VCM 을 위한 비디오 특징의 효율적인 표현 기법

윤용욱, 김재곤
한국항공대학교
yuyoon@kau.kr, jgkim@kau.ac.kr

Efficient representation of video features for VCM

Yong-Uk Yoon and Jae-Gon Kim
Korea Aerospace University

요 약

방대한 비디오 데이터의 지능형 분석을 수행하는 기계를 위한 비디오 부호화 기술의 필요성이 대두되면서 MPEG 에서는 VCM(Video Coding for Machine) 표준화를 시작하였다. VCM 은 지능형 머신(machine)의 임무 수행을 위한 비디오 또는 비디오 특징(feature)의 압축 표준 기술로 기술 탐색 단계의 표준화를 진행하고 있다. 본 논문에서는 머신비전(machine vision) 네트워크에서 추출되는 대용량의 특징 압축을 위한 전처리 단계로 보다 효과적인 특징 표현 방법을 제시한다. 제안하는 특징 표현 방법은 정규화, 양자화 과정을 거쳐 특징 데이터 크기를 감소시킨다. 실험에서 특징을 4 개의 값으로 양자화 했을 때, 원본 대비 16 배의 데이터 크기가 감소되지만 mAP 평가 성능은 35.4592 로 높은 수준으로 유지함을 확인하였다.

1. 서론

비디오 데이터가 폭증하면서, 스마트 시티, 자율주행, 사물인터넷(IoT) 등 수많은 응용에서 딥러닝 기반의 비디오 분석을 위해 대용량 비디오가 기계에 공급되고 있다. 이로 인해, 대용량의 비디오의 보다 효율적으로 전송 및 관리가 요구되고 있다. HVS(Human Visual System)에 기반으로 설계된 기존의 비디오 코덱은 기계를 위한 비디오 압축 기법으로 비효율적일 수 있으며, 또한, 비디오 압축으로 인한 정보의 손실은 기계가 지능형 분석의 임무 수행 정확도를 감소시킬 우려가 있다. 이에 따라 기계의 임무 수행 성능과 압축 효율성 측면에서 영상 및 비디오의 특징(feature)을 추출하고 이를 압축하는 접근방법이 고려되고 있다. 즉, 기존의 코덱으로 압축된 비디오 보다 적은 양의 데이터를 사용하여 다양한 지능형 비디오 분석 임무를 수행하는 것이다.

MPEG 에서는 이러한 지능형 분석을 수행하는 기계를 위한 비디오 부호화 기술의 필요성에 따라 VCM(Video Coding for Machine) 표준화를 활발히 진행하고 있다. 최근 제 132 차 MPEG 회의에서는 VCM 표준 기술 후보를 탐색하기 위한 CfE(Call for Evidence)[1]를 공표하였으며, 다가오는 회의에서 세계 유수 기관에서 다양한 VCM 후보 기술들이 제시될 것으로 기대되고 있다. VCM CfE 응답을 위해 실험조건, 평가방법 및 과정, 요구사항을 정의하였다[2].

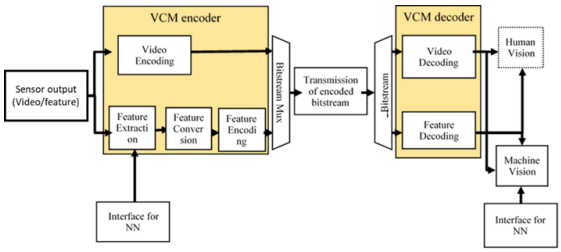


그림 1. VCM 시스템 구조

2. VCM 의 특징 압축

VCM 의 파이프라인(pipeline)은 그림 1 과 같이 부호화기, 복호화기, 머신비전(machine vision) 작업 분석기로 구성된다. 부호화기는 특징 추출, 특징 변환, 특징 부호화로 구성되며, 복호화기는 특징 압축 비트열을 입력으로 복호화를 수행하고, 머신비전을 통해 분석 작업을 수행한다. CfE 응답을 위한 평가기준을 정의하기 위해, 그림 2-(가)와 같은 구조로 최근 표준화가 완료된 VVC(Versatile Video Coding)을 이용하여 입력된 이미지/비디오를 압축하고 복호화된 영상을 머신비전 작업 분석을 수행하는 것을 기준(Anchor)으로 정의하였다. 그림 2-(나)는 특징 압축을 위한 구조로서, 특징 추출기로부터 특징을 추출하고, 해당 특징을 압축에 효율적인 형태로 변환하여 압축을 수행한다. 그림 2-(다)는 해당 구조의 예로 머신비전을 위한 특징 추출 네트워크로부터 추출된 특징을 이미지화 하여 특징맵(feature map)을 형성하고 이를 기존 비디오 코덱을 이용한 압축한다[3-4]. 하지만 높은 머신비전 성능을 위해 설계된 네트워크의 특징은 입력 영상 대비 방대한 크기의 특징 데이터를 포함하고 있으며, 많은 고주파 성분을 포함하고 있기 때문에 압축 효율이 좋지 않다. 따라서, 추출되는 특징 데이터 크기를 효과적으로 감소시키는 방법 또한 제시되고 있다[5-6]. 이에 본 논문에서는 VCM 에서 평가 방법으로 제시된 인공신경망으로부터 추출된 특징의 데이터를 감소시키기 위한 효율적인 특징 표현 방법을 제시한다.

표 1. 입력 이미지와 특징 데이터 크기

	Size (WxHxC)	Raw data size	Data size ratio (features/input)
Input images	2048x1024x3	6,144 KB (2048*1024*3*8bit)	-
Original features	512x256x256	174,592 KB (44,695,552*32bit)	28.41
	256x128x256		
	128x64x256		
	64x32x256		
	32x16x256		

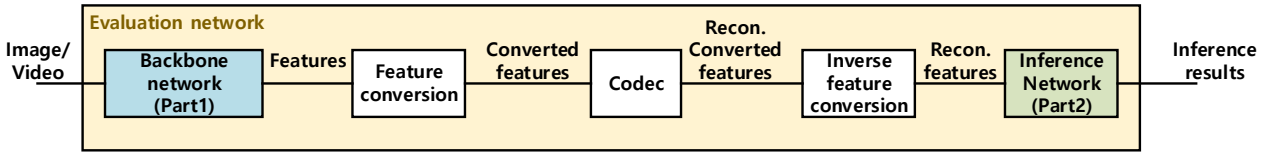
3. 제안하는 특징 표현 방법

일반적으로 영상으로부터 추출되는 특징은 머신비전의 높은 임무 수행 성능을 위해 입력 영상 보다 큰 데이터 크기를 갖는다. 본 논문에서는 VCM 에서 사용하는 평가 네트워크인 R50-FPN 및 Cityscapes 데이터셋을 사용하여 데이터 크기를 확인하였다[2]. R50-FPN 네트워크의 특징 추출 네트워크는 서로 다른 해상도의 256 채널 특징(P2~P6)을 생성한다. 표 1 은 입력 이미지와 추출된 특징의 데이터 크기를 보여준다. 입력 영상 크기 대비 추출된 특징 데이터가 약 28 배가 크다. 이러한 방대한 크기의 특징 데이터를 감소시키기 위해 32-비트 부동소수점 형태로 표현되는 특징 데이터를 보다 적은 비트로 표현할 수 있지만, 이 경우 데이터의 손실을 야기할 수 있으며 임무 수행 성능에 큰 영향을 줄 수 있다. 따라서, 머신비전 성능에 영향을 주는 데이터의 손실을 최소화할 수 있는 데이터 감소 방법이 필요하다.

제안하는 특징 데이터 표현 방법은 정규화, 양자화 두 단계로 이뤄진다. 정규화는 머신비전에 필요한 중요한 특징 정보를 구분하고 일정한 양자화 구간을 설정하기 위한 단계이며, 양자화는 정규화된 특징을 구간 분할하여 대표값을 부여함으로써 특징 데이터 크기를 감소시킨다.

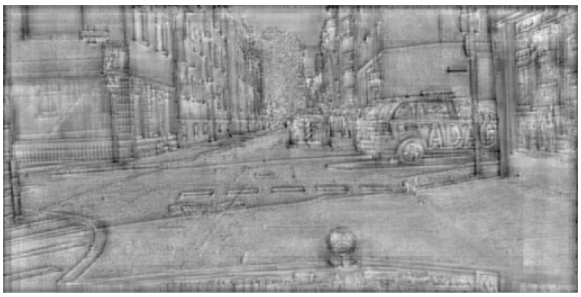
(가) 파이프라인 1

(나) 파이프라인 2



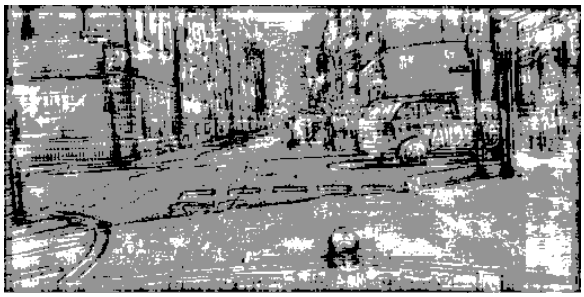
(다) 특징 압축을 위한 파이프라인의 예

그림 2. VCM 파이프라인



(가) 원본 특징

(나) 4 개의 심볼로 변환된 특징



(다) 3 개의 심볼로 변환된 특징

그림 3. R50-FPN 으로부터 추출된 특징 P2 의 256 개의 채널 중 첫번째 채널의 특징에 대한 예

(1) 특징 정규화(Feature normalization)

특징 추출기로부터 추출된 특징은 일반적으로 다채널의 32-비트 부동소수점 형태로 표현된다. 각 채널의 특징은 대체로 고유의 평균값을 갖도록 생성되기 때문에 특징 데이터 감소를 함에 있어 평균값을 유지하는 것이 중요하다. 따라서, 각 채널별(channel-wise)로 평균과 표준편차를 이용하여 식 (1)과 같이 정규화를 수행한다. 정규화된 특징은 평균이 0 이고 표준편차가 1 인 표준정규분포를 가지기 때문에, 모든 채널의 특징에 대해 동일한 구간으로 분할하는 균일 양자화가 가능하다.

$$z = \frac{x - \mu}{\sigma}$$
$$\mu: mean, \sigma: standard\ deviation \quad (1)$$

$$\begin{cases} interval1, & z < -1 \\ interval2, & -1 \le z < 0 \\ interval3, & 0 \le z \le 1 \\ interval4, & z > 1 \end{cases} \quad (2)$$

(2) 특징 양자화(Feature quantization)

정규화된 특징은 n 개의 구간으로 양자화되어 n 개의 심볼(symbol)로 표현된다. 여기서, n 은 양의 정수이다. 즉, 32-비트 부동소수점 형태의 특징이 ceil($\log_2$(n))-비트로 표현된다. 예를 들어, 특징이 4 개의 구간으로 양자화될 경우, 특징은 2-비트 형태로 구성된다. 따라서, 기존 32-비트부동소수점 형태의 특징은 2-비트 정수로로 변환되어 16 배 감소된다. 표 2 는 원본 영상 및 특징과 변환된 특징의 크기를 비교한 것이다. 양자화 과정의 예로 식 (2)와 같이 정규화된 특징을 임계값 z를 이용하여 4 개의 구간으로 양자화한다. 그림 3 은 정규화, 양자화를 거친 특징의 예를 보여준다. 특징을 영상으로 출력하기 위해, 32-비트 부동소수점 형태의 특징을 8-비트 정수 형태로 변환하였고, 양자화된 특징 또한 8-비트 정수 형태로 스케일링(scaling) 하였다.

(3) 특징 재구성(Feature reconstruction)

작은 크기로 변환된 특징 데이터의 성능 평가를 위해서 역양자화, 역정규화 과정이 필요하다. 역양자화는 정규화된 특징에서 분할된 구간의 평균값을 할당하고, 정규화에 사용된 평균과 표준편차를 이용하여 특징은 재구성된다. 다시 말해, 정규화, 양자화 과정을 거친 특징은 원본 특징의 평균과 표준편차를 갖는 n 개의 심볼로 표현되는 특징으로 복원된다.

표 2. 원본 이미지 및 특징과 변환된 특징의 데이터 크기 비교

	Raw data size	Data size ratio (features/input)
Input images	6,144 KB (2048*1024*3*8bit)	-
Original features	174,592 KB (44,695,552*32bit)	28.41
Converted features (8-bit image)	43,648 KB (44,695,552*8bit)	7.10
Converted 4 symbols features (2-bit)	10,912 KB (44,695,552*2bit)	1.77
Converted 3 symbols features (1~2-bit)	5,456~10,912 KB (44,695,552*(1~2bit))	0.88 ~ 1.77

4. 실험 결과

제안하는 특징 표현 방법의 성능을 평가하기 위해, CfE 에서 객체 분할(object segmentation) 작업의 평가 네트워크로 정의된 R50-FPN 을 사용하였고, 데이터셋은 Cityscapes 를 사용하였다[2]. 다양한 양자화에 따른 성능을 평가하기 위해 3 개 또는 4 개의 구간으로 양자화하고, 구간 분할을 위한 임계값 z 를 {0.5, 1.0, 1.5, 2.0}으로 구성하여 평가하였다. 표 3 은 제안하는 방법의 평가 결과를 보여준다. 4 개의 구간으로 양자화했을 때 데이터 크기는 원본 특징 대비 16 배가 감소되었지만, 평가 성능은 최대 35.4592 로 원본 특징의 평가 성능 36.4809 대비 미미한 성능 저하가 있음을 확인할 수 있다. mAP 성능 측면에서 양자화 구간의 수가 적을수록 성능이 감소하는 것을 보인다. 또한 임계값 z 에 따라 mAP 성능이 달라지기 때문에 최적의 양자화 경계를 찾는 방법도 고려할 수 있다. 데이터 크기 측면에서 특징 변환은 16 배의 크기를 줄임으로써, 입력 이미지 대비 약 1.7 배 크기의 특징을 압축할 수 있다.

변환된 특징을 VVC 와 같은 기존 비디오 코덱으로 압축할 수 있으나, 8-비트 또는 10-비트 심도의 비디오 입력에 대해 설계된 기존의 비디오 코덱은 소수의 심볼로 표현되는 변환된 특징을 압축하는데 비효율적일 수 있다. 따라서, 부호화 효율성 및 복잡성 측면에서 소수의 심볼로 표현된 특징을 압축하는데 적합한 새로운 코덱을 개발하는 것이 필요할 것으로 보인다.

표 3. 제안하는 방법의 성능평가(mAP)

Features	z threshold	mAP	Data size ratio (features/input)
Input image	-	36.4809	-
Original features (32-bit)	-	36.4809	28.41
Converted features (8-bit)	-	36.4617	7.10
Converted features (4 symbols, 2-bit)	0.5	23.3090	1.77
	1.0	33.8285	
	1.5	35.4592	
	2.0	34.1826	
Converted features (3 symbols, 1~2-bit)	0.5	22.8162	0.88 ~ 1.77
	1.0	32.2163	
	1.5	31.4846	
	2.0	21.6706	

5. 결론

본 논문에서는 정규화 및 양자화를 사용한 영상/비디오 특징 표현 방법을 제안하였다. 제안기법은 32-비트 부동소수점으로 표현되는 특징을 형태에서 2-비트 정수 형태로 변환하여, 특징 데이터의 크기를 16 배 줄였지만 머신비전 성능은 최대 35.4592 mAP 로 원본 특징 성능의 미미한 감소를 가져왔다. 따라서, 압축에 적합한 보다 효율적인 형태로 특징을 변환하는 특징 표현 기법에 대한 지속적인 연구가 필요하며, 또한 변환된 특징을 압축하는데 적합한 새로운 코덱의 개발이 필요할 것으로 보인다.

Acknowledgement

이 논문은 산업통상자원부 국가표준기술원에서 시행한 국가표준기술력향상사업[20011687]의 지원을 받아 수행된 연구임.

참 고 문 헌 (References)

[1] M. Rafie, L. Yu, Y. Zhang and S. Liu, “Draft of Call for Evidence for Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, N20, Online, Oct. 2020.

[2] M. Rafie, Y. Zhang and S. Liu, “Draft of Evaluation Framework for Video Coding for Machines,” ISO/IEC JTC 1/SC 29/WG 2, N19, Online, Oct. 2020.

[3] Y.-U. Yoon, D. Park, S. Chun and J.-G. Kim, “[VCM] Results of feature map coding for object segmentation on Cityscapes datasets,” ISO/IEC JTC 1/SC 29/WG 2, m55152, Online, Oct. 2020.

[4] S.-K. Kim, M. Jeong, H.-Y. Jin, H. K. Lee, H.-G. Choo, H. Lim and J. Seo, “[VCM] A report on intermediate feature coding for object detection and segmentation,” ISO/IEC JTC 1/SC 29/WG 2, m55243, Online, Oct. 2020.

[5] Y.-U. Yoon, D. Park and J.-G. Kim, “[VCM] Results of feature conversion for object segmentation,” ISO/IEC JTC 1/SC 29/WG 2, m55153, Online, Oct. 2020.

[6] H. Choi, M. Lee, J. Kim, Y. Lee, D. Sim, S. Oh, J. Do, H. Kwon and J. Seo, “[VCM] A result of feature data reduction using PCA for object detection,” ISO/IEC JTC 1/SC 29/WG 2, m55414, Online, Oct. 2020.