A flexible model for training action localization with varying levels of supervision

Guilhem Chéron* [1], Jean-Baptiste Alayrac* [1,2], Ivan Laptev[1] and Cordelia Schmid[1,3]

[1]INRIA/ENS/CNRS/PSL, Paris [2]DeepMind, London [3]Univerity Grenoble Alpes/INRIA/CNRS/Grenoble INP/ LJK, Grenoble

Goal and Overview

MOTIVATION

Spatio-temporal action localization is an important task that is typically addressed by **supervised learning** approaches.

Such methods rely on **exhaustive supervision** where each frame of a training action is manually annotated with a bounding box.

Manual annotation is **expensive** and often **ambiguous**.

Alternative methods using less supervision are needed.

OBJECTIVES

Compare various levels of supervision to understand what supervision is required for spatio-temporal action localization.

HOW?

We design a **unifying framework** for handling various levels of supervision.

Our model is based on **discriminative clustering** and integrates different types of **supervision** in a form of optimization **constraints**.

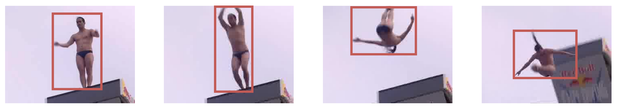
CONTRIBUTIONS

- a **flexible model** with ability to adopt and combine various types of supervision for **action localization**

- an **experimental study** demonstraing the strengths and weaknesses of a wide range of supervisory signals and their combinations

Preprocessing

Detection and tracking



From track to tracklets

Every **track** is divided into chunks (8 frames) that we call **tracklets**.

Feature extraction

For every detection d, we compute a visual feature ϕ_d (I3D ROI Pool).

For every tracklet m, we compute its feature x_m :

$$x_m = \frac{1}{|m|} \sum_{d \in m} \phi_d$$: average of the detection features belonging to the tracklet.

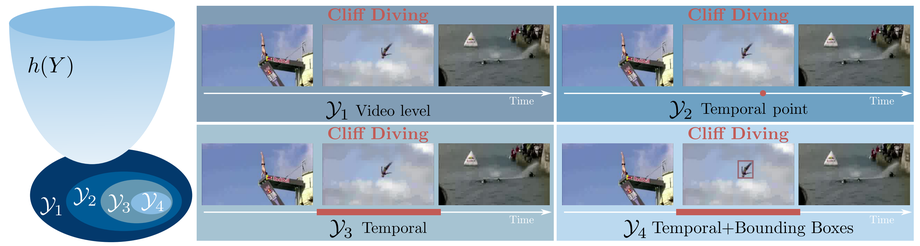
We stack all tracklet features x_m in the feature matrix X.

Project webpage

Check out our project webpage for our paper/code and more qualitative results!

`https://www.di.ens.fr/willow/research/weakactionloc/`

Our approach



A single flexible model: one cost function, different constraints.

$Y \Leftrightarrow$ **Assignment of tracklets to action classes**

$$\min_{Y \in \mathcal{Y}_s} h(Y)$$

Supervision levels $\Leftrightarrow$ Constraints $\mathcal{Y}_s$

Supervision levels considered in this work

Video level action labels: only video-level action labels are known.

Single temporal point: we know one frame of an action, but we do not have either the exact temporal extent or the spatial extent of the action.

One bounding box (BB): we are given the spatial location of a person at a given time inside each action instance.

Temporal bounds: we know the temporal interval of the action occurs (but not its spatial extent).

Temporal bounds with bounding boxes (BBs): combination of temporal bounds and bounding box annotation.

Fully supervised: annotation is defined by the bounding box at each frame of an action.

Transform supervision into constraints

Example: Temporal bounds $\Leftrightarrow$ *"At least one human track contained in the given temporal interval should be assigned to that action. Samples outside annotated intervals are assigned to the background class."*

Formalism

Exactly one class per tracklet

$$\forall\, m \in [1 \dots M], \quad \sum_{k=1}^{K} Y_{mk} = 1$$

NB: This could be extended to the multi class setting (e.g. for AVA [Gu et al.])

Strong supervision with equality constraints

Know what the tracklet is:

$$\forall (t,k) \in \mathcal{O}_s \quad Y_{tk} = 1$$

Know what the tracklet is not:

$$\forall (t,k) \in \mathcal{Z}_s \quad Y_{tk} = 0$$

Weak supervision with at-least-one constraints

A bag of tracklet that may contain an action:

$$\sum_{t \in \mathcal{A}_k} Y_{tk} \geq 1$$

$h(Y)$:Discriminative clustering

DIFFRAC framework [Bach and Harchaoui]:

$$h(Y) = \min_{W \in \mathbb{R}^{d \times K}} \frac{1}{2M} \|XW - Y\|_F^2 + \frac{\lambda}{2} \|W\|_F^2$$

Representation of tracklets ROI-pooled I3D features [M x d] matrix

Linear action classifier [d x K] matrix

Nice property:

W

The **action classifier** is learned during optim. and can be used at **test time**!

Intuition: Recover a **labeling** of the data (Y) so that this labeling can be easily recovered by a **linear classifier** (W) over some **features** (X).

Experiments

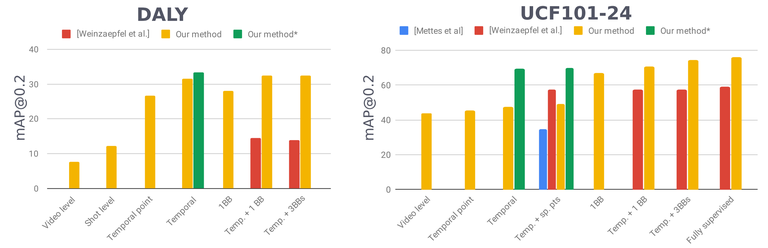
Datasets, metrics and implementation details

UCF101-24 and **DALY** datasets.

Evaluation metric: mean average precision

Faster-RCNN detectors (off-the-shelf or finetuned*), KLT tracker or online tracker, **Frank-Wolfe** optimization

Comparing levels of supervision



Conclusions: (i) dense spatial annotation is not always needed
(ii) 'Temporal point' results are very promising ('click' annotation)

Mixed supervision

Conclusion: on par with fully supervised with only **40%** of the fully annotated data!

Supervised baselines

UCF101-24

Video mAP	@0.2	@0.5
[Weinzaepfel et al.]	58.9	-
[Peng et al.]	-	35.9
[Singh et al.]	73.5	46.3
[Kalogeiton et al.]	76.5	49.2
[Gu et al.]	-	59.9
Our method	76.0	50.1

Conclusion: our simple method is competitive in the FS mode.

Qualitative results

References

Bach and Harchaoui. *DIFFRAC: A discriminative and flexible framework for clustering.* NIPS, 2007.
Gu et al. *AVA: A video dataset of spatio-temporally localized atomic visual actions.* CVPR, 2018.
Kalogeiton et al. *Action tubelet detector for spatio-temporal action localization.* ICCV, 2017.
Mettes et al. *Localizing actions from video labels and pseudo-annotations.* BMVC, 2017.
Peng et al. *Multi-region two-stream R-CNN for action detection.* ECCV, 2016.
Singh et al. *Online real time multiple spatiotemporal action localisation and prediction.* ICCV, 2017.
Weinzaepfel et al. *Human action localization with sparse spatial supervision.* arXiv, 2016.