Waseda_Meisei_SoftBank at TRECVID 2022: Activities in extended videos

Hideaki Okamoto,
Kazuya Ueki, Yuma Suzuki, Hiroki Takushima, Hayato Tanoue, Takayuki Hori
SoftBank Corp., Tokyo, Japan
hideaki.okamoto@g.softbank.co.jp

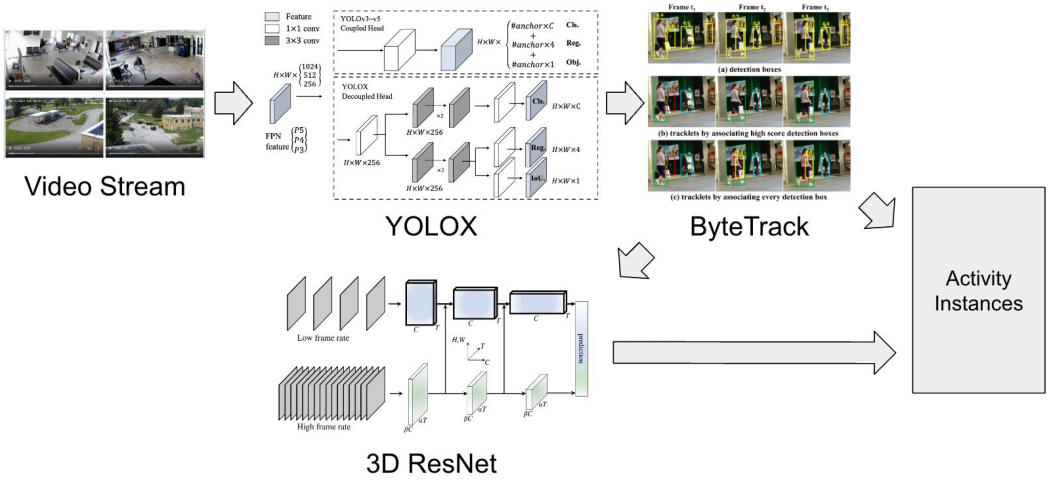
SoftBank

Agenda

- Overview
- Background
- Methods
- Experiments
- Results
- Discussion

- Participate in the ActEV task for the first time
- Propose a system that combines 3D ResNet training with YOLOX and ByteTrack trained models

System overview



Results

Activity and Object Detection (AOD)		
Pmiss @ 0.1Rfa	Nmd @ 0.1Rfa	nAUDC @0.2Rfa
0.9961	0.1080	0.9964

Activity Detection (AD)	
Pmiss @ 0.1Rfa	nAUDC @0.2Rfa
0.9829	0.9850

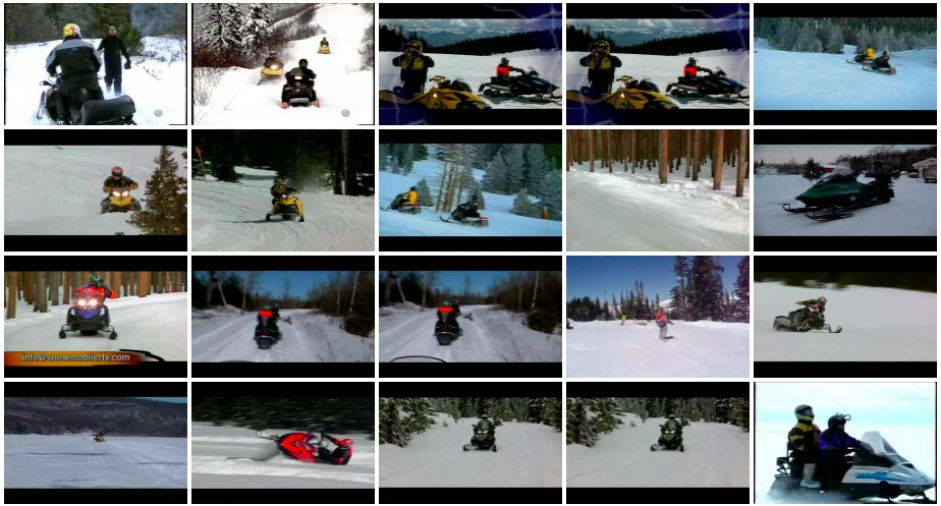
Engage in research to analyze, search, and understand content from videos
Participate only AVS until 2021, plus ActEV and VTT in 2022, at TRECVID

早稲田大学
WASEDA University

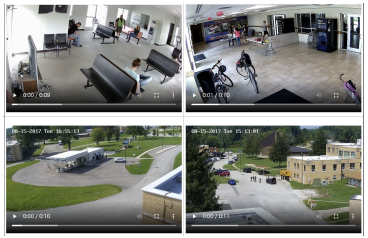
明星大学
MEISEI UNIVERSITY

SoftBank

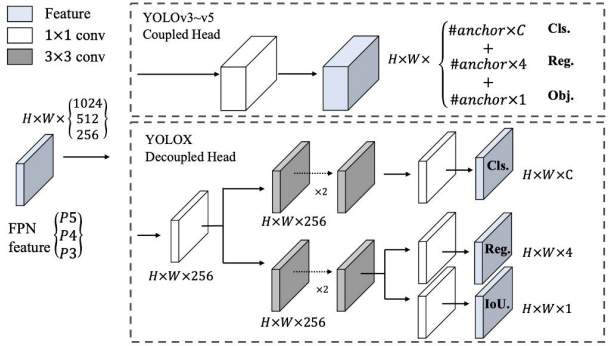
“one or more people driving snowmobiles in the snow”



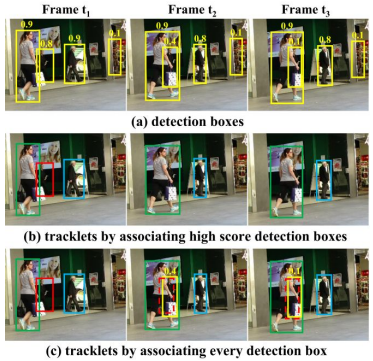
Methods - System overview



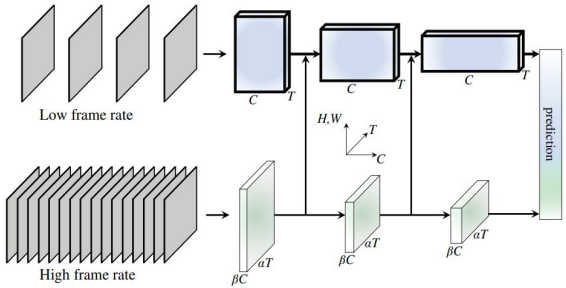
Video Stream



YOLOX



ByteTrack



3D ResNet

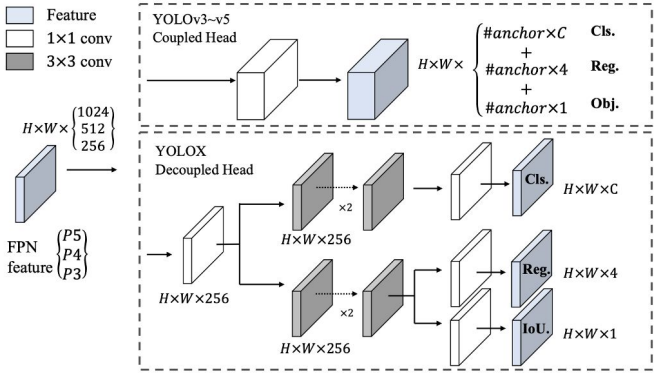
Activity Instances

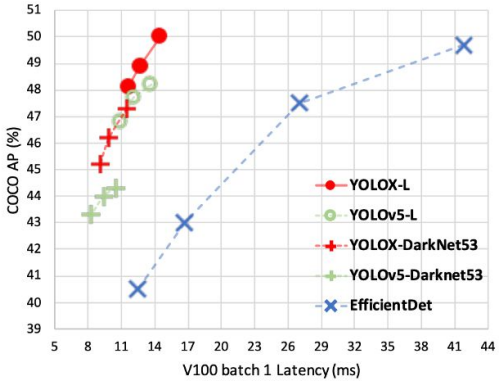
SoftBank

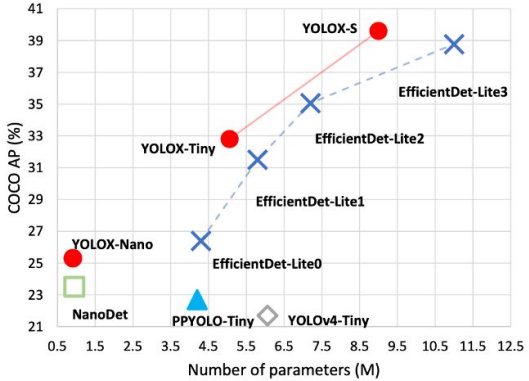
Conventional YOLO changed to anchor-free

Object detection model with decoupled head and SimOTA introduced

1st place in CVPR 2021 automous driving workshop

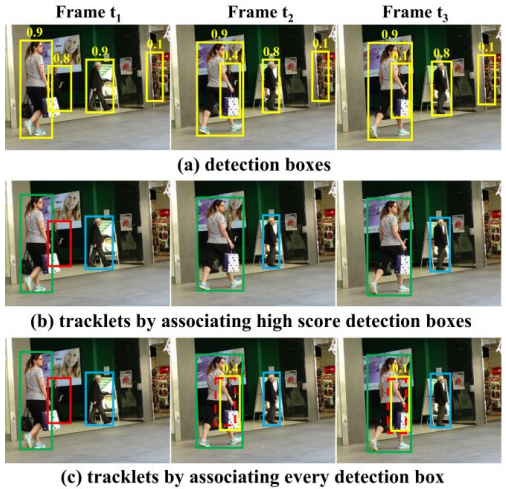


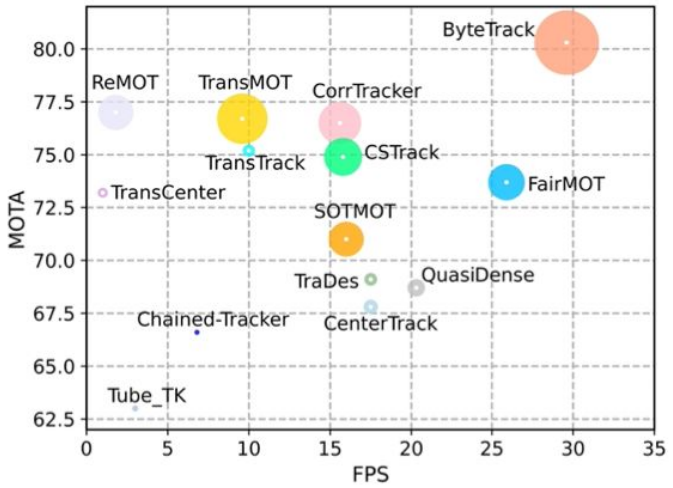




Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, ”Yolox: Exceeding yolo series in 2021,” arXiv preprint arXiv:2107.08430, 2021.

Motion model using a queue called tracklets that indicates the object being tracked
Eliminate the non-detection by considering bounding boxes with low confidence
Achieve SoTA beyond SiamMOT and transformer-based tracking models





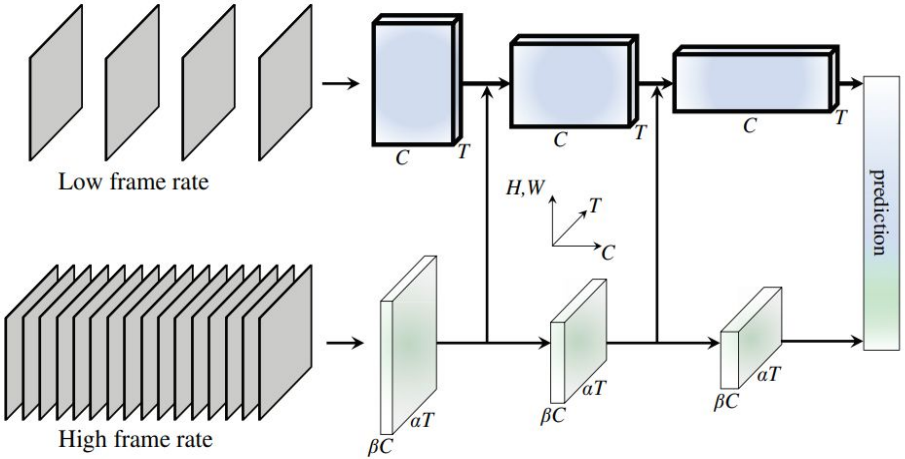
Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," arXiv preprint arXiv:2110.06864, 2021.

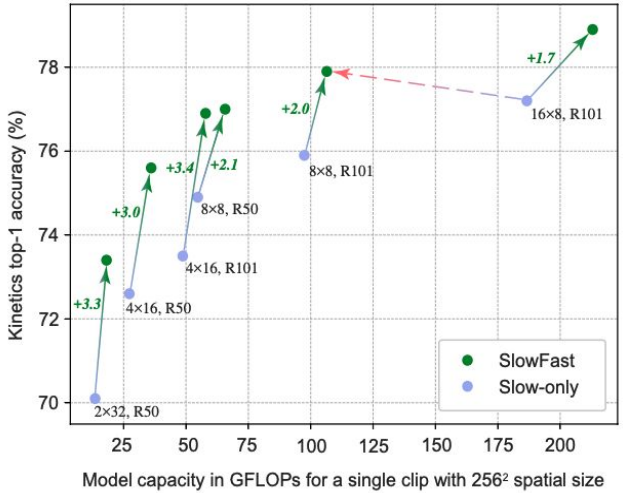
The tracking results by ByteTrack are joined together to form a single video
Then stretched and resized to produce a video with a 256 × 256 pixel resolution

Input Video

Output Videos

ResNet constructed using a 3D convolution to consider the time axis

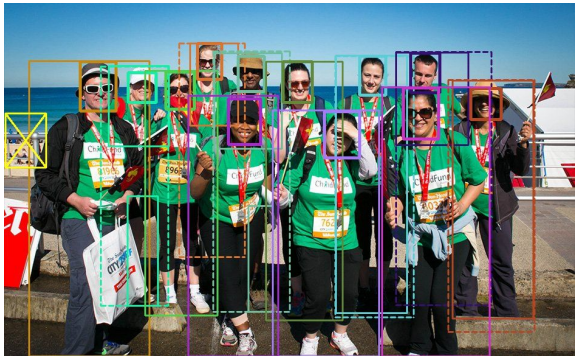


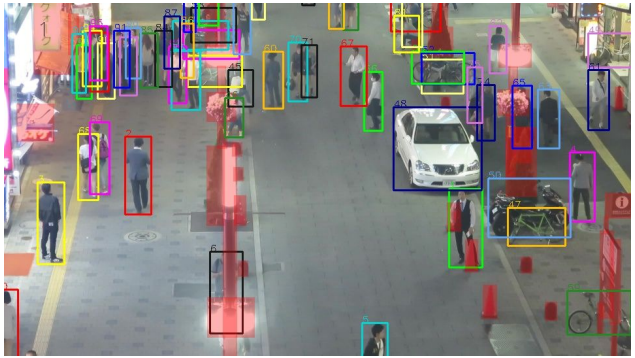


K. Hara, H. Kataoka, and Y. Satoh, ”Learning spatio-temporal features with 3d residual networks for action recognition,” In ICCV, 2017.

YOLOX and ByteTrack are trained on the CrowdHuman and MOT17 half-train sets
Use only the weights of the existing trained model labeled for only people

3D ResNet is trained on 53,027 square videos labeled for activity
created from annotations of the kitware-meva-training tracking





S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, ”Crowdhuman: A benchmark for detecting human in a crowd,” arXiv preprint arXiv:1805.00123, 2018.
A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, ” Mot16: A benchmark for multi-object tracking,” arXiv preprint arXiv:1603.00831, 2016.

3D ResNet is trained with the following settings

Train data: 53,027 square videos (256 × 256 pixels × 8 frames)

Preprocess: normalize
Data augmentation: random crop (224 × 224 pixels), random horizontal flip

Learning rate: 1e-5, Batch size: 8, Epoch: 20
Optimizer: momentum, Momentum: 0.9, Weight decay: 1e-4

Activity and Object Detection (AOD)		
Pmiss @ 0.1Rfa	**Nmd @ 0.1Rfa**	**nAUDC @0.2Rfa**
0.9961	0.1080	0.9964

Activity Detection (AD)	
Pmiss @ 0.1Rfa	**nAUDC @0.2Rfa**
0.9829	0.9850

Two major issues to be considered

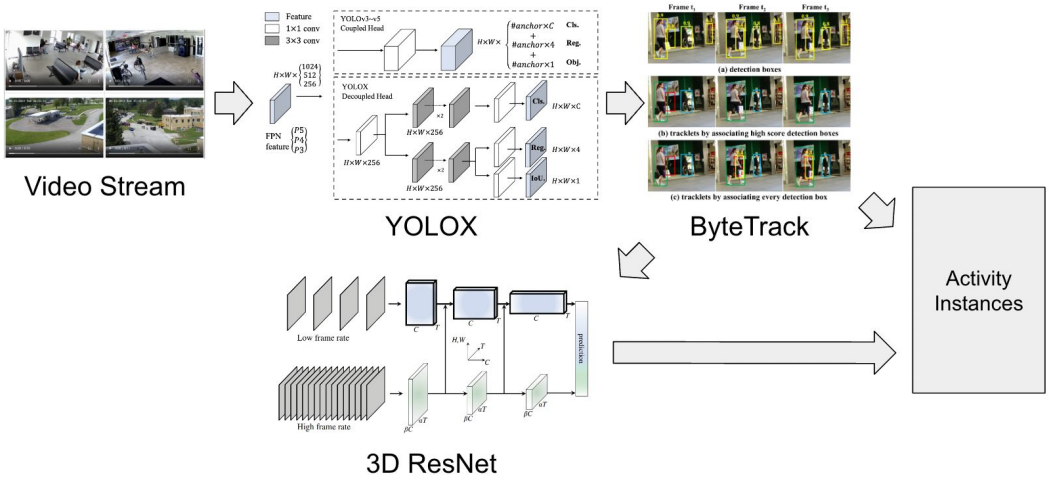
- YOLOX + ByteTrack only detects people and ignores other objects
  If the system misses anything in the first step, the mistake cannot be rectified
  Need to train to use the appropriate labels we omitted this time

- 3D ResNet is compressed down to 8 frames so much information is missed
  Essential to update to a model that can handle more frames
  The video treated at this time is 5 minutes in length

Good to be able to submit in spite of our first participation
Aim for further research development and business study

- Participate in the ActEV task for the first time
- Propose a system that combines 3D ResNet training with YOLOX and ByteTrack trained models

System overview



Results

Activity and Object Detection (AOD)		
Pmiss @ 0.1Rfa	Nmd @ 0.1Rfa	nAUDC @0.2Rfa
0.9961	0.1080	0.9964

Activity Detection (AD)	
Pmiss @ 0.1Rfa	nAUDC @0.2Rfa
0.9829	0.9850