Adaptive Spatial Re-sampling Method for Video Coding for Machines

Eun-Bin An*, Ayoung Kim*, Soon-heung Jung, Hyon-Gon Choo and Kwang-deok Seo*
* Yonsei University, Wonju, Korea
E-mail: {eunbin.an, aykim90, kdseo}@yonsei.ac.kr Tel: +82-33-760-2788
Electronics and Telecommunications Research Institute, Daejeon, Korea
E-mail: {zeroone, hyongonchoo}@etri.re.kr Tel: +82-42-860-6891

***Abstract*— As the performance of machine vision continues to improve, it is being used in various industrial fields to analyze and generate massive amounts of video data. Although the demand for and consumption of video data by machines has increased significantly, video coding for machines needs to be improved. Spatial re-sampling plays a critical role in video coding for machines because it reduces the volume of the video data to be processed while maintaining the shape of the data's features that are important for the machine to reference when processing the video. An effective method of determining the intensity of spatial re-sampling as an efficient coding tool for machines is still in the early stages. Here, we propose a method of determining an optimal scale factor for spatial re-sampling by collecting and analyzing information on the number of objects and the ratio of the area occupied by the object within a picture.**

I. INTRODUCTION

As deep-learning machine vision is now used widely in various industrial and research fields, the amount of video data being produced and distributed, not only by humans but also by machines, has been increasing rapidly. To manage these ever- growing volumes of data, many video codecs have been introduced, such as advanced video coding (AVC), high-efficiency video coding (HEVC), and versatile video coding (VVC) [1–2]. Research has also been progressing on perceptually optimized video coding. However, codecs based on human visual or sensory systems do not guarantee efficient compression of video data when used for deep-learning machine vision. There is therefore a need to study video coding that targets video data consumed by machines rather than humans [3-4]. An understanding of how machines interpret video data is needed to develop unprecedented coding technology for machines rather than for humans. However, a variety of machine vision systems have been designed, and how video data are under- stood differs among these system. It is essential to maintain as much as possible the contours and other common features that machines collect to interpret video while increasing compression efficiency. Spatial re-sampling can easily reduce the volume of video data while significantly maintaining the position or arrangement of objects and boundaries, which are common features in video data. In this paper, we apply a new spatial re-sampling method to each picture in the input video to maximize compression efficiency while maximizing retention of the common features that the machines need to interpret the video data. In the proposed method, the ratio of the area occupied by a detected object in a picture is calculated, and the optimal scale factor for spatial re-sampling for each picture is determined using an object occupancy distribution (OOD), which is generated based on the size and number of detected objects. Figure 1 is a flowchart for determining the optimal scale factor for spatial re-sampling based on the proposed OOD.

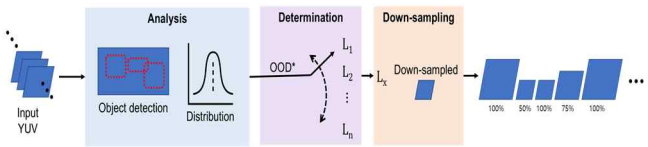

Fig. 1 Architecture of the proposed method for determining the optimal scale factor for spatial re-sampling based on OOD.

II. OOD FOR SPATIAL RE-SAMPLING

We used the test video sequences in the SFU-HW dataset [5], and explored the appropriate scale factors for spatial re-sampling for each video sequence through video data analysis and OOD generation.

A. Analysis of the impact of spatial re-sampling on machine vision Performance

Object detection was performed on the spatially up-sampled sequences at the same resolution as the original, and the performance results of the object detection accuracy (mAP) are organized by class and plotted in Figure 2. In Figure 2, several important phenomena are apparent. First, the scale factor for the spatial re-sampling increases as the object detection accuracy decreases because the boundaries of the image become increasingly blurred as the intensity of the spatial re-sampling strengthens.

Second, the machine accuracy drops significantly in a specific degree of re-sampling intensity depending on the class of the sequences, which are classified according to the resolution size.

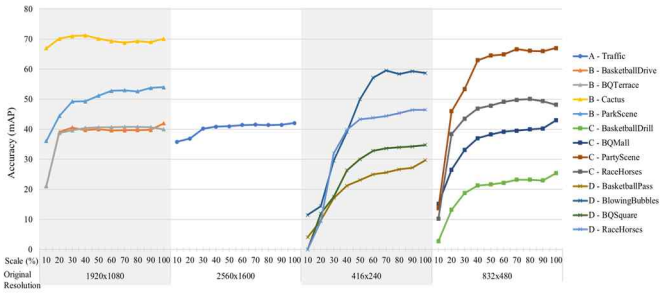

Fig. 2 Variation in object detection accuracy depending on scale factors for spatial re-sampling.

B. *OOD generation to determine an optimal re-sampling threshold*

In the analysis of Section A, the stronger the spatial re-sampling, the weaker the machine vision performance, and the object detection accuracy varies depending on the video resolution and the size and number of detected objects. The OOD is generated based on the detected object information including the size of detected objects within a single frame. The bounding box size of each detected object is divided by the original resolution size to obtain the object occupancy ratio (OOR). A histogram is then created depending on the OOR distribution. In simple terms, the OOR is the ratio of the detected object's size to the original resolution size, and the OOD is a histogram representing the distribution of the OOR. Figure 3 shows the experiment results related to OOD generation obtained by performing spatial down- sampling and object detection at increments of 10% from 10% to 100% of the original resolution for the 84th frame of the BasketballDrive sequence. Figure 3(a) depicts the results of object detection with bounding boxes for the spatial down-sampled pictures at various scales. Figure 3(b) depicts the generated OOD based on the detection results for each spatial down-sampled picture. It includes a histogram of object counts corresponding to various OOR ranges. Figure 3(c) is a graph expressing the kernel density estimation of Figure 3(b). Through Figure 3(c), the optimal re-sampling threshold can be explored visually. Figure 3(b) makes it clear that the distribution behaviors for the scale factors ranging from 50% to 100% are similar to each other. However, the distribution behaviors for scale factors below 50% are substantially different from those for the scale factors beyond 50%. We can therefore conclude that the optimal re-sampling threshold is 50%. Up to this threshold, we can keep the original OOD characteristics of the original picture with 100% resolution.

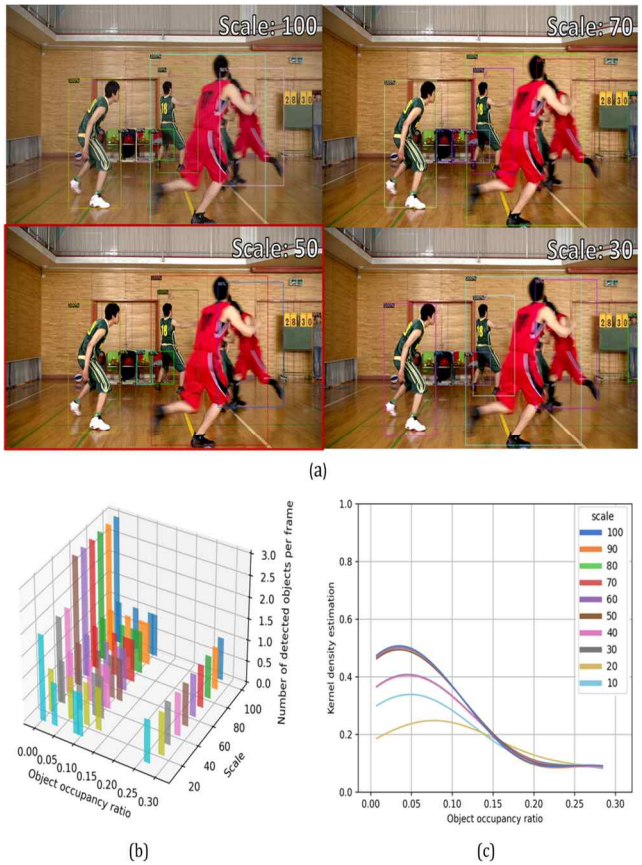

Fig. 3 Experiment results related to OOD generation for the 84th frame of the Bas- ketballDrive sequence: (a) Object detection results with bounding boxes for spatially down-sampled pictures with scale factors of 100%, 70%, 50%, and 30%, respectively, (b) Generated OODs and (c) Kernel density of the OOD.

III. DETERMINATION OF THE OPTIMAL SCALE FACTOR

Figure 4 is a flowchart of the process for determining the optimal scale factor for spatial re-sampling for each picture. The proposed process consists of four steps. First, the object detection network model generates object information, such as the bounding box size of detected object and the OOR for the original and spatial down- sampled pictures. Then, the OODs shown in Figure 4(b) are generated for all the candidate scale factors by using the object information. The next step is to analyze the correlation among the generated OODs to explore the similarity of the spatial down- sampled pictures to the original picture in terms of object detection performance. For analysis of similarity, we compared the correlation between the original picture with 100% resolution and the down-sampled picture with a scale factor x, where x is an element of the scale factor list SL which is {10%, 20%, 30%, . . . , 90%}. In this paper, the correlation was obtained using the OOD, which is the histogram representing the distribution of OOR.

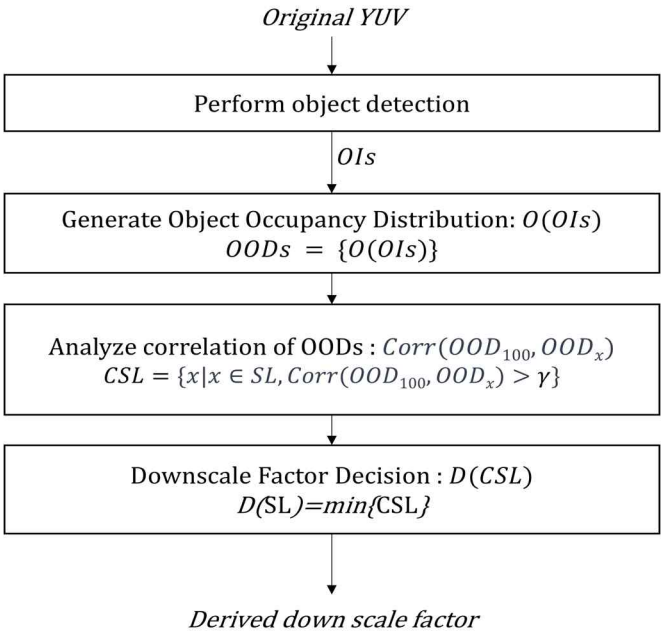

Fig. 4 Flowchart for determining the optimal scale factor for spatial re-sampling.

If the correlation exceeds the threshold value of $\gamma = 0.9$, the down-sampled picture is sufficiently similar to the original picture in terms of object detection. By gathering all the scale factors satisfying the threshold condition, we constitute the Correlated Scale- factor List (CSL). The CSL is a list of scale factors expected to have no significant difference in machine vision performance compared with the original picture. Among the CSL elements, the minimum scale factor is selected to be the optimal scale factor for re-sampling. If no element exists in the CSL, no definite optimal scale factor can be selected. In this case, the object detection performance even for the original picture with 100% resolution was found to be low. We applied the strongest spatial re-sampling ratio to achieve the utmost compression efficiency at the cost of slight machine performance loss.

Figure 5 is a conceptual diagram of an exemplary process for determining the opti- mal scale factor for spatial re-sampling among the candidate scale factors of 80%, 60%, 40%, and 20%. Because the OOD of the original picture and the OODs of the down-scaled pictures with 80% and 60% scale factors are estimated to be similar to each other, the scale factors of 80% and 60% are considered to be the candidates for an optimal scale factor for spatial re-sampling. Given the achievable compression efficiency, a scale factor of 60% corresponding to the lowest scale factor of the candidates was the optimal scale factor.

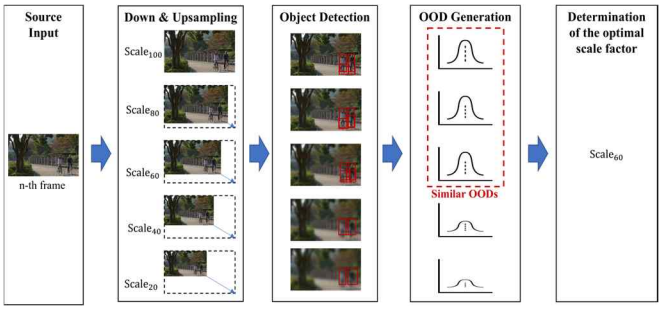

Fig. 5 Conceptual diagram of the proposed method for determining the optimal scale factor for frame-level spatial re-sampling.

IV. Experimental Results

This section describes the experimental verification of how much video data can be reduced and how much machine performance can be maintained by applying the proposed method. The experiment began by applying a re-sampling operation with a scale factor of {10%, 20%, 30%, . . . , 90%} to the original picture, after which object detection was performed using Faster-RCNN X101-FPN model. The proposed method determined the optimal scale factor for each picture. The machine accuracy error, ε, and the video data reduction ratio, C, were used to evaluate the performance of the proposed method. The machine accuracy error ε means the performance decrease in the object detection that occurs when the proposed spatial re-sampling with an optimal scale factor is applied compared with the object detection accuracy achieved for the original picture with 100% resolution (Equation 1). In this equation, $Accuracy_o$ represents the object detection accuracy for the original picture and $Accuracy_p$ is the object detection accuracy obtained for the down-sampled picture by the proposed method. The machine accuracy error ratio E_r is defined as the ratio between ε and $Accuracy_o$ as shown in Equation 2. C and C_r represent the data reduction ratio and rate, respectively, where C is the ratio of the original video data rate to the down-sampled video data rate, and C_r is the corresponding reduction rate expressed as a percentage. These are defined in Equations 3 and 4. From Equation 1 to 4 , the subscript o and p indicate the original picture and the down-sampled picture obtained by the proposed method, respectively

$$\varepsilon = \left| Accuracy_o - Accuracy_p \right| \geq 0 \quad (1)$$

$$E_r = \varepsilon / Accuracy_o \quad (2)$$

$$C = \text{datarat}\ \ e_o / \text{datarat}\ \ e_p \geq 1 \quad (3)$$

$$C_r = \left(1 - \frac{\text{datarat}\ \ e_p}{\text{datarat}\ \ e_o}\right) \times 100 \quad (4)$$

The data reduction–to–accuracy error ratio ($DRAER$) in Equation 5 is used to evaluate the achieved reduction in the video data volume against the obtained E_r.

$$DRAER = 10 * \text{b}\ g_{10}(C/E_r) \quad (dB) \quad (5)$$

where the E_r excludes the case of value 0 indicating no change in the resolution of the video data after applying the proposed method, that cannot happen in practice.

Table I is the number of video frames classified to each scale factor for spatial down- sampling. The behavior of the classification depends highly on the characteristics of the video frames in terms of the size and the number of objects detected.

Table I. Number of video frames classified to each scale factor for spatial down- sampling.

Sequence		Number of video frames per each scale factor									
Class	name	100	90	80	70	60	50	40	30	20	10
A	Traffic	0	0	0	0	0	1	1	0	3	28
B	ParkScene	1	4	2	4	2	4	5	4	3	4
B	Cactus	13	11	1	5	8	0	0	4	10	45
B	BasketballDrive	8	13	3	5	11	9	11	22	15	0
B	BQTerrace	0	0	0	0	0	0	0	0	27	102
C	RaceHorsesC	62	12	11	6	2	2	1	1	0	0
C	BQMall	62	23	14	12	8	5	3	2	0	0
C	PartyScene	27	19	14	7	4	7	8	8	3	0
C	BasketballDrill	20	12	7	5	4	5	3	5	4	0
D	RaceHorsesD	40	19	12	8	4	4	5	4	1	0
D	BQSquare	21	31	9	11	5	11	39	2	0	0
D	BlowingBubbles	28	24	21	10	8	5	0	0	1	0
D	BasketballPass	19	10	6	7	6	7	6	4	0	0

Table II describes the experimental results obtained using the proposed method for the SFU-HW dataset. It compares the mAP for the original video with 100% resolution and the mAP for the down-sampled video by the proposed method, and shows the value of E_r, as well as C. In most cases, the E_r related to the ability to keep the object detection accuracy is less than 10%, which is a satisfactory performance. Only the Traffic and BQTerrace sequences have E_r values exceeding 10% (16.97% and 37.44%, respectively). Except for a few sequences, the E_r is stays within 10%, indicating that the machine vision performance can be stable for most test sequences. As for the exceptional cases in which the Traffic and BQTerrace sequences were used for the test, most of the video frames are classified into the 10% scale factor.

Table II. Experimental results of the proposed method for the SFU-HW dataset.

Sequence Name	Orig. (mAP)	Prop. (mAP)	E_r	C	C_r	$DRAER$ (dB)
Traffic	42.13	34.98	16.97%	7.67	86.96%	16.55
ParkScene	54.07	49.14	9.11%	2.01	50.29%	13.44
Cactus	70.22	68.98	1.77%	2.44	59.07%	21.40
Basketball Drive	42.04	40.95	2.58%	1.91	47.71%	18.70
BQTerrace	40.04	25.05	37.44%	8.27	87.90%	13.44
RaceHorse sC	25.43	24.41	4.00%	1.09	8.60%	14.37
BQMall	43.02	41.39	3.80%	1.16	13.66%	14.84
PartyScene	67.05	65.67	2.06%	1.34	25.33%	18.13
Basketball Drill	48.22	49.77	3.23%	1.35	25.73%	16.20
RaceHorse sD	29.79	28.50	4.32%	1.21	17.30%	14.47
BQSquare	34.85	31.59	9.35%	1.46	31.42%	11.93
BlowingBu bbles	58.73	59.72	1.69%	1.20	16.42%	18.50
Basketball Pass	46.51	44.93	3.40%	1.34	25.61%	15.97

V. CONCLUSIONS

In this paper, we devised an OOD to determine the optimal scale factor for spatial re-sampling to optimize video compression efficiency for machines. The OOD was created by applying the area information occupied by the detected object in the picture. By analyzing the similarity of OODs for various candidate scale factors between the down-sampled and the original pictures, we were able to obtain the optimal scale factor for the frame-level re-sampling.

REFERENCES

[1] ISO/IEC: Information technology–coding of audio-visual objects–part 10: Advanced video coding. ISO/IEC 14496-10 (2022).

[2] ISO/IEC: Information technology–coded representation of immersive media–part 3: Versatile video coding. ISO/IEC 23090-3 (2022).

[3] H. Choi and I. Bajić, "Scalable image coding for humans and machines," IEEE Trans. Image Process. vol. 31, 2022, pp. 2739–2754.

[4] W. Yang, H. Huang, Y. Hu, L. Duan, and J. Liu, "Video coding for machines: Compact visual representation compression for intelligent collaborative analytics," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 46, July 2024, pp. 5174-5191.

[5] H. Choi, E. Hosseini and I. Bajić: SFU- HW-Objects-v1: Object labelled dataset on raw video sequences. https://doi.org/10.25314/7d8efc0a-3943-4738-b7a5-72badb04d765 (2020).