

특집논문 (Special Paper)

방송공학회논문지 제30권 제3호, 2025년 5월 (JBE Vol.30, No.3, May 2025)

<https://doi.org/10.5909/JBE.2025.30.3.301>

ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

## VCM 표준화에서의 성능 평가 방법 이슈 및 동향

정민혁<sup>a)</sup>, 이예지<sup>b)</sup>, 이희경<sup>b)</sup>, 이진영<sup>b)</sup>, 정순홍<sup>b)</sup>, 김상균<sup>a)†</sup>

### Issues and Progress in Performance Evaluation Methods for VCM Standardization

Min Hyuk Jeong<sup>a)</sup>, Yegi Lee<sup>b)</sup>, HeeKyung Lee<sup>b)</sup>, Jin Young Lee<sup>b)</sup>, Soon-heung Jung<sup>b)</sup>, and Sang-Kyun Kim<sup>a)†</sup>

#### 요약

MPEG의 비디오 그룹에서는 VCM이라는 이름으로 기계 기반 영상 인식 및 분석을 위한 비디오 압축 기술의 표준화를 진행하고 있다. 최근 143차 회의에서는 WD(Working Draft)의 초안이 작성되었으며, 150차 회의에서는 CD(Committee Draft) 단계로의 진입을 위한 논의가 계획되어 있다. 본 논문에서는 지난 약 2년 동안 VCM 표준화 과정에서 제기된 성능 평가 방법과 머신 성능 저하 관련 주요 이슈들을 종합적으로 분석하고 논의한다. 특히, 비디오 압축으로 인해 머신러닝 성능이 저하되는 문제를 검토하며, 이를 완화하거나 성능을 유지 및 향상시키기 위해 최근 연구된 다양한 보완 기술들을 소개한다. 이를 통해 향후 VCM 표준화 과정에서 고려해야 할 핵심 이슈들과 향후 발전 방향에 대한 실질적이고 명확한 통찰을 제공하고자 한다.

#### Abstract

The video group of MPEG is standardizing video compression technologies for machine-based video recognition and analysis under the name of VCM. At the 143rd meeting, a preliminary Working Draft (WD) was released, and discussions for progressing to the Committee Draft (CD) stage will be planned for the 150th meeting. This paper comprehensively reviews key issues related to performance evaluation methods and machine-learning performance degradation raised throughout approximately two years of VCM standardization. In particular, we deeply analyze performance degradation problems in machine tasks caused by video compression and introduce various recently proposed complementary technologies aimed at mitigating these issues while maintaining or enhancing machine-learning performance. Based on this analysis, we provide clear and practical insights regarding critical issues and future directions to be considered in the VCM standardization process.

Keyword : MPEG, VCM, Machine Performance, Video Compression, Common Test Condition

a) 명지대학교(Myongji University)

b) 한국전자통신연구원(ETRI)

† Corresponding Author : 김상균(Sang-Kyun Kim)

E-mail: goldmunt@gmail.com

Tel: +82-31-330-6443

ORCID: <https://orcid.org/0000-0002-2359-8709>

※ This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00011, Video Coding for Machine)

· Manuscript March 26, 2025; Revised April 27, 2025; Accepted April 30, 2025.

Copyright © 2025 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.”

## I. 서론

VCM은 MPEG에서 표준화를 진행 중인 기계 기반 영상 인식 및 분석을 위한 비디오 압축 표준 기술로서, 비디오 그룹에서 관련 논의를 진행하고 있다. VCM은 기존의 사람이 보는 영상을 위한 압축 방식과는 달리, 인공지능 및 머신러닝 시스템이 영상을 효율적으로 인식하고 처리할 수 있도록 최적화된 압축 기술을 개발하고 표준화하는 것을 목표로 하고 있다. VCM 표준은 회의 기간 중에는 VCM BoG(Breakout Group), 회의와 회의 사이에는 VCM AhG (Adhoc Group)에서 주로 논의가 진행되므로, 본 논문에서는 VCM BoG와 AhG를 VCM이라 칭한다.

VCM 표준은 143차 회의에서 WD(Working Draft) 초안 문서<sup>[1]</sup>가 작성되었으며, 현재 WD 단계에 있다. 지난 149차 회의에서 WG(Working Group) 04가 발간한 문서에 따르면, VCM 표준화 일정은 2025년 4월에 CD(Committee Draft) 단계 진입이 예정되어 있으며, 2025년 7월에 DIS(Draft International Standard), 2026년 1월에 FDIS(Final Draft International Standard), 최종적으로 2026년 7월에 IS(International Standard)의 발간이 계획되어 있다<sup>[2]</sup>.

VCM에서는 제안된 기술들의 성능 평가와 비교를 위해 테스트 데이터셋, 평가 신경망, 압축률, 성능 평가 방법 등을 명시한 CTC(Common Test Conditions) 문서를 매 회의마다 업데이트하고 있다. 각 기술 제안 기관은 가장 최근 회의에서 발표된 CTC 문서를 기반으로 성능 평가 결과를 보고하며, 이를 근거로 표준으로 채택될 기술이 결정된다.

인간 시각적 인지를 위한 영상 압축 기술은 지난 30년 이상 지속적인 연구와 표준화를 통해 발전해왔으나, 기계 기반 영상 인식 및 분석을 위한 압축 기술인 VCM은 비교적 최근인 2019년부터 MPEG에서 본격적으로 논의되기 시작하여 표준화 역사가 상대적으로 짧다. 이로 인해 지난 약 2년간 WG 04에서 VCM 표준화를 진행하면서 성능 평가 방법과 머신 성능 저하 등 다양한 이슈가 제기되었으며, 이와 관련된 다양한 논의가 진행되었다. 본 논문에서는 VCM 표준화 과정에서 논의된 성능 평가 방법과 머신 성능 저하 관련 주요 이슈들을 종합적으로 기술하며 분석한다. 특히 VCM의 핵심적인 문제 중 하나인 머신 성능 저하 현상을

검토하고, 이를 완화하거나 성능을 유지 및 향상시키기 위해 최근 제안된 보완 기술을 소개한다.

본 논문의 구성은 다음과 같다. 2장에서는 CTC에 서술된 테스트 데이터셋, 신경망 및 평가 방법 등을 서술한다. 3장에서는 CTC와 관련하여 논의된 이슈와 관련 기고들에 대해 서술하며, 4장에서는 VCM 성능저하 이슈 및 이를 보완하기 위해 제안된 기술에 대해 기술한다. 마지막으로, 5장에서는 본 논문의 결론을 도출하고, 향후 표준화 방향에 대한 서술을 하고자 한다.

## II. VCM 성능 평가 방법

VCM CTC 문서에는 VCM 그룹 내에서 제안된 기술의 공정한 평가와 객관적인 비교를 위해 테스트 데이터셋, 평가 신경망, 평가 방법, 기술 제안 시 필수 또는 선택적으로 제출해야 할 결과와 크로스체크 기준 등이 서술되어 있다. 또한 앵커 결과 및 엑셀 템플릿을 제공하고 있으며, 매 회의마다 갱신되고 있다.

VCM에서는 평가를 위한 테스트 데이터셋이 정의되어 있으며, 크게 비디오 데이터셋과 이미지 데이터셋으로 나누어져 있다. 비디오 데이터셋에는 SFU-HW 데이터셋<sup>[4]</sup>, TVD(Tencent Video Dataset)-video 데이터셋<sup>[5]</sup>, PandaSet<sup>[6]</sup>이 있다. SFU-HW 데이터셋은 14개 시퀀스로 이루어져 있으며, 다양한 해상도 및 프레임레이트로 구성되어 있다. 표 1은 CTC 문서에 기술되어 있는 해상도 및 프레임레이트 정보와 인코딩할 프레임 정보를 보여준다. 이 중 Kimono와 Cactus 시퀀스는 O 클래스로 분류되어 있으며 불안정한 평가 결과를 초래할 가능성이 있다는 의견으로 인해 필수 실험 시퀀스에서 제외되었다. TVD-video 데이터셋은 1920x1080 해상도의 총 10개 시퀀스로 구성되어 있으며, 각각 3개 시퀀스와 7개 시퀀스로 표 2와 같이 나누어 사용된다. PandaSet은 자율주행 연구 목적으로 설계된 고품질 비디오 데이터셋으로, 1920x1080 해상도의 총 106개 시퀀스 중 73개 시퀀스가 성능 평가를 위해 CTC에 포함되었다. PandaSet은 표 3에 명시된 것과 같이 총 6개의 클래스로 구성되어 있으며, 필수 데이터셋이 아닌 선택 제출 데이터셋으로 분류된다.

표 1. SFU-HW 데이터셋 구성

Table 1. SFU-HW dataset configuration

Class	Sequence	Total frames	Frame rate	Bit depth	Frames skipped	Frames coded
A	Traffic	150	30	8	117	33
B	ParkScene	240	24	8	207	33
B	BasketballDrive	500	50	8	403	97
B	BQTerrace	600	60	8	471	129
C	RaceHorsesC	300	30	8	235	65
C	BQMall	600	60	8	471	129
C	PartyScene	500	50	8	403	97
C	BasketballDrill	500	50	8	403	97
D	RaceHorsesD	300	30	8	235	65
D	BQSquare	600	60	8	471	129
D	BlowingBubbles	500	50	8	403	97
D	BasketballPass	500	50	8	403	97
O	Kimono	240	24	8	207	33
O	Cactus	500	50	8	403	97

표 2. TVD-video 데이터셋 구성

Table 2. TVD dataset configuration

Sequence	Frame count	Frame rate	Bit depth	Frames skipped	Frames coded
TVD-01-1	3000	50	8	1500	500
TVD-01-2	3000	50	8	2000	500
TVD-01-3	3000	50	8	2500	500
TVD-02-1	636	50	10	0	636
TVD-03-1	2334	50	10	0	500
TVD-03-2	2334	50	10	500	500
TVD-03-3	2334	50	10	1000	500

표 3. Pandaset 데이터 구성

Table 3. Pandaset dataset configuration

Class	Sequence ID	Frame count	Frame rate	Bit depth	Frames skipped	Frames coded
A1	57, 58, 69, 70, 72, 73, 77	80	10	8	0	0
B1	3, 11, 16, 17, 21, 23, 27, 29, 30, 33, 35, 37, 39, 43, 53, 56, 97	80	10	8	0	0
C1	88, 89, 90, 95, 109, 112, 113, 115, 117, 119, 122, 124	80	10	8	0	0
A2	64, 65, 66, 67, 71, 78, 149	80	10	8	0	0
B2	1, 5, 13, 15, 19, 24, 28, 32, 34, 38, 40, 41, 42, 44, 46, 52, 54, 139	80	10	8	0	0
C2	80, 84, 94, 101, 102, 103, 105, 106, 110, 116, 123, 158	80	10	8	0	0

이미지 데이터셋으로는 TVD-image 데이터셋<sup>[5]</sup>, Open-Images v6 데이터셋<sup>[7]</sup>, FLIR 데이터셋<sup>[8]</sup>이 있다. TVD-image 데이터셋은 1920x1080 해상도의 총 166개 이미지로 구성되어 있으며, OpenImages v6 데이터셋은 약 900만 장의 이미지 중에서 VCM 평가를 위해 5,000개의 이미지가 선별되어 사용된다. FLIR 데이터셋은 640x480의 300개 적외선

이미지로 구성된 데이터셋으로, 유일하게 비가시광 카메라로 촬영된 데이터셋이다. 이미지 데이터셋은 모두 선택적으로 제출되는 데이터셋으로 데이터셋 수집에 한계가 있어 현재 CTC에 서술되어 있지만 VCM에서는 비디오 코딩을 위한 영상 압축 표준을 타겟으로 개발되고 있기 때문에 기술 제한 시 이미지 데이터 결과가 제출되는 경우는 드물다.

이러한 데이터셋들은 압축 후 기계학습 성능을 평가하기 위해, 각 데이터셋이 가진 GT(Ground Truth)에 적합한 평가 네트워크를 정의하여 사용하고 있다. 구체적으로, SFU-HW 데이터셋은 객체 탐지에 대한 GT를 제공하며, 머신 성능 평가는 Detectron2의 Faster R-CNN X101-FPN<sup>[9]</sup>을 사용하여 머신 성능을 평가한다. TVD-video 데이터셋은 객체 추적에 대한 GT를 가지고 있으며, JDE-1088x608<sup>[10]</sup> 네트워크를 사용하여 성능 평가를 진행한다. PandaSet 데이터셋의 경우 의미적 분할에 대한 GT가 제공되어 있으며, 성능 평가는 Detectron2의 Panoptic FPN R-101-FPN 3x 신경망을 사용한다. 이미지 데이터셋의 경우 모두 객체 탐지에 대한 GT를 포함하고 있으며, Detectron2의 Faster R-CNN X101-FPN을 사용한다.

VCM CTC에서는 제안된 기술들의 성능을 평가하고 비교하기 위하여 총 6개의 인코딩 모드를 정의하고 있으며, 기술 제안 시에는 이 모든 모드에 대한 평가 결과를 필수적으로 제출해야 한다. VCM에서 사용하는 대표적인 인코딩 모드는 기존 비디오 압축에서 널리 활용되는 RA(Random Access), LD(Low Delay), AI(All Intra)가 있다. RA 모드는 임의의 접근을 위한 모드로서 GOP(Group of Pictures) 단위로 영상을 처리하며, LD 모드는 지연을 최소화하여 실시간 영상 처리를 목표로 하고, AI 모드는 모든 프레임은 독립적으로 처리하여 높은 압축 품질을 추구한다.

현재 VCM 표준화에서는 RA, LD, AI 모드는 Inner codec 내부에서만 각각의 조건에 맞추어 동작하는 Inner 모드와, VCM의 전체 파이프라인이 주어진 모드의 딜레이 조건을 만족하도록 설계된 E2E(End-to-End) 방식으로 구분되어 있다. 그 밖에도 CTC에는 기술 제안 시 성능 비교를 위해 사용할 수 있는 앵커 및 평가 파이프라인, 시퀀스별로 적용해야 하는 다양한 압축률 및 구성, 그리고 훈련을 위해 권장되는 데이터셋 및 사전 학습된 모델 등이 서술되어 있다.

### III. VCM의 평가 방법 관련 이슈 및 논의

#### 1. 테스트 데이터셋 및 평가 신경망 관련 이슈

VCM에서는 기술 제안 시 SFU-HW 데이터셋과 TVD-

video 데이터셋을 필수 제출 데이터셋으로 정의하고 있다. 이 데이터셋들은 각각 객체 탐지와 객체 추적에 대한 GT를 제공하며, 이를 바탕으로 압축된 영상의 머신 성능을 평가하고 있다. 그러나 VCM은 평가 네트워크에 독립적(Network Agnostic)이어야 하고, 멀티 태스크를 지원해야 한다는 요구사항에도 불구하고, 객체 기반 평가 네트워크만 주로 사용되고 있으며 표준 기술 개발 역시 이를 중심으로 진행되고 있다. 이는 다양한 머신 비전 태스크를 포괄적으로 지원하고자 하는 VCM의 본래 목적을 달성하는 데 한계로 작용하고 있다.

이러한 한계를 보완하고 평가의 다양성을 확보하기 위해, 146차 회의에서는 의미적 분할(Semantic Segmentation)을 위한 PandaSet 데이터셋이 제안<sup>[6]</sup> 되었으며, 이후 선택 제출 데이터셋으로 채택되었다. PandaSet 데이터셋의 추가는 기존의 객체 중심 평가에서 나아가 배경 영역 또한 머신 성능 평가에서 중요한 요소임을 시사하며, 향후 VCM에서 이를 고려할 가능성을 열어두었다. 그러나 현재 필수 제출 데이터셋이 아닌 선택적 제출 데이터셋으로 되어 있어 이에 관한 연구가 부족하며, 각 데이터셋이 다양한 태스크를 위한 여러 종류의 GT를 동시에 제공하는데 한계가 있어 멀티 태스크에 대한 평가를 진행할 수 없다. 따라서 현행 VCM 성능 평가 방식은 멀티 태스크 평가 측면에서 한계를 가지고 있다고 볼 수 있다. 더욱이 동일한 태스크에 대해서도 다양한 구조의 신경망으로 평가가 이루어져야 하지만, 현재는 하나의 데이터셋과 특정 태스크에 대해 단 하나의 신경망만을 사용하여 평가를 진행하고 있다.

테스트 데이터셋과 관련하여 기존에 있던 데이터셋에 대한 GT와 관련해서도 이슈가 있었는데 144차 회의에서 SFU-HW 데이터셋에 대한 GT의 정확성과 신뢰성에 대한 문제가 제기<sup>[9]</sup> 되었다. 특히 BasketballDrill, BasketballPass 등의 여러 시퀀스에서 GT 오류가 발견되었으며, 이는 SFU-HW 데이터셋을 활용한 모든 성능 평가 결과에 직접적인 영향을 미칠 수 있다는 우려가 제시되었다. 이러한 문제점을 해결하기 위해 147차 회의까지 해당 시퀀스들에 대한 GTS(Ground Truth Study)를 수행하여 GT 데이터의 정확성을 보장하는 작업이 이루어졌다.

VCM에서는 새로운 테스트 데이터셋이나 평가 네트워크에 대한 제안을 적극적으로 권장하고 있지만, 실제로는 라

이센스 제한, 저작권 문제 및 데이터 공개 이슈 등으로 인해 데이터셋과 평가 신경망의 다양성을 확장하는 데 현실적인 어려움이 존재한다. 하지만 위와 같이 이를 개선하려는 지속적인 시도가 있으며, VCM 그룹 내에서는 데이터셋의 다양성과 신경망 독립성을 높이기 위한 논의가 계속 이루어지고 있다. 향후 데이터셋 간의 협력적 구축, 멀티 태스크를 위한 다양한 GT 확보, 그리고 네트워크에 독립적인 평가 프레임워크 개발 등이 추가적으로 논의될 필요가 있을 것으로 예상된다.

## 2. 인코딩 모드 관련 이슈

146차 회의에서는 인코딩 모드와 관련한 중요한 이슈가 제기되었다. 140차 CFP(Call for Proposal)<sup>[10]</sup> 단계에서는 RA(Random Access) 모드만 존재했으나, 이후 WD(Working Draft) 작업이 본격화되면서 LD(Low Delay)와 AI(All Intra) 모드가 추가되었다. 그러나 이 과정에서 각 모드별 특성 및 딜레이 조건 등에 대한 명확한 논의가 충분히 이루어지지 않았다. 특히, 146차 회의까지 사용된 인코딩 모드는 기존 비디오 코딩 표준에서 일반적으로 사용하는 RA, LD, AI였으나 이는 Inner Codec 내에서만 적용된 조건이었고 RoI(Region of Interest), Spatial Resampling, Temporal Resampling과 같은 모듈에서는 인코딩 모드에 따른 딜레이 조건을 고려하지 않고 표준 개발이 진행되었다. 예를 들어, AI 모드에 경우 모든 프레임이 독립적으로 인/디코딩되어야 한다. 하지만 Temporal Resampling 모듈의 경우, 인코더가 일부 프레임을 삭제하고 디코더가 앞뒤 프레임을 이용한 비디오 보간을 통해 현재 프레임을 복원하게 되므로, 다른 프레임과의 종속성을 가지며 독립적이지 못하게 된다.

이러한 인코딩 모드에 대한 이슈가 146차 회의에서 제기되었으며<sup>[11][12]</sup>, 인코딩 모드의 특징 및 딜레이 조건에 대한 요구사항에 대해 논의를 하였다. 이에 따라 Inner Codec 내부에서 각 모드의 딜레이 조건을 명확히 준수하는 Inner 모드와 VCM 전체 파이프라인에서 각 모드의 딜레이 조건을 완벽히 충족하는 E2E 모드로 구분되어 RA\_inner, LD\_inner, AI\_inner, RA\_e2e, LD\_e2e, AI\_e2e로 총 6개의 인코딩 모드가 정의되었다.

## 3. 성능 평가 지표 관련 이슈

143차 회의까지 VCM CTC 문서에는 총 세 가지의 성능 평가 지표가 기술되어 있었다. 첫 번째 지표는 BD-rate(Bjontegaard Delta rate)<sup>[13]</sup>로, 머신 성능 대비 비트 감소를 나타내기 위해 사용된다. 이 지표는 VVC(Versatile Video Coding)와 같은 기존 영상 압축 표준에서 사용하는 평가 방법을 기반으로 하며, PSNR(Peak Signal-to-Noise Ratio) 대신 머신 성능 지표를 적용하여 평가하는 방식이다. 하지만 머신 성능과 kbps(kilobits per second)는 항상 단조 증가 또는 감소(monotonic)하는 상관관계를 가지는 것은 아니며, 만약 이들이 단조적이지 않으면 BD-rate를 계산할 수 없다는 문제점이 존재한다. 두 번째 지표는 Pareto-mAP(mean Average Precision)/MOTA(Multi-Object Tracking Accuracy)로, BD-rate 계산 시 비단조적(non-monotonic)인 RP(Rate Point)의 결과를 제외하고, 단조적인(monotonic) 결과만 추출하여 BD-rate을 계산하는 평가 지표이다. 이 지표는 143차 회의까지 자주 사용되었으나, 비단조적 결과를 임의로 제외함으로 인해 특정 Rate Point에서 결과 값이 편향될 수 있다는 문제점이 있었다. 마지막 지표는 BD-mAP/MOTA로, BD-rate와는 반대로 비트 감소를 대비 머신 성능의 향상 정도를 평가하기 위한 지표이다. 하지만 VCM 그룹 내에서는 BD-rate를 기반으로 한 평가 방식이 주로 사용되어 왔기 때문에 BD-mAP/MOTA 지표는 거의 활용되지 않았으며, 실제 평가에 적합한지에 대해서도 의문이 꾸준히 제기되어 왔다. 이와 같이 세 가지 VCM 성능 평가 지표가 존재하지만, BD-rate는 머신 성능과 kbps 간 비단조적 관계로 인해 계산이 불가능할 수 있다는 문제가 있고, Pareto-mAP/MOTA는 특정 Rate Point에서 결과 값의 편향이 발생할 가능성이 있으며, BD-mAP/MOTA는 실제 평가에서 잘 선호되지 않는다는 한계점이 있다.

이와 관련하여, 비단조적(non-monotonic)인 결과에서도 BD-rate를 계산할 수 있도록 하는 Python 라이브러리를 활용하는 방법이 기고<sup>[14]</sup>로 제안된 바 있으나, 코드의 신뢰성 문제로 인해 그룹 내에서 채택되지 않았다. 또한, Pareto-mAP나 BD-mAP/MOTA를 기본적인 성능 평가 지표로 활용하자는 제안도 있었지만, 이에 대해서도 그룹 내에서 충분한 합의를 얻지 못하였다. 결국 140차 회의부터 143차

회의까지는 기본적으로 BD-rate를 평가 지표로 사용하되, 머신 성능과 kbps 간의 관계가 비단조적인 시퀀스에 대해서는 Pareto BD-rate를 함께 사용하는 혼합 방식으로 결과를 제출하였고, 이를 기준으로 기술 채택 여부를 결정하였다.

이후 144차 회의에서는 머신 성능과 kbps 간의 관계를 커브 피팅(curve fitting)을 통해 강제로 단조적으로 만드는 방법이 새롭게 제안<sup>[15]</sup>되었으며, 현재는 이 방식을 모든 기술 평가 및 채택 여부 결정 시에 사용하고 있다. 그러나 커브 피팅 방식 역시 데이터 포인트 중간에서 값이 급격하게 변화하는 경우, 전체 성능 지표가 비정상적으로 높거나 낮아지는 문제점이 있다. 또한, 머신 성능을 인위적으로 단조적인 형태로 변형한다라는 점에서 이슈가 있으며, 원래 단조적인 결과에도 불필요하게 커브 피팅을 적용하여 실제 머신 성능과 평가 지표 간의 괴리를 발생시키는 문제가 지적되고 있다. 이러한 문제점들에 대한 논의가 현재까지 진행 중이지만, 이를 대체할 수 있는 뚜렷한 평가 방법이 없어 해당 방법이 계속해서 활용되고 있는 상황이다.

#### Ⅳ. 성능 저하 관련 이슈 및 개선 기술

3장 3절에서 기술한 바와 같이, 기술 평가와 비교를 위한 다양한 논의가 이루어졌음에도 불구하고, 현재 VCM 표준 기술의 채택 여부는 커브 피팅을 통해 머신 성능을 강제로 단조적으로 만든 후, 이를 바탕으로 계산한 BD-rate 최종 평가 지표로 사용하고 있다. BD-rate는 머신 성능 대비 비트 절감량을 나타내는 지표이므로, 머신 성능이 다소 저하되더라도 비트 절감률이 크게 증가하면 BD-rate 관점에서는 성능 향상으로 간주된다. 즉, 실제 머신 성능의 절대적인 값보다는 BD-rate 값만을 기준으로 기술 채택이 이루어지기 때문에, 회의가 거듭될수록 전반적인 머신 성능은 오히려 지속적으로 저하되는 경향이 나타나고 있다. 이러한 이슈는 MPEG 147차 회의에서 VCMRS v0.9의 RoI 전처리, Spatial resampling, Temporal resampling, Bit Depth Truncation과 같은 툴을 모두 켜둘 때와 모두 켜지 않을 때의 머신 성능 결과 비교를 통해 제기되었다<sup>[16]</sup>. 표 4는 SFU-HW에서 머신 성능의 차이를 보여주며, 평균 약 18%포인트의 머신 성능에 저하가 있었으며, 표 5와 같이 TVD-video에서는

AI 모드에서 일부 성능 향상이 있으나 평균 약 5%포인트의 MOTA 값의 감소가 있었다. 표 4는 SFU-HW에서 머신 성능 차이를 보여주며, 평균 약 18%포인트의 성능 저하가 나타났다. 표 5에서는 TVD-video에서 AI 모드로 일부 성능 향상이 있었지만, 평균 약 5%포인트의 MOTA 값 감소가 있었다.

표 4. SFU-HW 머신 성능 차이 결과(모든 툴 off-on)(단위: %포인트(mAP))  
Table 4. Machine performance difference result(all tools off-on) for SFU-HW(Unit: %p(mAP))

Sequence	SFU_RA	SFU_LD	SFU_AI
Traffic_2560x1600_30_val	-21.52	-17.40	-16.59
ParkScene_1920x1080_24_val	-25.67	-18.78	-29.99
Cactus_1920x1080_50_val	-52.67	-42.86	-10.53
BasketballDrive_1920x1080_50_val	-15.79	-15.37	-24.72
BQTerrace_1920x1080_60_val	-22.28	-19.79	-14.91
BasketballDrill_832x480_50_val	-18.34	-16.78	-15.92
BQMall_832x480_60_val	-16.88	-18.85	-15.53
PartyScene_832x480_50_val	-30.34	-23.90	-27.82
RaceHorses_832x480_30_val	-10.72	-10.92	-12.43
BasketballPass_416x240_50_val	-7.68	-8.26	-4.95
BQSquare_416x240_60_val	-10.10	-8.93	-8.44
BlowingBubbles_416x240_50_val	-25.00	-21.47	-13.43
RaceHorses_416x240_30_val	-7.76	-7.42	-5.13
avg.	-20.37	-17.75	-15.42

표 5. TVD-video 머신 성능 차이 결과(모든 툴 off-on)(단위: % 포인트(MOTA))

Table 5. Machine performance difference result(all tools off-on) for TVD-video(Unit: %p(MOTA))

Sequence	TVD_RA	TVD_LD	TVD_AI
TVD-01_1	-2.01	-5.70	3.02
TVD-01_2	-1.03	-4.95	1.85
TVD-01_3	1.63	2.61	2.66
TVD-02_1	-0.98	-1.62	-2.38
TVD-03_1	-12.71	-16.13	-2.53
TVD-03_2	-4.38	-9.43	-1.19
TVD-03_3	-7.70	-10.70	3.68
avg.	-6.14	-7.90	-1.71

이러한 머신 성능 저하 현상은 특히 자율주행 자동차, 감시 시스템 등과 같이 실제 응용 분야에서는 심각한 문제를 초래할 가능성이 크다. 그러나 현재 BD-rate를 대체할 수

있는 명확하고 객관적인 성능 평가 지표가 부재한 상황이며, 이 문제를 근본적으로 해결하는 데 어려움이 있다. 이후에는 머신 성능이 일정 기준 이상으로 크게 저하될 경우 BD-rate 수치가 양호하더라도 기술 채택 전 추가 논의가 이루어지고 있지만, 모든 제안 기술에 대해 적용되지는 않으며, 여전히 객관적이고 수치적인 기준이 없어 실질적인 개선에는 한계가 있다.

이러한 문제를 해결하고, 머신의 성능 저하를 방지함과 동시에 BD-rate 성능 개선을 도모하기 위한 시도 중 하나로 적응형 여백 확장 기법이 제안<sup>[17]</sup>되었다. 이 기법은 VCM-RS의 RoI 기반 전처리 과정에서, 객체 탐지 후 객체 영역을 제외한 배경을 회색으로 처리하는 방식에서 발생하는 머신 작업 성능 저하를 보완하기 위해 제안되었다.

적용 방식은 다음과 같다. 객체 탐지 후, RoI 영역의 크기가 사전에 정의된 임계값보다 작은 경우에 한해, 해당 RoI를 주변으로 확장시킨다. 이렇게 RoI를 확장함으로써, 작은 객체들도 주변 배경 정보를 함께 포함할 수 있게 되어, 전처리 과정에서의 정보 손실을 줄이고 모델의 전반적인 성능을 높일 수 있다.

여기서 사용하는 임계값(Threshold)은 영상의 해상도에 따라 식 (1)에서와 같이 동적으로 정의된다.

$$Threshold = \frac{Source\ width \times Source\ height}{1000} \quad (1)$$

RoI 확장은 해당 영역의 중심점을 기준으로 수행되며, 원본 RoI의 가로 및 세로 길이에 확장 비율을 곱하여 새로운

RoI 영역을 생성한다. 확장 비율은 식 (2)에서와 같이 계산되며, 실험적으로 최대 2배까지 확장되도록 제한하였다.

$$RoI\ dilation\ ratio = \frac{Threshold}{Original\ RoI\ area} \quad (2)$$

그림 1은 본 기법의 적용 예시를 보여준다. 그림에서는 풍선을 들고 있는 어린이 객체의 RoI가 임계값보다 작아 주변 배경까지 포함되도록 여백이 자동으로 확장된 모습을 확인할 수 있다. 반면, 골프를 치는 사람 객체는 RoI 크기가 충분히 커서 확장이 적용되지 않았다.

이 기법의 효과를 정량적으로 검증하기 위해, AI E2E 시나리오 상에서 객체 탐지 및 객체 추적 실험을 수행하고, 기존 방식과 비교하였다. 본 논문에서 사용하는 객체 탐지 성능(mAP)과 객체 추적 성능(MOTA)은 값이 클수록, 머신 성능 대비 비트 감소율인 BD-Rate는 값이 작을수록 성능이 향상된 것으로 해석한다.

표 6은 SFU-HW 데이터셋을 활용한 AI E2E 객체 탐지 실험 결과로, 본 기법 적용 전후의 비트레이트 변화와 객체 탐지 성능(mAP) 변화를 나타낸다. 모든 클래스(Class)에서

표 6. SFU-HW AI\_E2E 객체 탐지 비트레이트 및 mAP 변화량

Table 6. Bitrate and mAP variation for object detection on SFU-HW AI\_E2E

Class	bitrate variation	mAP variation
Class A	+20.18%	+2.57%
Class B	+10.71%	+3.96%
Class C	+0.40%	+0.32%
Class D	+2.44%	+0.47%



그림 1. Adaptive margin dilation 적용 예시  
Fig. 1. Example of Adaptive Margin Dilation

비트레이트가 소폭 증가하였으나, 이에 대응하여 객체 탐지 성능(mAP) 또한 전반적으로 향상되었다.

표 7은 SFU-HW 데이터셋을 활용한 객체 탐지 성능을 BD-Rate(mAP) 기준으로 정리한 결과이다. BD-Rate는 머신 성능 대비 비트 감소율로써 음수 Class B 시퀀스에서 -17.85%의 성능 개선이 나타났으며, 전체 평균 기준으로도 -4.67%의 향상이 확인되었다.

표 7. SFU-HW AI\_E2E 객체 탐지 성능 비교 (단위: BD-Rate(mAP))

Table 7. Performance comparison of object detection on SFU-HW AI\_E2E (Unit: BD-Rate(mAP))

SFU_AI_E2E(Object detection)	BD-Rate(mAP)
Class A	-0.83%
Class B	-17.85%
Class C	-1.00%
Class D	+0.58%
Average	-4.67%

표 8은 TVD 데이터셋을 활용하여 객체 추적 실험을 수행한 결과로, 본 기법 적용 전후의 비트레이트 변화와 객체

표 8. TVD AI\_E2E 객체 추적 비트레이트 및 MOTA 변화량

Table 8. Bitrate and MOTA variation for object tracking on TVD AI\_E2E

Class	bitrate variation	MOTA variation
TVD-01_1	+16.34%	+5.2%
TVD-01_2	+15.16%	+6.3%
TVD-01_3	+13.61%	+5%
TVD-02_1	+2.46%	+7.5%
TVD-03_1	+3.78%	+0.4%
TVD-03_2	+4.57%	-0.4%
TVD-03_3	+6.55%	+0.2%

추적 성능(MOTA)의 변화를 나타낸다. 모든 시퀀스에서 비트레이트가 증가하였으나, 이에 대응하여 MOTA 성능 또한 전반적으로 향상되었다.

표 9는 TVD 데이터셋을 기반으로 수행한 객체 추적 실험 결과를 BD-Rate(MOTA) 기준으로 정리한 것이다. TVD-02\_1 시퀀스에서는 -44.44%의 큰 성능 개선 효과가 나타났으며, 반면 일부 시퀀스에서는 성능이 다소 저하되기도 하였다.

표 9. TVD AI\_E2E 객체 추적 성능 비교 (단위: BD-Rate(MOTA))

Table 9. Performance Comparison of Object Tracking on TVD AI\_E2E (Unit: BD-Rate(MOTA))

TVD_AI_E2E (Object tracking)	BD-Rate(MOTA)
TVD-01_1	#####
TVD-01_2	#####
TVD-01_3	#####
TVD-02_1	-44.44%
TVD-03_1	-1.24%
TVD-03_2	5.10%
TVD-03_3	8.42%
Average	#####

TVD-01\_1, TVD-01\_2, TVD-01\_3 시퀀스의 경우, 기법 적용 전후의 성능 변화 폭이 커 BD-Rate 계산이 불가능하여 수치 기반 비교 대신 RD 곡선을 통해 성능 향상 여부를 정성적으로 분석하였다. 그림 2는 왼쪽부터 각각 TVD-01\_1, TVD-01\_2, TVD-01\_3 시퀀스의 RD 곡선을 나타낸다. 세 곡선 모두에서 제안 기법 적용 시 앵커 대비 더 우수한 RD 성능 곡선이 확인되며, 이는 기법의 효과성을 시각적으로 뒷받침한다.

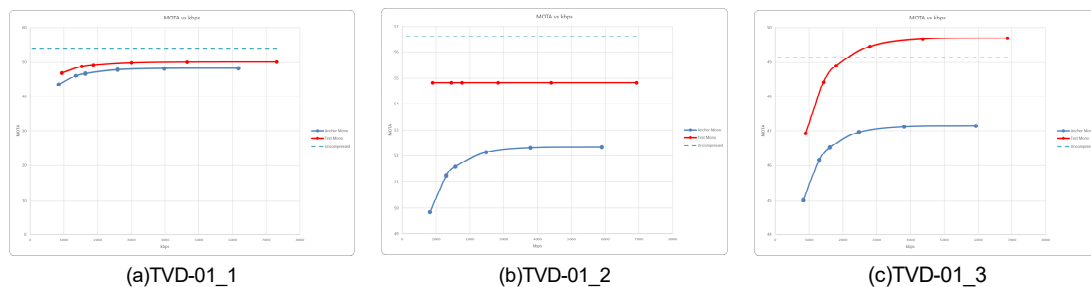


그림 2. TVD-01\_1, TVD-01\_2, TVD-01\_3 시퀀스의 RD 곡선

Fig. 2. RD curves for TVD-01\_1, TVD-01\_2, and TVD-01\_3 sequences



이러한 실험 결과는, RoI 크기에 기반한 적응형 여백 확장 기법이 VCM 환경에서의 머신 작업 성능 향상과 BD-rate 개선에 기여할 수 있음을 보여준다.

## V. 결 론

본 논문에서는 MPEG의 비디오 그룹에서 진행 중인 기계 기반 영상 인식 및 분석을 위한 비디오 압축 표준인 VCM 표준화 과정에서 나타난 주요 성능 평가 이슈와 머신 성능 저하 현상을 종합적으로 분석하였다. VCM 표준화의 역사가 비교적 짧고, 기존 영상 압축 표준의 평가 방식을 대부분 그대로 차용하면서 발생한 여러 한계점과 문제점들을 살펴보았다.

성능 평가 측면에서는 테스트 데이터셋과 평가 신경망이 객체 중심으로 제한되어 있어 네트워크 독립성 및 멀티 태스크 평가 측면에서 명확한 한계가 있음을 확인하였다. 인코딩 모드 측면에서는 각 모듈별 인코딩 모드의 특성과 딜레이 조건에 대한 명확한 정의와 논의가 부족하여 혼란이 있었으나, 최근 Inner 모드와 E2E 모드가 명확히 정의되면서 이러한 이슈가 일부 해소되었다. 그러나 Inner 모드가 반드시 필요하지에 대한 근본적인 의문은 여전히 존재한다. 성능 평가 지표 측면에서는 BD-rate 기반 평가 방식에서 나타나는 비단조적 관계의 문제점을 해결하기 위해 커브 피팅방식이 도입되었지만, 실제 머신 성능이 아닌 인위적으로 생성된 값을 기반으로 평가가 이루어지기 때문에 본질적인 문제가 아직 해결되지 않고 남아있다. 또한, BD-rate 값만으로 기술의 성능 비교와 채택 여부가 결정되는 현재의 평가 방식은 전반적인 머신 성능 저하 현상을 초래하고 있으며, 이는 자율주행 자동차, 감시 시스템 등 실제 응용 분야에서 매우 심각한 문제로 이어질 가능성이 크다.

이러한 다양한 문제점에도 불구하고, VCM 그룹에서는 지속적인 논의를 통해 문제를 인식하고 이를 개선하려는 노력을 기울이고 있다. 그러나 BD-rate 기반의 평가 방법이 머신 성능을 충분히 반영하지 못하는 한계를 근본적으로 극복하기 위해서는 보다 객관적이고 신뢰할 수 있는 성능 평가 지표의 개발이 시급하다. 향후에는 이러한 평가 지표를 통해 머신 성능 저하 문제를 체계적으로 해결하고, 자율

주행 및 감시 시스템 등 머신 성능이 중요한 실제 응용 분야에서 신뢰성과 효율성을 보장할 수 있는 실질적인 표준 기술 개발이 이루어지기를 기대한다.

## 참 고 문 헌 (References)

- [1] WG 04, "Preliminary WD 1 of video coding for machines," ISO/IEC JTC1/SC29/WG4 output document N00384, July 2023.
- [2] WG 04, "Recommendations of 18th Meeting," ISO/IEC JTC1/SC29/WG4 output document N00612, January 2025.
- [3] WG 04, "Common test conditions for video coding for machines," ISO/IEC JTC1/SC29/WG4 output document N00638, January 2025.
- [4] H. Choi, et. al., "A dataset of labelled objects on raw video sequences," Data in Brief, 2021, vol. 34, pp. 106701.  
doi: <https://doi.org/10.1016/j.dib.2020.106701>
- [5] W. Gao, et. al., "An Open Dataset for Video Coding for Machines Standardization," in 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 4008-4012.  
doi: <https://doi.org/10.1109/ICIP46576.2022.9897525>
- [6] H. Zhang, et. al., "[VCM] Semantic segmentation for VCM with PandaSet," ISO/IEC JTC1/SC29/WG4 input document m67677, April 2024.
- [7] A. Kuznetsova, et. al., "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," International journal of computer vision, 2020, vol. 128, no. 7, pp. 1956-1981.  
doi: <https://doi.org/10.1007/s11263-020-01316-z>
- [8] FLIR: Flir thermal dataset for algorithm training. <https://www.flir.in/oem/adas/adas-dataset-form>, (accessed March, 22nd, 2025).
- [9] S. Keating, et. al., "[VCM] Ground Truth data for SFU-HW dataset," ISO/IEC JTC1/SC29/WG4 input document m65718, Oct. 2023.
- [10] WG 02, "CfP response report for Video Coding for Machines," ISO/IEC JTC1/SC29/WG2 output document N00248, Oct. 2022.
- [11] Y. Lee, et. al., "[VCM] Comments on VCM test configurations," ISO/IEC JTC1/SC29/WG4 input document m67594, April 2024.
- [12] J. Lee, et. al., "[VCM] Additional test constraints on delay and latency," ISO/IEC JTC1/SC29/WG4 input document m67870, April 2024.
- [13] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," ITU-T SG16 standard VCEG-MM33, Apr. 2001.
- [14] Y. Lee, et. al., "[VCM] Issues on degradation of machine performance," ISO/IEC JTC1/SC29/WG4 input document m62969, April 2023.
- [15] D. Ding, et. al., "[VCM] A Curve Fitting Approach to Transform Non-Monotonic Test Data for BD-Rate Calculation," ISO/IEC JTC1/SC29/WG4 input document m65531, Oct. 2024.
- [16] Y. Lee, et. al., "[VCM] Issues on degradation of machine performance," ISO/IEC JTC1/SC29/WG4 input document m67870, July 2024.
- [17] S-K. Kim, et. al., "[VCM] Adaptive margin dilation based on RoI size," ISO/IEC JTC1/SC29/WG4 input document m70091, Oct. 2024.

---

— 저 자 소 개 —

---



**정 민 혁**

- 2009년 ~ 2016년 : 명지대학교 컴퓨터공학과 학사
- 2016년 ~ 2018년 : 명지대학교 일반대학원 컴퓨터공학과 석사
- 2018년 ~ 현재 : 명지대학교 일반대학원 컴퓨터공학과 박사과정
- ORCID : <https://orcid.org/0000-0001-6487-9219>
- 주관심분야 : Internet of Things, Virtual Reality, 4D media



**이 예 지**

- 2018년 2월 : 극동대학교 스마트모바일학과 졸업 (학사)
- 2020년 2월 : 건국대학교 스마트ICT융합과 졸업 (석사)
- 2025년 2월 : 건국대학교 컴퓨터공학과 졸업 (박사)
- 2025년 2월 ~ 현재 : 한국전자통신연구원 박사후연수연구원
- ORCID : <https://orcid.org/0000-0002-0292-160X>
- 주관심분야 : 영상처리, 인공지능, 컴퓨터비전



**이 희 경**

- 1999년 2월 : 영남대학교 공과대학 컴퓨터공학과 공학사
- 2002년 2월 : KAIST-ICC 정보통신공학부 공학석사
- 2002년 ~ 현재 : 한국전자통신연구원 실감미디어연구실 책임연구원
- ORCID : <https://orcid.org/0000-0002-1502-561X>
- 주관심분야 : 디지털방송 HCI, Gaze Tracking, VR/AR/MR



**이 진 영**

- 1998년 5월 : B.S. EECS Michigan State University
- 1999년 12월 : M.S. EECS Michigan State University
- 2008년 12월 : Ph.D. EECS Michigan State University
- 2004년 3월 ~ 현재 : 한국전자통신연구원
- ORCID : <https://orcid.org/0000-0002-8718-1961>
- 주관심분야 : AI기반 영상처리, 멀티미디어 시스템, 메타데이터 처리



**정 순 흥**

- 2001년 2월 : 부산대학교 전자공학 (학사)
- 2003년 2월 : KAIST 전기및전자공학 (석사)
- 2016년 2월 : KAIST 전기및전자공학 (박사)
- 2005년 4월 ~ 현재 : 한국전자통신연구원 책임연구원
- ORCID : <https://orcid.org/0000-0003-2041-5222>
- 주관심분야 : 실감미디어, 컴퓨터비전, 머신러닝, 영상부호화, 영상처리

---

저 자 소 개

---



김 상 균

- 1997년 : Computer Science, Univ. of Iowa, B.S.(1991), M.S.(1995), PhD(1997)
- 1997년 3월 ~ 2007년 2월 : Professional Researcher, Multimedia Lab. of Samsung Advanced Institute of Technology
- 2007년 3월 ~ 2016년 2월 : Professor of Computer Engineering, Myongji University
- 2016년 3월 ~ 현재 : Professor of Data Technology, Myongji University
- ORCID : <https://orcid.org/0000-0002-2359-8709>
- Research interests : Digital Content(image, video and audio) analysis and management, 4D media, Blockchain, VR, Internet of Things and multimedia standardization