CCNY at TRECVID 2015: Localization

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

[1]The Graduate Center, CUNY

[2]The City College of New York, CUNY

[3]NVIDIA Research

Task Description

- Concepts

Airplane

Anchorperson

Boat_ship

Bridge

Bus

Computer

Motorcycle

Telephone

Flags

Quadruped

- Determine the presence of the concept temporally within the shot

- For each frame that contains the concept, locate a bounding rectangle spatially

- Only one which is the most prominent among all submitted bounding boxes will be used in the judging.

Challenges

- How to locate object bounding box on each frame accurately?

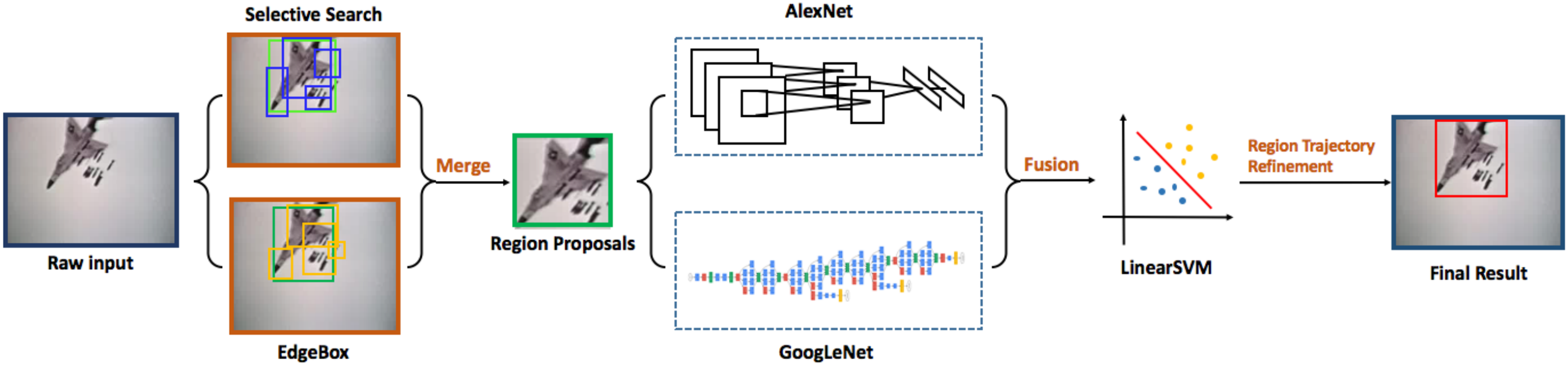
- How to extend the image-based object detection algorithms into the temporal domain?

Our solution:

Regions with Convolutional Neural Network Features(R-CNN)

Region Trajectory Algorithm

System Overview



- Apply improved image-based R-CNN algorithm on each frame independently.

- Propose a novel region trajectory algorithm to extend to temporal dimension.

Improved R-CNN

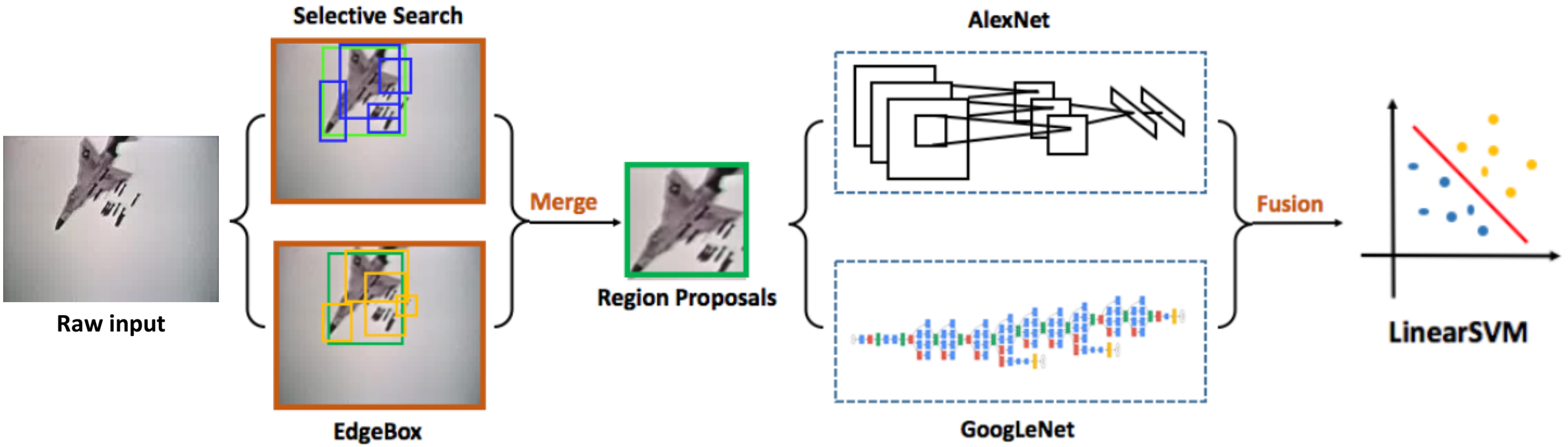


Image ➡ Region Proposals ➡ CNN Features ➡ Classification

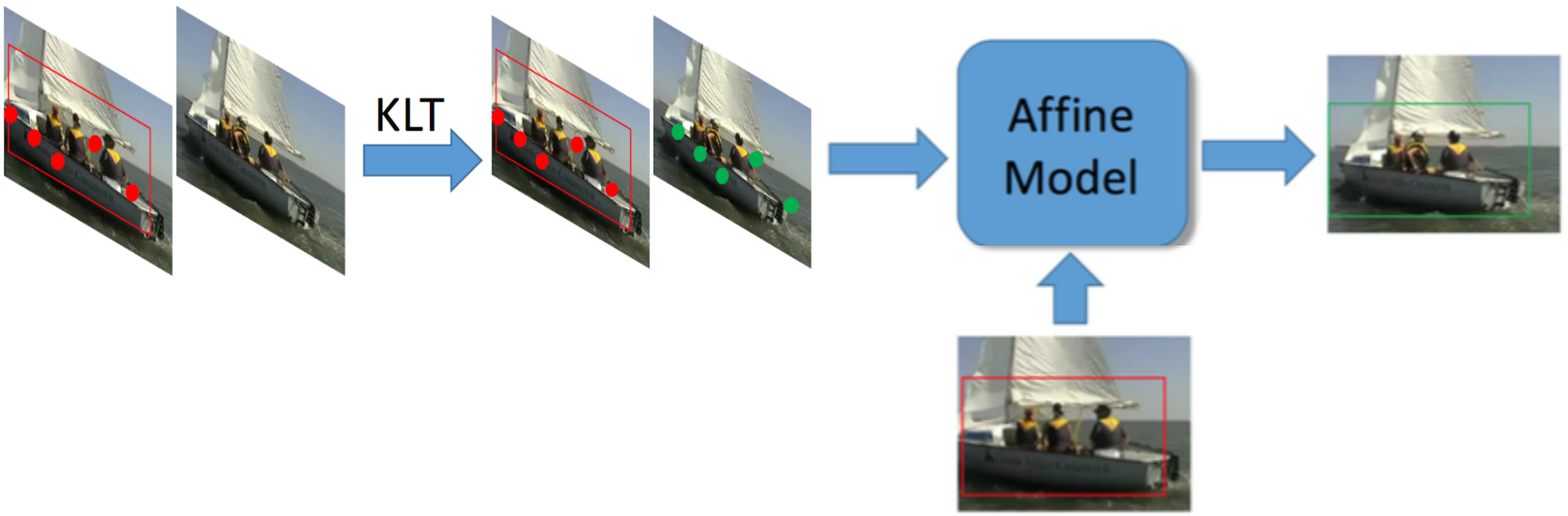
- Insufficient for object localization in videos

Input:

Output:

How to incorporate temporal info?

Region Trajectories



Set of detected regions → **Set of aligned trajectories**

However……

Input:

Output:

So many plausible trajectories are introduced!

Prune trajectories

- Threshold

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

Output after pruning:

Data

- Training data:
  - Internet Archive videos with Creative Commons licenses (IACC).
  - IACC.2.A, IACC.2.B
    - Totally 100 GB, 400h.
    - Size mostly 320 x 240.
    - Ranging from 10s to 6.4m.
    - Manual (temporal and spatial) annotations provided (.xml format).

- Auxiliary Data
  - AlexNet model is pre-trained on the PASCAL VOC 2007 dataset.
  - GoogLeNet model is pre-trained on the ILSVRC12 dataset.
- Testing data:
  - IACC.2.C:
    - A collection of 200h drawn randomly from the IACC.2 collection.
    - Size mostly 320 x 240.
    - 18 GB of Master I-Frames will be extracted for evaluation.

- Data Format:
  - I-frames: a sequence of key frames defines which movement the viewer will see, whereas the position of the key frames on the film, video, or animation defines the timing of the movement.
- Data Statistics

	airplane	anchor person	boat_ship	bridges	bus	computers	motorcycle	telephones	flags	quadruped
Positive I-frames	710	3482	7055	1380	860	4111	1835	3272	8429	6315
Negative I-frames	548	0	4156	1537	2288	0	2036	2064	3156	8595
Test I-frames	7047	14119	5874	6054	4774	15814	4165	5851	19092	13949

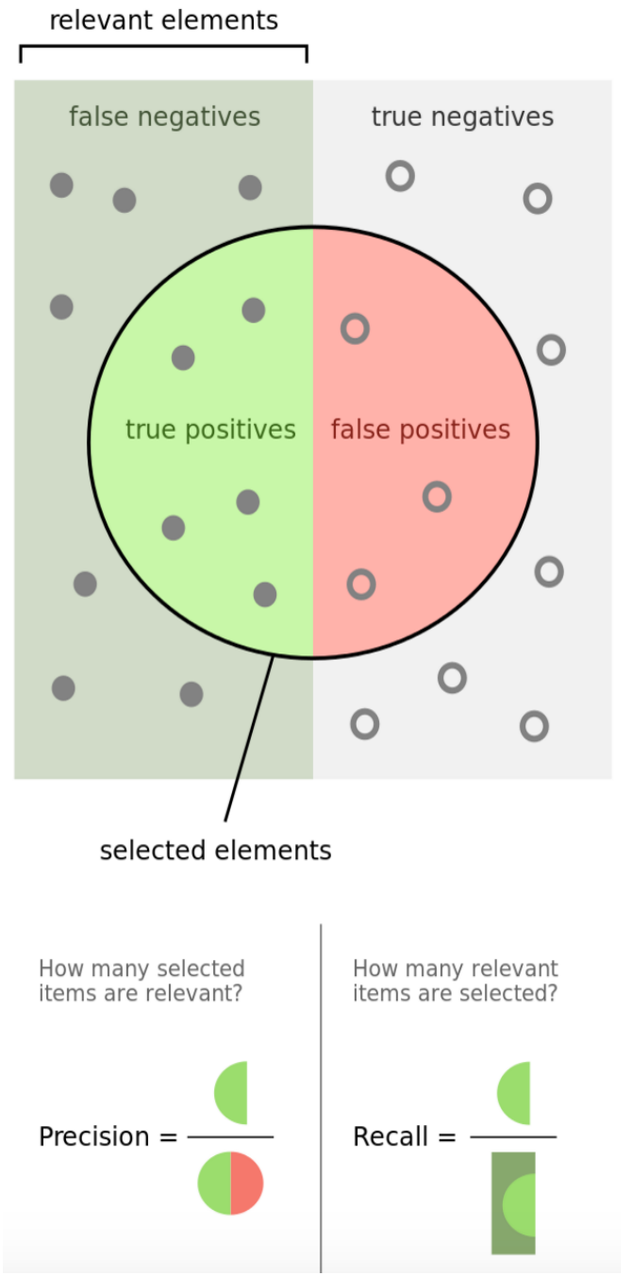
Evaluation Metrics

- Precision, Recall and F-Score are calculated based on temporal and spatial results respectively.

- Averages are computed for values of each concept.

- The computing units are frames (temporally) and pixels (spatially).

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



(from Wikipedia)

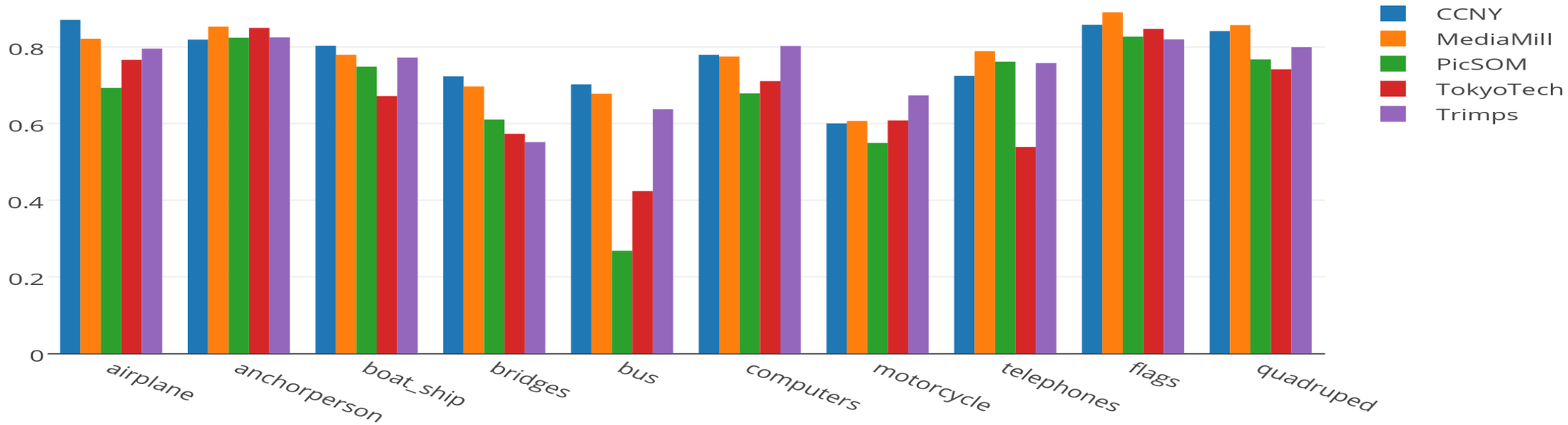
Results

- Mean_Per_Run

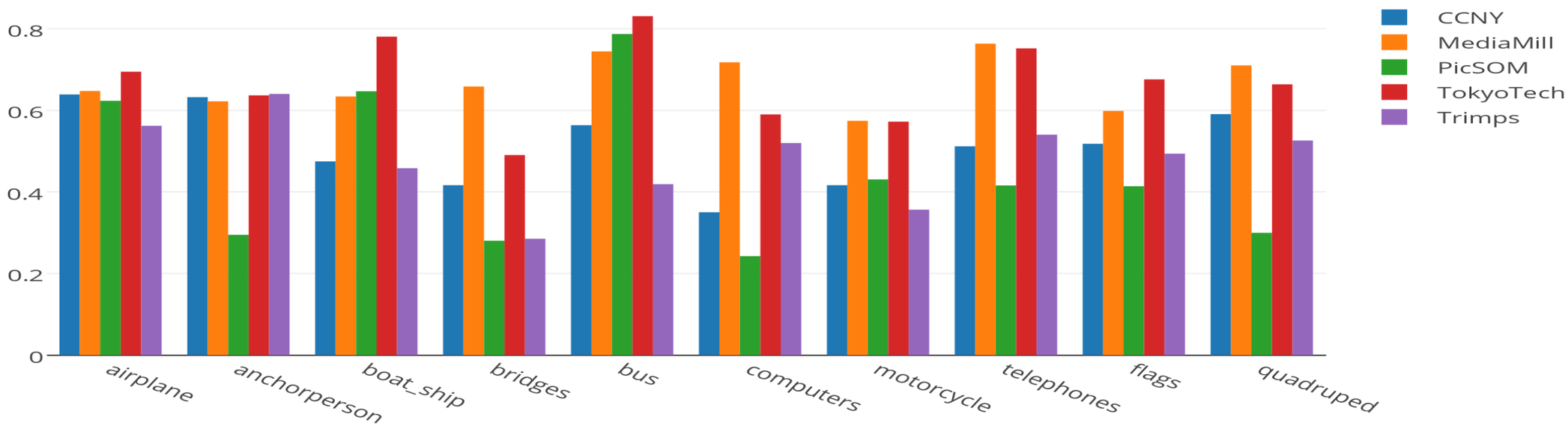
Run	iframe_fscore	mean_pixel_fscore
1	0.7447	0.4723
2	0.7682	0.4542
3	0.7309	0.5085
4	0.7661	0.4591
MediaMill*	0.7662	0.6557
PicSOM*	0.6643	0.3944
TokyoTech*	0.6699	0.6688
Trimps*	0.7357	0.4760

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

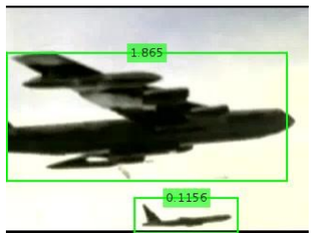


- mean_pixel_fscore per concept

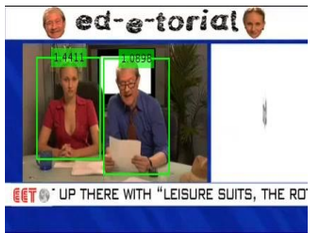


Results Visualization

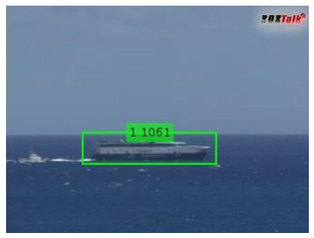
- Success Examples



Airplane

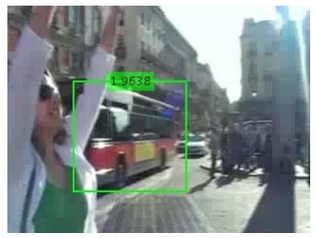


Anchorperson

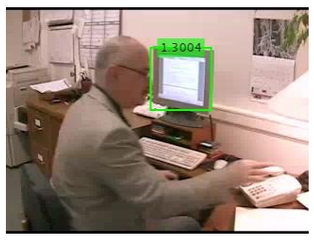


Boat_ship

Bridge



Bus

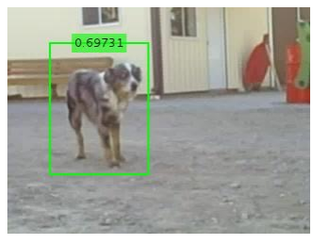


Computer

Motorcycle

Telephone

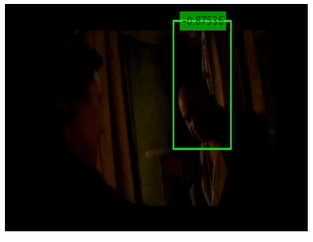
Flags



Quadruped

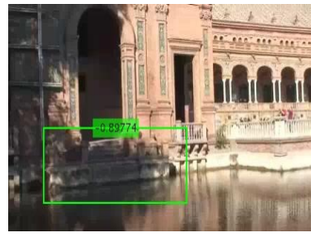
- Failure Examples

Airplane



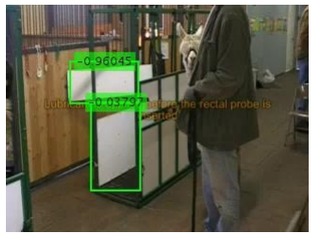
Anchorperson

Boat_ship

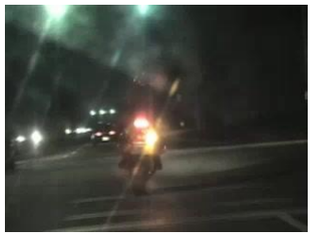


Bridge

Bus

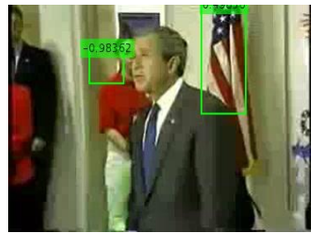


Computer

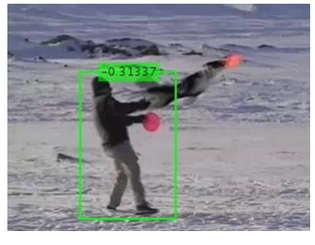


Motorcycle

Telephone



Flags



Quadruped

Conclusion

- By combining R-CNN and region trajectory algorithm, we propose a robust and effective system for video-based object detection task.

- Temporal information can make a contribution to the object detection task in videos.

- Among all participant teams, we rank 1st for the measurement of iframe_fscore, and 3rd for the measurement of mean_pixel_fscore.

Future Work

- Incorporate more accurate image-based object detection algorithms, e.g., Fast-RCNN.

- Improve the region trajectory algorithm for higher spatial accuracy.

- Adopt model ensembles to extract more discriminative features from region proposals.

Thank you