Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.918
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

An Advanced Multi-Scale Feature Compression (advMSFC) Method with Backbone Split Point and Its Performance Analysis

Gyu-Woong Han[a)], Yong-Uk Yoon[a)], and Jae-Gon Kim[a)‡]

Abstract

For more efficient video compression for machine vision applications, MPEG is developing a new video coding standard called Video Coding for Machines (VCM) optimized for videos consumed by machines. In this paper, we present an advanced Multi-Scale Feature Compression (advMSFC) method to compress features extracted at Backbone split point of machine vision networks and analyzes its performance in terms of vision task accuracy against bit rate. The proposed method demonstrates up to 98.16% Bjontegaard Delta-rate (BD-rate) gain compared to the Feature Compression for VCM (FCVCM) feature anchor in object tracking vision tasks. This performance indicates that, compared to the DN53 and ALT1 split points, the advMSFC performance at the Backbone split point exhibits a BD-rate gain of over 5%.

Keyword: Video Coding for Machines (VCM), Multi-Scale Feature Compression (MSFC), Selective Learning Strategy (SLS), Backbone Split Point, Feature Coding for Machines (FCM)

I. Introduction

Machine vision-based applications are widely utilized in various fields such as surveillance video, smart cities, the Internet of Things, and autonomous driving. These applications utilize artificial intelligence to analyze video collected by machines, enabling them to detect and recognize events or objects. To efficiently process and compress the large volume of video collected by machines, the Moving Picture Experts Group (MPEG) is currently developing a new video coding standard named Video Coding for Machines (VCM)[1]. This standard aims to achieve efficient coding of video consumed by machines in vision applications.

MPEG-VCM consists of two main groups: VCM and Feature Compression for VCM (FCVCM)[2]. VCM focuses on customized image and video compression methods tailored to enhance the performance of machine vision tasks, including end-to-end learning-based compression methods

a) Department of Electronics and Information Engineering, Korea Aerospace University
‡ Corresponding Author : Jae-Gon Kim
E-mail: jgkim@kau.ac.kr
Tel: +82-2-300-0414
ORCID: https://orcid.org/0000-0003-3686-4786

※This work was supported by National Standards Technology Promotion Program of Korean Agency for Technology and Standards (KATS) grant funded by the Korea government (MOTIE) (20011687).
· Manuscript November 12, 2023; Revised December 5, 2023; Accepted December 5, 2023.

Copyright Ⓒ 2023 Korean Institute of Broadcast and Media Engineers. All rights reserved.
"This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered."

and interest-region-based compression methods. On the other hand, FCVCM deals with feature extraction and compression methods specifically designed for machine vision tasks. This involves developing techniques to compress features extracted from machine vision networks.

Recently, in FCVCM, the feature compression method based on Multi-Sclae Feature Compression (MSFC)[3] has attracted significant attention due to its high compression performance[4]-[6]. In this paper, we apply the multi-scale feature compression (advMSFC) method, which utilizes a selective learning strategy (SLS) and QP-adaptive feature channel truncation (QACT)[7], to the compression of feature extracted at the Backbone split point for FCVCM. In addition, the details on its performance analysis are presented. The proposed advMSFC method compresses features of the Backbone split point using Versatile Video Coding (VVC)[8], the latest video coding standard, to achieve optimal performance in terms of task accuracy and bitrate. This is achieved using a single trained MSFC model rather than having individually trained models for each Quantization Parameter (QP). By using Backbone split points for advMSFC, relatively independent features can be extracted, and error propagation due to the hierarchical structure of the network can be prevented, resulting in improved performance compared to other split points such as DN53[11] and ALT1[12].

The proposed advMSFC was submitted as a response to the Call for Proposal (CfP) of FCVCM. In the MPEG meeting held in Oct. 2023, based on the evaluation of proposed technologies, the standardization of FCVCM was moved from Working Group 2 (WG 2) to WG 4, the standard name of FCVCM was changed to Feature Coding for Machines (FCM).

II. advMSFC with Backbone split point

MPEG-VCM defines an evaluation network and evaluation dataset to evaluate the performance of proposed technologies for object tracking tasks, as shown in Table 1. In order to specify different features to be compressed depending on the dataset, the feature output split points of the network are specified differently. m63045[9], which was recently announced at the 142nd MPEG meeting, presented a technique for compressing multi-scale features extracted from MSFC-based ALT1 split points, and showed performance improvement of BD-rate reduction of more than 90% compared to feature anchor performance[10]. In this paper, MSFC-based compression is performed by extracting features from split points other than those defined in MPEG-VCM.

Table 1. Split points according to dataset

Network	Dataset	Split point	Layer index
JDE-1088x608	TVD	DN53 [11]	[36, 61, 74]
	HiEve	ALT1 [12]	[75, 90, 105]

1. Backbone Split Point

Figure 1 briefly shows the structure of JDE-1088x608, an evaluation network for object tracking tasks defined in MPEG-VCM[13]. In the proposed method, the split point that extracts the compressed features is named the Backbone split point. The Backbone split point is a split point that can extract multi-scale features that are directly input to the predictor when the evaluation network is divided into backbone and predictor, and can be considered as the high-level feature finally extracted from the input image. Therefore the features extracted from the Backbone split point represent the highest-level features, and each scale feature is input independently to the predictor network. On the other hand, the features extracted from the DN53 and ALT1 split points exhibit relatively lower performance due to the possible propagation of compression errors, as the scale features have interdependencies that contribute to the generation of high-level features input to the predictor. The sizes of the three extracted multi-scale

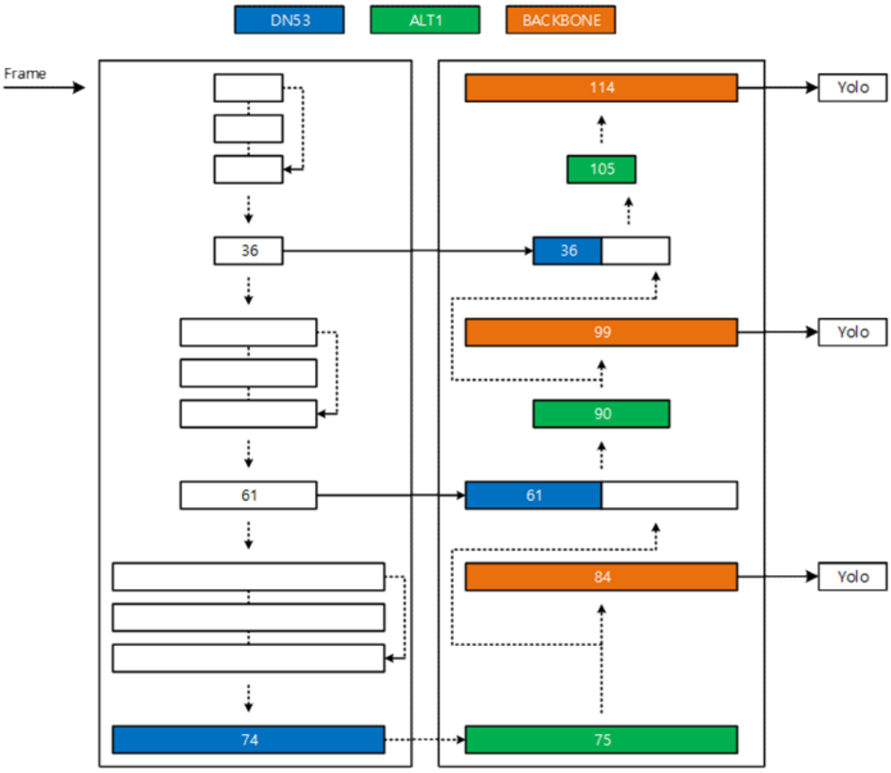


Fig. 1. Structure of JDE-1088x608 and Split points of DN53, ALT1, and Backbone

features can be found in Table 2. The advMSFC method is designed to accommodate the corresponding multi-scale features.

2. Overall Framework

The advMSFC Network consist of two modules: the multi-scale feature fusion (MSFF) module, which converts multi-scale features into a single-scale feature, and the multi-scale feature reconstruction (MSFR) module, which reconstructs the single-scale feature back into multi-scale features. Both modules consist of a few convolutional layers. Figure 2 illustrates overall framework of the proposed method. The advMSFC network is integrated into a

Table 2. Size of multi-scale feature and MSFF output according to split points

Split point	Layer index	Size (CxHxW)	The size of MSFF output (Single-scale feature)
DN53	36	256x76x136	1024x19x34
	61	512x38x68	
	74	1024x19x34	
ALT1	75	512x19x34	512x19x34
	90	256x38x68	
	105	128x76x136	
Backbone	84	536x19x34	536x19x34
	99	536x38x68	
	114	536x76x136	

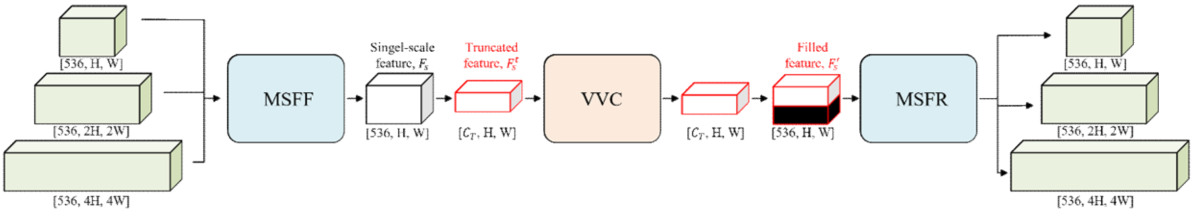


Fig. 2. Overall framework of the MSFC using SLS and QACT with Backbone Split point in object tracking

FCVCM machine vision task network[14], and the advMSFC network is trained along with a pretrained task network, which remains frozen during the training phase. During advMSFC network training, the SLS is applied, allowing the MSFF to output a single-scale feature arranged based on channel-wise importance. This approach enables adaptive adjustment of the data size of features to be compressed, obtained from a single trained MSFC model.

The MSFF multi-scale features extracted from the FCVCM task network's backbone as input. Through the MSFF module, these multi-scale features are converted into a single-scale feature, F_s, arranged based on channel-wise importance. The size of the converted feature can be adjusted through QP-adaptive channel truncation. The truncated feature, F_s^t, is then packed into a 2D frame feature map in a channel-wise manner. Figure 3 shows 2D frame packed single scale feature arranged based on the channel-wise importance to be compressed. This packed feature map is encoded using the VVC encoder, resulting in a single bitstream.

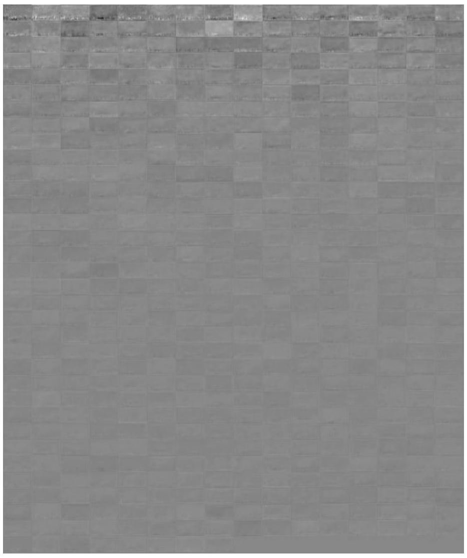


Fig. 3. 2D frame-packed single-scale feature of Backbone split point

The MSFR restores the single bitstream into multi-scale features in the following steps. The feature map decoded by the VVC decoder is unpacked into its original tensor form. If the unpacked feature is a truncated feature on the encoder side, the truncated channels are filled with zeros to restore the single-scale feature, $F_s^{'}$, to its original size before truncation. Finally, the single-scale feature is reconstructed back into multi-scale features through the MSFR module, and these multi-scale features are fed back into the predictor network.

Table 3. Pre-defined QP-adaptive channel numbers for each split points

Split point	Pre-defined QP-adaptive channel numbers
DN53	[1024, 768, 512, 364, 256, 192]
ALT1	[512, 256, 192, 128, 96, 64]
Backbone	[536, 256, 192, 128, 96, 64]

3. Selective Learning Strategy

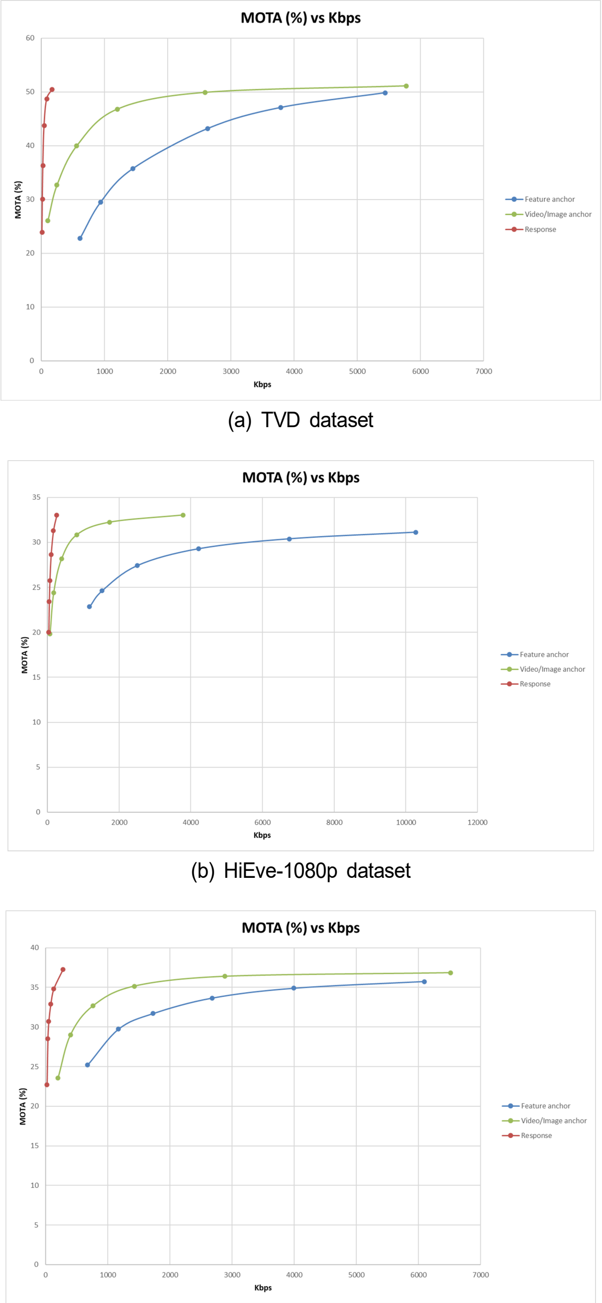
The advMSFC achieves efficient feature compression by adaptively truncating feature channels based on their importance according to QP. The MSFF module arranges multi-scale features into a single-scale feature, enabling the adjustment of its size through channel-wise truncation. By leveraging the SLS, the MSFC network is trained in an end-to-end manner, involving only a stochastically determined number of consecutive channels in each training step. Lower-index channels are involved more frequently, implying their higher importance. During training, the MSFR module is also trained to reconstruct multi-scale features from the masked single-scale feature, contributing to improved performance.

The MSFF module converts into a single-scale feature F_s, and then the channel-wise binary masking is applied using a uniform random variable, n, which determines the number of activated channels ranging from 0 to the total number of channels in F_s, resulting in a masked single scale feature F_s^m. By iteratively applying the SLS during training, low-index channels are involved more frequently in the MSFR module, influencing the feature map to be sorted based on channel importance, resulting in enhanced performance.

III. Experimental results

The advMSFC network trained specifically for the Backbone split point was evaluated on the TVD and HiEve datasets using the same network of object tracking. The proposed advMSFC network was trained along with the pre-trained task network which is frozen in the training. Experiments followed the FCVCM common test conditions (CTCs)[15] and utilized the feature anchor scripts[11][12]. The proposed advMSFC network performance was compared to that of the feature anchor. As shown, the advMSFC with Backbone split point gives 98.16%, 96.84%, and 96.31% BD-rate gain for TVD dataset, HiEce-1080p dataset, and HiEve-720p dataset, respectively. MOTA performance at all RP points meets the required MOTA range specified in the Call for Proposals (CfP)[10]. Figure 4 shows the RP curves of the results. The red line is the performance of the Backbone split points, and the other two are the performance of the feature anchor and video/image

anchor. Table 4 shows the performance comparison of the proposed advMSFC network according to the three types of split points. The performance of the Backbone split point is the best across all sequences.



(a) TVD dataset

(b) HiEve-1080p dataset

(c) HiEve-720p dataset

Fig. 4. RP curves of the advMSFC network with Backbone split point

Table 4. The performance comparison of advMSFC according to split points

Split point	Backbone			DN53	ALT1	
Sequence	TVD	HiEve-1080p	HiEve-720p	TVD	HiEve-1080p	HiEve-720p
BD-rate	-98.16%	-96.84%	-96.31%	-94.66%	-91.13%	-82.96%

Ⅳ. Conclusion

This paper presents an advanced Multi-Scale Feature Compression (advMSFC) method to compress features extracted at Backbone split point of machine vision networks and its performance analysis. The experimental results showed that the proposed method gives BD-rate gains of 98.16%, 96.84% and 96.31% for TVD, HiEve-1080p, and HiEve-720p, respectively. The proposed advMSFC was submitted as a response to the CfP of FCVCM and evaluated as one of the key technologies. Core Experiments (CEs) are scheduled to be conducted for further validation and investigation on the key proposed technologies including the proposed advMSFC.

Compared to the DN53 and ALT1 split points, the performance of advMSFC with Backbone split point shows a BD-rate gain of over 5%. This result indicates the importance of considering the feature characteristics based on the structure of the task network for efficient feature compression. For example, in the hierarchical structure of the JDE-1088x608, features extracted from the Backbone split point are independent when compared to feature extracted from DN53 and ALT1 split points. This can reduce the potential risk of compression error propagation. Furthermore, this suggests the need for further investigation on split points to compress features of various networks.

References

[1] Y. Zhang, “AHG on Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, m49944, Sep. 2019.

[2] C. Rosewarne and Y. Zhang, “AHG on Feature Compression for Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, m61552, Jan. 2022.

[3] Z. Zhang, M. Wang, M. Ma, J. Li, and L. Fan, “MSFC: Deep Feature Compression in Multi-Task Network,” in Proc. 2021 IEEE Int. Conf. Mult. Expo (ICME), Shezhen, China, 2021, pp. 1-6. doi: https://doi.org/10.1109/ICME51207.2021.9428258.

[4] H. Han, M. Choi, H. Choi, S.-H. Jung, S. Kwak, J. Lee, H.-G. Choo, W.-S. Cheong, and J. Seo, “[VCM] Response from Hanbat National University and ETRI to CfE on Video Coding for Machine,” ISO/IEC JTC 1/SC 29/WG 2, m60761, Oct. 2022.

[5] Y.-U. Yoon, D. Kim, J.-G. Kim, J. Lee, S. Jeong, and Y. Kim “[VCM] Response to VCM CfE: Multi-scale feature compression with QP adaptive feature channel truncation,” ISO/IEC JTC 1/SC 29/WG 2, m60799, Oct. 2022.

[6] C. Rosewarne and R. Nguyen, “[VCM Track 1] Response to CfE on Video Coding for Machine from Canon,” ISO/IEC JTC 1/SC 29/WG 2, m60821, Oct. 2022.

[7] Y. -U. Yoon, D. Kim, J. Lee, B. T. Oh and J. -G. Kim, “An Efficient Multi-Scale Feature Compression With QP-Adaptive Feature Channel Truncation for Video Coding for Machines,” IEEE Access, vol. 11, pp. 92443-92458, Aug. 2023. doi: https://doi.org/10.1109/ACCESS.2023.3307404.

[8] Versatile Video Coding (VVC), ISO/IEC 23090-3, 2020.

[9] “Experimental results of enhanced MSFC based on split point for HiEve in object tracking task”, ISO/IEC JTC 1/SC 29/WG 2, m63045, Apr. 2023

[10] “Call for Proposals on Feature Compression for Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, N00282, Apr. 2023.

[11] C. Rosewarne, and R. Nguyen, “TVD object tracking feature anchor update,” ISO/IEC JTC 1/SC 29/WG 2, m62504, Apr. 2023.

[12] C. Rosewarne, and R. Nguyen, “HiEve object tracking feature anchor,” ISO/IEC JTC 1/SC 29/WG 2, m62505, Apr. 2023.

[13] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, 2020, “Towards RealTime Multi-Object Tracking,” https://github.com/Zhongdao/TowardsRealtime-MOT (accessed Aug. 24, 2022).

[14] “Evaluation Framework for Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, N00220, Jul. 2022.

[15] “Common Test Conditions and Evaluation Methodology for Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, N00231, Jul. 2022.

Introduction Authors

Gyu-Woong Han

- 2023 : Korea Aerospace University, School of Electronics and Information Engineering (B.S.)
- 2023 ~ present : Korea Aerospace University, School of Electronics and Information Engineering (Pursuing Master)
- ORCID : https://orcid.org/0009-0009-5711-781X
- Research interests : video coding for machines, video coding, image processing

Yong-Uk Yoon

- 2017 : Korea Aerospace University, School of Electronics and Information Engineering (B.S.)
- 2019 : Korea Aerospace University, Dept. Electronics and Information Engineering (M.S.)
- 2023 : Korea Aerospace University, Dept. Electronics and Information Engineering (Ph.D)
- 2023 ~ present : Tencent Americas, Tencent Media Lab., Senior Researcher
- ORCID : http://orcid.org/0000-0002-5105-5437
- Research interests : video coding, immersive video, video coding for machines, image processing

Jae-Gon Kim

- 1990 : Kyungpook National University, Dept. Electronic Engineering (B.S.)
- 1992 : KAIST, Dept. Electrical and Electronic Engineering (M.S.)
- 2005 : KAIST, Dept. Electrical Engineering and Computer Science (Ph.D)
- 1992 ~ 2007 : ETRI, Broadcasting Media Research Group, Senior Researcher / Team Leader
- 2001 ~ 2002 : Columbia University, NY, Dept. Electrical Engineering, Staff Associate
- 2015 ~ 2016 : UC San Diego, Video Signal Processing Lab., Visiting Scholar
- 2007 ~ present : Korea Aerospace University, School of Electronics and Information Engineering, Professor
- ORCID : http://orcid.org/0000-0003-3686-4786
- Research interests : video compression, video signal processing, immersive video