Localization using Faster R-CNN and Multi-Frame Fusion

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

Outline

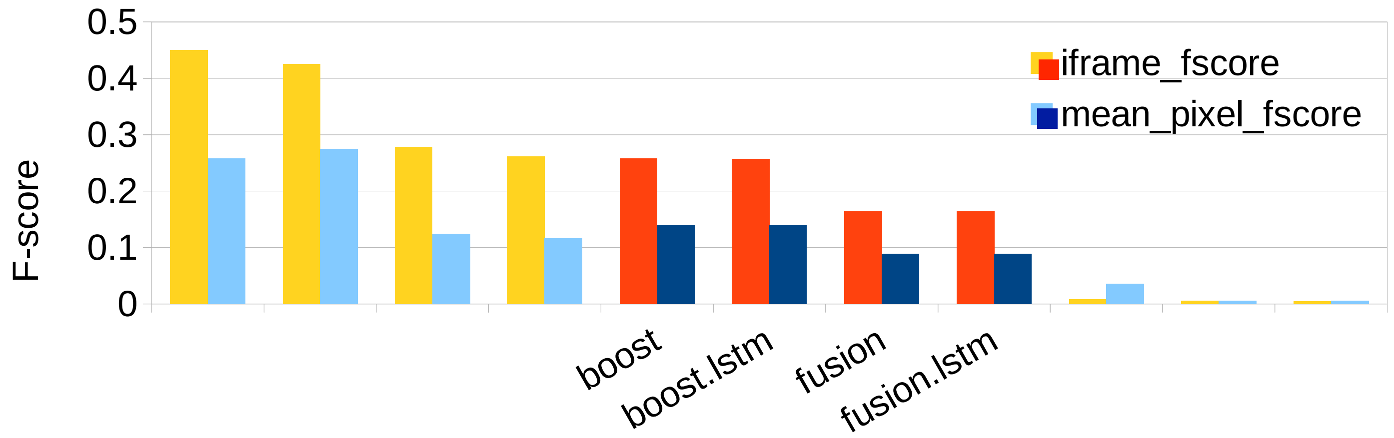
Motivation: detect an action concept "SittingDown"

Our method: Faster R-CNN + LSTM + Re-scoring

Annotation: Frame-wise annotation for SittingDown,
Key-frame annotation for other concepts

Results:

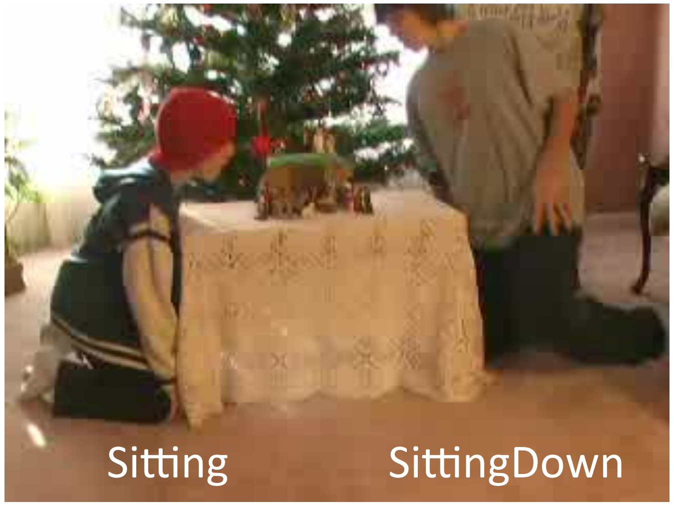
2nd among 3 teams, best result at SittingDown



Motivation

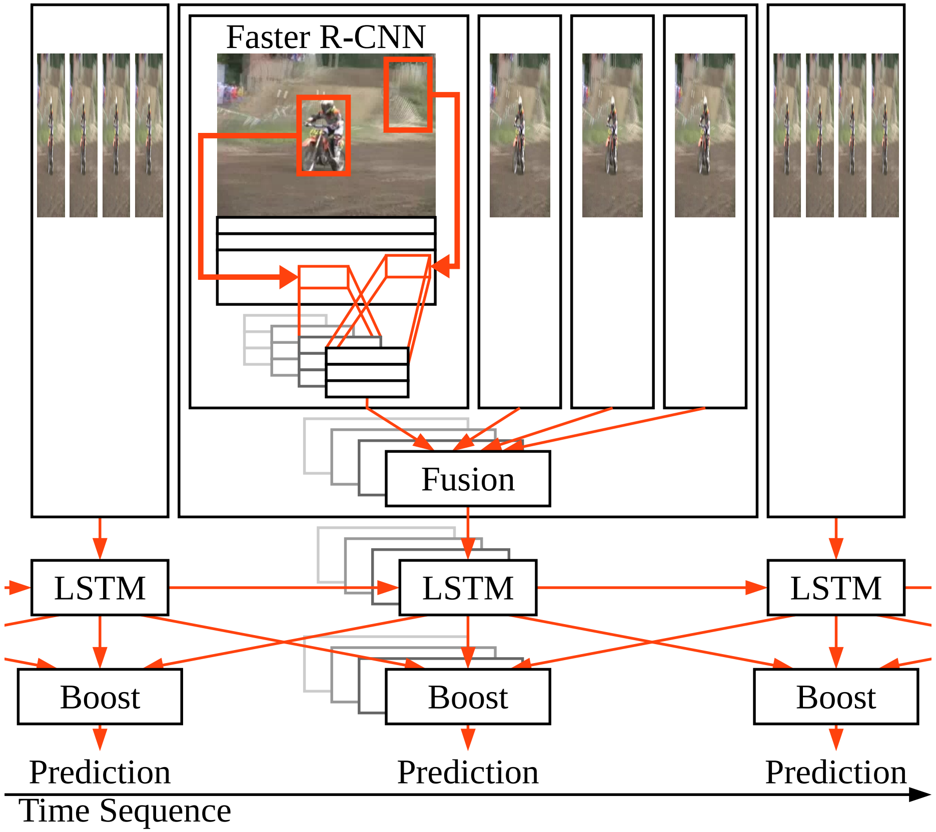
- Localization task focuses not only on static objects, but also on action concepts
- We focus on SittingDown, one of action concepts
- How to distinguish between Sitting and SittingDown?

→ Dynamic information is important for precise detection



Our Method

- Faster-RCNN (Ren 2015)
  - Efficient object localization
- LSTM (Donahue 2015)
  - Precise action localization
  - Applied to SittingDown
- Re-scoring (Yamamoto 2015)
  - Multi-frame Score Fusion
  - Multi-Shot Score Boosting



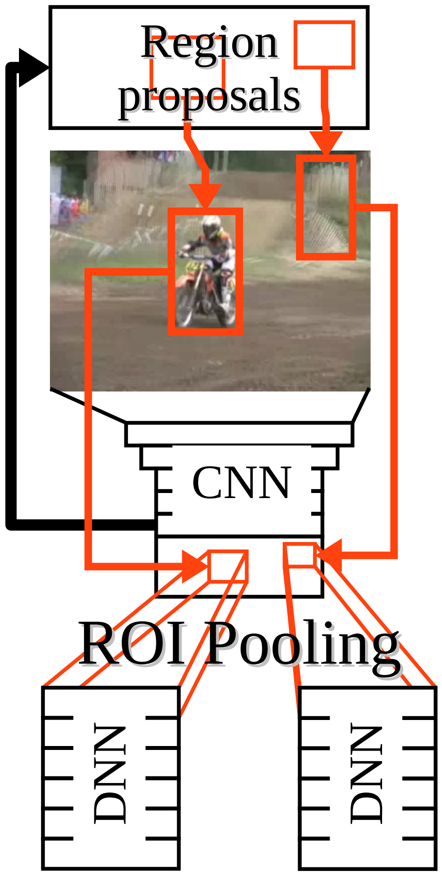
Faster R-CNN (Ren 2015)

Efficient End-to-End object localization

1. Generate region proposals by a network
2. Predict scores for each region by using CNN features

Example CNNs:

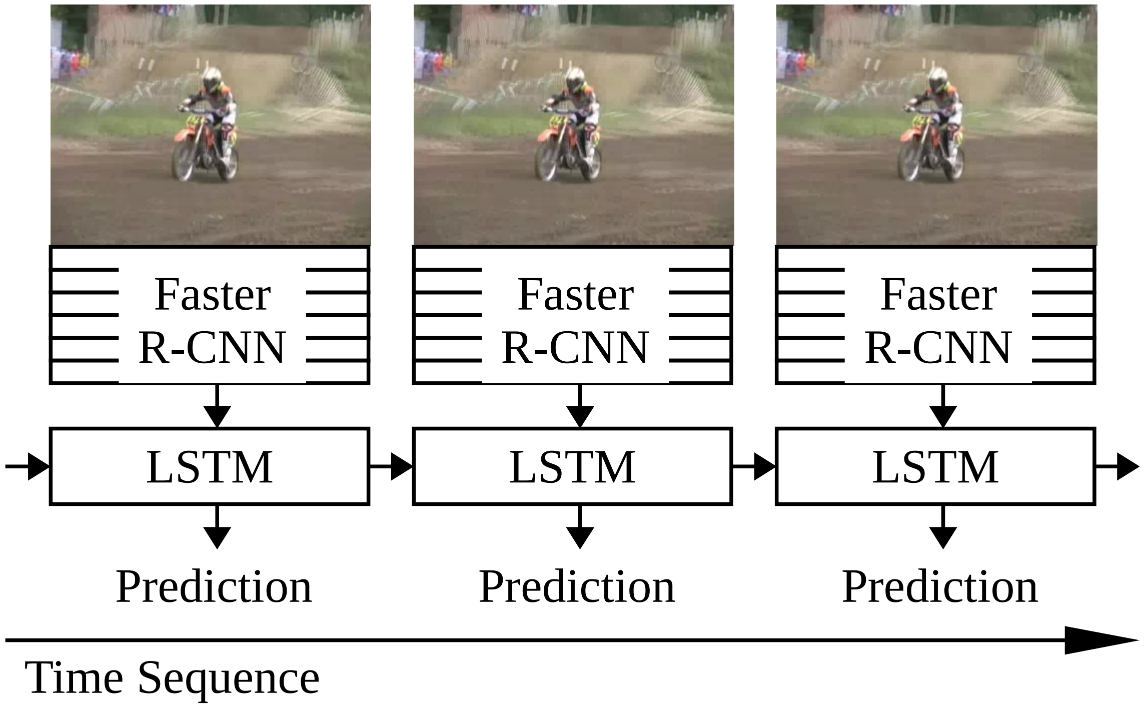
- ZF Net (Zeiler 2014) ← we use
- VGG-16 (Simonyan 2014)
- GoogLeNet (Szegedy 2015)
- ResNet (He 2016)



Long Short-Term Memory (LSTM)

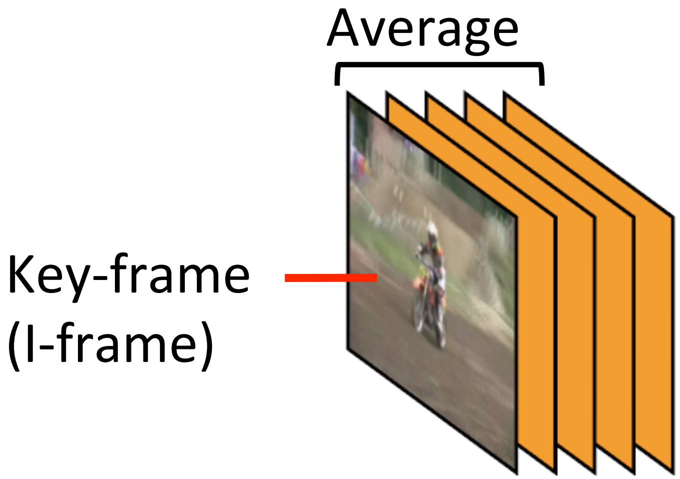
An LSTM layer is introduced to Faster R-CNN

- memorize long and short term information
- applied only to SittingDown



Multi-Frame and Multi-Shot (Yamamoto 2015)

- **Multi-Frame Score Fusion**
  Average pooling of scores over 5 frames in a shot
- **Multi-Shot Score Boosting**
  Add adjacent shot scores



$$\mathrm{score}^{\mathrm{boost}}(r_i^t) = \mathrm{score}(r_i^t) + \beta \max_j \frac{r_i^t \cap r_j^{t\pm1}}{r_i^t \cup r_j^{t\pm1}}$$

r_i^t: ith region in time t; β: multiplier

Key-Frame Annotations

Bounding-box annotation on the representative key-frame for each shot labeled as positive in collaborative annotation

Concept	# frames	# boxes	Concept	# frames	# boxes
Animal	11,545	9,155	Inst.Musician	4,923	7,229
Bicycling	599	1,355	Running	945	1,394
Boy	1,848	2,492	SittingDown	-	-
Dancing	2,118	5,199	Baby	898	895
ExplosionFire	2,483	2,402	Skier	320	521

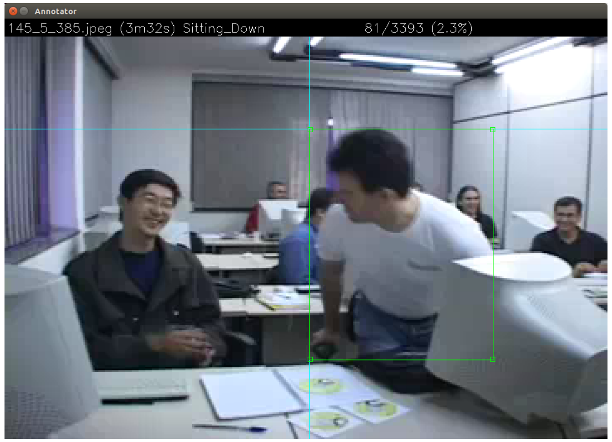
I-Frame Annotations for SittingDown

- I-Frame annotation for SittingDown to train LSTM
- Annotation results

  # shots = 92

  # frames = 481

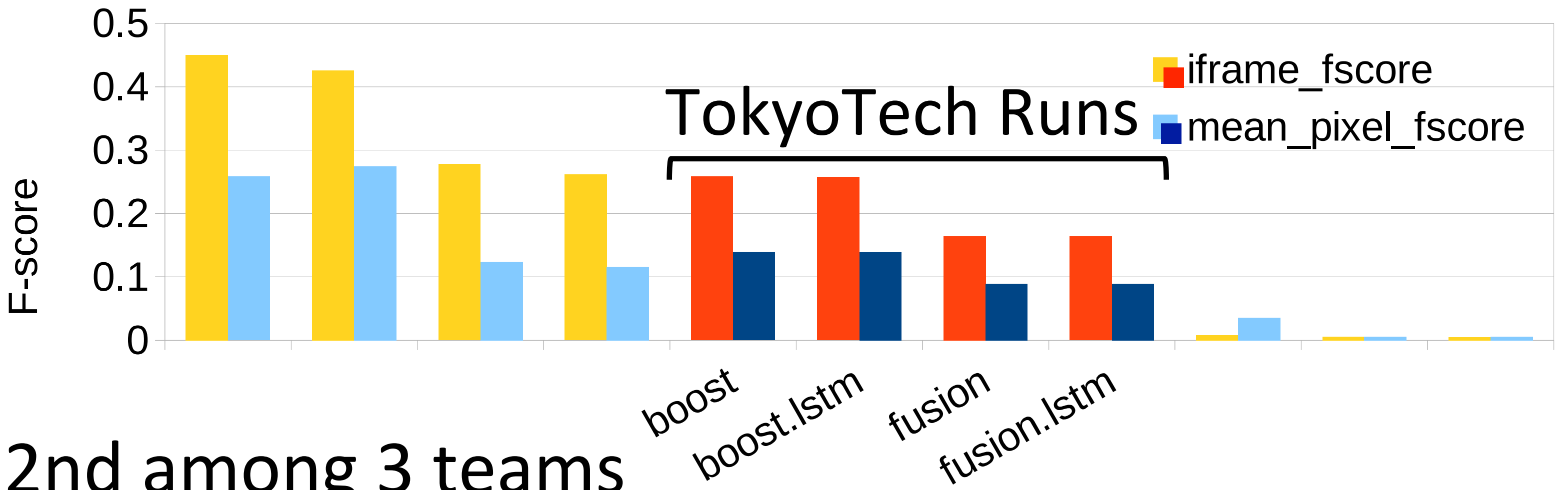
  # bounding-boxes = 515



* We found SittingDown in only 92 shots in the 3K shots labeled as positive in collaborative annotation

Results

ID	Method	RunID
1*	Faster R-CNN + Multi-Frame Score Fusion	fusion
2*	1 + Multi-Shot Score Boosting	boost
3*	1 + LSTM(4096units) for SittingDown	fusion.lstm
4*	2 + LSTM(4096units) for SittingDown	boost.lstm
5	2 + LSTM(64units) for SittingDown	(post exp.)



- 2nd among 3 teams

Results for SittingDown

Best result for SittingDown with run #2
LSTM with 4096 units (run #4) did not work
→ LSTM with 64 units (run #5) avoided over-fitting
and worked in post submission experiment

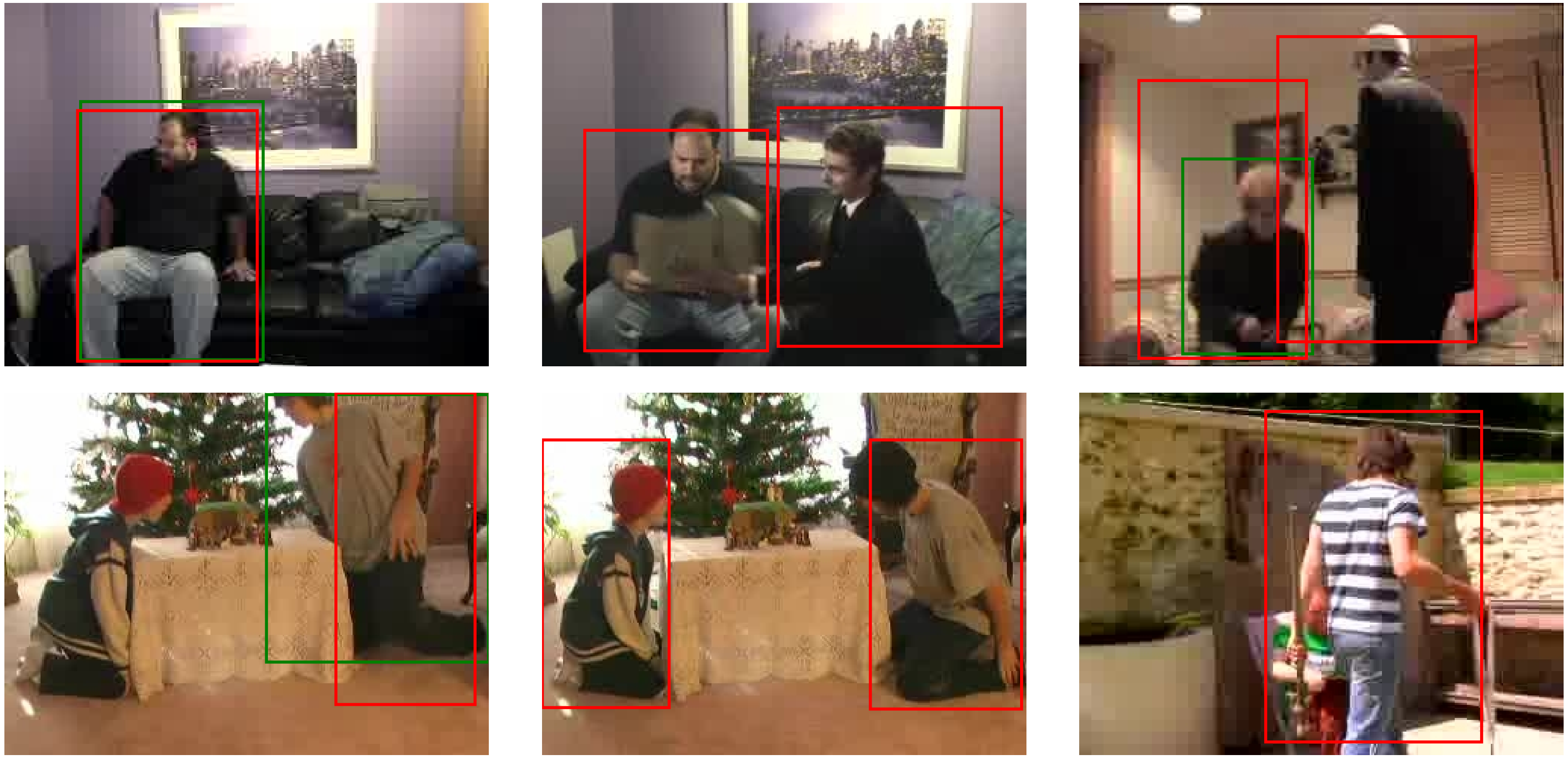
ID	Method	I-Frame F-score	Pixel F-score
2*	Fusion + Boosting	0.63	0.22
4*	2 + LSTM (4096units)	0.00	0.00
5	2 + LSTM (64units)	**11.96**	**4.51**

SittingDown

Re-trained network with LSTM 64 units

System output
Ground truth

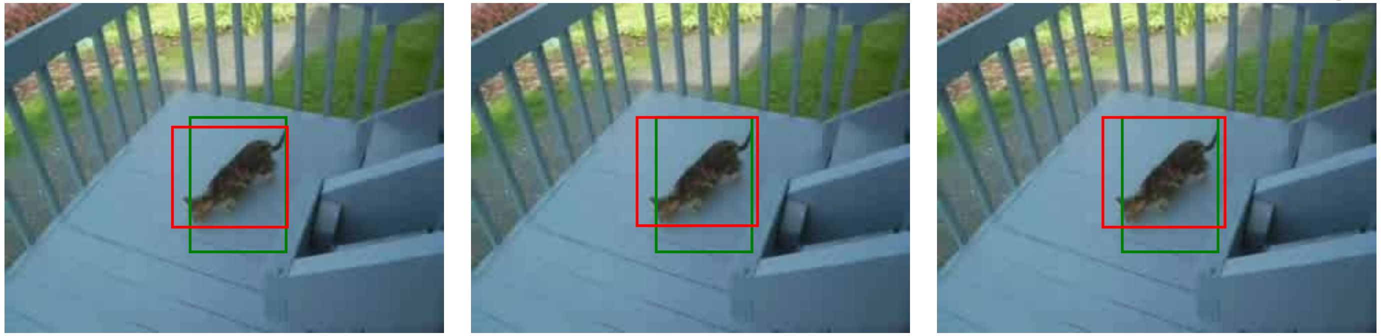
Good cases | Bad cases



Sitting down | Moving but not sitting down | Moving around a chair

Animal, Good Results

Faster R-CNN | Score Fusion | Score Boosting



Cat (no movement)

Dog (walking)

System output Ground truth

Animal, Bad Results

Faster R-CNN | Score Fusion | Score Boosting

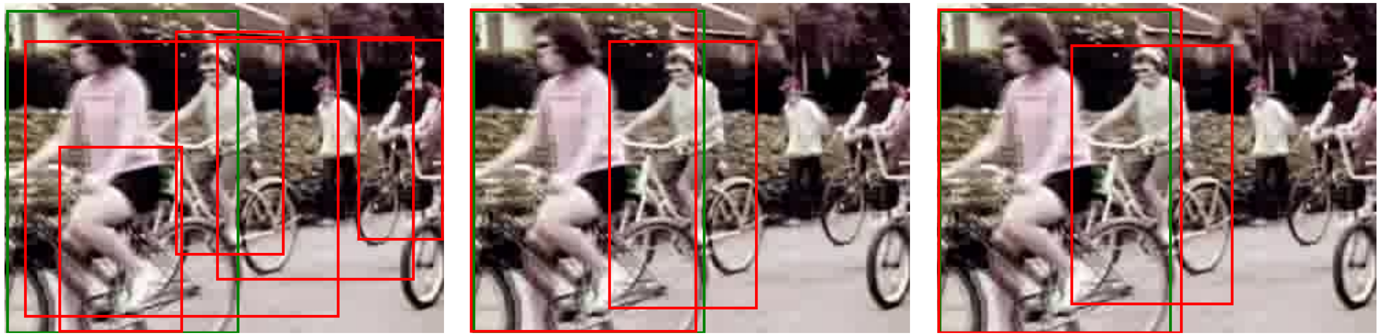
Many animals

Bird (flying fast)

System output   Ground truth

Others

Faster R-CNN | Score Fusion | Score Boosting



Bicycling

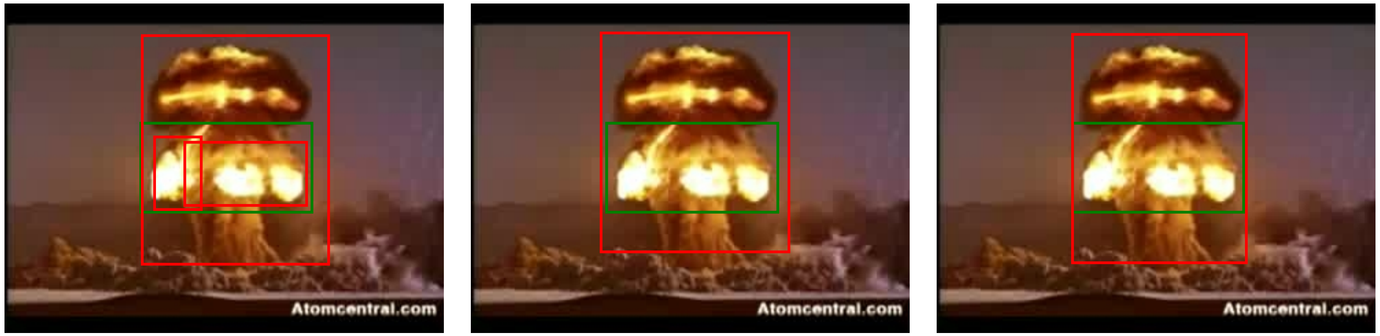
Boy

System output Ground truth

Others

Faster R-CNN | Score Fusion | Score Boosting

Dancing



ExplosionFire

System output   Ground truth

Others

Faster R-CNN | Score Fusion | Score Boosting

InstrumentalMusician

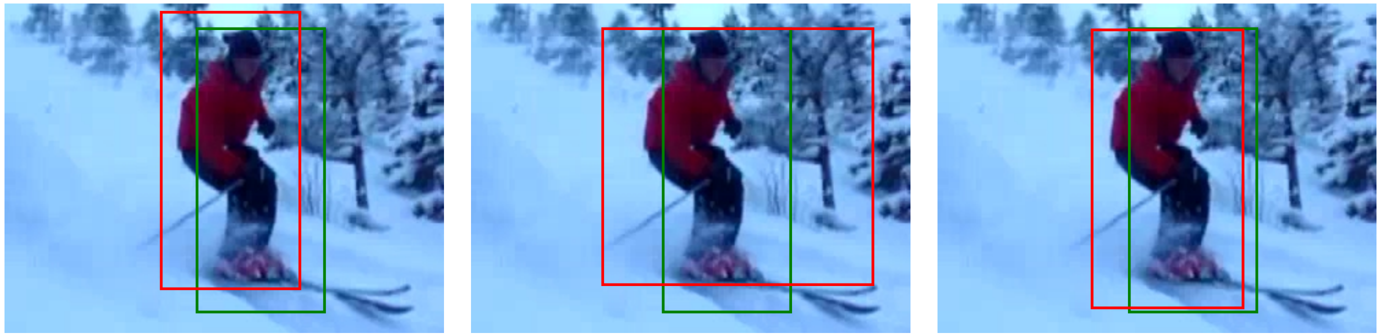
Running

System output Ground truth

Others

Faster R-CNN | Score Fusion | Score Boosting

Baby



Skier

System output Ground truth

Conclusion & Future Work

- We proposed a localization system
  - Faster R-CNN + LSTM + Re-scoring
- Manual annotation
  - 31K bounding boxes
- Results
  - 2nd among 3 teams, best result at SittingDown
  - LSTM with 64 units was effective for SittingDown
- Future work
  - Find a better way to localize action