Video Object Detection Method Using Single-Frame Detection and Motion Vector Tracking

Masato Nohara, Hiroaki Nishi

Graduate School of Science and Technology
Keio University

Introduction

- Increasing of video traffic on the Internet
- Growing demand for video analysis
- Demand for edge computing
  - Load balancing, bandwidth reduction, and network delay reduction
  - Computational resources are scarce, and high-performance GPUs are not always available
  - Network-transparent processing eliminates the need to change end-device and server configurations
    - Network-transparent : Add new services without changing network settings

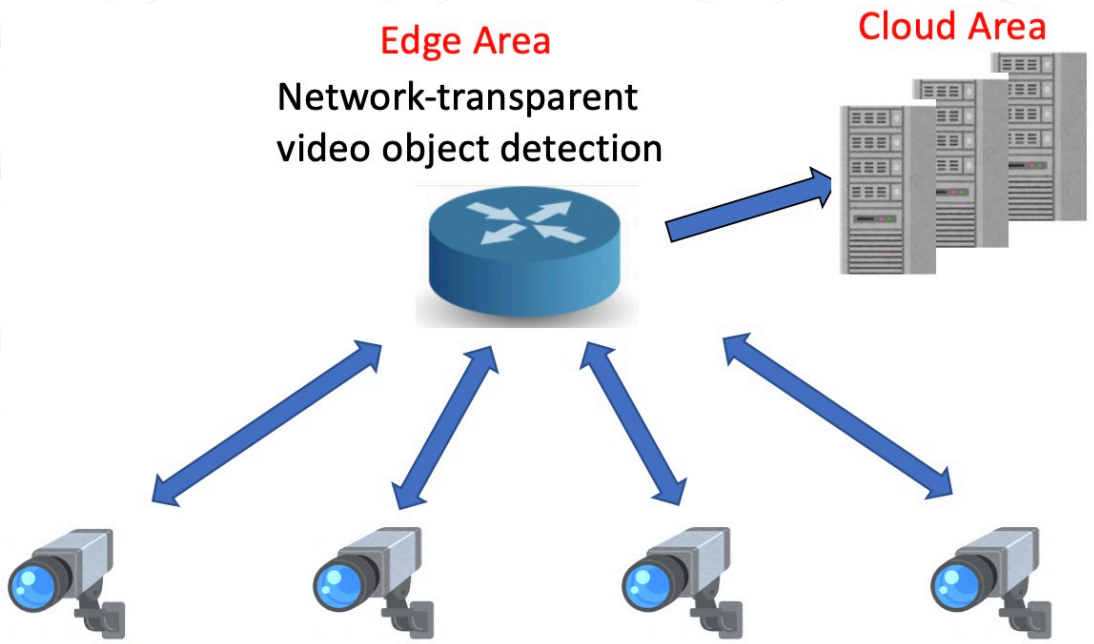
- High-throughput and memory-saving object detection method
- Network-transparent mechanism for acquiring video in edge area

Motion Compensation

- Compression technology used in every video codec
- Defines motion vectors for each macro blocks
- Three types of frame
  - I-frame : All information of the image
  - P-frame : Forward-predicted motion vectors
  - B-frame : Forward- and backward-predicted motion vectors

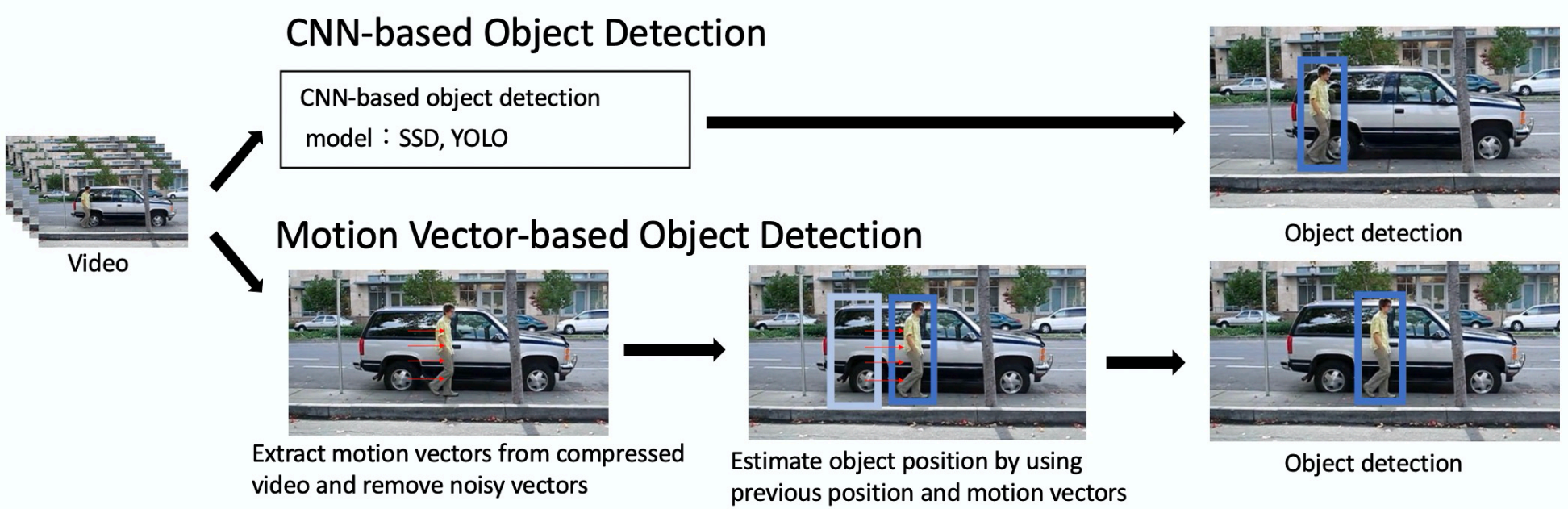
Assumed Environment

- Multiple surveillance cameras send video data to the cloud
  - Cameras for monitoring roads around residential areas
  - Cameras for monitoring production lines in a factory
- Network-transparent video object detection in edge area



Proposed Video Object Detection Method

- Different processes depending on frame types
  - I-frame
    - CNN-based object detection
  - P-frame, B-frame
    - CNN-based object detection
    - Motion vector-based object detection



Proposed Video Object Detection Method

- Reuse motion vectors in compressed video
  - No need to recalculate motion information
  - Motion vectors contain noisy vectors
- Remove noisy motion vectors
  - Apply median filter (filter size : 5 × 5 × 5) for maintaining spatiotemporal vector consistency

$$mv[x,y,t] = median\{mv[x',y',t'], x',y',t' \in w\} \quad (1)$$

x, y : position of the object
t : time axis
w : filter size

- Estimate next position of the object

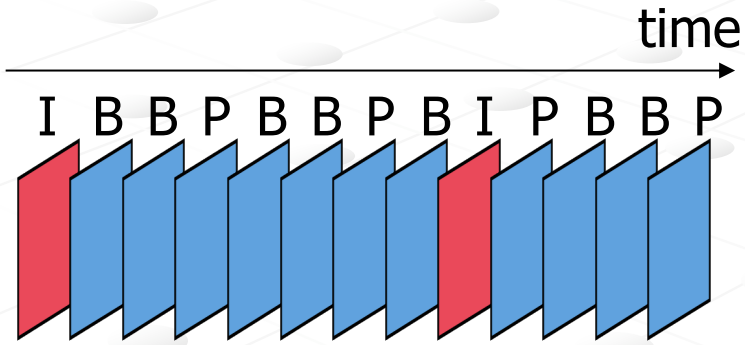
$$P_{now} = P_{previous} + \overline{mv} \quad (2)$$

$$\overline{mv} = average\{mv[w,h], w,h \in P_{previous}\} \quad (3)$$

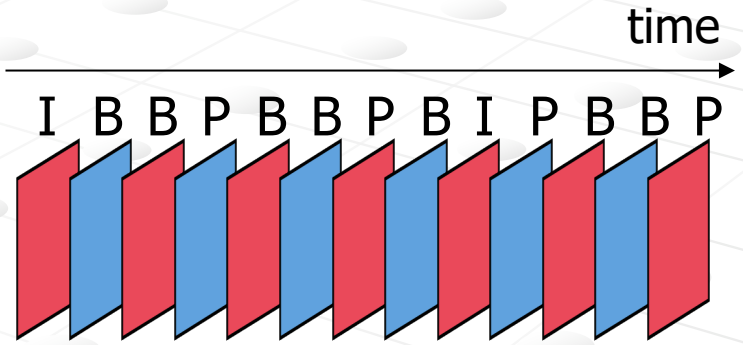
P_{now} : position of the object on current frame
$P_{previous}$: position of the object on previous frame

Trade-offs between Throughput and Accuracy

CNN-based object detection : Red
Motion vector-based object detection : Blue



- Throughput : High
- Accuracy : Low

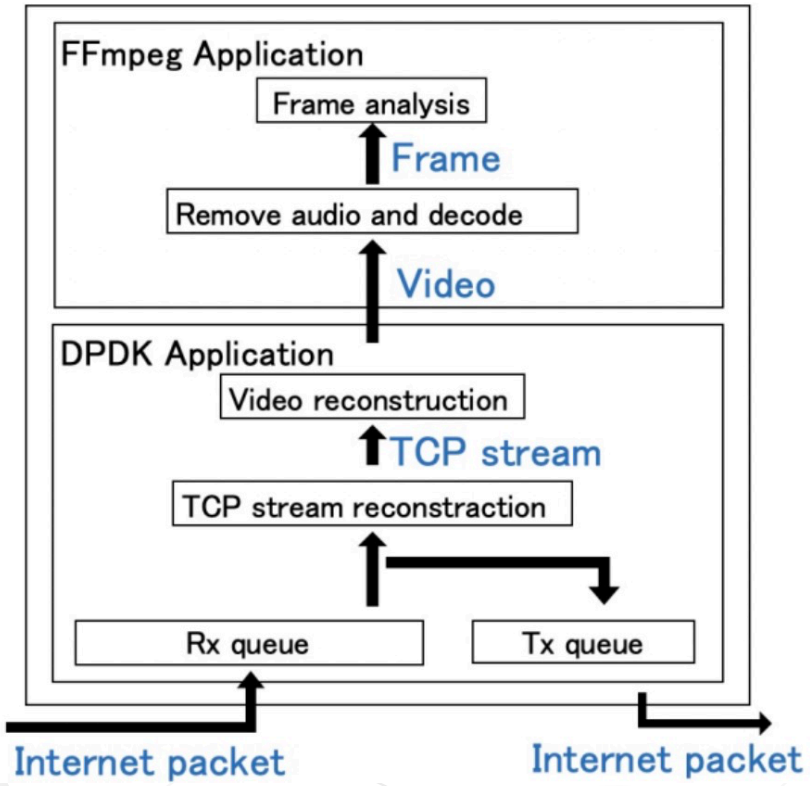


- Throughput : Low
- Accuracy : High

- The proposed method responds to various requirements in the trade-offs between throughput and accuracy

Implementation

- Used libraries
  - Data Plane Develop Kit (DPDK) : Set of libraries and drivers for fast packet processing using kernel bypass technology
  - FFmpeg : Complete cross-platform solution for recording, converting, and streaming audio and video
- DPDK Application
  - Reconstruct TCP stream
  - Reconstruct video
- FFmpeg Application
  - Remove audio data
  - Decode video data
  - Analyze frame data



Experimental Environment

- Video analysis device gets video network-transparently
- H.264 compressed video is used, which is captured by video distribution device

- Device Specifications
  - Small device for edge region was used

Device	Intel NUC	Shuttle DH310
OS	Ubuntu 18.04.1 LTS	Ubuntu 18.04.1 LTS
CPU	Intel(R) Core(TM) i3-6100U CPU@2.30GHz	Intel(R) Core(TM) i7–8700CPU@3.20GHz
Memory	8 G	32 G
Size	102 x 102 x 28 mm	190 x 165 x 43 mm

Experimental Results

Moving Object Detection


(a)


(b)


(c)


(d)


(e)


(f)

Video Object Detection

(a)

(b)

(c)

(d)

(e)

(f)

Evaluation : Accuracy and Throughput

- Comparison method
  - Video object detection : CNN-based object detection for all frames
- Accuracy (mean Average Precision 50 (mAP50))

	SSD	YOLOv3
CNN-based object detection for all frames (a)	0.796	0.774
Proposed method (b)	0.703	0.705
Accuracy ratio (b/a)	0.88	0.91

- Throughput (Frame Per Second (FPS))

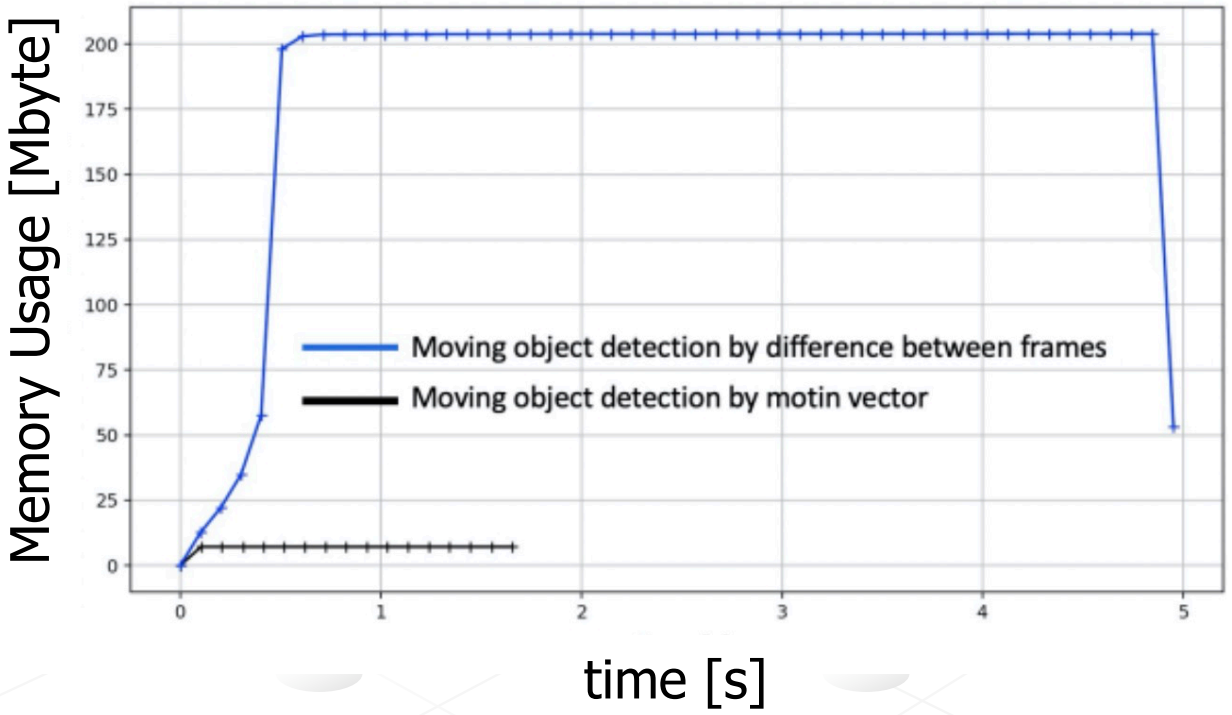
	SSD	YOLOv3
CNN-based object detection for all frames (a)	1.3	4.7
Proposed method (b)	58.3	75.5
Throughput ratio (b/a)	44.5	16.1

Throughput of 50 FPS or more is achieved by using motion vectors

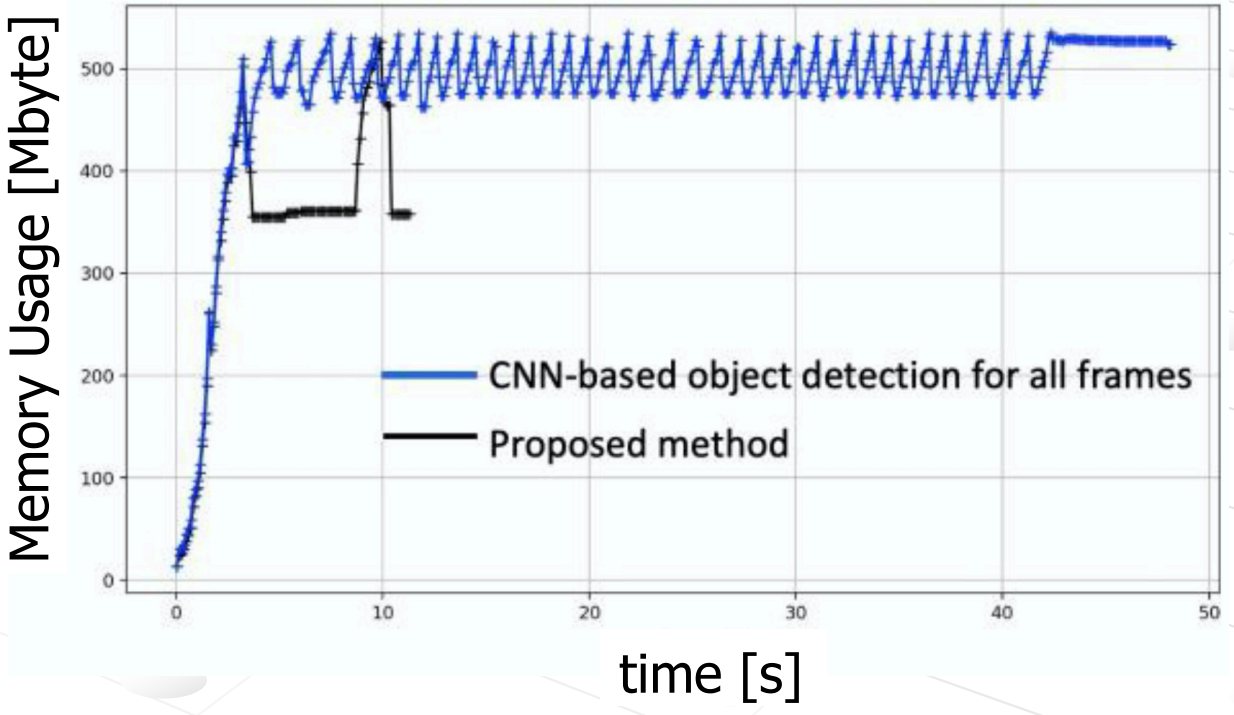
Evaluation : Memory Usage

- Time variation of memory usage

Moving object detection

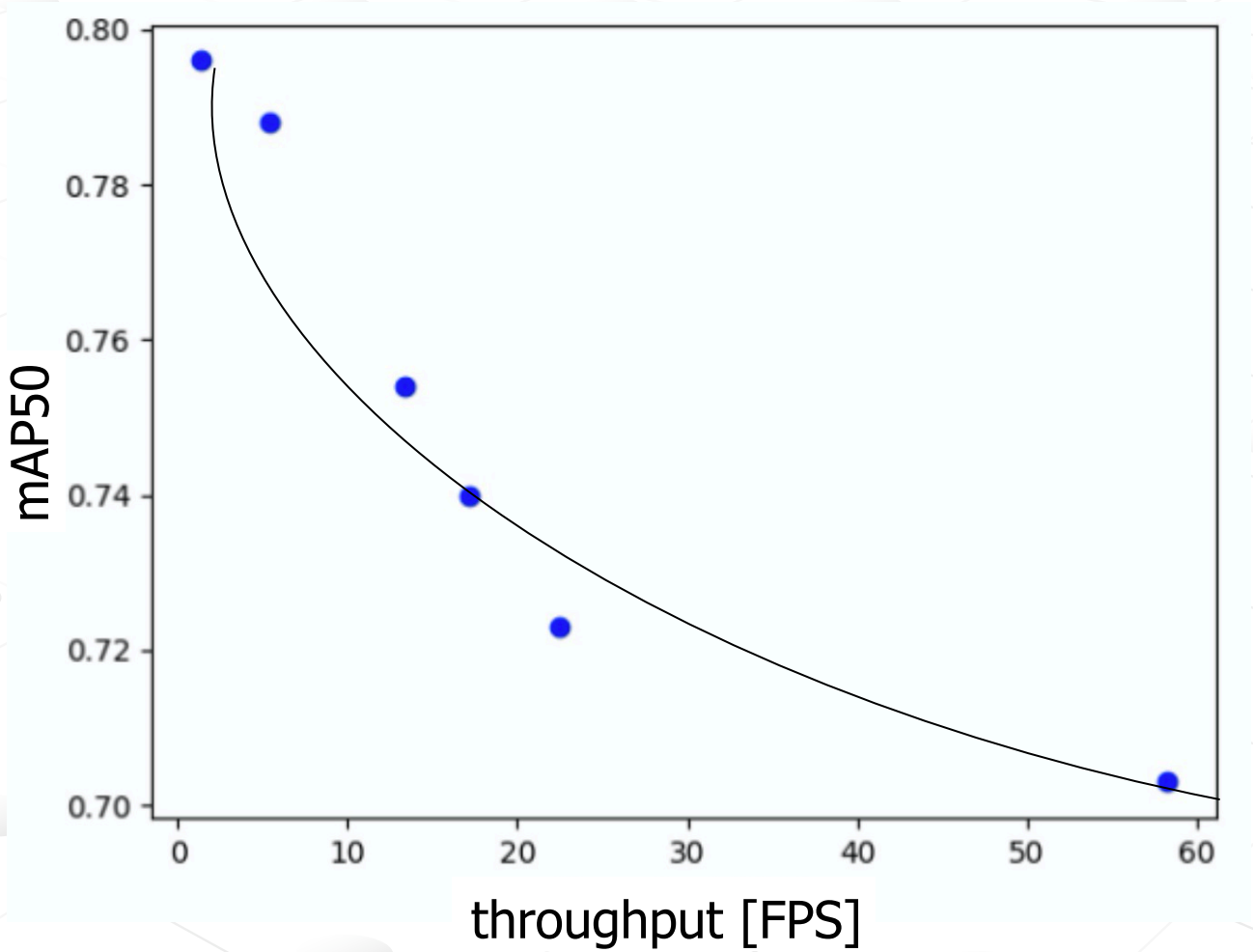


Video object detection



Trade-offs between Throughput and Accuracy

- Trade-offs between throughput and accuracy when using SSD



Conclusion

- A video object detection method was proposed for memory-saving and high-throughput analysis without the use of high-performance computing nodes like GPU

- A network-transparent mechanism for acquiring video was constructed in edge areas

- High-throughput and memory saving object detection in edge area where computational resources are scarce

Thank you for your attention

References

- [1] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper - Cisco. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/s ervice-provider/visual-networking-index-vni/white-paper- c11-741490.html. [Accessed: 08-Feb-2020].
- [2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 779– 788, 2016, doi: 10.1109/CVPR.2016.91.
- [3] J. Redmon and A. Farhadi, YOLOv3: An Incremental Improvement,2018.
- [4] W. Liu *et al.*, SSD: Single shot multibox detector, *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3- 319-46448-0_2.
- [5] R. Girshick, Fast R-CNN, *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.
- [6] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, Selective search for object recognition, *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013, doi: 10.1007/s11263-013- 0620-5.
- [7] S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017, doi: 10.1109/TPAMI.2016.2577031.
- [8] S. Ranjbar Alvar, H. Choi, and I. V. Baji, Can You Tell a Face from a HEVC Bitstream?, *Proc. - IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, vol. 1, pp. 257–261, 2018, doi: 10.1109/MIPR.2018.00060.

References

- [9] K. S. Devi, N. Malmurugan, and H. Ambika, Moving region segmentation from compressed video using global motion estimation by macroblock classification and markov random field model, *2013 IEEE Int. Conf. Emerg. Trends Comput. Commun. Nanotechnology, ICE-CCN 2013*, no. Iceccn, pp. 163–167, 2013, doi: 10.1109/ICE-CCN.2013.6528484.
- [10] R. C. Moura, E. M. Hemerly, and A. M. Cunha, Temporal Motion Vector Filter for Fast Object Detection on Compressed Video, *J. Commun. Inf. Syst.*, vol. 29, no. 1, pp. 12–24, 2014, doi: 10.14209/jcis.2014.1. [11] S. Biswas and R. V. Babu, Anomaly detection in compressed H.264/AVC video, *Multimed. Tools Appl.*, vol. 74, no. 24, pp. 11099–11115, 2015, doi: 10.1007/s11042-014-2219-4.
- [12] S. Gul, J. T. Meyer, C. Hellge, T. Schierl, and W. Samek, Hybrid video object tracking in H.265/HEVC video streams, *2016 IEEE 18th Int. Work. Multimed. Signal Process. MMSP 2016*, 2017, doi: 10.1109/MMSP.2016.7813363.
- [13] S. R. Alvar and I. V. Bajic, MV-YOLO: Motion vector-aided tracking by semantic object detection, *2018 IEEE 20th Int. Work. Multimed. Signal Process. MMSP 2018*, 2018, doi: 10.1109/MMSP.2018.8547125.
- [14] R. Morishima and H. Nishi, Network Transparent Fog-based IoT Platform for Industrial IoT, pp. 920–925, 2020, doi: 10.1109/indin41052.2019.8972178.
- [15] FFmpeg. [Online]. Available: http://ffmpeg.org/. [Accessed: 13-Feb-2020].
- [16] DPDK. [Online]. Available: https://core.dpdk.org/. [Accessed: 13-Feb-2020].

Acknowledgements

- This work was supported by JST CREST Grant Number JPMJCR19K1, and the commissioned research by National Institute of Information and Communications Technology (NICT, Grant Number 22004) , JAPAN.