SEARCH VIDEO ACTION PROPOSAL WITH RECURRENT AND STATIC YOLO

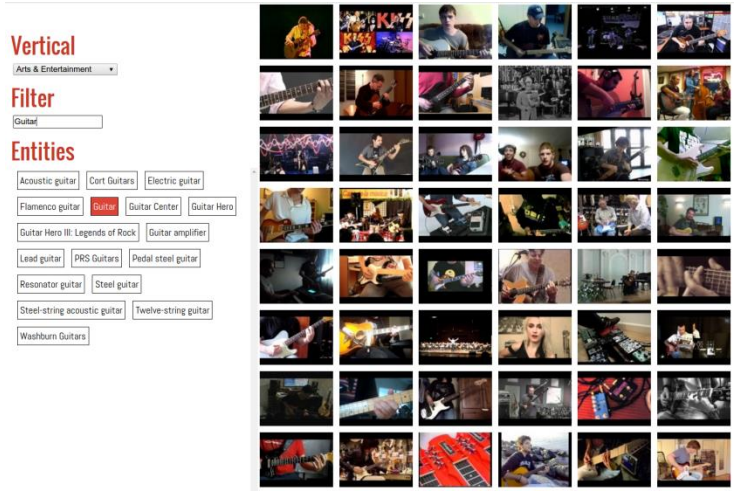
Romain Vial, Hongyuan Zhu, Yonghong Tian, Shijian Lu

Introduction

Video Understanding Tasks

Classification

Prediction of the overall category of the video (e.g. Football, Concert)



[Abu-El-Haija *et al.*, arXiv 2016]

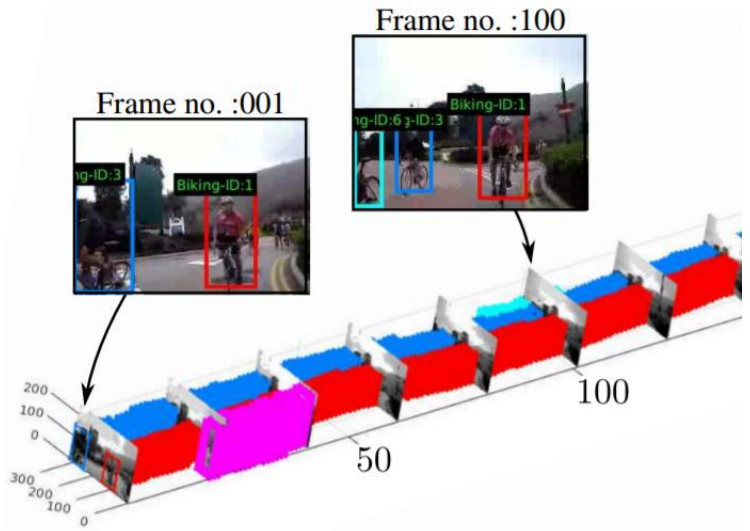
Temporal Detection

Prediction of the temporal location of the action in the video

[Heilbron *et al.*, CVPR 2015]

Tubes Detection

Localize both spatially and temporally the action in the video



[Saha *et al.*, BMVC 2016]

Introduction
Action Proposal

Def.: Produce a set of candidate spatio-temporal tubes that are likely to contain a human action.

Why? To reduce the computational complexity of further tasks (e.g. classification) by trimming the video in highly discriminative sections.

Related work

- Segmentation of point trajectories based on optical flow [T. Brox *et al*., ECCV 2010]
- Supervoxel segmentation and hierarchical clustering [D. Oneata *et al.*, ECCV 2014]
- Dense Trajectories extraction with clustering [J. C. Gemert *et al*., BMVC 2015]

  - Low-level hand-crafted features
  - Only local temporal context

- Human-centric RPN with optical flow motion estimation [N. Li *et al*., ACCV 2016]

  - Frames are processed individually
  - Not end-to-end

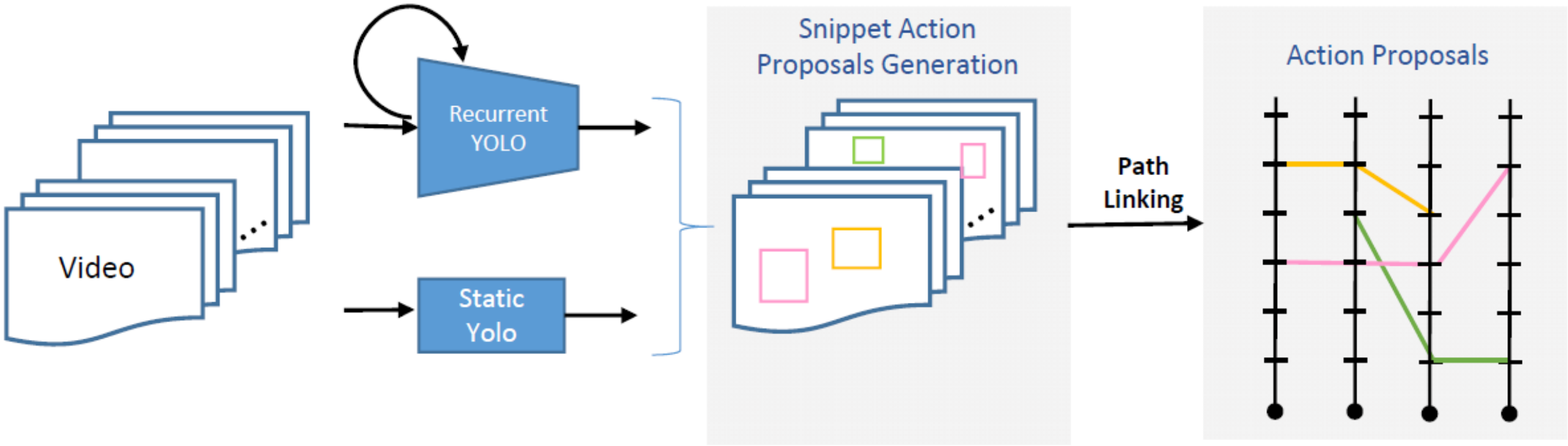
Our model for action proposal
Motivations

Based on previous approaches limitations, we want:

- to handle long-term temporal relationship for bounding box regression
- in an end-to-end framework
- with seamless integration of the bounding box linking method

Our model for action proposal

Overall architecture

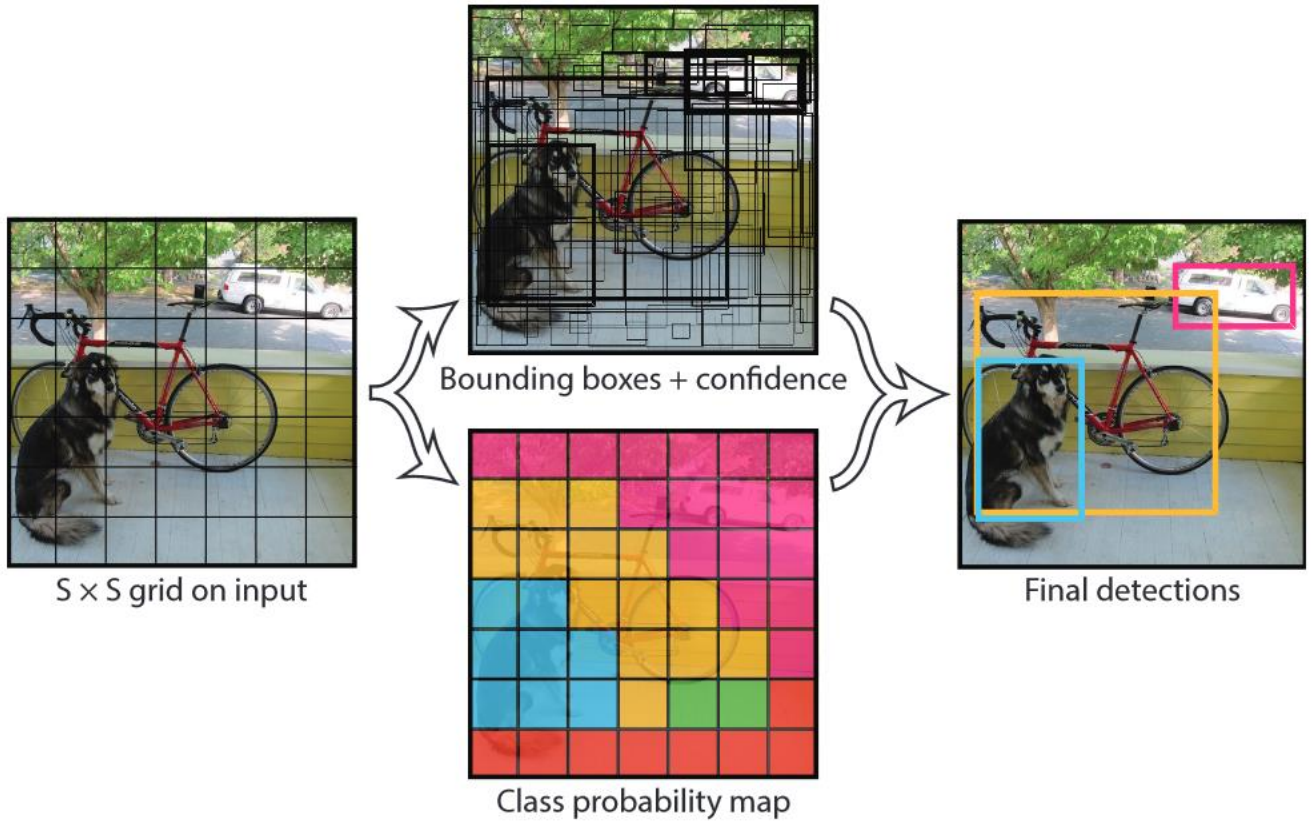


Our model for action proposal

YOLO detector [J. Redmon et al., CVPR 2016]

Frame processing:

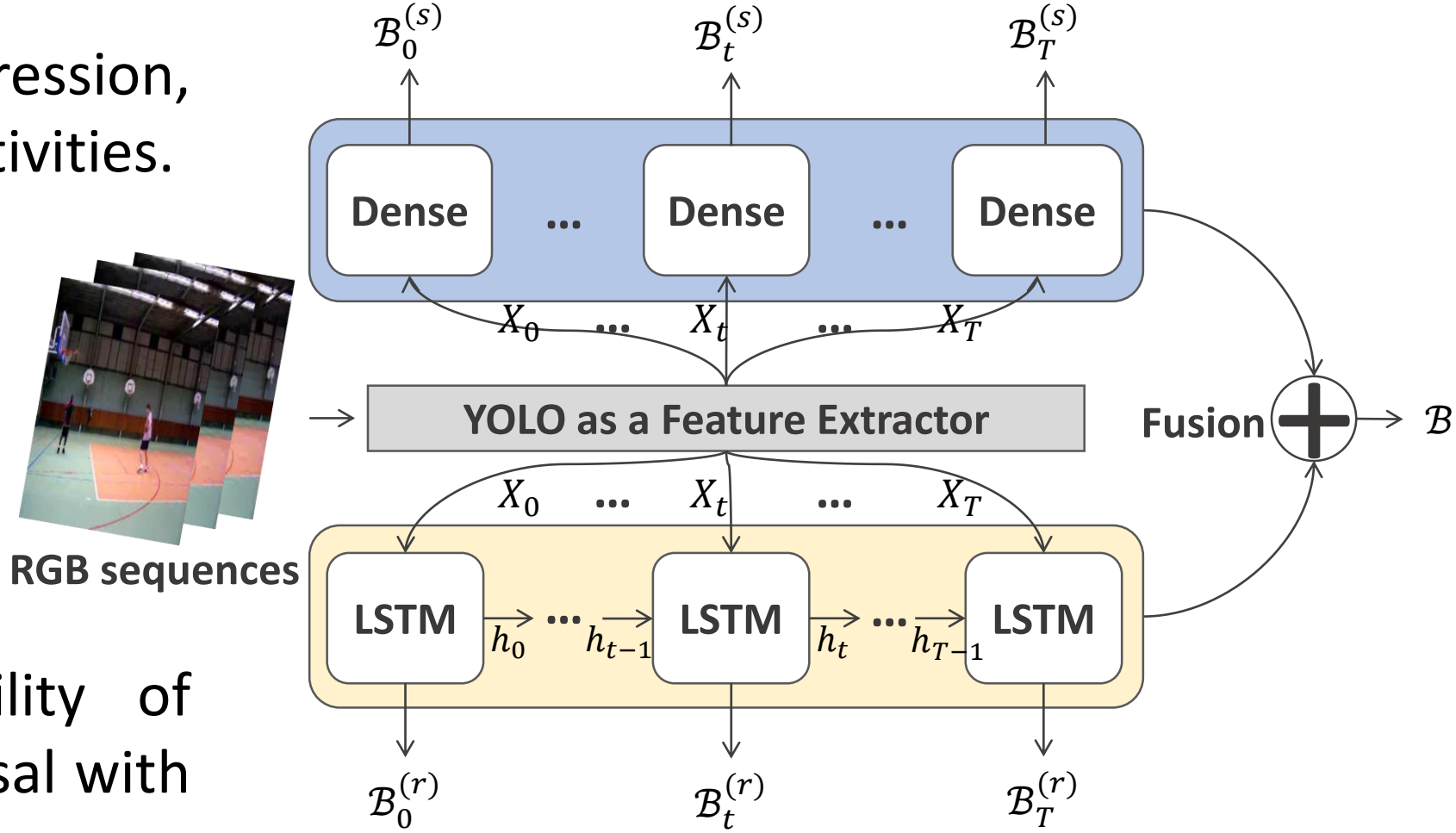
- Divide the image in a $S \times S$ grid
- Predict $B = 2$ bounding boxes per grid cell with a confidence score C (human estimator)
- Predict one actionness (s_{ac}) and backgroundness (s_{bc}) score per grid cell (motion estimator)



Our model for action proposal

Temporal enhancement with LSTM

Use frame-by-frame regression, useful for motionless activities.



Use regression capability of LSTM to produce proposal with long-term relationship.

Our model for action proposal

Path generation

From the frame proposal method we have a set of bounding boxes:

$$\mathcal{B} = \{\mathcal{B}_t = \left\{b_t^1 \dots b_t^{|\mathcal{B}_t|}\right\}, \qquad t \in [1 \dots T]\}$$

We want to output a set of proposal paths:

$$\mathrm{P} = \{p_i = \{b_{s_i}, b_{s_i+1}, \dots, b_{e_i}\}, \qquad i \in [1 \dots |\mathrm{P}|]\}$$

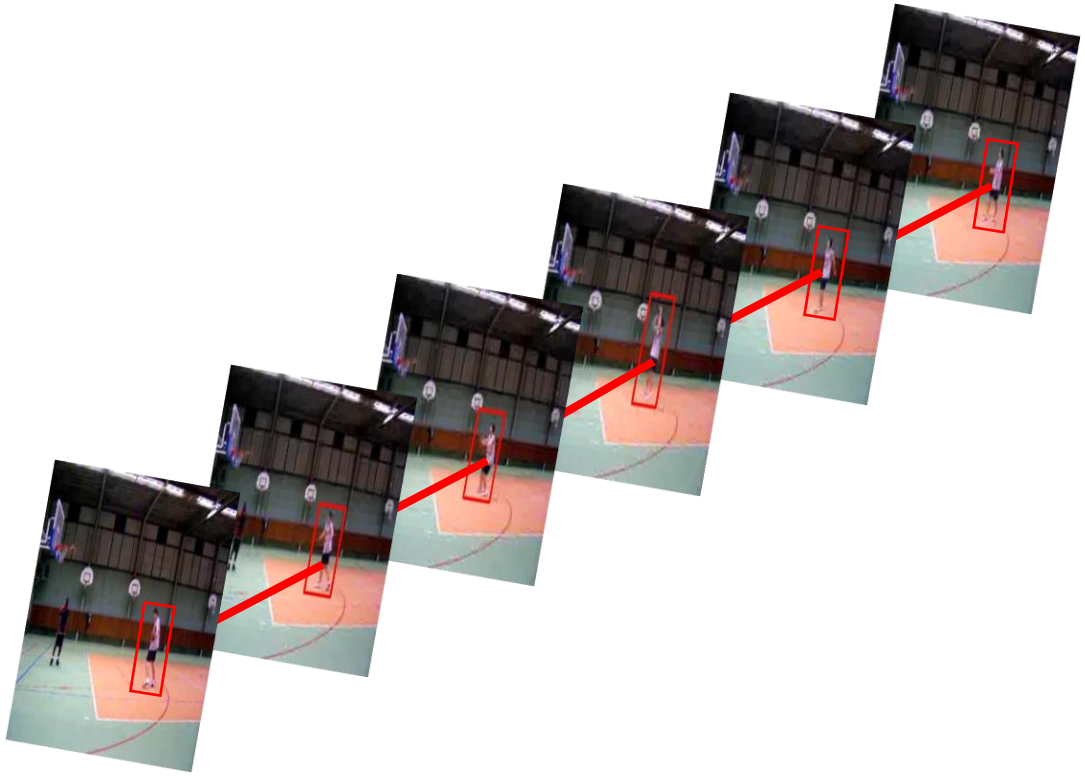
Our model for action proposal

Path generation

Path Linking:

High confidence, coherent paths that span the entire video duration

$$S(p) = \underbrace{\sum_{i=1}^{T} C(b_i)}_{unary} + \lambda_0 \times \underbrace{\sum_{i=2}^{T} IoU(b_i, b_{i-1})}_{pairwise}$$



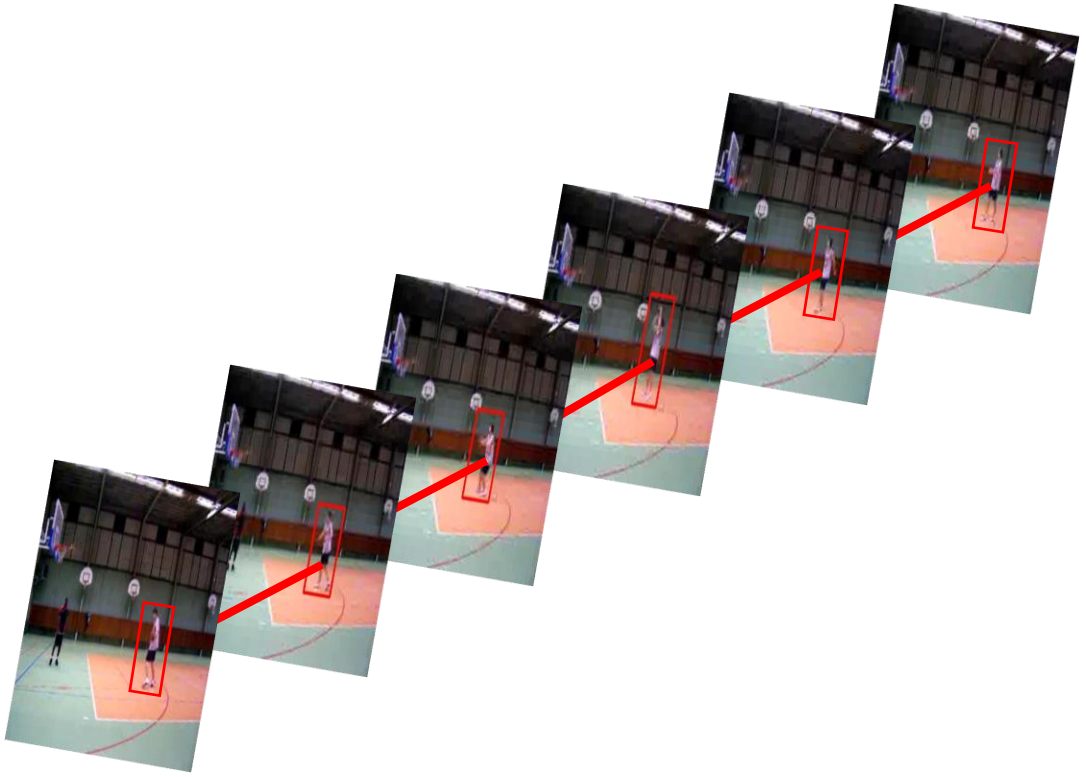
Our model for action proposal

Path generation

Path Trimming: [S. Saha *et al.*, BMVC 2016]

Trimmed paths with high human action likelihood

$$S(p) = \underbrace{\sum_{i=1}^{T} s_{l_i}(b_i)}_{unary} - \lambda_1 \times \underbrace{\sum_{i=2}^{T} 1_{\{l_i \neq l_{i-1}\}} \times \alpha_{l_i}}_{pairwise}$$



where $l_i \in \{ac, bg\}, \quad \alpha_{l_i} > 0$

Experiments

Dataset

UCF 101 [K. Soomro *et al.*, CRVC-TR-12-01]

- Widely used dataset (600+ citations) released in 2012
- 13k+ videos in 101 action categories
- We use a subset of 24 classes with bounding box annotations of human
  [Y.-G. Jiang *et al.*, ICCV Action Workshop 2013]

Experiments

Metrics

Intersection over Union:

$$IoU(\boldsymbol{p}, \boldsymbol{g}) = \frac{1}{|\boldsymbol{p} \cup \boldsymbol{g}|} \times \sum_{i \in \boldsymbol{p} \cap \boldsymbol{g}} \frac{b_i^{\boldsymbol{p}} \cap b_i^{\boldsymbol{g}}}{b_i^{\boldsymbol{p}} \cup b_i^{\boldsymbol{g}}}$$

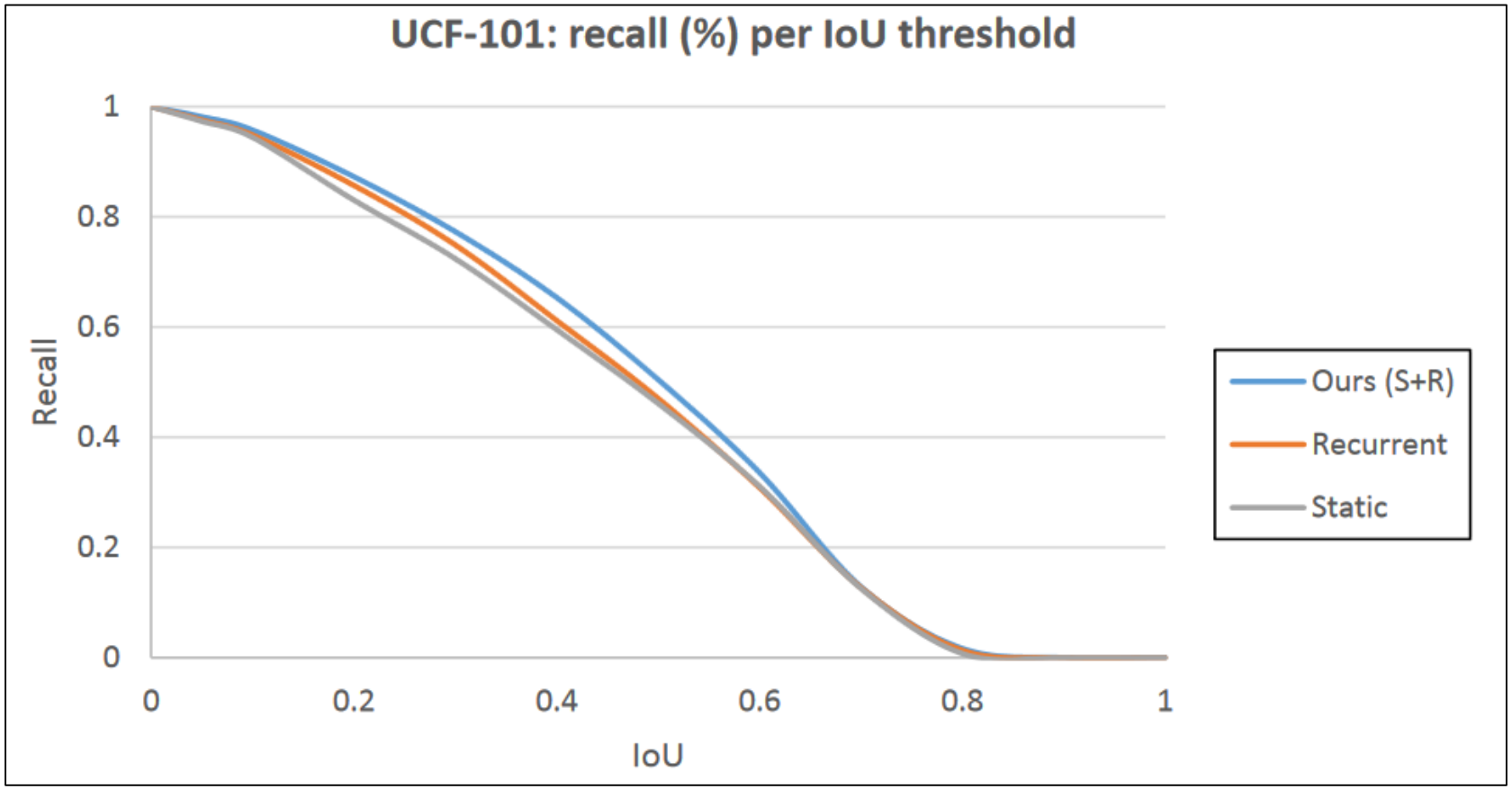
Average Best Overlap:

$$ABO(P, \boldsymbol{G}) = \frac{1}{|\boldsymbol{G}|} \sum_{\boldsymbol{g} \in \boldsymbol{G}} \max_{\boldsymbol{p} \in P} IoU(\boldsymbol{p}, \boldsymbol{g})$$

$$MABO = \frac{1}{|\boldsymbol{C}|} \sum_{c \in \boldsymbol{C}} ABO(P, \boldsymbol{G^c})$$

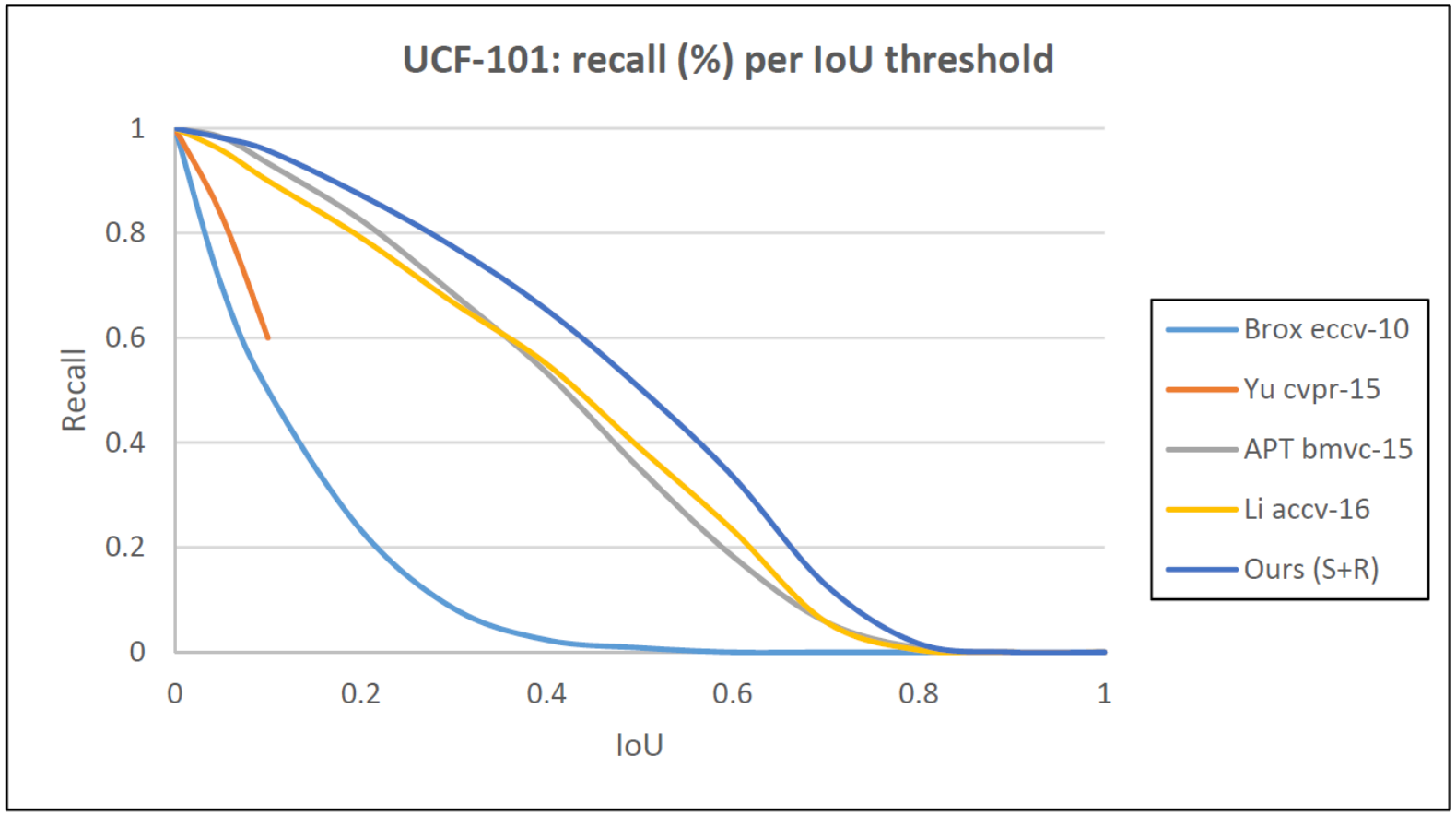
Experiments
Quantitative results

- Recurrent version is slightly better than static one
- +8% in recall by ensembling static and recurrent versions at 0.5 IoU

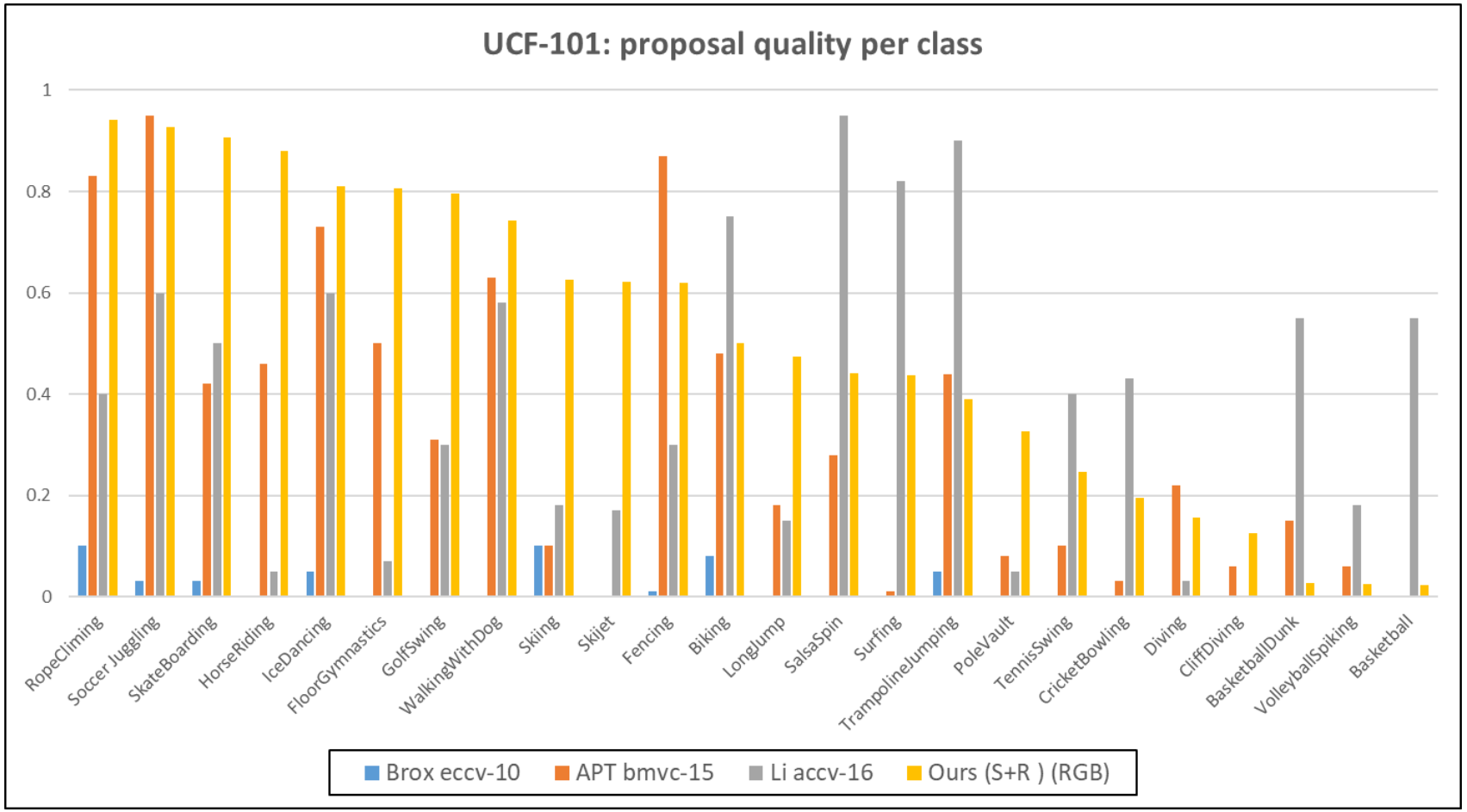


Experiments
Quantitative results

UCF101	ABO	MABO	Recall	#
Brox et al.	13.2	12.82	1.40	**3**
Yu et al.	n.a	n.a	0.0	10k
APT	40.8	39.97	35.45	2k
Li et al.	**63.8**	40.84	39.64	18
Ours	47.5	**47.72**	**50.46**	35

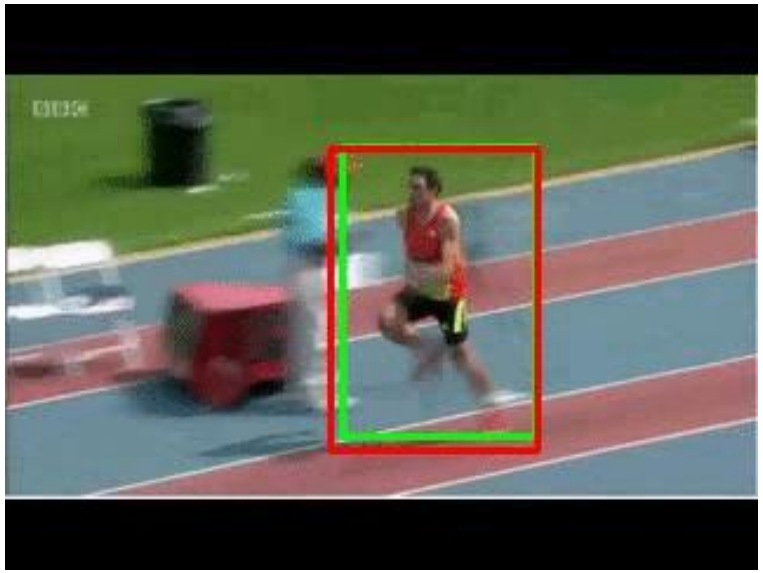


Experiments
Quantitative results



- RopeCliming (Easiest)
- Soccer Juggling
- SkateBoarding
- HorseRiding
- IceDancing
- FloorGymnastics
- GolfSwing
- WalkingWithDog
- Skiing
- Skijet
- Fencing
- Biking
- LongJump
- SalsaSpin
- Surfing
- TrampolineJumping
- PoleVault
- TennisSwing
- CricketBowling
- Diving
- CliffDiving
- BasketballDunk
- VolleyballSpiking
- Basketball (Challenging)

Experiments
Qualitative results



Conclusion

- We handled **long-term temporal relationship with LSTM** for regressing bounding boxes in an **end-to-end architecture**
- We formulated the path generation as an **energy-maximization problem** which considers both **actionness measure and temporal overlap**
- We validated our approach on **UCF-101 dataset** and proved that our method achieves **state-of-the-art performance**

Thank you!

- Email: romain.vial@mines-paristech.fr
- Website: http://romainvial.xyz