기계를 위한 비디오 부호화 표준화 동향

Standardization Trends in Video Coding for Machines

권형진 (H.J. Kwon, kwonjin@etri.re.kr) 미디어부호화연구실 책임연구원
정세윤 (S.Y. Cheong, jsy@etri.re.kr) 미디어부호화연구실 책임연구원
최진수 (J.S. Choi, jschoi@etri.re.kr) 미디어부호화연구실 책임연구원
이태진 (T.J. Lee, tjlee@etri.re.kr) 미디어부호화연구실 책임연구원/실장
서정일 (J.I. Seo, seoji@etri.re.kr) 실감미디어연구실 책임연구원/실장

ABSTRACT

An increase in high-quality video service continually leads to the standardization of high-performance video codecs such as the versatile video coding standard. Although such codecs have improved coding efficiency in terms of high fidelity, a tremendous increase in the amount of video data is required for more efficient compression, especially for efficiently recognizing and analyzing the target within the millions of objects/events captured every day, such as those by surveillance systems. Therefore, newly established MPEG standardization efforts have studied the new generation of video compression standards for machine vision-oriented video. This paper presents the standardization trends in video coding for machines and discusses further directions for improvement.

KEYWORDS 비디오 부호화, 기계 시각, 기계를 위한 비디오 코딩

Ⅰ. 서론

사물인터넷, 스마트 시티, 자율 주행 차 등 다양한 응용 환경에서 수집되는 비디오 데이터의 양은 기하급수적으로 증가하고 있으며, 이를 기반으로 비디오의 객체나 이벤트를 인식하고 이를 분석하여 활용하는 서비스 요구 역시 지속적으로 증대되고 있다. 이에 덧붙여 방대한 비디오 데이터를 사람이 직접 감시 및 분석하는 것이 한계에 달함에 따라 사람 대신 기계가 비디오 내의 영상 정보를 분석하여 다음에 발생할 상황을 예측하여 사람에게 알려주거나 직접 능동적으로 대처하는 지능화, 자동화 요구사항이 점진적으로 증대되고 있다.

한편, 비디오를 수집하는 영상 획득 장치와 이를 분석, 활용하는 임무 수행하는 장치가 분리된 경우, 통상적으로 비디오 코덱을 사용하여 부호화

* DOI: https://doi.org/10.22648/ETRI.2020.J.350509
* 본 연구는 미래창조과학부의 정보통신 · 방송 기술개발사업의 일환으로 수행하였음[2020-0-00011, 기계를 위한 영상 부호화 기술 개발].

하여 전송하고 이를 수신한 장치는 복호화한 비디오를 다양한 임무에 활용하는 방법을 취한다. 하지만, 비디오 데이터의 압축률을 높이면 영상 정보의 손실이 늘어나 임무 수행 성능이 떨어지게 된다. 특히 기존 비디오 코덱의 경우 HVS(Human Visual System) 특성을 고려하여 설계되었기 때문에 기계의 임무 수행에 필요한 영상 정보가 아님에도 인간 시각 인지에 중요한 특성 정보는 유지하는 비효율성이 존재한다. 따라서 기계에 임무 수행 성능을 유지하면서 비디오 데이터를 효율적으로 압축하기 위해서는 기계의 임무 수행에 초점을 맞춘 비디오 부호화 기술이 필요하다. 기존 영상 부호화 기술과 같이 사람을 위한 부호화와 새롭게 표준화가 진행되는 기계를 위한 부호화의 비교는 표 1과 같다.

임무 수행 성능을 유지하며 다중 임무 수행을 위한 비디오 또는 비디오로부터 추출된 특징(Feature)의 압축 비트스트림의 기술표준 제정을 목표로 ISO/IEC JTC1 SC29 WG11(MPEG)에서는 2019년 7월 VCM(Video Coding for Machine) AHG이 결성되었다. 본 고에서는 이러한 MPEG VCM 기술에 대한 표준화 추진 현황 및 주요 이슈들에 대해 기술하고자 한다.

표 1 사람을 위한 영상 부호화와 기계를 위한 영상 부호화 비교

비교항목	사람을 위한 영상 부호화	기계를 위한 영상 부호화
영상신호의 목적	사람에게 영상을 통한 정보전달	기계의 임무수행을 위한 정보전달
영상부호화 목표	인지화질을 유지하며 압축성능을 최대화	임무성능을 유지하며 압축성능을 최대화
영상부호화 복잡도와 전력소모량	압축성능 향상이 목표이므로 상대적으로 중요하지 않음	기계 간 실시간 연동을 위해 매우 중요
스마트시티 등을 위한 중앙집중처리	높은 복잡도로 인해 부적합	고압축률/저복잡도 구현이 가능하므로 적합

II. MPEG VCM 표준 기술 동향

1. 기존 유사 MPEG-7 표준

MPEG VCM의 목표와 유사한 기존 표준 기술로서 영상으로부터 특정 비전 임무 수행을 위해 영상 특징 정보를 추출하여 효율적으로 표현하는 기술인 MPEG-7 CDVS(Compact Descriptors for Visual Search)/CDVA(Compact Descriptors for Video Analysis) 표준이 제정되었다. 이 CDVS 기술의 목적은 비주얼 검색 임무를 위하여 영상으로부터 효율적인 서술자 추출과정(특징 정보의 추출/분석/비교/검색)을 표준화하는 것이다. 표준화 진행 당시 후보 기술의 성능 평가는 주어진 질의 영상과 일치하는 영상을 찾아내는 검색의 정확도와 영상으로부터 추출한 특징 정보의 압축률을 나타내는 서술자의 크기의 2가지 척도를 사용하였다. 하나의 영상마다 나타나는 서술자의 크기는 512Byte에서 최대 16KByte까지 지원한다. CDVS는 2015년에 표준화가 완료되어 ISO/IEC 15938-13:2015: "Part 13: Compact descriptors for visual search"란 문서로 공표되었다.

이 CDVS의 후속 표준인 CDVA는 실시간으로 비디오 검색 수행을 목표로 하여 시간축상의 중복성을 제거하여 CDVS보다 효율을 증대시켰다. 보다 상세하게는 비디오를 우선 시각적으로 균일한 temporal segment로 나눈 후 각 segment별로 CDVA를 생성한다. 하나의 segment의 CDVA 서술자는 이를 대표하는 참조 프레임으로 추출한 참조 서술자와 이 참조 서술자로부터 예측하여 생성한 나머지 서술자로 구성된다. 또한 하나의 CDVA 서술자의 크기는 1초당 2~4KByte 범위이며 서술자를 추출하는 데 1초 분량의 비디오당 약 0.7초 소요되어 실

시간 처리가 되도록 설계하였다. CDVA는 2019년에 표준화가 완료되어 ISO/IEC 15983-15: 2019: "Part 15: Compact descriptors for video analysis"란 이름의 문서로 공표되었다.

상기 두 가지 표준 기술은 서버-클라이언트 구조의 통신 패러다임을 바꾸었다. 기존엔 영상을 압축하여 서버에 전송하는 대신 추출한 특징의 효과적인 서술자만을 전송하여 압축률을 높였다. 또한 이 서술자는 영상의 압축된 표현이고, 다른 영상의 서술자와의 빠른 정합을 통해 유사도의 계산이 가능하여 대규모 DB에서의 검색에 매우 적합하다. 하지만, 비주얼 검색 임무에만 한정되어 있어 다른 다양한 컴퓨터 비전 임무에 사용되기에는 한계가 있다.

2. MPEG VCM기반 응용 서비스 및 요구사항

최근 인공지능 기술의 급격한 발전으로 인하여 객체 분류, 검출, 분할, 추적 등의 다양한 컴퓨터 비전 임무의 성능이 비약적으로 향상됨에 따라 물체 인식 기술의 실용화가 현실이 되어가고 있다. 또한 영상 센서로부터 수집되는 영상 데이터의 양이 폭증함에 따라 기계에 의존한 영상 분석의 시대가 도래하였다. 따라서 사람이 개입되지 않고 기계의 시각기반 임무 수행하기 위하여 기계가 임무를 수행하는 성능을 유지하면서 최대한 영상데이터를 압축하는 새로운 패러다임의 필요성을 2019년 7월 MPEG 회의에서 중국 회사들의 발의로 기계를 위한 비디오 부호화 논의가 시작되었다. 이후 2019년 10월 제128차 MPEG 회의부터 MPEG VCM AHG이 결성되어 2020년 7월 제131차 MPEG 회의까지 각종 기술 및 제반 사항들에 대한 논의가 진행 중이다.

MPEG VCM AHG에서는 서비스 시나리오, 주요 임무 선정, 요구사항, 시스템 아키텍처, 성능 평가 방법 등을 논의해 왔으며, 주요 내용의 동향에 대해 차례대로 기술한다.

VCM 기술이 사용될 수 있는 응용분야는 다양하다. 2020년 7월까지 정의한 문서에 따르면 다음과 같은 서비스로 정리할 수 있다[1].

- Surveillance
- Intelligent Transportation
- Smart City
- Intelligent Industry
- Intelligent Content

또한 VCM은 이 5가지 서비스 시나리오에서 필요한 기능을 도출하여 영상 화질 향상 또는 영상 복원과 같은 저수준 임무부터 객체 인식 또는 비디오 장면 이해와 관련된 고수준 임무까지 다양한 범위의 주요 시각 임무16가지를 선정하였다[1].

한편, VCM를 이용한 서비스는 그 용도와 사용 범위가 다양하지만, 몇 가지 공통된 요구사항으로 정리하면 다음과 같다.

- 비트스트림의 효율적 압축: 비슷한 임무 수행 성능을 보이는 VVC 압축 비트스트림보다 더 높은 압축률을 가져야 한다.
- 단일 또는 다중 임무를 모두 지원하는 비트스트림: 단일 임무와 다중 임무를 지원하는 시나리오에 대해 다르게 최적화를 수행하도록 지원하는 것이 목표이다. 이때 단일 임무를 위해 최적화한 비트스트림은 다중 임무용 비트스트림보다 더 작거나 더 작은 부호화 복잡도와 같은 장점을 가져야 한다.
- 다중 임무 수행 시 다양한 성능 정도 지원: 특정 임무가 다른 임무보다 우선순위가 높은 경우, 혹은 지연(Latency), 전송속도(Bandwidth)와 같은 특정 요구사항으로 인한 임무별 다

양한 부호화가 필요한 경우를 위하여 임무별 다양한 수준의 품질을 지원해야 한다.

- 사람의 모니터링을 위한 복원 가능한 비트스트림: 사람의 모니터링이 필요한 서비스를 위하여 부가적인 비트스트림을 지원해야 한다. 이때 부가 비트스트림의 비트율은 비슷한 PSNR에서 VVC 비트스트림의 비트율보다 작아야 한다.

VCM 서비스 시나리오와 여기서 파생된 시각 임무, 그리고 압축을 위한 요구사항을 살펴보았다. 다양한 요구사항이 있지만, 가장 중요한 요구사항은 압축률이다. 그런데 압축률기반으로 성능을 평가하는 방법이 잘 정립되어 있는 비디오 코덱과 다르게 VCM에서는 평가 방법의 각 세부 항목마다 많은 논의가 있었으며, 중요한 사항이므로 별도의 절을 할당하여 Ⅲ장에서 자세히 기술한다.

3. MPEG VCM 시스템 및 후보 파이프라인

MPEG VCM AHG의 기계를 위한 비디오 부호화 목표를 달성하기 위해 기존 비디오 압축 기술과 기계 시각 기술이 모두 필요하다. 전자는 기존 표준인 HEVC/VVC와 같이 비디오를 입력받아 인코더를 통하여 압축 스트림을 생성하고, 이 압축 스트림을 입력받아 디코더를 통하여 복원 비디오를 생성한다. 후자는 CDVS/CDVA와 같이 비디오를 입력받아 특징을 뽑아내고 이로부터 효율적인 서술자 비트스트림을 생성한다. 이 두 기술 분야의 확장으로서 VCM 파이프라인을 다양하게 설계할 수 있는데, 표준화 초창기이므로 특정 방향으로 한정하지 않고 열린 구조를 지향하고 있다. 이러한 VCM 구조의 한 가지 후보 예는 다음과 같다. VCM 부/복호화기는 크게 인간을 위한 비디오 부/복호화기와 기계를 위한 특징 부/복호화기로 구성될 수 있다. 여기서 특징 부호화기는 특징 추출, 특징 변환, 특징 부호화 모듈로 구성된다. 또한 기계 시각에 덧붙여 인간 시각 응용까지 고려하면 특징 부호화/복호화 모듈의 출력을 비디오 부호화/복호화 모듈의 입력으로 고려할 수 있다. 이처럼 다양한 시나리오에 따라 파생될 수 있는 VCM 부/복호화기의 구조를 명확히 하기 위하여 3가지 대표 파이프라인이 제시되었다[2].

VCM의 요구사항인 다수의 기계 시각 임무의 지원을 만족시키기 위한 비디오 압축하는 방법은 비디오 코덱의 파이프라인을 재사용하는 방법과 다중의 임무에 공유되어 사용될 수 있는 특징의 추출 및 압축하는 방법으로 크게 구분된다.

비디오 코덱의 파이프라인을 재사용하는 방법은 먼저 비디오 코덱을 사용하여 압축/복원한 후 기계 시각 임무를 수행하는 방식이다. 이 방식은 비디오 코덱으로 기존 비디오 코덱을 사용하는 그림 1(a)과 신경망으로 대표되는 학습가능한 비디오 코덱을 사용하는 그림 1(b)로 구분할 수 있다. 그림 1(a)의 방법은 VCM 부/복호화기로 비디오 부/복호화기를 재사용함으로써 사람과 기계를 위한 부호화를 동일한 구조를 사용한다는 장점이 있고, 특히 오랜 기간 동안 세부 모듈 최적화가 이루어져 화면 간 부호화 효율이 매우 좋은 비디오 코덱 기술의 장점을 그대로 활용할 수 있다. 게다가 복원된 비디오를 입력으로 사용하면 수많은 기계 시각 임무를 지원하기 위하여 VCM 부/복호화기의 추가 작업이 필요 없게 된다. 그러나 기존 비디오 코덱을 그대로 사용하면 기계 시각 임무의 수행에 최적화되지 않았기 때문에 비효율적일 수 있으며, 특히 손실된 시각 임무를 위한 특징 정보로 인한 성능 저하가 클 수 있다는 단점도 존재

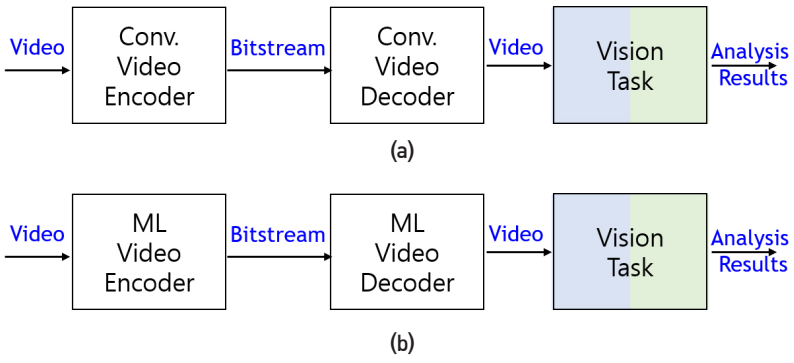

그림 1 VCM 파이프라인 예(비디오 부호화)

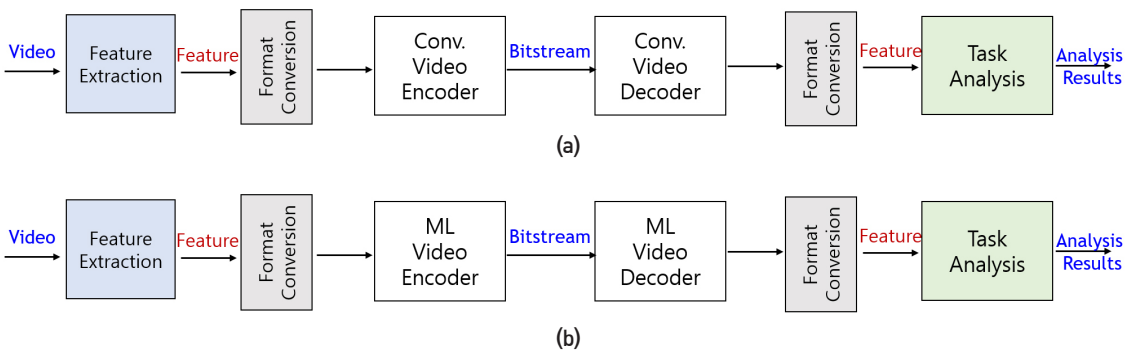

그림 2 VCM 파이프라인 예(특징 정보 부호화)

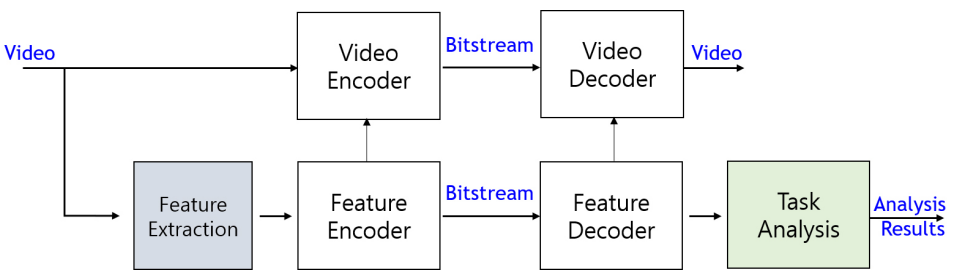

그림 3 VCM 파이프라인 예(하이브리드 부호화)

한다. 이러한 단점을 극복하기 위하여 제안되는 그림 1(b)의 방법은 학습 가능한 심화 신경망을 기존 비디오 코덱 대체하는 구성이다. 기존 비디오 코덱에서의 율–왜곡 비용 함수로 학습하여 정지 영상의 경우 기존 코덱 성능을 달성하였으며, 이를 동영상으로 확장하는 연구가 활발히 진행 중이다.

다중 임무에 공유되는 특징을 추출 및 압축하는 방법은 그림 1(a) 및 (b)의 기계 시각 임무 네트워크를 특징 추출하는 서브 네트워크와 이 서브 네트워

크의 출력 특징 맵을 사용하여 임무 수행하는 서브 네트워크로 분할한 후, 그 사이에 특징 압축 부호화하는 압축 네트워크를 사용하는 방식이다. 이 압축 네트워크로 기존 비디오 부호화기를 사용하는 그림 2(a)와 학습가능한 비디오 부호화기를 사용하는 그림 2(b)로 구분할 수 있다. 그림 2(a)의 방법은 전통적인 비디오 코덱의 효율을 그대로 사용할 수 있는 장점이 있으나, 영상 데이터의 특성에 최적화된 기존 비디오 코덱이 특징 정보의 특성에도 여전히 효율적인 방법인지 검증이 필요하다. 그림 2(b)의 방법은 기계학습 기반 압축 네트워크를 최적의 특징 정보의 압축을 수행하도록 학습할 수 있는 장점이 있다. 또한, 두 방법 모두 그림 1(a), (b)와 비교하여 추가적으로 코덱에 입력하기 전후의 특징 정보의 효율적인 처리 방법을 탐색해야 한다. 이 두 가지 방식은 기존 CDVS/CDVA와 같이 특정 임무에 최적화된 특징 표현 및 압축 방법의 다른 기계 시각 임무로의 scalable하게 확장되지 않는 단점을 개선하여 다중 임무 수행에 공유되는 특징을 추출할 수 있는 구조를 사용하는 것이 핵심이다. 하지만, 특징 추출 공유 아키텍처를 사용하더라도 각 임무에 사용할 특징을 어떻게 선정할 것인가 하는 문제가 남아 있다. 예를 들어, 심층 신경망을 사용한 구조로부터 임무마다 최적의 특징을 어떤 레이어의 출력을 사용하느냐를 결정해야 한다.

그림 3은 기계 시각 임무에 추가적으로 인간 시각 임무가 필요한 경우 기계 시각을 위해 이미 부호화한 정보를 사용하여 인간 시각을 위한 부호화의 성능을 개선할 수 있도록 기계 시각 임무를 위한 경로와 이 경로의 부호화 정보를 추가로 입력받아 인간 시각 임무를 위한 경로로 구성된 파이프라인이다. 앞서 요구사항에서 언급한대로 인간 시각을 위한 부가 비트스트림은 VVC 대비 비트율 효율적이어야 한다.

마지막으로, 그림 1과 2로 대표되는 2가지 압축 파이프라인 중에 어떤 방법이 본격적인 표준화의 참조 모델의 파이프라인으로 정해질지 미리 예측하는 것은 어렵다. 다만, 이 구조로부터 파생되는 하이브리드 시각 처리, 다중임무 처리 등을 고려하여 하나의 구조로 결정될 가능성이 높으며, 결국 실험 결과 성능과 다중 임무로의 확장성 등의 논의를 통해 결정될 것이며, 앞으로 관련 표준화 논의가 활발할 것으로 기대된다.

III. VCM 제안 기술의 성능 평가

이 장에서는 VCM의 표준 기술의 목표인 비전 임무 수행 성능을 급격히 저하시키지 않으면서 효율적인 압축을 달성하기 위해 각 회원사가 제안할 기술의 성능 평가를 위한 절차에 대해 기술한다. 보다 상세하게는 3가지 주요 기계 시각 임무의 성능 측정을 위해 사용되는 데이터셋들에 대해 살펴본 뒤, 기계 시각 임무 기술의 질적 수준을 평가하는 데 사용되는 평가 방법을 알아보고, 손실 압축을 수행했을 때의 부호화기의 효율을 비교하기 위한 참조 모델로서 Anchor를 생성하는 절차를 기술하고, 마지막으로 향후 표준화에서 쟁점으로 예상되는 이슈 사항에 대해 논의한다.

1. Anchor 정의

MPEG VCM AHG에서는 2020년 10월 CfE 발간 예정이며, 이 CfE에 응답하는 기고에서 성능 평가 베이스라인으로 사용할 Anchor 생성 방법에 대하여 2020년 7월 회의까지 논의하였다. 회의에서 정의된 Anchor 생성 절차는 그림 1(a)의 파이프라인에 따라 VVC 부/복호화기를 통해 압축 복원

한 영상을 객체 검출, 분할, 추적 3가지 기계 시각 임무를 수행하는 신경망 네트워크에 개별 입력하여 추론 결과에 대해 각각의 임무 수행 성능을 측정하는 것이다.

Anchor 생성을 위해 사용되는 주요 임무는 객체 인식 관련된 임무로서 정지 영상에 대해 객체 검출 및 분할하는 임무와 동영상에 대해 객체 분할 및 추적 임무로 나뉜다.

정지영상을 입력으로 하는 객체 검출 그리고 객체 분할 기술은 객체 검출 및 분할 챌린지[3,4]에서 제공한 데이터셋과 챌린지에서 제공한 성능평가 방법을 그대로 도입하며, 비교적 벤치마킹 데이터셋과 평가 방법이 일찍 정리된 분야이다. 2015년 챌린지가 시작한 이후로 대부분의 메이저 학회에 공개되는 논문들은 MS COCO[3] 데이터셋에 대해 성능평가를 수행하였다. Cityscape[4] 데이터셋은 자율주행을 위한 도로환경 데이터로 자동차 주행영상과 각 영상의 영역 분할 정보를 함께 제공함으로써 기계가 자동차 주행 중에 앞에 있는 물체가 무엇인지 정확하게 인식하여 얼마 후에 물체가 어디에 위치할지 정확하게 예측하여 자동차의 주행 계획이나 제어와 같은 기계의 후속 임무 수행에 적용하기 위한 목적으로 생성되었다.

80개의 객체 클래스를 다루는 MS COCO 2017[3] 데이터셋은 검출 및 분할 관련하여 정지영상을 학습용으로 118,000여 개, 검증용으로 5,000여 개, 테스트용으로 41,000여 개의 데이터를 제공하는데, 테스트영상에 대한 정답을 제공하지 않아 학습영상으로 학습한 모델을 검증영상으로 성능을 평가한다.

Cityscape[4] 데이터셋은 자율주행 차가 객체 인식을 위해 실제 차량이 도시 거리를 주행하여 얻은 영상의 장면을 분석하여 8개의 의미 그룹(평지, 사람, 차량, 공사 중, 객체, 자연, 하늘, void)으로 구성된 총 30개의 클래스를 다루며, 25,000여 개의 주석이 달린 영상을 제공한다.

동영상을 입력으로 하는 동영상 객체 분할 기술은 동영상 프레임 안의 객체 영역을 배경으로부터 분할하는 기술로서 사용자가 제공하는 정보의 정도에 따라 비지도/준지도/인터렉티드 동영상 객체 분할의 세 분류로 나뉜다. 본 고에서는 VCM에서 고려하는 준지도 동영상 객체 분할에 대해서만 다룬다. 준지도 동영상 객체 분할은 첫 프레임에서 제공한 타겟 객체에 대한 정확한 분할 영역을 이용하여 이후 프레임에서의 타겟 객체 분할을 수행한다. 동영상 객체 분할 기술은 DAVIS[5] 데이터셋을 주로 사용한다. DAVIS 데이터셋은 동영상마다 하나의 타겟 객체만을 대상으로 다루는 DAVIS 2016과 여러 객체들을 대상으로 다루도록 확장한 DAVIS 2017 데이터셋이 존재하는데 VCM에서는 2가지 데이터셋들을 모두 이용한다.

마지막으로 동영상 객체 추적 기술은 동영상 프레임 안의 타겟으로 하는 객체를 포함하는 영역 박스를 추적하는 기술로서 첫 프레임에서 제공한 타겟 객체에 대한 정확한 영역 박스를 이용하여 이후 프레임에서의 타겟 객체 추적을 수행한다. 동영상 객체 추적 기술은 매프레임 객체의 영역 박스를 출력한다는 면에서 객체 검출 기술과 유사하다고 볼 수 있으나, 하나의 프레임 내의 영역 박스 내의 객체를 다른 프레임의 객체 영역 박스 내의 객체와 동일한 객체임을 식별하여야 한다는 점에서 더 어려운 기술이다. 동영상 객체 추적 기술은 응용에 따라 다양한 데이터셋이 있는데, 특히 다중 객체 추적 기술은 MOT(Multiple Object Tracking) Challenge[6]에서 2015년부터 제공한 벤치마킹 데이터셋과 평가 방법을 제공하였으며, VCM에서는 MOT 20[7] 데이터셋을 사용한다.

2. 임무 수행 성능 평가 방법

객체 검출 임무의 성능 평가를 위해서 특정 IoU 값에 대해 객체 클래스별 AP(Average Precision)를 평균하여 계산한 mean AP를 주로 사용한다. COCO 챌린지[3]에서 도입한 평가 방법은 0.5부터 0.95까지 0.05 간격의 매 IoU값에 대해 mean AP를 평균한 mAP값을 사용한다. 정지영상의 객체 분할 임무의 성능 평가를 위한 메트릭 역시 객체 검출 임무와 동일한 mAP를 사용한다. IoU를 계산에 사용되는 데이터 형태가 전자는 예측한 객체 영역 직사각형 박스이고 후자는 영역 박스 내 각 픽셀마다 해당 객체에 포함 여부를 나타내는 이진값인 임의의 모양의 분할맵으로 인한 차이만 있을 뿐 계산하는 방식은 동일하다.

동영상 객체 분할 임무의 성능 평가를 위한 메트릭은 분할 영역 유사도를 측정하는 J score와 객체 경계선 정확도를 측정하는 F score를 평균한 J&F mean을 사용한다.

동영상 객체 추적 임무의 성능 평가를 위한 메트릭은 MOTA(MOT Accuracy)를 사용한다. 이 객체 추적의 메트릭은 객체의 영역 박스는 잘 검출하였으나 해당 객체의 ID를 오인하거나 새로운 객체로 인식한 경우도 해당 프레임의 추적을 틀린 것으로 간주하며, 이처럼 잘못 추적한 프레임을 배제하여 맞게 추적한 추적궤도의 프레임만으로 정확도를 계산하도록 설계되었다.

3. 부호화 효율 평가 방법

2절에서 대표 기계 시각 임무의 성능 평가의 방법에 대해 기술하였다. MPEG VCM에서는 기존 성능 평가 방법에 비디오의 압축률까지 고려한 성능 평가 방법을 제시한다.

일반적으로 비디오 코덱의 부호화 효율은 압축 비트스트림의 비트율과 복원된 비디오의 화질의 트레이드오프(Tradeoff)로 결정된다. 또한 대부분의 응용에서 비트율이 주어졌을 때, 최상의 복원 화질로 부호화하고자 한다. 따라서 서로 다른 비디오 부호화기의 성능을 비교하기 위해서는 복원된 비디오의 비트율과 화질을 측정하는 방법을 정해야 한다. 현재까지 MPEG 비디오 부호화 표준에서의 부호화기의 객관적 화질 평가 방법은 BD-PSNR과 BD-rate[8,9]를 사용하는데, 이는 두 부호화기로 4개 이상의 QP(Quantization Parameter) 값을 변화시키며 PSNR을 측정하고, 이러한 동작점들로 획득한 2개의 rate-distortion(R-D) curve를 사용하여 평균 PSNR 및 비트율 차이를 계산하여 부호화기의 성능을 비교하는 방법이다.

이와 같은 비디오 코덱의 부호화 효율 평가하는 방법을 기계 임무 수행 성능을 고려하여 수정하여 VCM 부호화 효율 평가를 한다. 수정 사항은 비트율 변화시키며 손실 압축으로 인한 복원한 비디오의 왜곡 대신에 타겟 임무 수행 성능을 측정하여 rate-performance(R-P) curve 그래프로 나타내고, 비교하는 2개의 부호화기의 성능은 대응되는 R-P curve를 사용하여 평균 비트율 차이를 BD-rate로 계산하여 비교한다.

지금부터는 부호화 효율 평가 방법을 위한 베이스라인 Anchor를 생성하는 과정에 대해 기술한다. 그림 1(a)의 파이프라인을 따라 Anchor 생성하는데, 보다 상세하게는 크게 영상을 다운샘플링, 컬러 포맷 변환, VVC 부호화, VVC 복호화, 컬러 포맷 역변환, 업샘플링, 시각 임무 수행, 성능 평가의 세부 파이프라인 순서로 진행된다. 다른 비트율로 인한 임무 성능 변화를 측정하기 위하여 6개의 QP 값(22, 27, 32, 37, 42, 47)을 사용하여 부호화하여 생성한 복원 영상을 임무 수행한 후 성

능을 측정한다. 그 밖에 상기 세부 파이프라인에 따라 생성된 Anchor의 성능의 재현이 가능하도록 파이프라인에 사용되는 SW의 버전(전후 처리는 FFmpeg 4.2.2, VVC 부/복화기는 VTM 8.2)이나 SW의 configura-tion 등의 상세한 사항은 다음 문서를 참고하면 된다[2].

4. 향후 주요 이슈 사항

아직 많은 논의가 이루어지지 않은 사항은 기계와 인간 모두를 고려하는 하이브리드 시각을 위한 압축 프레임워크로서, 손실 압축으로 인한 복원된 화질의 왜곡과 임무 수행 성능의 저하를 어떻게 다룰 것인지가 큰 문제이다. 3가지 항목의 트레이드오프를 다뤄야 하는데, 하나의 비트스트림으로 특성이 다른 2가지 목적을 효율적으로 지원할 수 있을지 검토해야 한다. 만약 인간 시각을 위한 영상 복원을 위해서 추가적인 비트스트림이 필요하다고 한다면, 기계 시각 임무를 인간 시각보다 우선순위를 두도록 정한다하더라도 각 항목에 대응되는 비트스트림의 비트율 배분문제가 여전히 남아있다. 향후 효율적인 VCM기반 하이브리드 시각 서비스를 생각한다면 표준화에서 연구되어야 할 아이템이다.

또한 논의가 많이 이루어지지 않은 주요 사항은 다중 임무 수행 시 성능 평가 방법에 대한 것이다. 상기 하이브리드 시각의 경우 2가지 목표로 인해 고려사항이 늘어났는데, 2개 이상의 다중 임무 수행을 위해 각각의 임무 성능을 다르게 지원할 수 있는 방법론을 제시해야 한다.

게다가 응용 서비스에 따라 성능 요구사항이 달라지는데 이를 어떤 식으로 지원할지 논의가 필요하다. 예를 들어 자율주행차를 대표되는 서비스와 같이 저지연, 실시간 처리가 요구될 경우를 타겟팅한 아키텍처나 파이프라인을 별도로 지정할 수도 있고, 아니면 하나의 아키텍처에서 프로파일링을 통해 요구사항별 프로파일과 그 세부 레벨을 미리 정하는 방식으로 지원할 수도 있다.

VCM이 기계 시각을 위하여 비디오 또는 비디오로부터 추출된 특징을 압축할 때 기존 비디오 코덱에 쓰였던 기술을 응용하여 적용할 것으로 보인다. 또한, 최근 비약하는 심층 신경망 기술에도 이러한 기존 코덱 기술과 접목하는 새로운 시도가 출몰할 것으로 기대된다. 관련 표준화가 이제 시작단계이나 충분한 기술로 성숙된다면, 다양한 사물인터넷 응용의 사물 간 통신 시나리오에서 주요한 미디어 전달 기술로 자리 잡을 것으로 기대한다.

Ⅳ. 결론

MPEG VCM은 기계가 비디오를 처리하는 환경에서 현재 인간이 소비하는 환경에서 최고수준인 VVC보다 비디오를 효율적으로 전송하고 저장하기 위한 표준화 작업이 새로 시작되어 1년 정도 진행되었으며, 2020년 10월 CfE 목표로 하고 있다. 본 고에서는 이러한 MPEG VCM 표준화 현황을 간략하게 알아보고, CfE를 위한 anchor 생성, 파이프라인 등 주요 이슈 사항들에 대해 기술하였다.

이와 같은 기계를 위한 비디오 부호화의 최신 도입은 실제 시장의 수요와 필요에 의하여 제기되어 논의되고 있는 것으로서, 기존 비디오 부호화 전문가 그룹과 MPEG-7 산하 모바일 비주얼 검색 임무를 위한 서술자 추출 전문가 그룹이 주도적으로 각자 전문분야의 경험을 확장하여 기계 시각 임무를 위한 압축 기술을 제안할 것으로 예상된다. 국내에서도 기업, 연구소, 대학 등에서 VCM 표준화에 초창기부터 적극 참여하고 있어 VVC 표준과 CDVA 표준에서의 경험과 성과를 바탕으로 VCM 표준화

전략 및 IPR 확보 전략을 수립하여 활발한 표준화 활동을 지속한다면 VCM 표준에 대한 국내 기술 및 IPR 반영 가능성이 상당히 높다고 판단된다.

용어해설

기계 시각 기계에 인간이 가지고 있는 시각과 판단 기능을 부여한 것으로 사람이 인지하고 판단하는 기능을 하드웨어와 소프트웨어의 시스템이 대신 처리하는 기술

MPEG(Moving Picture Experts Group) 1988년에 설립되어 350여 명의 다양한 산업 및 학계 전문가가 참여하고 있는 멀티미디어 표준화 전문 그룹

VCM(Video Coding for Machine) MPEG Requirements 서브그룹 산하 AHG으로 기계를 위한 비디오 부호화 표준화를 목표로 다양한 표준화 준비 작업을 진행하고 있으며, 현재 2020년 10월 CfE 발간 예정임. 이를 통해 VVC 대비 성능 개선되는 다양한 기술 제안을 모집하여 표준화 진행 충분한지 확인하고자 함

약어 정리

AHG	Ad-Hoc Group
BD	Bjontegaard Delta
CfE	Call for Evidence
CfP	Call for Proposal
IoU	Intersection over Union
PSNR	Peak Signal-to-Noise Ratio
VCM	Video Coding for Machine
VVC	Versatile Video Coding

참고문헌

[1] ISO/IEC JTC1/SC29/WG11/w19506, "Use cases and requirements for Video Coding for Machines," July 2020.

[2] ISO/IEC JTC1/SC29/WG11/w19507, "Draft Evaluation Framework for Video Coding for Machines," July 2020.

[3] Lin TY et al., "Microsoft COCO: Common Objects in Context," ECCV, 2014.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," CVPR, 2016.

[5] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," CVPR, 2016.

[6] https://motchallenge.net/

[7] P. Dendorfer et al., "MOT20: A benchmark for multi object tracking in crowded scenes," arXiv:2003.09003, 2020.

[8] Gisle Bjontegaard, "Calculation of Average PSNR Differences between RD Curves," ITU-T SG16/Q6, 13th VCEG Meeting, Austin, Texas, USA, Doc. VCEG-M33, Apr. 2001.

[9] Gisle Bjontegaard, "Improvements of the BDPSNR Model," ITU-T SG16/Q6, 35th VCEG Meeting, Berlin, Germany, Doc. VCEG-AI11, 16-18 July, 2008.