Localization with Spatio-Temporal Selective Search and SPPnet

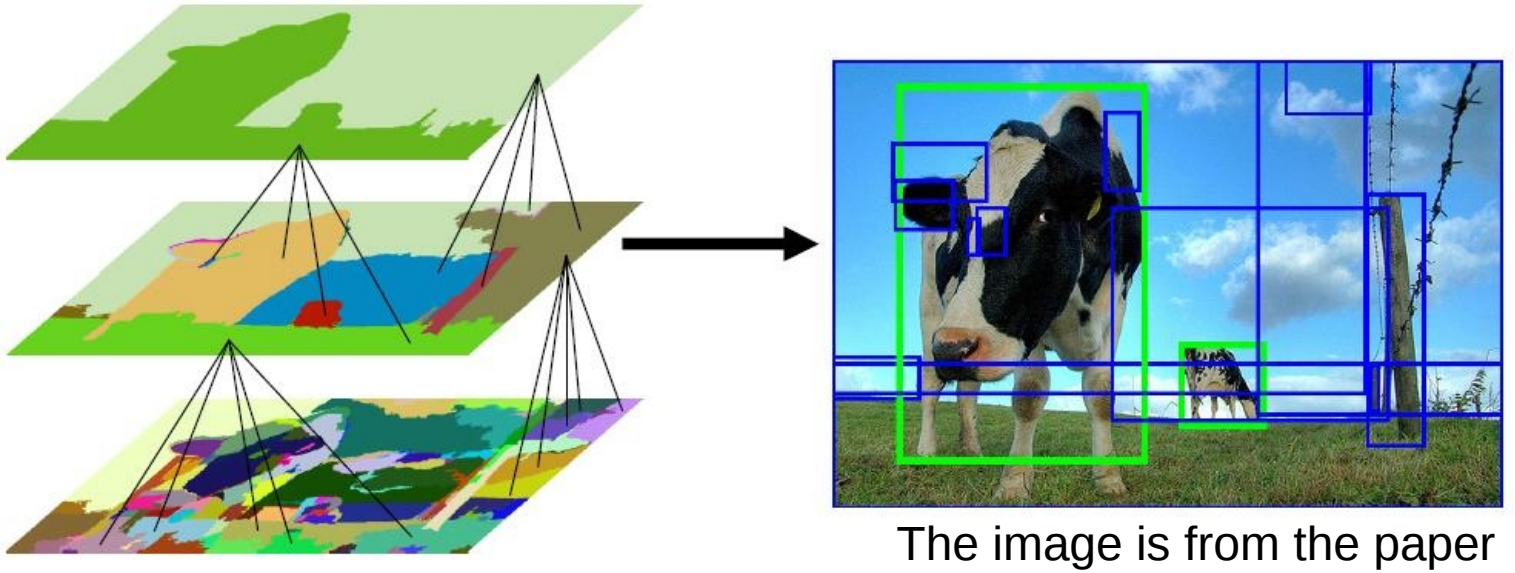
Ryosuke Yamamoto, Nakamasa Inoue, Koichi Shinoda
Tokyo Institute of Technology

Outline

- Previous works
  - Selective Search
  - Spatial Pyramid Pooling (SPP) net
- Our Methods
  1. Spatio-Temporal Selective Search
  2. Multi-Frame Score Fusion
  3. Neighbor-Frame Score Boosting
- Experiments, Results and Conclusion

Selective Search

- Selective Search produces a large number of object region proposals from an image
  - Use several strategies including useless ones

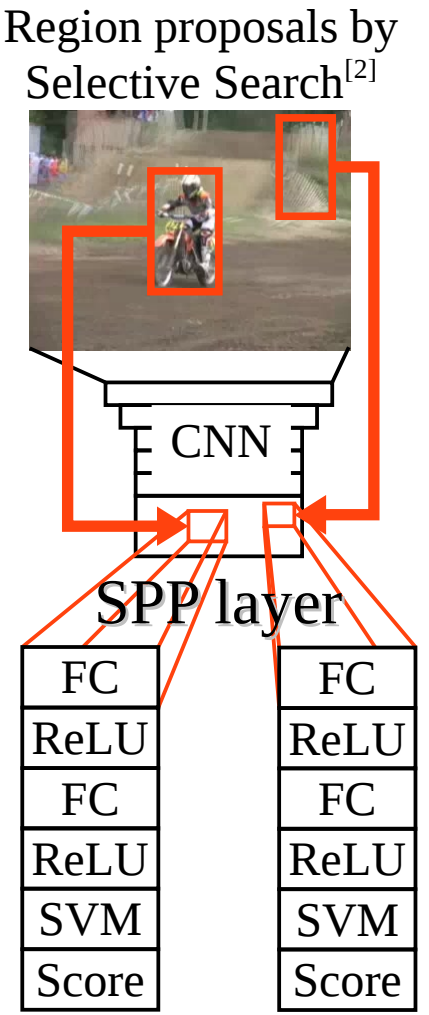


The image is from the paper

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders, Selective search for object recognition. In IJCV, vol.104, pp.154-171, 2013

Spatial Pyramid Pooling (SPP) net

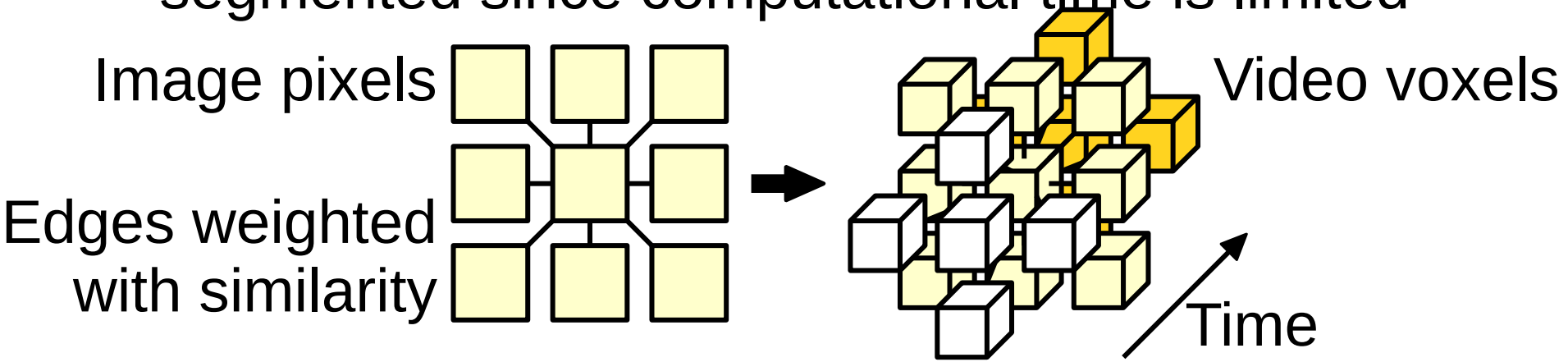
- An efficient method to extract CNN scores from a large number of object regions of an image
  - CNN layers shared among all regions
  - SVMs computed for each region
  - Selective Search is used for region proposals



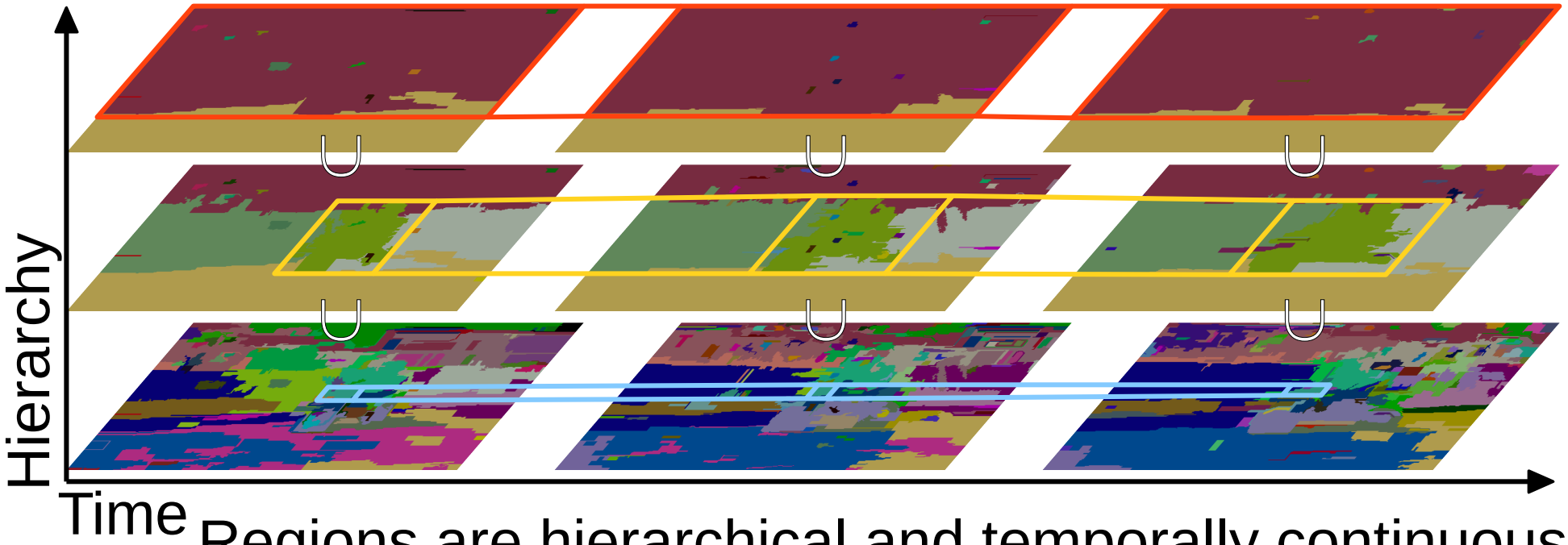
K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1904-1916, 2015

1.Spatio-Temporal Region Proposals(1)

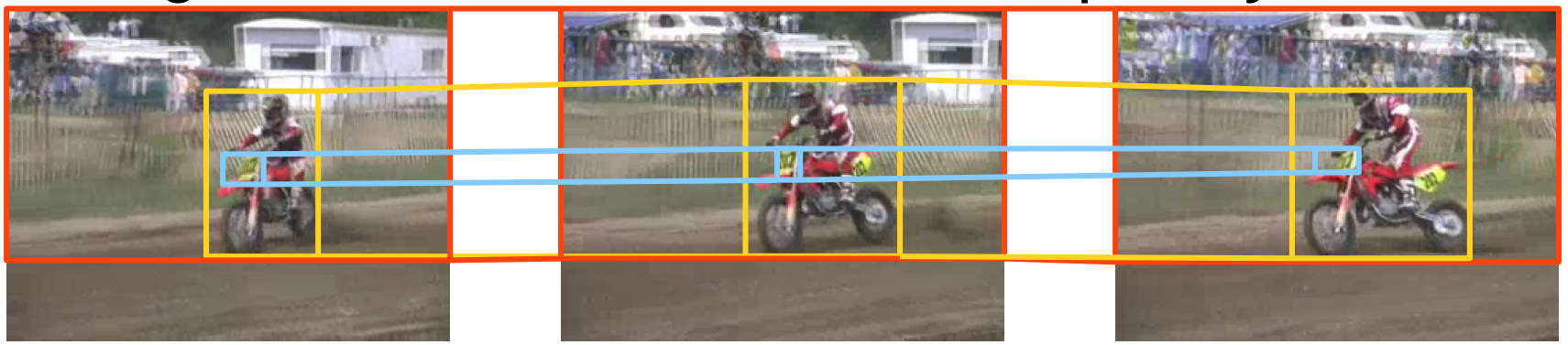
- Selective Search with temporal dimensional extended region proposals
  - Produce temporally continuous regions
  - Contains a large number of meaningless regions
  - Each video is separated at each I-frame and segmented since computational time is limited



1.Spatio-Temporal Region Proposals(2)



Regions are hierarchical and temporally continuous

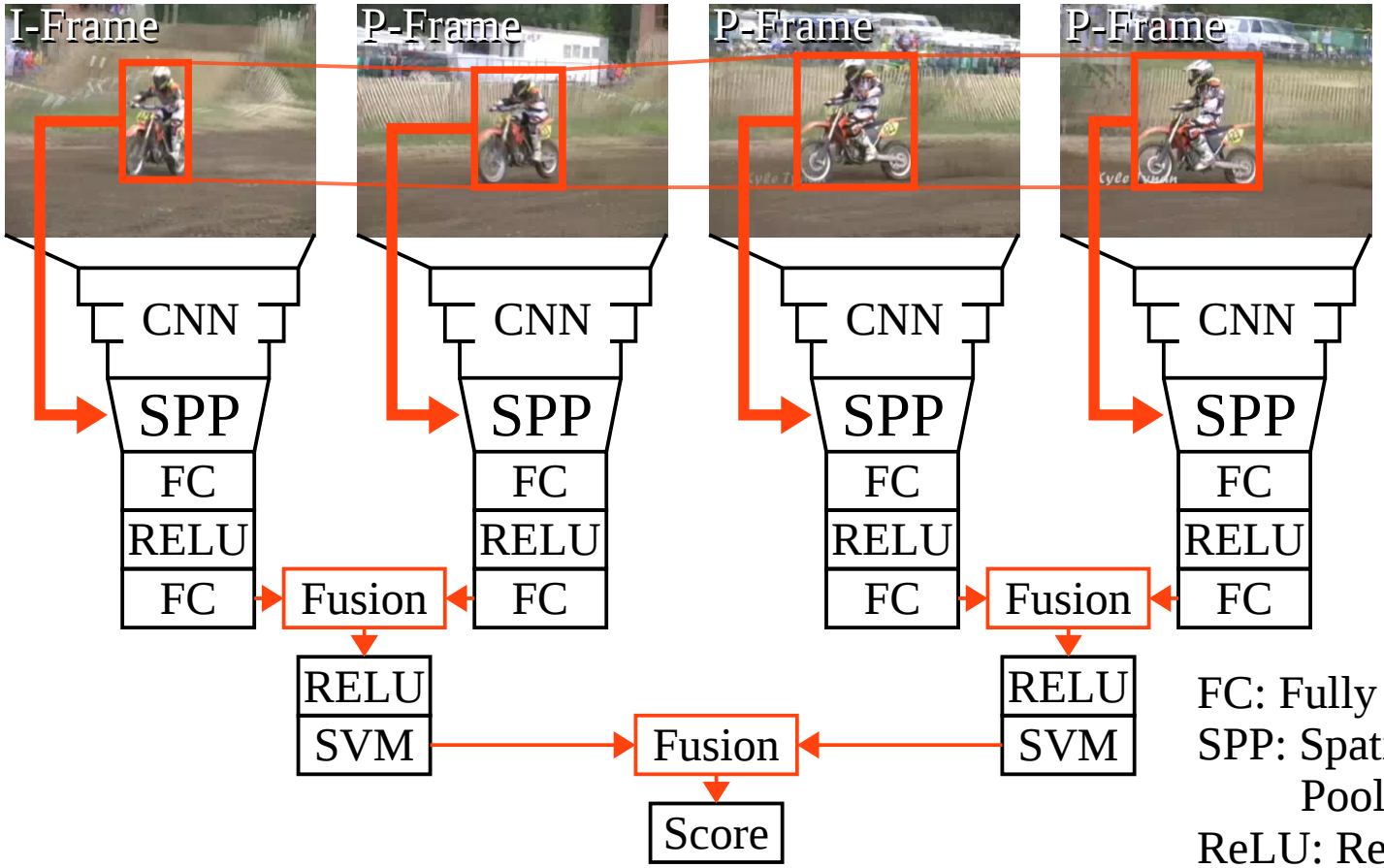


2.Multi-Frame Score Fusion (1)

- Basic idea
  - Some frames contain **noise** or **object deformation** making detection harder
  - Results of ST-Region Proposals contain many meaningless region proposals
  - ➔ Information of neighbor frames provides robustness
- Fuse feature maps among several frames
  - This requires region proposals temporal continuous – ST-Region Proposals adopted

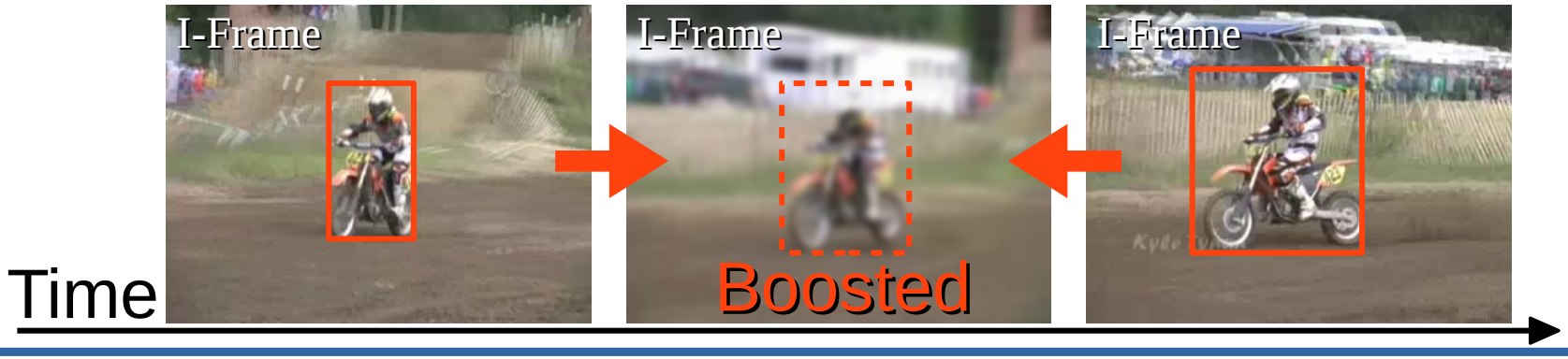
2.Multi-Frame Score Fusion (2)

- In experiments, we concluded late fusion is the best



3.Neighbor-Frame Score Boosting

- Basic idea
  - Based on same aspect of previous score fusion
  - Objects will appear in several continuous frames
  - ➔ Information of neighbor frames provides robustness
- Boost scores of I-frames between positives by Increase their scores by a constant



Experiments – Manual Annotations

- Airplane, Boat_Ship, Bridges, Bus, Motorcycle, Telephones, Flags, Quadruped – provided
- Anchorperson – annotated 12k I-frames
- Computers – annotated 7k I-frames

Experiments – Training

- Deciding the threshold and the fusion method
  - Used last year's dataset and concepts
  - Train: IACC_2_A
  - Val: IACC_2_B
- Submitted runs
  - Train: IACC_2_A including additional annotations, IACC_2_B
  - Test: IACC_2_C

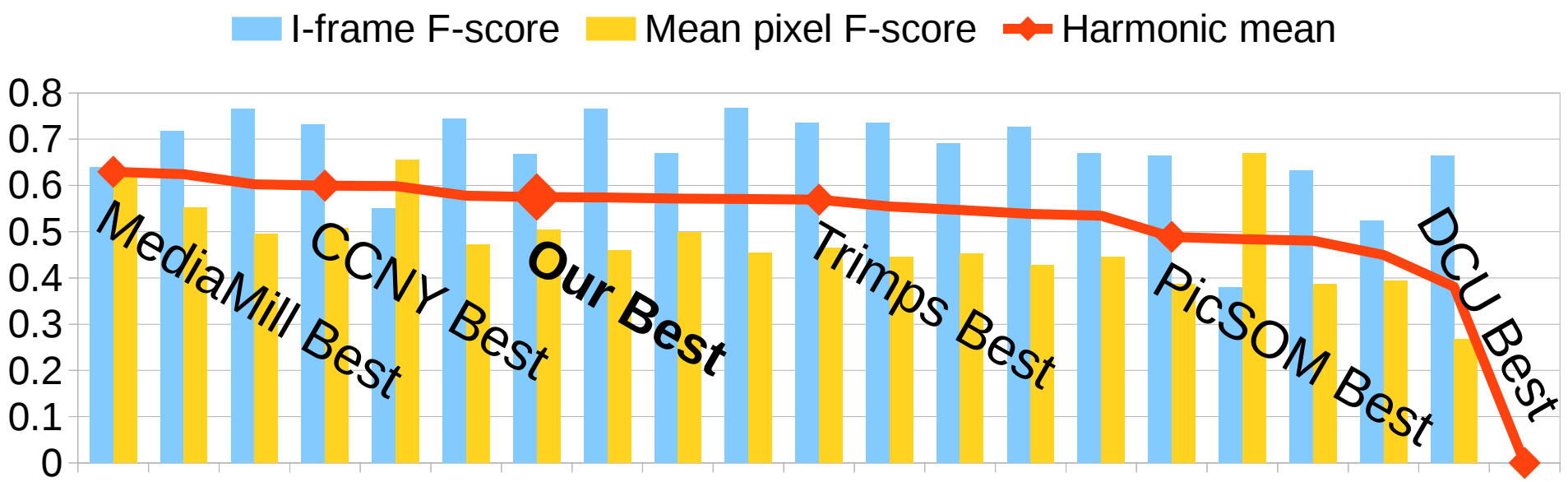
Results

- Multi-Frame Score Fusion and Neighbor-Frame Score Boosting improved the score
- We archived 3rd place among all teams with harmonic mean of F-scores

		Harm. Mean of F-scores	
Run ID	Method	Val	Test
(Base)	Selective Search + SPPnet	0.4481	0.5656
Multiple	+ ST-Region Proposals, Multi-Frame Score Fusion	0.4518	0.5716
Multiple_Aug3	+ Neighbour-Frame Score Boost	**0.4569**	**0.5750**

Results

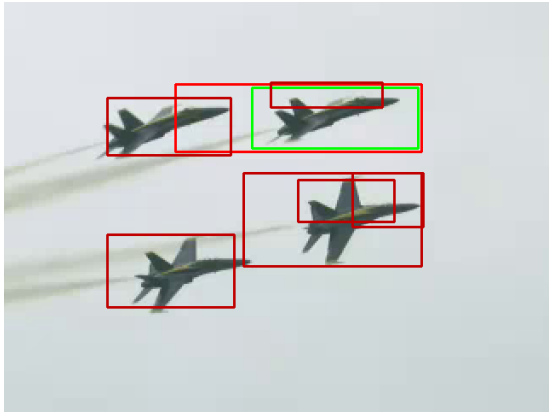
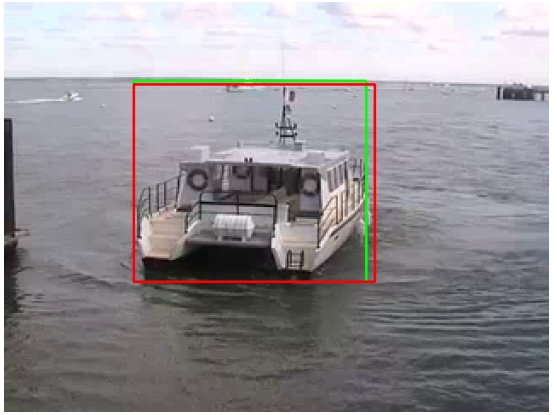
- Multi-Frame Score Fusion and Neighbor-Frame Score Boosting improved the score
- We archived 3rd place among all teams with harmonic mean of F-scores

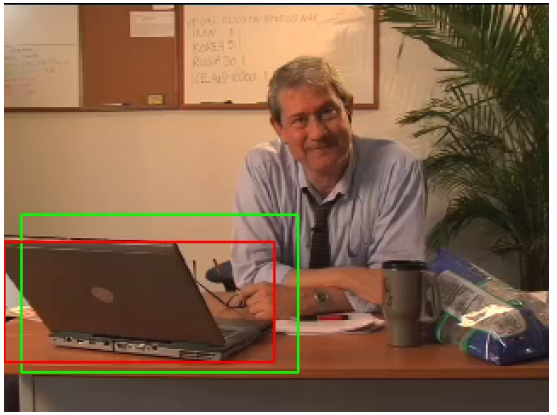
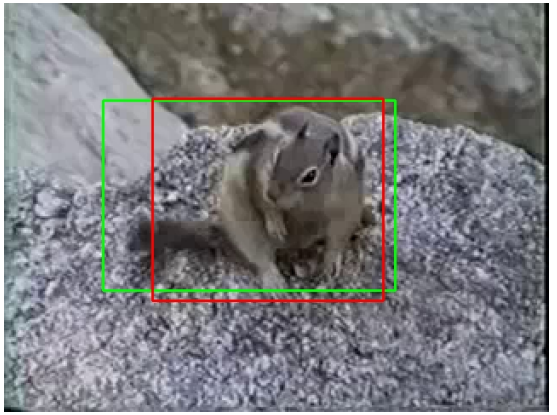


Results – Examples

- Sometimes better than GT

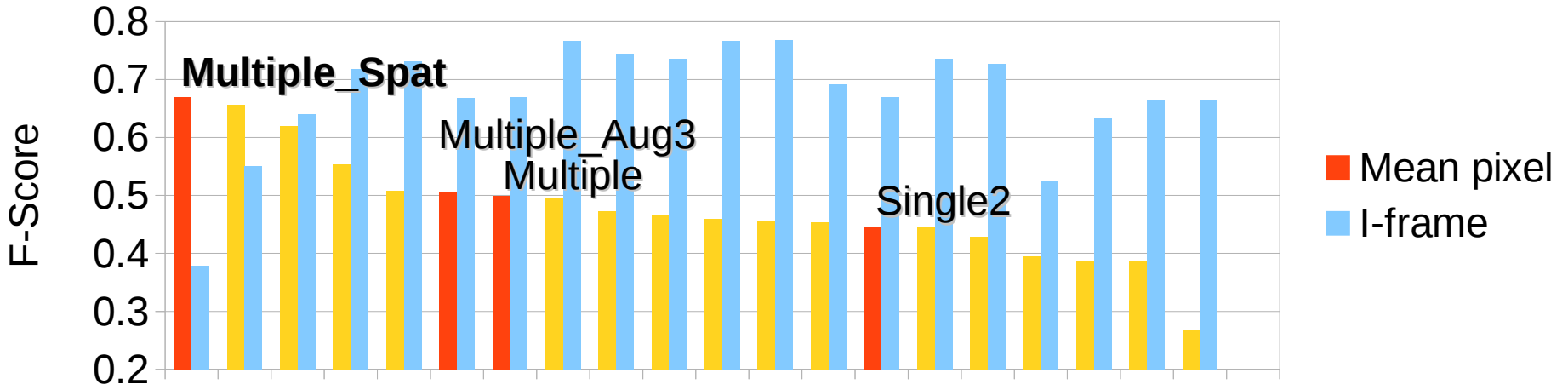
System output
Ground truth

Results – Spatial Score

- We achieved 1st place in Mean Pixel F-score by throttling a number of positives to reduce FPs
  - Of course I-frame F-score is not good



- Mean Pixel F-score is calculated from true positive and false positive I-frames, not intuitive

Conclusion

- We developed a localization system using ST-Region Proposals and CNN with SPP-net
- Multi-Frame Score Fusion with ST-Region Proposals and Neighbor-Frame Score Boosting improved the score
- Problem: The detection results strongly depend on quality of ST-Region Proposals
  - Improve ST-Region Proposals quality
  - Localization without region candidates