# REGION-OF-INTEREST-BASED VIDEO CODING FOR MACHINES

*Olgierd Stankiewicz, Tomasz Grajek, Sławomir Maćkowiak ,Jakub Stankowski, Sławomir Różek, Mateusz Lorkiewicz, Maciej Wawrzyniak, Marek Domański*

Poznan Univeristy of Technology, Poland
{olgierd.stankiewicz, tomasz.grajek, slawomir.mackowiak, jakub.stankowski, slawomir.rozek, mateusz.lorkiewicz, maciej.wawrzyniak, marek.domanski}@put.poznan.pl

## ABSTRACT

The paper is within the scope of Video Coding for Machines (VCM), where video compression not only caters to human viewers but also acts as input to systems engaged in diverse machine vision tasks like object detection and tracking. A novel compression technique is proposed, which exemplifies the proficient utilization of RoIs-based video coding. It includes significant modifications to the state-of-the-art technique employing image retargeting and also addresses the related problem of resolution change. In the novel approach presented in this work, due to the proposed application of image padding, video frames encoded by an Inner Encoder (e.g. a VVC encoder) remain of constant size. Experimental evaluation demonstrates significant average bitrate reduction, up to 57% with respect to the current version of VCM, while maintaining consistent quality across various machine vision tasks and encoding scenarios.

**Index Terms—** Video Coding for Machines, VCM, Regions of Interest, RoI, retargeting video, image padding.

## 1. INTRODUCTION

Traditional video coding technologies such as Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC), or Versatile Video Coding (VVC) [1,2,3,4,5,6] are primarily designed with human perception in mind. However, in today's rapidly evolving landscape, where machine processing algorithms are advancing at an unprecedented rate, a significant portion of video applications is being processed by machines [7,8]. This shift highlights a notable misalignment between the intended objectives of conventional video coding standards and the emerging demands dictated by the necessity of machine vision processing. Consequently, there arises a pressing need for the development of specialized video coding tools precisely attuned to the intricacies of machine vision processing requirements, commonly referred to as Video Coding for Machine (VCM).

In light of the dynamic interplay between technological advancements and application demands, it becomes evident that the conventional paradigm of video coding is encountering limitations in meeting the evolving needs of machine-centric applications. The disparity between the optimization for human perception and the burgeoning reliance on machine processing underscores the necessity for a paradigm shift in video coding methodologies. Therefore, the emergence of VCM represents a pivotal step towards addressing this incongruity, offering tailored solutions that align with the intricacies of machine vision processing (Fig. 1). This necessitates a concerted effort towards the development of specialized coding tools that can effectively bridge the gap between traditional video coding standards and the burgeoning requirements of machine-centric applications. Such an effort is currently undertaken within ISO/IEC MPEG which aims at the development of a new VCM standard.

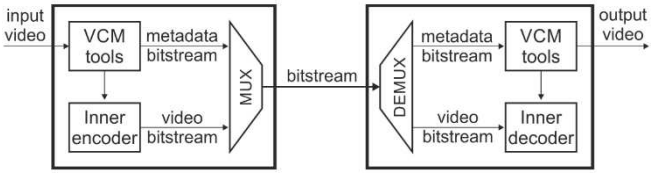


**Fig. 1.** The considered general architecture of Video Coding for Machines (VCM).

The proposal in this paper is inspired by the technique presented by Różek et.al in [9] for VVC codec. As described in detail in Section 2, this technique has a drawback consisting of resolution change of the encoded frames, which we address. The main novelties of our work are:

- The solution to the problem of resolution change by the usage of padding.
- Improvement and simplification of the technique.
- Adaptation to MPEG VCM codec.

The proposal solves the problem of a resolution changes of the encoded frames in the Inner Encoder and still achieves a significant reduction in the video bitrate. Section 3 presents the general idea of the proposed method whereas experimental results are presented in Section 4.

## 2. STATE-OF-THE-ART

In recent years, there has been a notable increase in academic interest towards optimizing image and video compression for machine vision multi-tasks [10,11,12,13,14,15]. Recently, there has been a surge of research aimed at developing solutions for video coding that optimize the efficiency of machine vision tasks while simultaneously minimizing the bitrate for the video. These efforts included exploration of various approaches, including refining parameter selection [16], rate-distortion optimization [17], and bit allocation within existing codecs [18]. Additionally, researchers have delved into the creation of end-to-end compression networks tailored to machine vision tasks, incorporating specific constraints to ensure optimal performance [19, 20].

The necessity for a standardized solution in Video Coding for Machines has been acknowledged by ISO/IEC MPEG. In July 2019, the MPEG Video Coding for Machines Ad-Hoc group commenced the development of video coding standards devoted to highly efficient compression and representation in intelligent machine-vision or hybrid machine/human-vision applications [21]. Additionally, in 2021, JPEG AI issued a call for proposals on learning-based coding standards [22]. By the same year, MPEG VCM had outlined initial requirements and use cases for various application domains such as surveillance, intelligent transport, smart city, and smart industry, anticipating a growing reliance on machine processing algorithms [23, 24]. The complexity and broad range of applications prompted experts to split the MPEG VCM standardization process into two tracks: one focusing on enhancing classical video compression methods, and the other on visual features compression approaches. A Call for Proposals (CfP) with updated requirements and use cases was released in April 2022, with responses from renowned research centers and universities worldwide [25]. The ensuing scientific and technical contributions led to the establishment of MPEG WG4 activity, fostering further advancements through collaboration and competition among researchers. As a result, ongoing research in this field largely centers around the efforts of the MPEG VCM group, positioning them at the forefront in this field.

In general, the MPEG approach to VCM is codec-agnostic, e.g. it separates VCM-specific tools from the general video codec, called Inner Codec, which may be e.g. HEVC or VVC. VCM architecture (Fig. 2) comprises key components crucial for adapting video content for machine-based applications: spatial and temporal resampling, region of interest (RoI) encoding, and bit depth truncation [26].

Temporal downsampling adjusts the video frame rate by omitting certain frames during encoding and interpolating them during decoding. Essentially, the encoder may process every second, fourth, or eighth frame from the video sequence. On the decoder side, the missing frames are interpolated using a dedicated neural network.

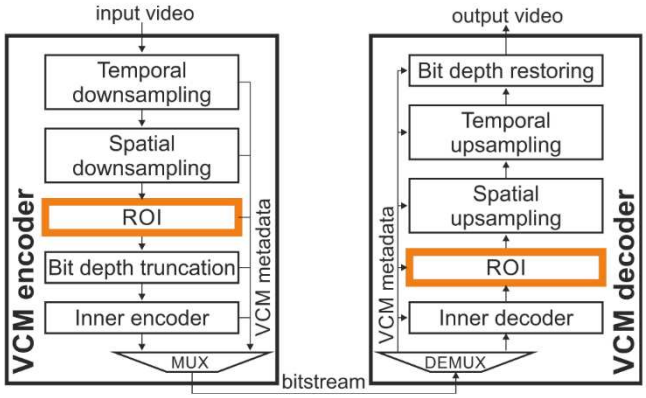


**Fig. 2.** The current architecture of MPEG VCM, with the scope of the paper - RoI-based tools – marked in dotted lines.

Spatial downsampling aims at decreasing the resolution of video frames, striving to strike a balance between minimizing bit rate and preserving adequate detail for effective machine vision recognition. This process also serves to reduce the computational complexity of the entire encoding procedure. Note: The spatial downsampling tool was not defined in MPEG VCM specification at the time of research. Further works in this area focused on adaptation of Reference Picture Resampling (RPR) technique from VVC to reinforce RoI tool considered in this paper.

Video coding techniques based on Regions of Interest (RoI) coding prioritize [27] areas within a video frame that are relevant to machine-based tasks. This ensures that these areas are encoded with higher fidelity compared to the less relevant areas. The RoI tool adapted to MPEG VCM, used as a starting point for research in this paper, employs the detection of RoIs, which are encoded without modification, and graying out of the remaining regions.

Another tool, bit depth truncation, adaptively reduces the dynamic range of the luma component samples, such as reducing from 10 bits to 9 bits at the encoder side.

Ultimately, the processed image or video sequence is fed into the Inner Encoder, which can be any video encoder. In the current MPEG experiments, considered is VVC [3,4] and its modifications.

To sum up, VCM aims to optimize video encoding for machine-learning tasks by minimizing redundant information during encoding and reconstructing it during decoding. Thus, the VCM architecture aims to optimize the efficiency of video encoding for usage in machine vision processing tasks.

The proposal presented in this paper is inspired by the technique introduced by Różek et.al in [9] for VVC codec. The technique from [9] is based on the concept of retargeting video frames, which involves processing them based on Regions of Interest (RoI). During retargeting, the content of video frames is adaptively scaled in rectangular regions, in order to allocate relatively greater area (and thus more bits) to the Regions of Interest and to allocate relatively smaller

area (and thus fewer bits) to regions of lesser importance. Consequently, content within ROIs is encoded with high fidelity, while the remaining regions in a frame may be downsized or even removed. As a result, frames with reduced resolution are created. In the decoder, the original dimensions of RoIs are restored through inverse processing using additional information transmitted within the bitstream.

The described approach [9] results in a situation where a sequence of frames of different resolutions will be fed to the Inner video Encoder. In the worst case, each frame may have a different resolution. This can be perceived as an advantage, as reduced resolution leads to decreased encoding and decoding time. However, in practice, for standard video coding, it is typically assumed that a video sequence has a fixed spatial resolution. Changing the resolution of individual frames during encoding is not a trivial task and represents a disadvantage of the mentioned technique. In this work, we address this problem.

## 3. THE PROPOSED SOLUTION

We propose a novel RoI-based encoding tool (Fig. 2) for VCM which builds upon the technique described in [9] with vital modifications. Firstly, the tool is adapted to the architecture of MPEG VCM to ensure interoperability with other coding tools, e.g. the same RoI detection framework is used. Secondly, the tool is improved by the utilization of RoI-based filtering, aimed at discarding redundant information from the processed video. Thirdly, RoI aggregation is employed to simplify processing and limit the amount of transmitted information. Lastly, in order to address the issue of resolution change, we introduce a padding algorithm to ensure that frames of the video sequence inputted to the Inner Encoder have an unchanged spatial resolution relative to the original content.

The main steps of the proposed modified RoI-based tool (Fig. 3) are as follows:

- Detection of RoIs - The detection process utilizes a specialized neural network within the encoder, customized for the specific machine vision task executed after decoding. The RoI descriptions primarily entail bounding boxes, though, for effective encoder control, they may also encompass class identification and importance level metrics. This step is identical to the reference solution currently adopted in MPEG VCM.
- Filtering – A low-pass filtration of areas of the image outside RoIs. For the experiments, a Lanchos filter with a normalized bandwidth of 0.01 is used.
- RoI aggregation - To create a region of interest that is broader and more consistent in time, detected objects are aggregated in the spatial and time domains, i.e. inside of a particular group of pictures (GOP). Thus, the final RoI is the smallest RoI that contains all detected RoIs within all images in a given GOP. Therefore, the information required to describe RoI is common for the entire GOP.
- Retargeting - In most cases, the final RoI will have a lower resolution than the original image, and pixels outside this RoI are completely irrelevant from the machine vision task point of view. Therefore they are discarded and are not encoded.


**Fig. 3.** Illustration of the idea of the proposal: original image (a), detection of RoIs and filtering (b), RoI aggregation (c), retargeting (d), and padding (e).

- Padding - Having in mind that the video will finally be fed to the Inner Encoder and to keep the resolution of the video as the original one, the final RoI is shifted to the left-top corner of the image. Then outside the RoI (on the right and below) padding is applied. Namely, if the horizontal padding size is smaller than 25% of the original frame width, the last column is duplicated. Otherwise, the padding area is grayed out. Similarly, if the vertical padding size is smaller than 25% of the original frame height, the last row is duplicated. Otherwise, the padding area is grayed out.

For the sake of experimental verification, the idea of the proposal has been implemented in Video Coding for Machines Reference Software (VCM-RS) version 0.7 [28]. In particular, implementation has been done on top of the existing RoI tool [29], which involves RoI detection. In order to provide a fair evaluation of the gains of the proposal, the already existing processing steps have been left unchanged. In particular:

- RoI detection. The same neural networks are used, i.e. JDE-1088x608 [30,36] for object tracking and Detectron2 (Faster R-CNN) [31] for object detection.
- All already adopted VCM techniques and tools enabled, e.g. Temporal Resampling and Bit Truncation tool.
- VVC [3,4] is used as the Inner Codec.

## 4. EXPERIMENTAL EVALUATION

### 4.1. Evaluation

The proposed technique was evaluated according to Common Test Conditions (CTC) [32] defined by ISO/IEC MPEG for the development of VCM. The anchor (reference) of the experiments, as designated in CTC, is the performance of the original, not modified VCM-RS version 0.7 [28].

**Table 1.** Datasets used in evaluation.

| Dataset Name | Class | Number of seq. | Frame rate | Resolution | Bit Depth | Machine Task |
|---|---|---|---|---|---|---|
| SFU [33] | A | 1 | 30 | 2560×1600 | 8 | Object Detection |
| | B | 4 | 24, 50 or 60 | 1920×1080 | 8 | |
| | C | 4 | 30, 50 or 60 | 832×480 | 8 | |
| | D | 4 | 30, 50 or 60 | 416×240 | 8 | |
| | O | 1 | 24 | 1920×1080 | 8 | |
| TVD [35] | ----- | 7 | 50 | 1920×1080 | 8 or 10 | Object Tracking |

The evaluation procedure includes two tasks: object detection and object tracking. The object detection case uses SFU-HW-objects-v1 (SFU) [33] test sequences and Detectron2 [34] software set to use the Faster R-CNN [31] X101-FPN model. The tracking case uses the Tencent Video Dataset (TVD) [35] and JDE-1088x608 [36] network. Details about the test dataset are presented in Table 1.

In both evaluation cases, three coding scenarios are used: AI – All Intra, LD – Low Delay, and RA – Random Access. Moreover, six different quantization parameter (QP) values are used in order to evaluate different quality/bitrate points.

### 4.2. Coding performance

Table 2 showcases the results for the object detection task across all configurations in terms of bitrate reduction (BD-RATE) [37,38] while maintaining a constant mean Average Precision (mAP) [39]. The bitrate reductions are estimated with respect to the current VCM technology of MPEG [26].

Notably, in the AI configuration, the presented method achieves the most significant results, that is the 19.15% averaged bitrate reduction (over constant mAP). Moreover, improvements over the anchor method are observed across all dataset classes in this scenario. In configurations employing Inter coding, such as RA and LD, the proposed method continues to demonstrate improvements. Specifically, the RA configuration yields an average bitrate reduction of 2.45% (over constant mAP), while the LD scenario observes a reduction of 0.07% bitrate over constant mAP. Class-specific analysis reveals that the proposed method reports slightly inferior results in two instances (LD in class B and RA in class D), with a bitrate increase (over constant mAP).

The evaluation of the proposed method for the tracking task is detailed in Table 3, where it is presented as bitrate reduction BD-RATE relative to the same task performance measured in Multiple Object Tracking Accuracy (MOTA) [40]. Notably, the proposed method exhibits even more substantial improvements in efficiency compared to the anchor. In the AI scenario, an average 57.65% reduction in bitrate is achieved for the same tracking result. Furthermore, improvements are observed across the remaining CTC configurations as well. The RA configuration demonstrates an average bitrate reduction of 31.94% (over constant MOTA), while the LD configuration benefits from a 7.45% reduction in bitstream (over preserved MOTA quality). It is worth noting that only one sequence (TVD-02-1), for only one scenario (LD), was found to be encoded worse than the anchor. This is an indicator of generality of the proposal, as typically tools offer arbitrarily gains and losses, which is a subject to complex encoder control mechanisms.

The benefits of the proposed method are most evident in AI scenario, whereas the gains for RA are smaller, and occasionally negative for LD. This can be attributed to the fact that while the proposed technology is versatile, the

results are compared against already well-developed technology, which includes tools optimized for LD and RA scenario.

**Table 2.** Object detection - comparison of delta Bitrate (BD-Rate) and mAP change (BD-mAP). Coding scenarios included: AI – All Intra, LD – Low Delay, and RA – Random Access. Class O is not mandatory, hence is not included in average (All).

| SFU seq. class | End-to-End BD-Rate [%] over mAP | | | End-to-End BD-mAP | | |
|---|---|---|---|---|---|---|
| | AI | LD | RA | AI | LD | RA |
| A | -22.23% | -0.20% | -17.04% | 2.62 | -1.55 | 0.47 |
| B | -22.93% | 12.46% | -4.48% | -0.11 | -1.14 | 1.66 |
| C | -23.74% | -6.48% | -4.84% | 4.02 | 0.83 | 0.95 |
| D | -10.02% | -6.15% | 5.61% | 1.31 | 0.61 | -0.81 |
| **All** | **-19.15%** | **-0.07%** | **-2.45%** | **1.81** | **-0.03** | **0.59** |
| O | -10.34% | -4.97% | -40.67% | 4.27 | 2.14 | 10.50 |

**Table 3.** Object tracking - comparison of delta Bitrate (BD-Rate) and MOTA change (BD-MOTA). Coding scenarios included: AI – All Intra, LD – Low Delay, and RA – Random Access.

| TVD seq. | End-to-End BD-Rate [%] over MOTA | | | End-to-End BD-MOTA | | |
|---|---|---|---|---|---|---|
| | AI | LD | RA | AI | LD | RA |
| 01-1 | -67.79% | -21.54% | -58.89% | 17.08 | 4.34 | 12.44 |
| 01-2 | -64.11% | -5.38% | -28.84% | 20.80 | 1.67 | 4.56 |
| 01-3 | -84.22% | -31.16% | -64.87% | 27.81 | 5.02 | 11.97 |
| 02-1 | -26.06% | 14.73% | -2.05% | 2.17 | -1.07 | -0.25 |
| 03-1 | -53.13% | -1.78% | -9.99% | 6.65 | 0.38 | 2.37 |
| 03-2 | -44.73% | -2.39% | -19.97% | 8.77 | 0.48 | 3.59 |
| 03-3 | -63.48% | -4.66% | -38.94% | 18.28 | 0.65 | 5.17 |
| **All** | **-57.65%** | **-7.45%** | **-31.94%** | **14.51** | **1.64** | **5.69** |

## 5. CONCLUSIONS

A novel VCM compression technique has been presented, which utilizes RoIs-based video coding. It has evolved from the state-of-the-art RoI-based preprocessing and retargeting tools originally proposed in [9]. The devised technique includes vital modifications and involves RoI detection, filtration, aggregation, and retargeting. Also, the problem of resolution change, related to technique [9] is addressed. This is achieved with the utilization of image padding which ensures that the spatial resolution of encoded video frames remains unchanged.

The proposed technique has been adapted to the architecture of MPEG VCM which allows interoperability with existing coding tools. The experimental evaluation demonstrates a significant improvement in coding performance with respect to the VCM Reference Software version 0.7 [28]. In most cases, significant average BD-Rate bitrate reductions were reported (even up to 57%) maintaining roughly unchanged the average mean precision of object detection and tracking.

Additionally, due to the reduced size of the useful information within frames encoded by the Inner Encoder, encoding and decoding times are significantly reduced. Exact numbers for these reductions are not cited, as the time measurement methodology settled within the VCM group is susceptible to computing environment variations.

ACKNOWLEDGEMENT: The work was supported by the Ministry of Education and Science of the Republic of Poland under the subvention for research SBAD.

## 6. REFERENCES

[1] G. J. Sullivan, J. Ohm, W. J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard”, in IEEE Trans. on Circuits Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[2] ITU-T Rec. H.265 | ISO/IEC IS 23008-2, “High efficiency coding and media delivery in heterogeneous environment – Part 2: High efficiency video coding”.

[3] J. Chen, Y. Ye, S. Kim, “Algorithm description for Versatile Video Coding and Test Model 3 (VTM3)”, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JVET L1002, Macao, October 2018.

[4] ISO/IEC DIS 23090-3 (2020) / ITU-T Recommendation H.266 (08/2020), “Versatile video coding”.

[5] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, “Overview of the versatile video coding (VVC) standard and its applications,” in IEEE Trans. on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3736–3764, 2021.

[6] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the h. 264/AVC video coding standard,” in IEEE Trans. on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 560 576, 2003.

[7] Xu, D., Chellappa, R., Van Gool, L. et al. Guest Editorial: Special Issue on Deep Learning for Video Analysis and Compression. Int J Comput Vis 129, 3171–3173 (2021). https://doi.org/10.1007/s11263-021-01530-3.

[8] Ning Xu , Weiyao Lin , Xiankai Lu , Yunchao Wei "Video Object Tracking: Tasks, Datasets, and Methods", Springer Synthesis Lectures on Computer Vision (SLCV), 2024.

[9] S. Różek, O. Stankiewicz, S. Maćkowiak and M. Domański, "Video Coding for Machines using Object Analysis and Standard Video Codecs", 2023 IEEE Int. Conference on Visual Communications and Image Processing (VCIP), pp. 1-5, Jeju, Korea, 2023.

[10] J. Chao, E. Steinbach, "Keypoint encoding for improved feature extraction from compressed video at low bitrates", IEEE Trans. on Multimedia 18(1), 25–39, 2016.

[11] L. Galteri, M. Bertini, L. Seidenari, A. Del Bimbo, "Video compression for object detection algorithms",

24th Int. Conference on Pattern Recognition (ICPR), pp. 3007–3012, 2018.
[12] S. Ma, X. Zhang, S. Wang, X. Zhang, C. Jia, S. Wang, "Joint feature and texture coding: Toward smart video representation via frontend intelligence", IEEE Trans. on Circuits and Systems for Video Technology, vol. 29, no. 10, pp. 3095–3105, 2019.
[13] L. Duan, J. Liu, W, Yang, T. Huang, W. Gao, "Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics", IEEE Trans. on Image Processing vol. 29, pp. 8680–8695, 2020.
[14] K. Fischer, F. Brand, C. Herglotz, A. Kaup, "Video Coding for Machines with Feature-Based Rate-Distortion Optimization", 22nd Int. Workshop on Multimedia Signal Processing (MMSP), 2020.
[15] Y. Lee, S. Kim, K. Yoon, H. Lim, S. Kwak, H.-G. Choo, "Machine-attention-based Video Coding for Machines", 2023 IEEE Int. Conference on Image Processing (ICIP), pp. 2700-2704 (2023).
[16] X. Li, J. Shi, and Z. Chen, "Task-driven semantic coding via reinforcement learning," IEEE Transactions on Image Processing, vol. 30, pp. 6307–6320, 2021.
[17] Q. Zhang, S. Wang, and S. Ma, "A novel visual analysis oriented rate control scheme for hevc," in 2020 IEEE Int. Conference on Visual Communications and Image Processing (VCIP). IEEE, 2020, pp. 491–494.
[18] Z. Huang, C. Jia, S. Wang, and S. Ma, "Visual analysis motivated rate-distortion model for image coding," in 2021 IEEE Int. Conference on Multimedia and Expo (ICME). IEEE, 2021, pp. 1–6.
[19] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Towards image understanding from deep compression without decoding," arXiv preprint arXiv:1803.06131, 2018.
[20] Y. Bai, X. Yang, X. Liu, et al., "Towards end-to-end image compression and analysis with transformers," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 104–112.
[21] "Conclusions of 127th MPEG meeting," ISO/IEC JTC 1/SC 29/WG 11, MPEG doc. N18540, July 2019.
[22] J. Ascenso, E. Upenik, "White Paper on JPEG AI Scope and Framework", ISO/IEC JTC 1/SC 29/WG1, MPEG doc. N90049, 2021.
[23] ISO/IEC JTC1/SC29/WG2, "Use cases and requirements for Video Coding for Machines", MPEG doc. N18, October 2020.
[24] ISO/IEC JTC1/SC29/WG2, "Use cases and requirements for Video Coding for Machines", MPEG doc. N0043, January 2021.
[25] ISO/IEC JTC 1/SC 29/WG 2, "Call for Proposals for Video Coding for Machines", MPEG doc. N191, April 2022.
[26] ISO/IEC JTC 1/SC 29/WG 4, "Algorithm description of tools in VCM reference software", MPEG doc. N418, December 2023.
[27] H. Chen, Y. Xu, "Video Coding for Machines Based on Motion Assisted Saliency Analysis", Lecture Notes in Computer Science book series, Springer LNCS, volume 14357, 2023
[28] "Video Coding for Machines Reference Software", https://mpeg.expert/software/MPEG/Video/VCM/
[29] S.-K. Kim, M. H. Jeong, J. Y. Lee, H.-K. Lee, H.-G. Choo, S.-H. Jung, "[VCM] CfP response: Region-of-Interest based video coding for machine", ISO/IEC JTC1/SC29/WG2 m60758, October 2022.
[30] J. Redmon, S. Divvala, R. Girshick, et al. "You Only Look Once: Unified, Real-Time Object Detection" In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 779-788), 2016.
[31] S. Ren, K. He, R. Girshick, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Trans. on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149, 2016.
[32] ISO/IEC JTC 1/SC 29/WG 04, "Common test conditions for video coding for machines", MPEG doc. N419, October 2023.
[33] H. Choi, E. Hosseini, S. R. Alvar, R. A. Cohen, I. V. Bajić, A. Karabutov, Z. Yin, E. Alshina, "[VCM] Object labelled dataset on raw video sequences," ISO/IEC JTC1/SC29/WG11 MPEG doc. m54737, July 2020.
[34] Y. Wu, A. Kirillov, F. Massa, et al. "Detectron2" https://github.com/facebookresearch/detectron2
[35] X. Xu, S. Liu and Z. Li, "A Video Dataset for Learning-based Visual Data Compression and Analysis," in 2021 Int. Conference on Visual Communications and Image Processing (VCIP), Dec. 2021.
[36] Z. Wang, L. Zheng, Y. Liu, et al. "Towards real-time multi-object tracking", in European Conference on Computer Vision (ECCV), p. 107-122, 2020.
[37] S. Akramullah, "Video Quality Metrics" in "Digital Video Concepts, Methods, and Metrics", Apress, Berkeley, CA, SpringerLink open access, 978-1-4302-6713-3, 2014.
[38] A. M. Tourapis, D. Singer, Y. Su, K. Mammou, "Bd-rate/BD-PSNR excel extensions", ISO/IEC JTC1/SC29/WG11 M41482, 2017.
[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge", Int. Journal of Computer Vision, 88(2), 303-338. doi:10.1007/s11263-009-0275-4, 2010.
[40] K. Bernardin, R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics", EURASIP Journal on Image and Video Processing, 2008(1), 1-10, 2008.