CCNY at TRECVID 2015: Localization

Yuancheng Ye[1], Xuejian Rong[2], Xiaodong Yang[3], and YingLi Tian[1,2]

[1]Graduate Center and [2]City College, City University of New York

[3]NVIDIA Research

Introduction

- **What we did?** We present a novel video-based object localization system based on R-CNN (Regions with CNN features).
- **Main Contributions:** We propose the region trajectory algorithm which can keep tracking possible object regions while pruning false detections.
- **Results:** Our system ranks
  - **First** in the temporal measurement.
  - **Third** in the spatial measurement.

Task Description

- Determine the presence of the concept temporally within the shot.
- For each frame that contains the concept, locate a bounding rectangle spatially.

Challenges

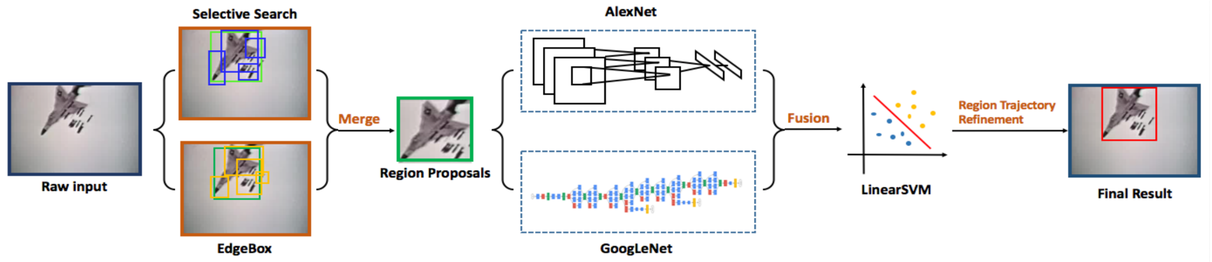
- How to locate object bounding box on each frame accurately?
- How to extend image-based object detection algorithms into the temporal dimension?

Our solution:

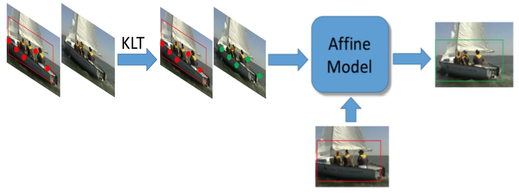
Regions with Convolutional Neural Network Features

Region Trajectory algorithm

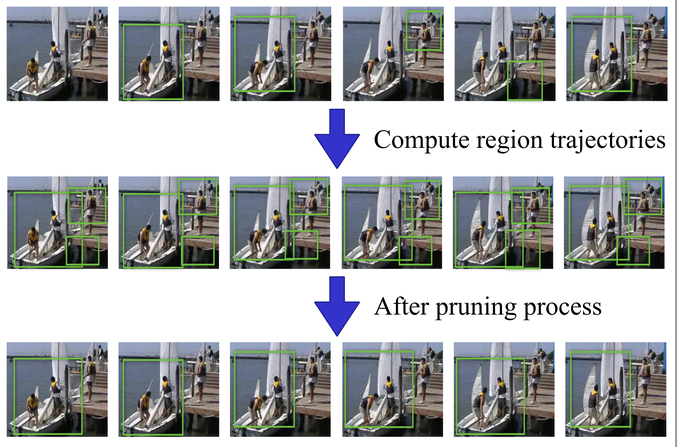
System Overview



Region Trajectory Algorithm



Set of detected regions → Set of aligned trajectories



Prune trajectories:

$$ratio = \frac{number\ of\ regions\ detected\ by\ R\text{-}CNN}{total\ number\ of\ regions\ in\ the\ trajectory}$$

Evaluation Metrics

- Precision, Recall and F-Score are calculated based on temporal and spatial results respectively.
- Averages are computed for values of each concept.
- Computing units are frames (temporally) and pixels (spatially).

$$F\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Experiments

- Testing Concepts

1003 Airplane	1005 Anchorperson	1015 Boat_ship	1017 Bridges
1019 Bus	1031 Computers	1080 Motorcycle	1117 Telephones
1261 Flags	1392 Quadruped		

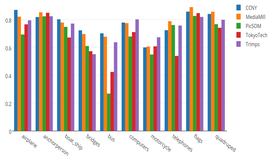
- Auxiliary Data
  - AlexNet model is pre-trained on the PASCAL VOC 2007 dataset.
  - GoogLeNet model is pre-trained on the ILSVRC12 dataset.
- Data Statistics

	1003	1005	1015	1017	1019	1031	1080	1117	1261	1392
Positive frames	710	3482	7055	1380	860	4111	1835	3272	8429	6315
Negative frames	548	0	4156	1537	2288	0	2036	2064	3156	8595
Test frames	7047	14119	5874	6054	4774	15814	4165	5851	19092	13949

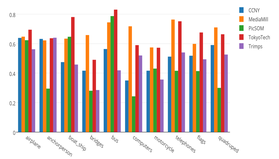
- Results:

Run	iframe_fscore	mean_pixel_fscore
1	0.7447	0.4723
2	0.7682	0.4542
3	0.7309	0.5085
4	0.7661	0.4591
MediaMill*	0.7662	0.6557
PicSOM*	0.6643	0.3944
TokyoTech*	0.6699	0.6688
Trimps*	0.7357	0.4760

Table 1: The results of Mean_Per_Run for four submitted runs. * indicates the best results of other teams among all their submitted runs.



iframe_fscore per concept



mean_pixel_fscore per concept