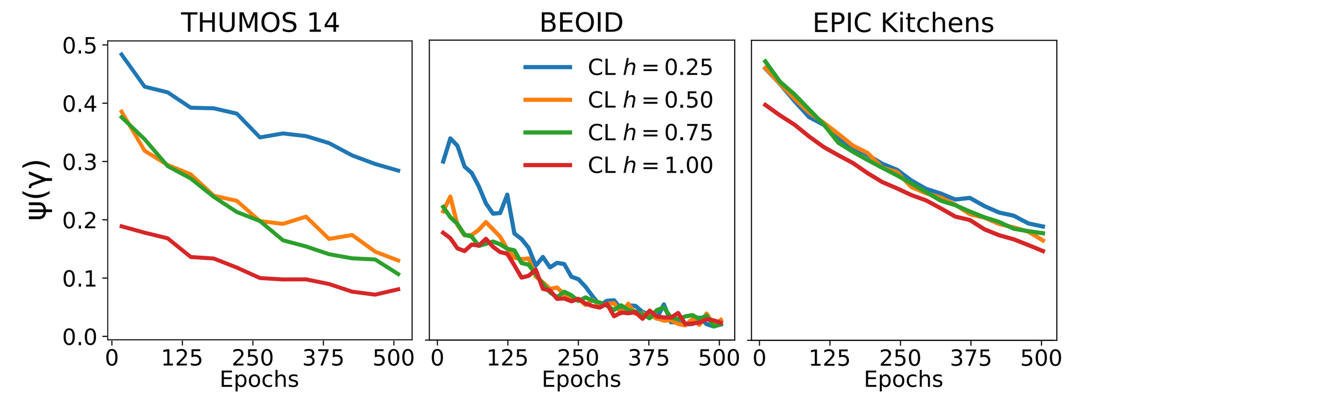



**Figure 3:** Average confidence of selected update proposals over training epochs.