M2M 영상통신을 위한 Video Coding for Machines 표준화 현황

서정일, 이동민, 김준수*, 추현곤*
동아대학교, 한국전자통신연구원*

요 약

오늘날, 진화하는 딥러닝 기술과 머신 비전이 연계되면서 머신 비전의 정확도가 기존보다 크게 진보했고 이에 기계에 의한 영상 처리 수요도 증가하고 있다. 하지만 기존의 비디오 부호화 기술은 기계가 아닌 인간만을 위해 최적화 되어있다. 이를 고려하여 MPEG에서는 인간이 아닌 기계에 적합한 명상 부호화 기술에 대한 표준화를 진행하고 있다. 본고에서는 MPEG에서 진행되고 있는 VCM(Video Coding for Machine) 표준화 현황에 대해 알아본다.

Ⅰ. 서 론

머신 비전이란 기계에 인간의 시각에 따른 판단 기능을 부여한 것으로, 인간이 눈으로 인지하고 판단하는 기능을 기계가 가지고 있는 하드웨어와 소프트웨어의 시스템이 대신 처리하는 기술이다. 최근 딥러닝 기반 기술이 머신 비전에 적용됨으로써 머신 비전은 기존의 정확도보다 발전했음을 넘어, 인간의 정확도 범위까지도 넘어서고 있다. 이를 바탕으로 머신 비전은 산업 자동화라는 기존의 응용범위를 넘어 자율주행 자동차, 스마트시티, 영상 감시, 보안 및 안전 등으로 점점 응용범위를 확장하고 있다. 또한 모바일 기기, 통신 기술, 사회 문화의 변화로 인해 비디오 데이터 또한 전세계적으로 폭증하고 있어, 앞으로 기계를 통한 비디오 처리에 대한 수요 또한 늘어날 것으로 예측되는 상황이다.

이미 영상 및 이미지 데이터가 인터넷 트래픽 전체의 80% 이상을 차지하고 있는 오늘날, 이러한 데이터 용량을 줄이기 위한 부호화 효율 향상에 대한 요구도 진행중이다. 그러나 기존 비디오 부호화 기술은 인간에게 최적화된 영상 품질을 제공한다는 점이 걸림돌로 작용한다. 인간에게 최적화 되어있다는 것은 달리 말하면 기계의 관점에서는 현재의 비디오 부호화 기술이 최적이 아닐 수 있다는 의미이며, 이는 기계 입장에 최적화된 비디오 부호화 기술이 필요함을 의미한다. 이러한 새로운 패러다임은 차량 간 연결, IoT 기기, 대규모 비디오 보안 감시 네트워크, 스마트 시티 및 품질 검사 등의 출현이 주도하고 있는 상황이다. 이들은 모두 대기 시간과 규모에 대한 엄격한 요구사항을 가지며, 머신 비전을 목표로 하는 이미지와 비디오 부호화 솔루션을 요구하고 있다

이러한 머신 비전에 대한 요구사항은 기존과 다른 새로운 연구 방향과 접근 방식에 영향을 끼쳤다. 예를 들어, 최근 다양한 분류와 추론 작업에 대한 딥러닝 기술이 진보함에 따라, 머신 비전 알고리즘의 적절한 압축 표현과 효과 사이의 관계에 대한 연구와 컴팩트한 특징에 대한 추가 연구를 유발하였다.

멀티미디어 표준화 단체인 MPEG에서는 딥러닝을 포함한 머신 비전 응용을 위한 새로운 영상 부호화 방식인 VCM(Video Coding for Machines) 표준 기술에 대한 논의를 진행 중이다. 본고에서는 현재까지의 VCM의 표준화 활동에 대해 알아보고, 향후 VCM 표준화 진행 방향에 대해서 예측한다.

Ⅱ. 본 론

본고는 크게 3개의 장으로 VCM과 이의 표준화 동향에 대해 알아본다. 첫번째 장은 VCM의 개요, 두 번째 장은 VCM Track 1에 대한 내용, 그리고 마지막 장은 VCM Track2에 대한 내용을 설명한다.

1. VCM의 개요

기계를 위한 비디오는 사람이 개입되지 않고 기계가 비디오를 이해하여 임무를 수행하는 것을 목적으로 한다. 그리고 기계를 위한 비디오 부호화(VCM)는 기계가 임무를 수행하는 데 필요한 성능을 유지하면서 비디오 데이터를 최대한 압축하는 기술이다. 이러한 VCM 기술의 표준화에 대한 논의가 현재 MPEG에서 진행중에 있다. MPEG에서는 VCM Track1, VCM Track2로 트랙을 두 가지로 나누어 표준화를 진행하고 있다. 본고에서는

Track1과 Track2의 현재까지의 활동과 앞으로의 방향성에 대해 논하기 전에 MPEG에서 정의하고 있는 VCM의 표준화 범위, 사용 예시, 요구사항, 성능 평가 지표, 대략적인 표준화 진행사항을 알아본다.

가. VCM 표준화 범위

MPEG에서 정의된 가장 최근의 표준화 범위는 2022년 4월에 정해진 아래와 같다.

"MPEG-VCM aims to define a bitstream from encoding video, descriptors or features extracted from video that is efficient in terms of bitrate and performance of a machine task after decoding."

상기 표준화 범위에서 정의된 바와 같이, VCM 코덱의 입력 및 출력은 머신 비전 업무를 수행하기에 필요한 Video, Feature, Descriptor 와 같은 다양한 형태가 될 수 있다. 이러한 다양한 입력과 출력 조합에 따른 VCM 구조의 예는 〈그림 1〉에서와 같다.

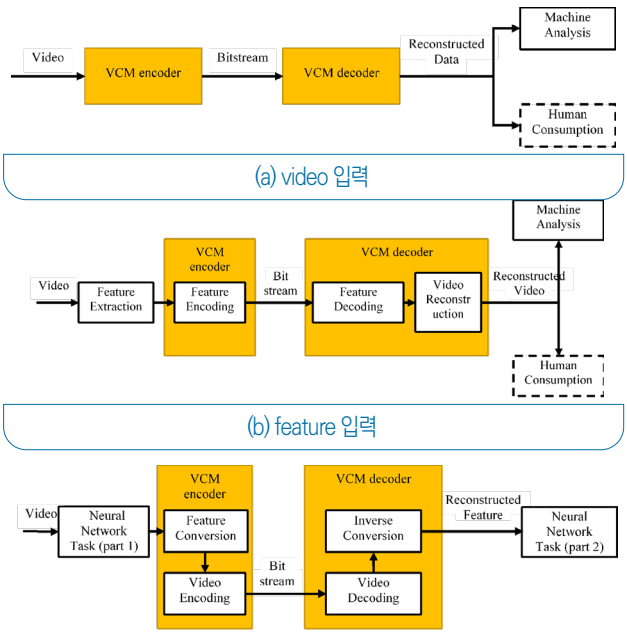


그림 1. VCM 코덱 구조의 예.

나. Use Case

MPEG에서는 VCM의 Use Cases로 〈그림 2〉에 나타나 있는 대표적인 일곱 가지를 제시하고 있다.

첫 번째로 감시 분야이다. 최근 영상 감시 시스템은 객체감지, 추적 등의 다양한 임무를 수행하기 위하여 인공 신경망을 도입하고 있으며, 시스템에 이용되는 카메라 수도 폭발적으로 증가하고 있어 카메라로부터 얻은 비디오를 전송하고 신경망을 통해

그림 2. VCM의 Use Case

처리하기 위해서 많은 대역폭과 계산 성능을 요구하고 있는 상황이다.

두 번째는 지능형 교통 시스템이다. 스마트 교통 시스템에서 자동차들은 각자의 임무를 수행하기 위하여 자동차들간 또는 다른 센서와의 통신이 필수적이며, 이러한 통신 대상에는 비디오가 다수 포함되어 있다. 송수신되는 비디오의 최종 소비주체가 자동차인 만큼, 이에 맞는 부호화 방법을 요구하고 있다.

세 번째는 스마트 시티이다. 스마트 시티의 경우 교통 모니터링, 밀도 감지 및 예측, 교통 흐름 예측 및 리소스 할당 등 스마트 시티 어플리케이션을 위해 서로 다른 노드 센서와 장치 간에 높은 수준의 상호 연결이 이루어진다. 이런 상호 연결을 통해 일어나는 작업을 최적화하기 위해선 장치 간 통신의 효율화와 수신 장치가 필요로 하는 정보만을 부호화 할 수 있는 방법이 필요한 상황이다.

네 번째는 지능형 산업 시스템이다. 자동화된 생산 환경에서, 제품 불량 검사 등과 같이 비디오 송수신 및 분석이 지속적으로 필요한 작업의 효율화를 위한 부호화 방법이 필요한 상황이다.

다섯 번째는 지능형 컨텐츠 분야이다. 최근 모바일 기술의 발달로 인해 엄청난 양의 이미지/비디오 콘텐츠가 생성되고 있다. 이 중 부적절한 컨텐츠로부터 특정 집단을 보호하는 것 또한 큰 이슈가 되고 있는 상황이다. 라이브 이미지/비디오, 짧은 비디오 및 소셜 미디어 등을 머신 비전 기술을 이용하여 효율적으로 처리할 수 있도록 필요한 정보만 부호화하는 방법이 필요하다.

여섯 번째는 소비자 가전 분야다. 최근 상황 인식 서비스와 정보를 사용자에게 제공하기 위해 신경망을 사용하는 가전 제품이 증가하고 있다. 해당 작업을 수행하기 위한 정보를 얻거나, 얻은 정보를 클라우드와 같은 외부 서버로 전송하기 위한 통신에서 네트워크 대역폭 부담을 줄이기 위한 부호화 방법이 필요한 상황이다.

일곱 번째는 디스크립터를 이용한 다중임무 수행이다. 하나의 비디오 스트림이 다양한 목적을 위한 다양한 종류의 기계 작업에 사용될 수 있으며, 각각의 기계 작업들은 병렬, 직렬 또는 혼합 방식으로 이루어질 수 있다. 이때 공통적으로 이루어지는 기계 작업을 전송 전에 수행하여, 디스크립터로써 출력이 다른 기

계 작업들의 입력에 전송하면 비디오 전송에 필요한 대역폭과 디코딩 후 기계 작업의 계산 시간을 크게 줄일 수 있다. 따라서 비디오 대신 디스크립터를 전송하거나 둘을 동시에 전송하는 데 있어서 효율적인 부호화 방법이 필요한 실정이다.

다. 요구사항

현재 MPEG 표준화 회의에서 논의된 VCM의 요구사항[2]은 2022년 4월에 정의된 17개로, 〈표 1〉과 같다. 17개의 요구사항 중 반드시 만족해야 하는 의무적 요구사항(shall)이 16개로, 대부분을 차지하고 있다. VCM에 대한 요구사항은 크게 VCM 전반에 대한 요구사항, Track 1에 해당하는 Feature Coding의 요구사항, Track 2에 해당하는 Video Coding의 요구사항, 그리고 Feature/Video Coding 모두에 적용되는 요구사항으로 구별되며, 각 요구사항이 어디에 해당하는 지도 명시되어 있다.

라. 성능 평가 지표

VCM을 위한 성능 평가 지표는 평가 구조 문서[3] 및 공통 실험 환경 문서[4]에 정의되어 있다. VCM 부호화기의 성능 평가를 위해 사용되는 성능 평가 지표는 동일하거나 유사한 데이터에서의 각각의 머신 비전 성능에 따라 정해진다.

객체 검출(Object Detection)의 성능 평가를 위해서는 특정한 범위의 IoU(Intersection over Union)값에 대해 객체 클래스별 AP(Average Precision)를 평균하여 계산한 mAP(mean AP)를 사용한다.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

표 1. VCM 표준 요구사항

No.	Requirements
1	VCM shall support video coding for machine task consumption purposes.
2	VCM shall support feature coding. (Feature coding)
3	VCM shall support a coding efficiency improvement for at least 30% BD-rate over the VVC standard on machine vision tasks. (Video coding)
4	VCM shall support a broad spectrum of encoding rates.
5	VCM shall support various degrees of delay configuration. (Video coding)
6	VCM shall be agnostic to network models. (Video/Feature coding)
7	VCM shall be agnostic to machine task types. (Video/Feature coding)
8	VCM shall provide description of the meaning or the recommended way of using the decoded data. (Feature coding)
9	VCM should support the use and inclusion of information such as descriptors in its bitstream. (Feature/Video coding)
10	A single VCM bitstream shall support any number of instances of machine tasks. (Video/Feature coding)
11	VCM shall support at least the following colour formats; monochrome, RGB, and YUV (YCbCr). (Video coding)
12	VCM shall support at least the following input bit depths: 8-bit and 10-bit. (Video coding)
13	VCM shall allow for feasible implementation within the constraints of the available technology at the expected time of usage.
14	VCM shall support rectangular picture format up to 7680x4320 pixels (8K).
15	VCM shall support fixed and variable rational frame rates for video inputs.
16	VCM shall support any input source from video or image.
17	VCM shall support privacy and security.

여기서 IoU란, 정답 사각형 영역과 예측 사각형 영역 사이의 일치하는 비율을 의미한다.

$$Precision(T_{IoU}) = \frac{TP(T_{IoU})}{TP(T_{IoU}) + FP(T_{IoU})}$$

Precision은 예측된 결과 전체 대비 정답의 비율을 의미한다. VCM 성능 평가에는 0.5부터 0.95까지 0.05 간격의 매 IoU값에 대해 얻어진 AP값 (AP@[0.5:0.95])들의 평균한 값을 사용한다.

객체 분할(Instance Segmentation)의 경우, 정지영상과 동영상의 성능 평가 지표가 다르다. 정지영상의 객체 분할 임무의 성능 평가 지표의 경우, 객체 검출 임무와 동일한 mAP를 사용한다. 다만 객체 검출에서 객체가 포함된 사각형 영역으로 IoU를 계산했다면, 객체 분할의 경우 객체의 모양에 따라 이진 맵을 생성한 후 이진 맵들 간의 영역의 중복성을 이용하여 계산한다. 동영상 객체 분할의 경우, 분할 영역 유사도를 측정하는 J score (Jaccard Index Score)와 객체 경계선 기반 유사도를 측정하는 F score를 평균한 J&F means를 사용한다.

동영상 객체 추적(Object Tracking)의 성능 평가 지표는 MOTA(Multiple Object Tracking Accuracy)를 사용한다. t번째 프레임에서 GT를 Ground Truth라고 할 때, MOTA는 다음과 같이 계산할 수 있다.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t (GN_t)}$$

여기서 FN은 False Negative 에러, FP는 False Positive 에러, IDSW는 Identity Switch 에러로 두 개의 물체가 겹쳐질 경우 동일하게 나타나지 않는 에러를 뜻한다. 즉 MOTA는 객체의 ID를 오인하거나 새로운 객체로 잘못 인식한 경우에 대한 에러까지 고려함으로써, 정확하게 추적된 프레임 내의 물체의 궤도의 정확도를 계산하도록 고안되었다.

하이브리드 비전과 같이 인간이 보는 화질 평가의 경우, 기존의 비디오 부호화에서 사용한 PSNR과 SSIM값을 이용한다. 단 기존의 비디오 부호화에서 사용한 주관적 화질 평가의 경우는 VCM에서 사용하지 않는다. 또 비디오 부호화 표준화에 있어서 계산 복잡도 평가를 위해 부복호화 수행 시간도 평가에 고려될 수 있다.

마. 표준화 진행사항

2019년 7월, 스웨덴 예테보리에서 열린 MPEG 회의에서 기계가 시각 기반 임무를 수행할 때 머신 비전의 성능을 유지하면서 영상 데이터를 압축하는 비트스트림 표현에 대한 기술에 관한 논의가 시작되어, 2019년 9월에 Video Coding for Machines(VCM)란 이름의 새로운 비디오 부호화 방식에 대한 논의를 위해 MPEG VCM AhG(Adhoc Group)가 구성되었다.

2021년 4월에 열린 회의에선 VCM 표준화 기술에 대한

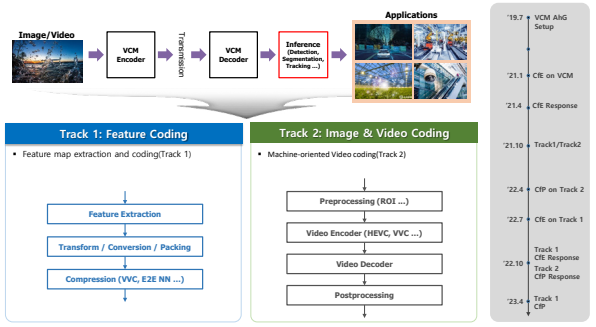


그림 3. VCM 표준화 진행 현황

CfE(Call-for-Evidence)[5]를 진행하여 이미지 부호화에 있어서 가능성을 검증하였으며, 이를 바탕으로 2022년 4월에 Image/Video Coding 트랙(Track 2)에 대한 CfP(Call-for-Proposal)[6]가 발간되어 10월에 이에 대한 기술 제안을 진행하였다. 2022년 7월에는 Feature Coding 트랙(Track 1)에 대한 CfE[7]를 발간하고, 10월에 기술 평가를 진행하였다. 그리고 2023년 4월에 열린 142차 회의에서 Feature Coding 트랙의 CfP 가 발간되었다.

2. [Track 2] VCM Image and Video Coding

본 장에서는 MPEG VCM의 두 번째 트랙인 Image and Video Coding 트랙에 대해서 살펴본다. Image and Video Coding 트랙은 〈그림 4〉와 같이 영상 또는 비디오 입력을 받아서 압축된 비트스트림을 생성하고, 이를 다시 원래의 비디오와 같은 형태의 비디오로 복원한 후 복원된 영상을 이용하여 머신 비전 네트워크의 입력으로 사용하는 형태의 부호화기를 의미한다.

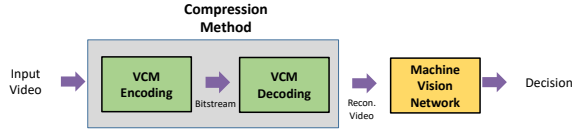


그림 4. Image and Video Coding Track의 Coding Pipeline

Image and Video Coding 트랙의 경우, 2021년 4월에 CfE에 대한 평가를 진행했고, 2022년 10월에 CfP에 대한 제안 평가를 진행하였다. CfP에서의 평가는 객체 검출, 객체 분할, 객체 추적의 세 개의 임무를 대상으로 성능을 비교하였으며 각 임무에 대한 데이터셋은 〈표 2〉와 같이 정해져 있다.

Image and Video Coding 트랙에 제안된 기술은 크게 전처리를 기반으로 한 영역기반 접근방식(ROI-based Approach)와

표 2. VCM 표준 요구사항

Target Tasks		Object Detection	Instance Segmentation	Object Tracking
Network		Faster-RCNN	Mask R-CNN	JDE-1088
Dataset	Image	FLIR, TVD, OpenImages v6	TVD, OpenImages v6	-
	Video			TVD
Performance Metric		mAP/Bpp	mAP/Bpp	MOTA/Bpp

End-to-End 딥러닝 네트워크 방식으로 구분할 수 있다.

가. 영역기반 접근방식

VCM에서 ROI(Region-of-Interest)는 머신 비전에서 주요 특징을 추출하는 데 사용되는 영역을 의미한다. 〈그림 5〉와 같이 영역기반 접근방식(ROI-based-Approach)은 입력 비디오에 대해 머신 비전에 사용되는 네트워크 또는 네트워크 일부를 사용하여, 비디오 내의 주요 객체가 존재하는 영역 또는 프레임을 검출하고 해당하는 영역 또는 프레임만을 부호화 하거나, 해당 영역 또는 프레임에 더 많은 데이터를 부가하도록 하는 방법을 의미한다.

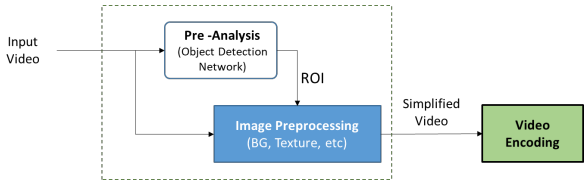


그림 5. 영역기반 접근방식

기장 기본적인 영역기반 부호화 방법은 ROI를 찾아서 해당하는 부분만을 전송하는 방법이다. 포츠난 대학에서는 ROI를 검출하여 배경을 단순화한 영상을 기존의 비디오 부호화를 이용하여 압축하는 방식[8][9][10]을 제안하였으며, ETRI와 명지대에서는 머신 비전을 이용하여 얻어진 결과에 대한 서술자와 디스크립터, 객체 정보 영상을 이용하여 다른 머신 비전 임무인 객체 추적에 적용하는 기술[11]에 대해 제안하였다. 알리바바와 홍콩시립대학에서는 Yolo v7 기반의 객체 검출 네트워크를 이용하여 물체의 경계를 추출하고, 이를 바탕으로 시간적, 공간적으로 정보를 줄여서 비트율을 낮추는 방식[12]을 제안하였으며, Florida Atlantic 대학과 OP solutions에서는 머신 비전을 이용하여 물체가 포함된 영역을 찾고, 이 영역만을 이용하여 영상을 재구성하여 전송하고, 원래의 영상으로 복원하는 방식의 부호화 기술[13]에 대해 제안하였다.

ROI 영역만을 보내지 않고, ROI에 대해 서로 다른 Bitrate를 부여하는 방법도 제안되었다. Ericsson에서는 머신 비전을 이용하여 물체가 포함된 영역을 찾고, 이 영역에 대한 정보를 기존의 VVC Encoder에 입력하여 각 영역의 화질인자(QP, Quality Parameter)를 조절하여 부호화 효율을 높이는 방법[14]을 제안하였다. ETRI와 건국대는 〈그림 6〉에서와 같이 객체 인지 중요도에 따라 서로 다른 성능 지표를 할당하여 전송하는 방법을 제안하였다. CfE 기술 제안에는 배경과 전경을 구분한 후, 이를 두 개의 스트림으로 나누어서 전송하는 부호화 방식[15]을 제안하였으며 CfP 기술 응답에서는 배경과 물체를 새로운 프레임으로 합성한 후, 서로 다른 화질로 부호화 하는 방안[16]을 제안하였다.

비디오 영상에 대한 샘플링을 위한 방법도 제안되었다. CAS-ICT와 China Telecom에서는 시간적으로 Frame sample만을 이용한 비디오 부호화 기술[17]을 제안하였다. 이 방법은 복호화기에서 기존의 프레임을 생성하기 위한 Interpolation 네트워크를 이용하여 중간 프레임을 재생성하였다. Tencent와 우한대학은 시공간 샘플링 기술을 이용하여 데이터의 용량을 줄이고, 복호화 시 심층신경망을 기반으로 한 Post Filtering을 이용한 솔루션에 대해 기술[18][19]을 제안하였다.

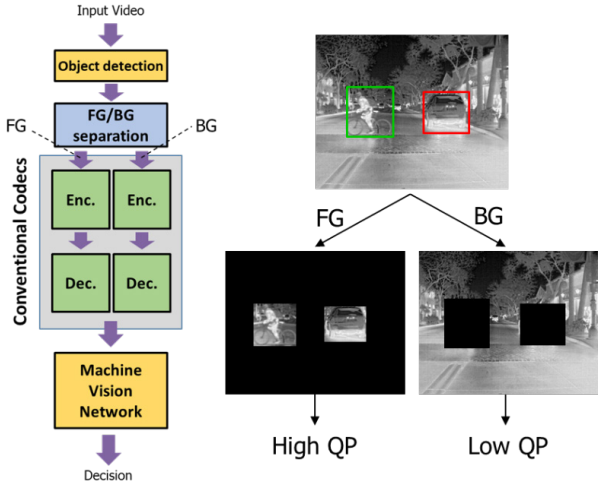


그림 6. 중요도에 따라 서로 다른 화질로 전송하는 방식의 예

나. 딥러닝 네트워크 기반 압축 방식

최근 딥러닝 기술의 발전으로 인해, 비디오를 비롯한 다양한 멀티미디어를 딥러닝 네트워크를 이용하여 압축하는 기술도 활발히 연구되고 있다. 딥러닝 기반 압축 기술은 심층신경망에 이미지 또는 비디오를 입력하여 제한된 형태의 은닉벡터를 추출하여 부호화한다.

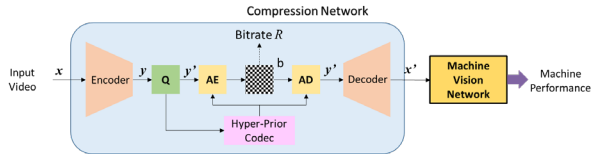


그림 7. 딥러닝 네트워크 기반 압축방식

일반적인 영상 압축의 경우, 압축 효율을 높이기 위해 심층신경망은 복원 영상의 화질은 높이면서 은닉벡터가 적은 비트로 표현될 수 있도록 학습된다. VCM은 딥러닝 네트워크를 이용하여 부호화 시 머신 비전 네트워크의 에러 함수를 같이 사용한다. 이를 통해 네트워크가 머신 비전에 더 유리한 방향으로 훈련이 될 수 있도록 한다. 예를 들어 객체 검출 네트워크를 대상으로 할 경우 압축 네트워크에 대한 비용함수(Loss)는 다음과 같이 비트율(R)과 영상의 에러, 그리고 검출 네트워크의 비용을 조합하여 다음과 같이 표현할 수 있다.

$$L_{overall} = R + \lambda_{mse} L_{mse} + \lambda_{detect} L_{detect}$$

중국의 저장대학에서는 Key frame Encoder(MIKEnc)와 Key frame Decoder(MIKDec)로 구성된 심층신경망 기반 부호화 기술[20]을 제안하였다. 이 방법에서 부호화기인 MIKEnc는 영상의 크기 조절 및 정규화를 수행하는 전처리 네트워크인 PreP, 이에 대한 특징을 추출하는 NN-FE 및 다시 특징을 압축하는 NN-FC 모듈로 구성되었으며, 복호화기인 MIKDec는 특징을 복호화하는 NN-FR, 특징벡터로부터 원래의 영상을 재구성하는 NN-IR, 그리고 영상의 크기 조절 및 정규화를 수행하는 후처리 네트워크로 구성되어있다.

Tencent와 우한대학에서는 Variable-rate Intra Coding과 Scale space flow Coding[23]을 이용한 End-to-end Learning based Solution을 제안하였다[21]. Intra Coding Network를 위해서 기존의 Cheng2020 with attention 모델[22]에 가변 bitrate 지원을 위한 Scalingnet을 추가하였다. 노키아에서는 Intra-frame Coding에는 딥러닝 기반의 압축 부호화 기술을 이용하고, Inter-frame Coding에는 VVC 코딩을 이

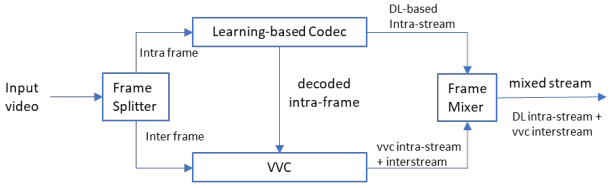


그림 8. 노키아의 Hybrid Codec 구조

용하는 하이브리드 형태의 부호화 기술[24][25]을 제안하였다.

Track 2 CfP에 대한 평가 결과는 CfP test report 문서와 CfP response report에 제공되어 있다[26][27]. 제안 결과 보고서에 기재되어 있는 각 임무 별 최고 성능은 〈표 3〉과 같다.

표 3. CfP 최고 성능

Task	Dataset	BD-rate change
Object tracking	TVD (videos)	-57%
Instance segmentation	OpenImages	-51%
	TVD	-57%
Object detection	OpenImages	-47%
	FLIR	-53%
	TVD	-65%
	SFU (videos)	-36%

임무별로 VVC 대비 성능을 비교한 결과, 객체 추적 임무에서 57%, 객체 분할 임무에서 45%, 객체 검출 임무에서 39%의 성능 향상이 있었음을 확인할 수 있다. 〈표 3〉의 결과는 개별 임무에 대한 최고 성능에 대해 나타난 것이기 때문에 통합된 성능의 경우 위의 결과보다 성능이 다소 떨어질 것으로 예상된다.

〈그림 9〉는 제안된 기술을 이용하여 영상을 복원한 예시이다.

그림 9. CfP 제안기술을 이용한 영상 복원 결과

다. Track 2의 표준화 동향

VCM은 〈그림 10〉과 같은 표준 모델 초안을 바탕으로 테스트 모델(Test model)을 만들고, 테스트 모델 내의 주요 알고리즘의 비교 평가를 위한 실험 절차(CE, Core Experiment)를 진행했다. 표준 모델의 초안에 표기된 Inner Codec(Encoder, Decoder)은 AVC, HEVC, VVC 등 기존의 비디오 코딩 기술 또는 그와 유사한 기능을 하는 비디오 부호화 기술을 뜻한다.

2022년 10월을 기준으로 다섯 개의 CE가 만들어졌고, 2023

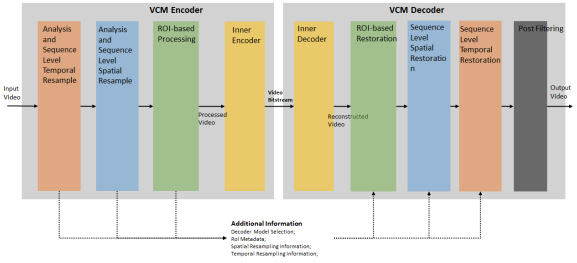


그림 10. VCM 표준 모델 초안

년 4월에는 CE 1,2,3은 지속하고 CE 4,5는 다른 CE에 병합하기로 결정되었다.

CE 1은 RoI Based Coding Methods로, 관심 영역을 추출하고 이를 통해 각 프레임을 재구성하여 비디오를 처리하는 방식에 대한 비교 실험이다. CE 1은 Evaluation Metric의 문제 및 Candidate Selection 방법의 문제로 CE 결과를 도출하지 못해, 다음 회기까지 CE를 계속 진행하기로 했다. 4월 회의에 5개의 추가 Proposal이 Hybrid Codec CE에 포함되었으나, 현재 VTM CE와 Hybrid Codec CE가 각각의 Anchor를 사용하고 있어 성능 비교가 불가능한 상황이다.

CE 2는 Neural Network Based Intra Frame Coding이다. 심층신경망을 이용한 Intra Frame 코딩 알고리즘에 대한 비교 실험이다. CE 2의 경우, 전반적인 Hybrid Coding 방식의 활용 가능성에 대해서는 긍정적인 기류이나 Learned Image Codec 방식과 관련하여 Training에 관해 검증이 필요하다는 의견이 존재했다. LIC의 경우에는 Training Dataset및 Network에 의존성이 강하여 Generalization 가능성에 대한 문제점이 제기되었다. 성능 개선 가능성에 대해서는 이의가 없었으나 Hybrid Codec 방식을 TuC(Technologies under Consideration, 기술 고려사항)에 넣는 것에 대해서는 합의되지 않았다. 이에 CE 2는 다음 회의에서도 계속 진행하기로 결정됐다.

CE 3은 Frame Level Spatial Resampling으로 각각의 프레임을 Resampling을 통해 크기를 줄여서 부호화하고, 복호화 시 Upsampling하는 방법에 대한 비교 실험이다. 지난 141차 회의 이후 Spatial Resampling과 관련해 4개의 테스트가 진행되었고, 3개의 실험 결과가 제출되었으며 2개의 실험 결과에 대해 Crosscheck가 완료되었다. 실험 결과, Spatial Resampling이 어느 정도의 성능 향상을 보이고 있어 다음 회의까지 CE를 계속 진행하기로 하였다.

CE 4는 Temporal Resampling으로 관심영역 또는 타겟의 유무 등의 기준 또는 그 외 필터링 기술을 사용하여 일정 프레임을 삭제하여 부호화 후, 복호화 시 Interpolation 네트워크 등의 방식으로 삭제된 프레임을 채워 넣는 알고리즘에 대한 비교 실험

이다. CE 4는 지난 141차 회의 이후에 3개의 테스트가 진행되었고, 2개의 테스트에서 실험 결과가 정상적으로 제출되었다. 실험 결과 의미 있는 성능 향상을 보여주었고, 두 실험이 같은 접근 방식을 가지고 있으므로 Ref.SW와 candidate WD에 채택하기로 결정되었다. CE 결과가 Ref.SW에 채택됨에 따라 CE 4의 대한 논의는 마무리되어 더 이상 진행하지 않기로 결정되었다.

CE 5는 Post Filtering으로, 복호화된 비디오에서 각 프레임의 해상도 및 화질을 높이기 위한 도구에 대한 비교 실험이다. CE 5는 2개의 테스트가 진행될 예정이었으나, 하나의 테스트만 진행되었다. 또한 CE 5는 CE 2에 병합되어 추가 실험을 진행하기로 결정되었다.

3.VCM Feature Coding Track

본 장에서는 MPEG VCM 그룹의 첫 번째 트랙인 Feature Coding 트랙에 대해서 살펴본다. Feature Coding 트랙은 〈그림 11〉과 같이 영상 또는 비디오로부터 1차적인 머신 비전 네트워크의 출력을 받아, 이를 Feature Map과 데이터를 입력받아 부호화하는 기술을 의미한다.

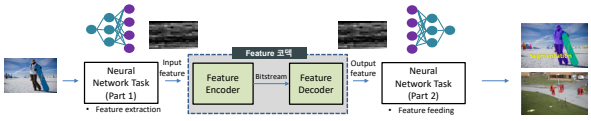


그림 11. Feature Coding의 파이프라인

MPEG VCM Track 1에서는 2022년 7월, CfE[7]를 발행하였다. Track 1에서는 CfE를 위한 필수 임무로 객체 추적과 객체 분할이 지정되었으며, 선택 임무로 객체 검출이 지정되었다. 한밭대학과 ETRI는 피라미드 형태의 심층신경망에서 각 계층의 특징을 입력받아 계층을 결합하는 결합 네트워크(MSFF : Multi-scale feature fusion)와 결합된 특징맵의 차원을 줄여서 데이터를 줄이는 특징 부호화 네트워크(SSFC : Single-stream feature codec), 복원된 특징을 원래의 피라미드 형태의 심층신경망의 특징 계층으로 복원(MSFR : Multi-scale feature reconstruction)하는 복수 계층 특징 압축(MSFC : Multi-scale feature compression) 프레임워크를 기반으로 하는 파이프라인 구조[28]를 제안하였다. VVC 코덱과의 연동을 위해, SSFC 부호화기에서 출력되는 특징을 패킹, 양자화 후 VVC 부호화기의 입력에 맞는 포맷으로 변경하는 과정이 존재하며, VVC 복호화기의 출력을 포맷 변경, 역양자화, 언패킹 과정을 거쳐 SSFC 복호화를 수행 후 이를 복수 계층의 특징으로 복원하여 임무를 수행할 수 있도록 하였다.

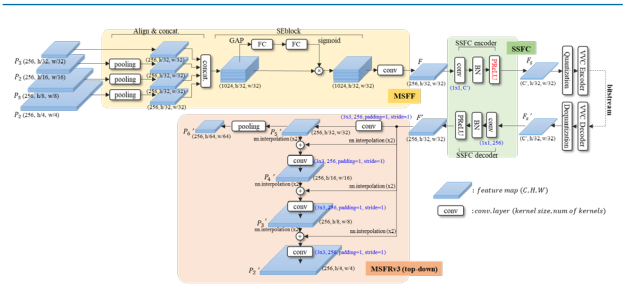


그림 12. MSFC 기반 VCM Feature Codec

항공대학과 ETRI는 MSFC 프레임워크를 기반으로 MSFF 후 출력되는 특징을 중요도 순서대로 정렬하여 MSFC-transformed-features를 구성하여 VVC 부호화를 수행하는 방식[29]를 제안하였다. MSFC-transformed-features는 VVC 부호화기에서 사용하는 비트율에 따라 적응적으로 선택되어 압축하게 되며, 복호화 과정에서 제거된 성능을 예측하는 방식으로 특징을 복원한다. Canon에서는 MSFC로부터 얻어진 특징을 PCA를 이용하여 차원을 줄여서 특징을 압축하는 방식[30]을 제안하였으며, 광운대학과 ETRI는 부호화 대상이 되는 특징에 대해서 PCA 기법을 사용하여 기저(basis)와 평균을 사전에 구하여 코덱에 사용하는 방식의 변환 기반 특징 압축 기술[31]을 제안하였다. Tencent와 우한대학에서는 Learning-based Feature Conversion 모듈을 사용하여 부호화 효율을 개선하는 방식[32]을 제안하였다. Feature Encoding/Decoding으로 이미지에서 얻어진 특징 부호화 네트워크[22]를 사용하였고, 비디오에서 얻어진 특징 부호화 시 VVC 부호화기를 사용하였다.

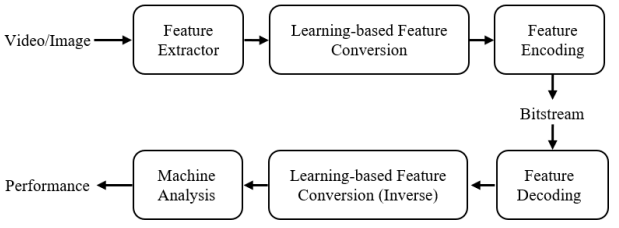


그림 13. Learning-based Feature Codec

CfE 제안 결과는 〈표 4〉[33]와 같다. 기존의 VVC 부호화기를 이용해 만든 Feature anchor 대비 객체 추적에서 97.58%, 객체 분할에서 98.60%, 객체 검출에서 98.34%의 성능 개선을 확인할 수 있었다. 한편, VVC anchor 대비로는 객체 추적에서 87.44%, 객체 분할에서 93.04%, 객체 검출에서 94.46%의 성능 개선이 있는 것으로 나타났다. 성능 향상이 확인됨을 바탕으로 MPEG에선 2023년 4월 CfP를 발간했다.

표 4. 객체추적 및 객체 분할에 대한 결과

제안 문서번호	Object Tracking		Instance Segmentation		Object Detection	
	BD-rate over Video	BD-rate over Feature	BD-Rate over Image	BD-Rate over Feature	BD-Rate over Image	BD-Rate over Feature
m60761	-87.44%	-97.58%	-79.21%	-95.56%	-81.11%	-94.15%
m60788	63.69%	-74.43%	-47.46%	-89.48%	-54.51%	-85.06%
m60799	-80.18%	-97.09%	-93.04%	-98.60%	-94.46%	-98.34%
m60803/ m60802	218.93%	-33.01%	-19.35%	-83.38%	-	-
m60821	-77.40%	-95.84%	-78.11%	-95.84%	-70.39%	-91.14%
m60925	-64.94%	-92.17%	-69.08%	-92.30%	-	-

CfP anchor 관련 기고서로는 anchor 실험 결과 2건, crosscheck 실험 결과 4건, split point 관련 2건이 검토되었다. 제시된 anchor 실험 결과 2건은 모두 ETRI에서 제출한 것으로, 각각 OpenImages를 사용한 객체 검출, 객체 분할 task에 대한 CfP anchor를 제시한 것이었으며 둘 다 feature anchor로 반영되었다.

Qualcomm에선 Amazon cloud를 사용하여 SFU 객체 검출 feature anchor에 대한 crosscheck 결과를 제시했다. Crosscheck시 발생하는 차이 값을 줄이기 위한 방안으로 제시되었지만, 지금까지 차잇값을 줄인 방법으로 CfP를 진행하는 데 문제가 없을 것으로 예상되어 Cloud를 사용한 crosscheck는 선택할 수 있는 옵션으로 CfP 문서에 명시하는 것으로 결정되었다. canon에선 ETRI에서 제출한 anchor에 대한 crosscheck 실험 결과를 제시했고, 차이가 threshold를 벗어나지 않는 수준으로 통과했음을 확인했다. China Telecom은 단독으로 HiEve 객체 추적 feature anchor에 대한 crosscheck 결과를 제시했고, 중국 저장대학과 협력해 TVC 객체 추적 feature anchor에 대한 crosscheck 결과를 제시했다. 두 결과 모두 차이가 threshold를 벗어나지 않는 수준으로 통과했다.

Sharp에선 JED에 대한 대체 split point를 사용한 실험 결과를 제시했고, 여러 split point 적용을 통해 과최적화 방지에 도움을 줄 수 있을 것을 주장했으나 새로운 split point 사용을 위해선 VTP 적용 결과와 split point의 특성에 대한 평가가 정의될 필요가 있는 것으로 정리되었다. Sharp에서 제시한 대체 split point에 대한 crosscheck 결과를 제시한 canon의 기고는 새로운 split point의 도입에 대해 결정된 사항이 없으므로, 검토되지 않았다.

한편 2023년 4월 회의에서는 Track 1에 대해 5건의 기술 기고가 제출되었다.

China Telecom에선 지난 회의 기고인 m60257에서 제안한 DCT based encoder/decoder를 feature extractor와 feature reconstructor와 결합하여, JDE network의 FPN 직전 layer를 사용한 video object tracking task 성능을 제시했고, Feature anchor대비 object tracking에서 92%의BD-rate 성능 개선을 보였다. ETRI에선 지난 회의 기고 m60761에서 제안한 기술을 새롭게 추가된 dataset인 HiEve dataset과 새로운 split point에 적용한 video object tracking task 성능을 제시했으며 Feature anchor대비 1080p sequence에서 93.38%, 720p sequence에서 92.09%의BD-rate 성능 개선을 보였다. 중국 베이항대학은 지난 회의에서 기고한 m61973 및 m61980를 업데이트해 Compression network, Enhancement network, VVC의 세 part를 learnable codec으로 대체한 기술 기고를 제출했고, 기고에 따르면 OpenImages dataset의 instance segmentation task에서 91.84%, object detection task에서 94.70%의 BD-rate 성능 향상이 이루어졌다. 상하이교통대학, 저장대학, 레노버에선 Channelwise dynamic pruning 방법을 사용한 object detection task에서의 성능 개선 방법 제시했다. 이 방법은 Feature channel 중 중요하지 않은 채널을 제거함으로써 압축하는 구조로, pruning-ratio design, pruning-mask design의 두가지 제거 방식이 제시되었다. 이 방법은 image anchor대비 object detection에서 BD-rate 성능이 개선됨을 그래프로 보여주었으나 BD-rate 값은 제시하지 않았고, JDE network에서의 실험 결과는 얻지 못했다.

Ⅲ. 결 론

MPEG VCM은 현재 Feature Coding (Track 1)과 Image and Video Coding (Track 2)로 세부 분야를 달리하여 표준화를 진행하고 있다. 두 트랙 모두 제안된 기술이 기존의 VVC 부호화를 이용하는 경우보다 높은 성능을 보여주었다. Feature Coding (Track 1)의 경우는 2023년 4월에 Call-for-Proposals 문서가 발간되었으며, 2023년 10월에 CfP에 대해 제출된 기술 제안에 대한 평가가 이루어질 예정이다.

Imane and Video Coding 트랙의 경우, 2022년 10월에 만들어진 다섯 개의 CE에 대해 2023년 4월 회의 결과, CE 1(RoI based coding methods), CE 2(neural network based intra frame coding), CE 3(frame level spatial resampling)은 지속하기로 결정됐고, CE 4(Temporal resampling)은 기술 고려사항에 반영하고 Reference Software에 병합되었고 CE 5(Post filtering)는 CE 2에 병합되는 것으로 결정되었다.

한편, VCM Reference Software의 경우 2023년 4월 ver 0.41까지 개발이 진행되었으며, MPEG-gitlab을 통해 관리되고 있다.

Acknowledgement

본 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2020-0-00011, 기계를 위한 영상 부호화 기술)

참 고 문 헌

[1] 추현곤, 정원식, 서정일(2023), 기계를 위한 비디오 부호화 표준화 동향, 방송과 미디어, 28(1),38-52.

[2] Use cases and requirements for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2 N190, 2022.04

[3] Evaluation Framework for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2 N193, 2022.04.

[4] Common Test Conditions and Evaluation Methodology for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2 N193, 2022.04.

[5] Call for Evidence for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2 N42, 2021.01.

[6] Call for Proposals for Video Coding for Machines, ISO/IEC JTC 1/SC 29/WG 2 N191, 2022.04

[7] Call for Evidence on Video Coding for Machines, ISO/IEC JTC 1/SC29/WG2 N215, 2022.07.

[8] Marek Domanski et. al., [VCM] Poznan University of Technology Proposal A in response to CfP on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60727, 2022. 10.

[9] Marek Domanski et. al., [VCM] Poznan University of Technology Proposal B in response to CfP on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60728, 2022. 10.

[10] Marek Domanski et. al., [VCM] Poznan University of Technology Proposal C in response to CfP on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60729, 2022. 10.

[11] Sang-Kyun Kim et. al., [VCM] CfP response: Region-of-interest based video coding for machine, ISO/IEC JTC1/SC29/WG2/m60758, 2022. 10.

[12] S. Wang et. al., [VCM] Video Coding for Machines CfP Response from Alibaba and City University of Hong Kong, ISO/IEC JTC1/SC29/WG2/m60737, 2022. 10.

[13] Hari Kalva et. al., [VCM] Response to VCM CfP from the Florida Atlantic University and OP Solutions, ISO/IEC JTC1/SC29/WG2/m60743, 2022. 10.

[14] Christopher Hollmann et. al., [VCM] Response to Call for Proposals from Ericsson, ISO/IEC JTC1/SC29/WG2/m60757, 2022. 10.

[15] Yegi Lee et. al., [VCM] Response to CfE: Object detection results with the FLIR dataset, ISO/IEC JTC1/SC29/WG11/M56572, 2021.04

[16] Yegi Lee et. al., [VCM Track2] Response to VCM CfP: Video Coding with machine-attention, ISO/IEC JTC1/SC29/WG2/m60738, 2022. 10.

[17] Jianran Liu et. al., [VCM] Video Coding for Machines CfP Response from Institute of Computing Technology, Chinese Academy of Sciences (CAS-ICT) and China Telecom, ISO/IEC JTC1/SC29/WG2/m60773, 2022. 10.

[18] ZIzheng Liu et. al., [VCM] Response to VCM Call for Proposals - an EVC based solution, ISO/IEC JTC1/SC29/WG2/ m60779, 2022. 10..

[19] ZIzheng Liu et. al., [VCM] Response to VCM Call for Proposals from Tencent and Wuhan University - an ECM based solution, ISO/IEC JTC1/SC29/WG2/m60780, 2022. 10..

[20] Ke Jia et. al., [VCM] Response to the CfP on Video Coding for Machine from Zhejiang University, ISO/IEC JTC1/SC29/WG2/m60741, 2022. 10.

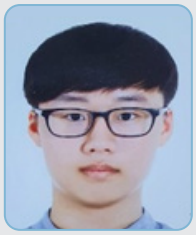
[21] Wen Gao et. al., [VCM]Response to VCM Call for Proposals from Tencent - an End-to-end Learning based Solution, ISO/IEC JTC1/SC29/WG2/m60777,

2022. 10.
[22] Cheng, Z., et. al., Learned image compression with discretized gaussian mixture likelihoods and attention modules. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7939-7948), 2020
[23] E. Agustsson, et. al., Scale-space flow for end-to-end optimized video compression, IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020)
[24] Honglei Zhang et. al., [VCM] Response to the CfP of the VCM by Nokia (A), ISO/IEC JTC1/SC29/WG2/m60753, 2022. 10.
[25] Honglei Zhang et. al., [VCM] Response to the CfP of the VCM by Nokia (B), ISO/IEC JTC1/SC29/WG2/m60754, 2022. 10.
[26] C. Rosewarne, [VCM Track 2] CfP test report, ISO/IEC JTC1/SC29/WG2/m61010, 2022. 10..
[27] CfP response report for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/N248, 2022. 10.
[28] Heeji Han et. al.,, [VCM] Response from Hanbat National University and ETRI to CfE on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60761, 2022.10.
[29] Yong-Uk Yoon et. al.,, [VCM] Response to VCM CfE: Multi-scale feature compression with QP-adaptive feature channel truncation, ISO/IEC JTC1/SC29/WG2/m60799, 2022.10.
[30] C. Rosewarne, R. Nguyen, [VCM Track 1] Response to CfE on Video Coding for Machine from Canon, ISO/IEC JTC1/SC29/WG2/m60821, 2022.10.
[31] Minhun Lee, et. al.,, [VCM Track 1] Response to CfE: A transformation-based feature map compression method, ISO/IEC JTC1/SC29/WG2/m60788, 2022.10.
[32] Yong Zhang et. al.,, [VCM] Response to VCM Call for Evidence from Tencent and Wuhan University - a Learning-based Feature Compression Framework, ISO/IEC JTC1/SC29/WG2/m60925, 2022.10.
[33] CfE response report for Video Coding for Machines, ISO/IEC JTC 1/SC29/WG2 N247, 2022.10

약 력

서정일

1994년 경북대학교 공학사
1996년 경북대학교 공학석사
2005년 경북대학교 공학박사
1998년~2000년 LG반도체 선임연구원
2000년~2023년 한국전자통신연구원 실감미디어연구실장
2023년~현재 동아대학교 컴퓨터공학과 부교수
관심분야: 멀티미디어, 오디오/비디오 부호화, 딥러닝, 머신비전



이동민

2020년 동아대학교 컴퓨터공학과 입학
관심분야: 머신러닝, 딥러닝, 영상부호화

김준수

2012년 서울대학교 전기컴퓨터공학부 학사
2017년 서울대학교 전기정보공학부 박사
2017년~현재 ETRI 실감미디어연구실 선임연구원
관심분야: Light field, VR/AR, 컴퓨터 비전, 기계를 위한 영상 부호화

추현곤

1998년 한양대학교 전자공학과 (공학사)
2000년 한양대학교 전자공학과 (공학석사)
2005년 한양대학교 전자통신전파공학과 (공학박사)
2015년~2017년 한국전자통신연구원 디지털홀로그래피연구실장
2017년~2018년 Warsaw University of Technology, Poland 방문연구원
2023년~현재 한국전자통신연구원 실감미디어연구실실장
관심분야: Computer vision, 3D imaging and holography, 3D depth imaging, 3D broadcasting system