일반논문 (Regular Paper)
방송공학회논문지 제26권 제4호, 2021년 7월 (JBE Vol. 26, No. 4, July 2021)
https://doi.org/10.5909/JBE.2021.26.4.441
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

VVC 인코더에서 합성 곱 신경망의 어텐션 맵을 이용한 휘도 매핑 함수 생성 방법

권 나 성[a], 이 종 석[b], 변 주 형[b], 심 동 규[b]‡

Luma Mapping Function Generation Method Using Attention Map of Convolutional Neural Network in Versatile Video Coding Encoder

Naseong Kwon[a], Jongseok Lee[b], Joohyung Byeon[b], and Donggyu Sim[b]‡

요 약

본 논문에서는 VVC의 LMCS에서 휘도 신호 매핑 방법의 부호화 효율을 향상시키기 위한 휘도 신호 매핑 함수 생성 방법을 제안한다. 본 논문에서 제안하는 방법은 기존 LMCS에서 지역적 특징을 반영하기 위하여 사용하는 지역적 공간 분산에 합성 곱 신경망의 어텐션 맵을 곱하여 인지 지각적 특징을 추가적으로 반영한다. 제안하는 방법의 성능 평가를 위하여 AI (All Intra) 조건에서 VVC 표준 실험 영상의 A1, A2, B, C, D 클래스를 이용하여 VTM-12.0과 BD-rate 성능을 비교한다. 실험 결과로서 본 논문에서 제안하는 방법이 VTM-12.0 대비 BD-rate 성능 관점에서 휘도 성분이 평균 -0.07%의 성능 향상을 보이고, 부/복호화 시간은 거의 동일하다.

Abstract

In this paper, we propose a method for generating luma signal mapping function to improve the coding efficiency of luma signal mapping methods in LMCS. In this paper, we propose a method to reflect the cognitive and perceptual features by multiplying the attention map of convolutional neural networks on local spatial variance used to reflect local features in the existing LMCS. To evaluate the performance of the proposed method, BD-rate is compared with VTM-12.0 using classes A1, A2, B, C and D of MPEG standard test sequences under AI (All Intra) conditions. As a result of experiments, the proposed method in this paper shows improvement in performance the average of -0.07% for luma components in terms of BD-rate performance compared to VTM-12.0 and encoding/decoding time is almost the same.

Keyword : VVC, Encoder, Luma mapping with Chroma Scaling, CNN

Ⅰ. 서 론

최근 비디오 산업이 빠르게 발전함에 따라 더 높은 해상도 및 고화질 영상에 대한 수요가 증가하고 있으며 VR/AR, 360도 영상 등의 새로운 애플리케이션에 대한 수요 또한 증가하고 있다. 이에 따라 ITU-T (International Telecommunication Union Telecommunication Standardization Sector)의 VCEG (Video Coding Experts Group)와 ISO/IEC (International Organization for Standardization/International Electrotechnical Commission Joint Technical Committee)의 MPEG (Moving Picture Experts Group)은 JVET (Joint Video Experts Team)을 구성하고, HEVC (High Efficiency Video Coding)[1] 대비 2배 이상의 압축 효율 향상을 목표로 VVC (Versatile Video Coding)[2]의 표준화 진행을 2020년 10월 완료하였다. 차세대 비디오 압축 표준인 VVC는 부호화 효율을 높일 수 있는 다양한 기술이 채택되었다[3]. 그 중 VVC의 인-루프 필터[4]기술은 기존 HEVC에 존재했던 DF (Deblocking Filter), SAO (Sample Adaptive Offset)와 VVC에 새롭게 채택된 ALF (Adaptive Loop Filter), LMCS (Luma Mapping with Chroma Scaling)[5]기술로 구성되어 있다. 새롭게 추가된 LMCS 기술은 2019년 1월 회의부터 채택되어 VVC의 참조 소프트웨어인 VTM (VVC Test Model)[6]에 포함되어 있다. VVC는 SDR (Standard Dynamic Range) 영상뿐만 아니라 HDR (High Dynamic Range)[7] 영상, WCG (Wide Color Gamut) 영상 특성을 지원하는 비디오 압축 표준으로, LMCS 기술을 채택하여 HDR과 SDR 영상의 부호화 성능을 향상시켰다. LMCS는 입력 영상의 특성에 따라 SDR 영상, HLG (Hybrid Log Gamma) HDR 영상, PQ (Perceptual Quantizer) HDR 영상에 대해 각각 다른 알고리즘을 적용한다. 본 논문에서는 SDR 입력 영상을 기준으로 작성하였다. LMCS는 화소 값의 동적 범위를 부분 구간 선형 함수를 통해 변경함으로써 인코더가 효율적으로 부호화를 수행하게 하여 부호화 효율을 향상시키는 기술이다. 예를 들어 ITU- R BT.2100-2[8]에 따르면 narrow range video의 경우, 휘도 값의 범위가 10비트 기준으로 64부터 940까지 범위만 사용되는데, LMCS를 적용하면 부/복호화기에서 0부터 63과 940부터 1023 범위의 휘도 값을 추가적으로 사용할 수 있게 된다.

최근 비디오 압축 분야에서 인-루프 필터 등 다양한 기술에 합성 곱 신경망 (Convolutional Neural Network; CNN)[9]을 적용하여 부호화 효율을 향상시키고자 하는 시도가 다양하게 진행되고 있다[10-13]. 하지만 합성 곱 신경망을 적용함에 따라 부/복호화 복잡도가 증가하여 현실적으로 적용하기 어렵다는 문제점이 있다. 또한 현재 VVC에 새롭게 추가된 LMCS의 성능을 향상시키기 위해 합성 곱 신경망을 적용한 선행 연구는 거의 시도되지 않고 있다. 따라서 본 논문은 처음으로 LMCS에 수용 가능한 복잡도를 지닌 합성 곱 신경망을 적용하는 방법을 제안한다.

기존 LMCS는 휘도 값을 총 16개의 구간으로 나누고 영상의 특징을 분석하여 코드 워드를 각 구간에 할당한다. 이후 구간별로 할당된 코드 워드를 이용하여 구간 선형 함수를 생성하여 휘도 값의 동적 범위를 변경한다. 휘도 성분의 코드 워드를 할당할 때, 사람의 인지 시각 특성을 반영하여 영상의 복잡도에 따라 코드 워드를 다르게 할당할 수 있다. 기존 LMCS는 영상의 복잡도 특징으로 지역적 공간 분산 값을 사용하고 있다. 하지만 지역적 공간 분산 정보만을 이용하게 되면 분산이 클 때, 그 영역이 윤곽선과 같은 중요한 특징을 포함하는 영역인지 노이즈가 포함된 영역인지에 대한 분별이 어렵다는 문제가 있다. 따라서 본 논문에서는 인지 시각적 특성을 반영한 합성 곱 신경망을 사용하여 중요한 특징이 있는 부분과 노이즈가 있는 부분에 대한 정보를 기존 지역적 공간 분산에 가중치로 정보를 제공함으로써

a) 광운대학교 소프트웨어학부(School of Software, Kwangwoon University)
b) 광운대학교 컴퓨터공학과(Dept. of Computer Engineering, Kwangwoon University)
‡ Corresponding Author : 심동규(Donggyu Sim)
E-mail: dgsim@kw.ac.kr
Tel: +82-2-940-5470
ORCID: http://orcid.org/0000-0002-2794-9932
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학 지원사업(2017-0-00096) 및 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업(IITP-2021-2016-0-00288)에 의해 연구되었음.
※ This research was supported by the MSIT(Ministry of Science and ICT), under the National Program for Excellence in SW (2017-0-00096), supervised by the IITP(Institute for Information & communications Technology Promotion) and the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2021-2016-0-00288) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).
· Manuscript received May 12, 2021; Revised June 16, 2021; Accepted June 16, 2021.

기존 LMCS 성능을 향상시키고자 한다. 제안하는 방법으로 코드 워드를 재할당하여 휘도 성분을 효율적인 동적 범위로 재매핑시킴으로써 부호화 효율을 향상시킬 수 있다. 제안하는 방법은 기존 LMCS와 성능을 비교했을 때, 동일 화질에서 평균 -0.07%의 적은 비트율로 부호화를 수행할 수 있으며, 합성 곱 신경망을 사용함에도 불구하고 부호화 및 복호화 시간은 거의 동일하다.

이후 본 논문의 구성은 다음과 같다. 2장에서는 합성 곱 신경망과 VGG-16[14]에 대한 소개와 VVC에 존재하는 기존 LMCS의 기술 소개 및 등장 배경을 설명한다. 3장에서는 제안하는 휘도 신호 매핑 함수 생성 방법을 설명한다. 4장에서 제안하는 방법의 성능을 평가하고 5장에서 결론을 맺는다.

Ⅱ. 관련 이론

1. Convolutional Neural Network 및 VGG-16

합성 곱 신경망은 컴퓨터 비전 분야의 대표적인 신경망으로, 합성 곱 층을 통해 입력된 이미지의 특징을 추출하고 이를 기반으로 물체 분류, 객체 인식, 화질 개선 등을 수행하는 딥러닝 기반의 알고리즘이다. 합성 곱 연산을 통해 출력된 특징 맵은 이미지의 특징 정보를 담고 있다.

합성 곱 신경망 중 하나인 VGG-16은 이미지넷 이미지 인식 대회인 ILSVRC (ImageNet Large Scale Visual Recognition Challenge) 2014에서 우수한 성능을 보인 이미지 분류 모델로, 이미지 특징을 추출하는 기본 네트워크 모델로 활발히 활용되고 있다[15]. VGG-16 네트워크 구조는 합성 곱 계층 13개, 최대 풀링 (max pooling) 5개, 전결합층 (Fully Connected Layer) 3개로 구성되어 있다. 합성 곱 계층은 모두 필터 커널 크기를 3×3 크기로 고정하여 구성하였다. 5×5 필터로 합성 곱 연산을 한번 수행하는 것에 비해 3×3 필터로 합성 곱 연산을 두 번 수행할 경우, 성능은 비슷하지만 연산량이 더 적다는 이점이 있다.

2. LMCS 기술 개요

VVC의 인-루프 필터 기술 중 하나인 LMCS의 구성 요소는 휘도 신호 매핑과 색차 성분 잔차 신호 스케일링으로 구성되어 있다. 휘도 신호 매핑은 구간 선형 모델을 통해 휘도 신호의 기존 동적 범위를 부호화 성능이 향상할 수 있는 효율적인 휘도 신호의 동적 범위로 휘도 값을 매핑하는 역할을 수행한다. 색차 성분 잔차 신호 스케일링은 복원된 주변 VPDU (Virtual Pipeline Data Unit) 샘플 라인의 평균 휘도 값에 따라 색차 성분의 잔차 값을 스케일링하는 역할을 수행한다.

그림 1은 LMCS 기술 내 기존 휘도 신호 매핑 방법 블록도이다. LMCS 기술 내 기존 휘도 신호 매핑 방법은 그림 1과 같이 휘도 코드 워드를 초기화하여 할당한 후, 휘도 성분 프레임의 지역적 공간 분산을 계산하여 휘도 코드 워드를 재할당한다. 그리고 휘도 코드 워드를 사용하여 순방향 휘도 신호 매핑 함수를 구성해 휘도 신호를 효율적인 동적 범위로 매핑시킨다. 휘도 신호 매핑 함수는 순방향 매핑과 역방향 매핑을 수행해 휘도 신호를 각각 효율적인 동적 영역과 기존 동적 영역으로 매핑시키는 역할을 한다. 순방향 매핑 함수는 휘도 신호의 동적 범위 각 구간에 대한 선형 모델을 통해 코드 워드를 재분배하는 함수로, 다음 식과 같이 휘도 신호 Y를 휘도 신호 Y'으로 재매핑한다. 재매핑에 사용되는 구간 선형 함수는 수식 (1)과 같다.

여기서 i는 휘도 신호 샘플을 포함하는 동적 영역 구간을

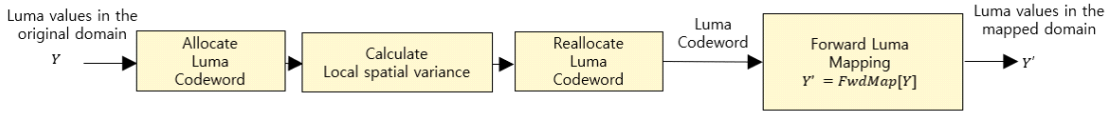


그림 1. 기존 휘도 신호 매핑 방법
Fig. 1. The existing luma signal mapping method block diagram

$$Y' = \frac{MappedPivot[i+1] - MappedPivot[i]}{InputPivot[i+1] - InputPivot[i]} \times (Y - InputPivot[i]) + MappedPivot[i] \quad (1)$$

의미하고, Y는 기존 휘도 신호를 의미한다. $InputPivot[i]$는 i 값과 초기화된 코드 워드를 곱한 값을 의미하고, $MappedPivot[i]$는 i 번째 영역의 재할당된 코드 워드 누적 값을 의미한다. 초기화된 코드 워드는 비트 심도가 10비트일 때, [0,1023] 동적 범위를 유효한 구간으로 분할하여 구간 당 코드 워드가 동일하게 할당되는 방식으로 구해진다. 역방향 매핑 함수는 매핑된 동적 영역에서 기존 동적 영역으로 휘도 신호를 매핑하는 함수로, 순방향 매핑 함수를 통해서 구할 수 있다. 순방향 매핑 함수와 역방향 매핑 함수는 위 수식을 통해 계산하거나 LUT (Look Up Table)로 구현할 수 있다.

색차 잔차 신호 스케일링은 휘도 신호와 대응하는 색차 잔차 신호 사이의 상관관계에 따라 색차 잔차 신호를 보정한다. 색차 잔차 신호 순방향 스케일링은 휘도 신호를 이용하여 유도된 색차 스케일 계수와 색차 잔차 신호를 곱해 스케일링된 색차 잔차 신호 $C_{ResScale}$을 구하는 기술이다. 색차 잔차 신호 역방향 스케일링은 전송받은 색차 잔차 신호와 색차 스케일 계수를 곱해 색차 잔차 신호 C_{Res}를 구하는 과정이다. $C_{ResScale}$과 C_{Res}를 구하는 식은 각각 다음 식 (2)와 (3)과 같이 구할 수 있다.

$$C_{ResScale} = C_{Res} \times C_{Scale} = \frac{C_{Res}}{C_{ScaleInv}} \tag{2}$$

$$C_{Res} = \frac{C_{ResScale}}{C_{Scale}} = C_{ResScale} \times C_{ScaleInv} \tag{3}$$

역방향 스케일 계수 $C_{ScaleInv}$는 휘도 신호 평균값의 역방향 매핑을 수행한 후, 동적 범위에서 예측 블록의 평균값과 deltaCRS[16]에 따라 결정된다. deltaCRS는 LMCS APS (Adaptation Parameter Set)에서 전송된 색차 보정을 위한 색차 신호 스케일링 오프셋이다. 역방향 스케일 계수 $C_{ScaleInv}$는 다음 식 (4)와 같이 구할 수 있다.

휘도 신호의 코드 워드는 영상의 통계적 특성을 반영하여 코드 워드를 재할당한다. 휘도 신호의 코드 워드 재분배 과정은 다음과 같다. 10비트 영상을 기준으로 하여 [0, 1023] 휘도 신호의 범위를 유효한 구간으로 나눠 동일한 코드 워드인 $binCW[i]$로 초기화한다. $binCW[i]$는 식 (5)와 같이 구할 수 있다. $totalCW$는 전체 코드 워드 수를 의미하고, $startIdx$와 $endIdx$는 각각 유효한 구간의 첫 번째 인덱스와 마지막 인덱스를 의미한다.

$$binCW[i] = round(\frac{totalCW}{endIdx - startIdx + 1}) \tag{5}$$

원본 영상에서 각 휘도 픽셀을 중심으로 $WinSize \times WinSize$ 크기의 영역에 대한 지역적 공간 분산 $pxlVar$를 구한다. 입력 영상의 크기에 따른 $WinSize$는 다음 식 (6)과 같이 구할 수 있다.

$$WinSize = Floor(\frac{\min(width, height)}{240} \times 2 + 1) \tag{6}$$

휘도 신호 범위의 16개 구간 중 i 번째 구간에 해당하는 화소의 지역적 공간 분산의 로그를 취한 합의 평균인 $binVar[i]$는 식 (7)과 같이 구할 수 있다.

$$binVar[i] = \frac{\sum_{bin} log_{10}(pxlVar + 1.0)}{binCnt[i]} \tag{7}$$

$binCnt[i]$는 i번째 구간에 해당하는 픽셀의 수를 의미한다. $binVar[i]$를 정규화한 $normVar[i]$는 식 (8)과 같이 구할 수 있다. $meanVar$는 전체 평균 분산을 의미한다.

$$normVar[i] = \frac{binVar[i]}{meanVar} \tag{8}$$

휘도 코드 워드는 $normVar[i]$에 따라 재할당한다. 복잡한 영상에 비해 평탄한 영상일 때 변화에 더 민감하게 반응한다는 인지 시각적 관점에 따라 휘도 성분의 지역적 공간

$$C_{ScaleInv}[i] = \frac{InputPivot[i+1] - InputPivot[i]}{(MappedPivot[i+1] - MappedPivot[i]) + deltaCRS[i]} \tag{4}$$

$$\begin{aligned} &\text{if } \ normVar[i] < 1.0, \qquad binCW[i] = \begin{cases} binCW[i] + \ delta1[i], & 0.8 \le normVar[i] < 0.9 \\ binCW[i] + \ delta2[i], & normVar[i] < 0.8 \end{cases} \\ &else \text{ if } \ normVar[i] > 1.0, \ binCW[i] = \begin{cases} binCW[i] - \ delta1[i], & 1.1 < normVar[i] \le 1.2 \\ binCW[i] - \ delta2[i], & normVar[i] > 1.2 \end{cases} \end{aligned} \tag{9}$$

분산이 작을수록 휘도 코드 워드를 많이 할당하고, 클수록 휘도 코드 워드를 적게 할당한다. 휘도 성분 코드 워드 $binCW[i]$는 다음 식 (9)와 같이 구할 수 있다.

$delta1$, $delta2$는 각 구간에 해당하는 원본 이미지 영역의 히스토그램 값에 비례하여 결정된다. $delta1[i]$와 $delta2[i]$는 각각 i 번째 구간의 상대 도수에 10과 20을 곱하여 반올림하여 구한다.

다음 장에서는 제안하는 방법에 대하여 보다 자세히 설명한다.

Ⅲ. 제안하는 어텐션 맵을 이용한 휘도 매핑 함수 생성 방법

LMCS의 휘도 신호 매핑 함수는 코드 워드를 사용하여 구간 선형 모델을 생성한다. 코드 워드를 구하는 과정에서 기존 LMCS는 코드 워드를 할당할 때, 영상의 복잡한 영역과 평탄한 영역을 구분하여 코드 워드를 다르게 할당하는데, 이 때 지역적 공간 분산만을 사용하여 판단한다. 하지만 노이즈한 정보가 많은 영역일 때, 지역적 공간 분산 정보만으로 판단하게 되면 코드 워드가 적게 할당된다. 그로 인해 휘도 신호가 비효율적인 동적 범위로 구성될 수 있다는 문제점이 있다. 이 문제를 해결하기 위해, 본 논문에서는 휘도 신호의 코드 워드를 구할 때, 합성 곱 신경망의 어텐션 맵을 지역적 공간 분산에 가중치로 곱해 중요한 특징이 있는 영역과 노이즈한 영역을 분별하여 적응적으로 휘도 신호를 매핑할 수 있게 하는 휘도 신호 매핑 방법을 제안한다. 제안하는 방법은 SDR 입력 영상을 기준으로 휘도 신호 매핑 함수를 생성한다.

그림 2는 제안하는 휘도 신호 매핑 방법 블록도이다. 노란색 블록은 기존 휘도 신호 매핑 과정이고, 녹색 블록은 제안하는 방법을 추가한 블록이다. 그림 2와 같이 휘도 코드 워드를 초기화하여 할당한 후, 휘도 성분 영상을 특징 추출 네트워크에 입력으로 하여 출력된 특징 맵들을 채널 간 제곱 평균 제곱근 (Root Mean Square; RMS) 풀링 과정을 수행한다. 계산된 어텐션 맵을 구하여 지역적 공간 분산에 어텐션 맵을 가중치로 곱한 값을 사용하여 휘도 코드 워드를 재할당한다. 그 후, 휘도 코드 워드를 이용하여 순방향 휘도 신호 매핑 함수를 구성하여 휘도 신호를 효율적인 동적 범위로 재매핑한다.

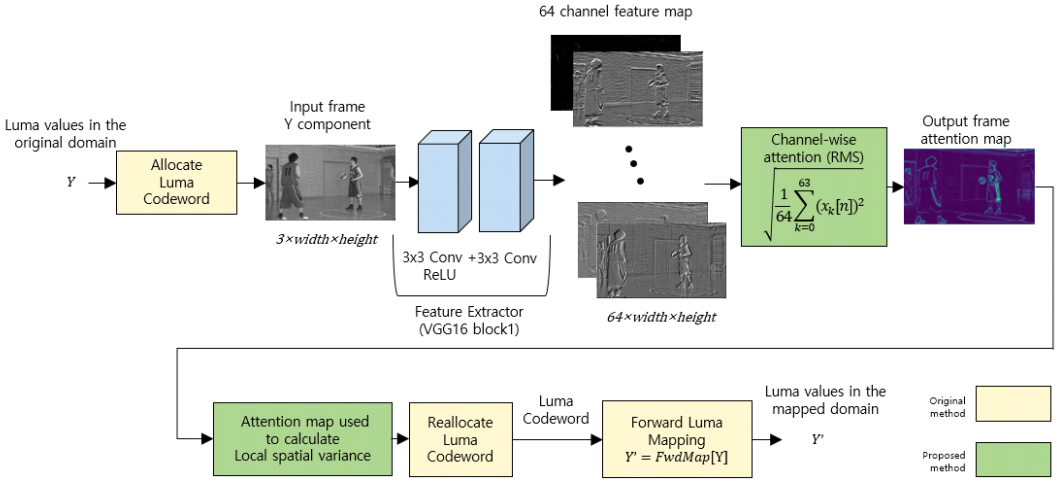


그림 2. 제안하는 휘도 신호 매핑 방법
Fig. 2. The proposed luma signal mapping method block diagram

어텐션 맵 생성을 위한 네트워크는 이미지넷 데이터셋[17]을 이용해 사전 학습된 VGG-16 모델의 일부를 사용한다. 이미지의 공간적인 정보를 잃지 않고 계산 복잡도를 최소화하기 위해 VGG-16 네트워크에서 2개의 합성 곱 층과 합성 곱 층 사이에 포함된 ReLU (Rectified Linear Unit) 활성화 함수로 이루어진 첫 번째 블록만 사용한다. 첫 번째 블록은 윤곽선 검출기와 같이 이미지의 저수준 특징을 추출하는 블록이다. 기존 사전 학습된 VGG-16 네트워크 같은 경우, 사전 학습 시 RGB 채널을 입력으로 하고, [0,1]로 정규화시켜 학습되었기 때문에, 어텐션 맵을 구하는 네트워크의 입력은 1채널의 원본 영상 휘도 신호를 3채널로 복사하고, 10비트 영상 기준 [0,1023] 범위를 [0,1] 범위로 정규화시킨다. 네트워크 출력으로 나온 64개 특징 맵에서 동일한 영역에 위치한 픽셀들의 제곱 평균 제곱근을 구해 하나의 어텐션 맵을 생성한다. 제곱 평균 제곱근은 변화하는 값의 크기에 대한 통계적 척도로, 제곱한 값들의 평균값의 양의 제곱근을 의미한다. 어텐션 맵 $attention map$은 다음 식 (10)과 같이 구할 수 있다.

$$\text{attention map}[n] = \sqrt{\frac{1}{64}\sum_{k=0}^{63}(x_k[n])^2} \quad (0 \le n < width \times height) \tag{10}$$

영상에서 i 번째 위치에 있는 휘도 픽셀을 중심으로 $WinSize \times Winsize$ 크기의 영역에 대한 분산, $pxl\,Var$를 구한 후, $pxl\,Var$와 같은 위치에 있는 $attention map$을 곱한 $pxl\,Var_f$를 구한다. 어텐션 맵은 영상의 주요한 특징 정보를 가지고 있어 지역적 공간 분산 값에 어텐션 맵을 곱하여 가중치 역할을 한다. 어텐션 맵은 주요한 특징이 있는 영역의 값은 크고, 노이즈한 영역의 값은 작다는 특징이 있다. 이에 따라 지역적 공간 분산 값에 어텐션 맵을 가중치로 곱하여 평탄한 영역의 노이즈한 부분이 있는 경우 지역적 공간 분산을 작게 만들어 코드 워드가 많이 할당되도록 한다. 이는 노이즈한 영역의 높은 지역적 공간 분산으로 인해 노이즈가 포함된 평탄한 영역에 코드 워드가 적게 할당되는 문제를 해결할 수 있다. $Pxl\,Var_f$를 사용해 16개 구간 중 i번째 구간에 해당하는 화소의 지역적 공간 분산의 평균인 $bin\,Var_f[i]$를 구한다. $bin\,Var_f[i]$는 다음 식 (11)과 같이 구할 수 있다.

$$binVar_f[i] = \frac{\sum_{bin} log_{10}(PxlVar_f + 1.0)}{binCnt[i]} \tag{11}$$

기존 논문과 같은 방식으로 16개 각 구간의 $bin\,Var_f[i]$를 정규화시킨 $norm\,Var_f[i]$를 구한 후, $norm\,Var_f[i]$에 따라 코드 워드를 재할당한다. 재할당된 휘도 코드 워드를 사용하여 휘도 신호를 재매핑하고, 휘도 신호에 대응하는 색차 신호 사이의 상관관계에 따라 색차 신호 값을 스케일링한다.

Ⅳ. 실험 환경 및 결과

1. 실험 환경

본 논문에서 제안하는 방법의 성능을 평가하기 위하여 VVC CTC (Common Test Condition)[18]의 AI (All Intra) 조건에서 실험을 진행하였다. 양자화 파라미터 (Quantization Parameter, QP)는 22, 27, 32, 37을 사용하였고, VVC 표준 실험 영상 중 A1, A2, B, C, D 클래스를 사용하여 전체 프레임에 대하여 실험을 진행하였다. 실험은 VVC 참조 소프트웨어 VTM-12.0의 결과를 기준으로 PSNR (Peak Signal-to-Noise Ratio) 기반 BD-rate[19]로 비교하였다.

VTM에서 합성 곱 신경망을 사용하기 위해 ONNX Runtime(Open Neural Network eXchange Runtime)[20]을 사용하였다. ONNX는 신경망 프레임워크 간의 상호 연동을 위한 딥러닝 모델의 개방형 표준 포맷이다. 실험에서 사용한 ONNX는 1.4.0 버전을 사용하였고, ONNX opset은 안정화된 9 버전을 사용하였다.

2. BD-rate를 이용한 객관적 화질 평가 및 부/복호화 시간 비교

표 1은 VTM-12.0 대비 제안하는 방법의 A1, A2, B, C, D 클래스 PSNR 기반 BD-rate 성능 및 부/복호화 시간을 나타낸다. 제안하는 방법은 VTM-12.0 대비 휘도 성분의 평균 BD-rate가 A1 클래스의 경우 -0.25%, A2 클래스의 경우

표 1. 기존 VTM-12.0 대비 제안하는 방법을 사용한 경우의 A1, A2, B, C, D 클래스의 BD-rate (PSNR) 성능 및 부/복호화 시간
Table 1. BD-rate (PSNR) performance and coding time of A1, A2, B, C, D classes of proposed method compared to VTM-12.0

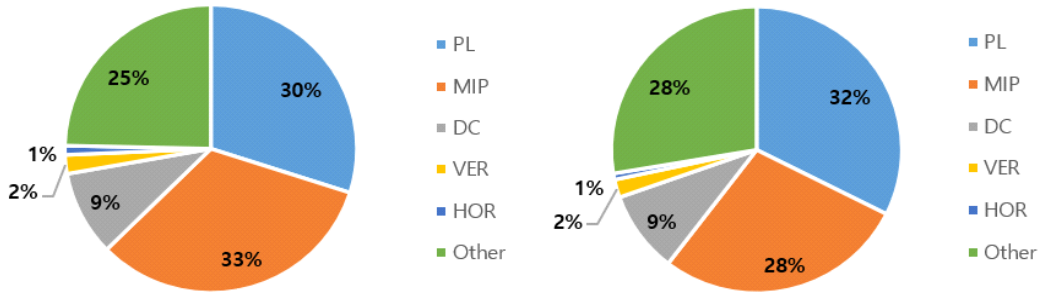
Class	Sequence	Proposed method				
		BD-rate(PSNR)			Enc Time	Dec Time
		Y	U	V		
Class A1 (3840×2160)	Tango2	-0.55%	1.42%	1.43%	102%	102%
	FoodMarket4	-0.20%	0.77%	0.84%	107%	102%
	Campfire	0.00%	0.00%	0.00%	103%	102%
Class A2 (3840×2160)	CatRobot	-0.01%	-0.03%	0.03%	103%	101%
	DaylightRoad2	0.02%	0.03%	-0.07%	102%	101%
	ParkRunning3	-0.20%	0.24%	0.26%	102%	103%
Class B (1920×1080)	MarketPlace	0.01%	0.51%	0.27%	103%	100%
	RitualDance	-0.11%	-0.21%	-0.26%	102%	100%
	Cactus	0.00%	0.00%	0.00%	102%	100%
	BasketballDrive	-0.17%	0.93%	1.07%	104%	102%
	BQTerrace	0.01%	0.06%	0.01%	103%	101%
Class C (832×480)	BasketballDrill	-0.12%	1.31%	1.31%	103%	102%
	BQMall	0.01%	-0.02%	-0.06%	101%	101%
	PartyScene	0.00%	0.06%	0.17%	101%	100%
	RaceHorsesC	-0.10%	0.39%	0.30%	102%	100%
Class D (416×240)	BasketballPass	0.05%	-0.03%	0.27%	99%	99%
	BQSquare	0.02%	-0.08%	0.19%	101%	100%
	BlowingBubbles	0.00%	-0.15%	-0.13%	99%	98%
	RaceHorses	-0.07%	0.40%	0.30%	100%	102%
Class A1 Average		-0.25%	0.73%	0.76%	105%	102%
Class A2 Average		-0.06%	0.08%	0.08%	102%	102%
Class B Average		-0.05%	0.26%	0.22%	103%	101%
Class C Average		-0.05%	0.44%	0.43%	102%	101%
Class D Average		0.00%	0.04%	0.16%	100%	100%
All Class Average		-0.07%	0.29%	0.31%	102%	101%

-0.06%, B 클래스의 경우 -0.05%, C 클래스의 경우 -0.05%, D 클래스의 경우 0.00% 성능을 보이고 거의 동일한 부/복호화 시간을 갖는다. 실험 영상의 크기에 따른 휘도 성분의 평균 BD-rate 결과를 비교해 보았을 때, 영상의 크기가 커짐에 따라 성능 향상이 증가함을 확인할 수 있다.

3. 화면 내 예측 모드 발생 비율 및 블록 분할 결과 비교

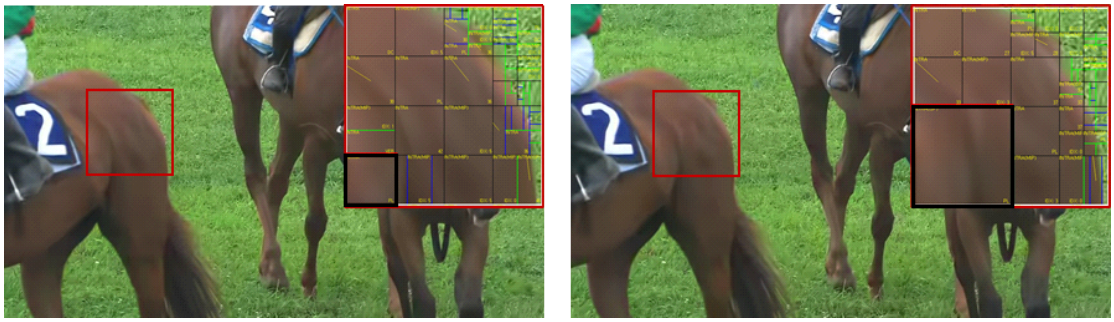
그림 3은 휘도 성분의 화면 내 예측 모드를 비교한 그림으로, AI 환경에서 QP=37로 부호화된 RaceHorsesC 시퀀스의 13번째 프레임에서 비교하였다. VVC에서 CU 크기는 최소 4×4부터 최대 64×64 크기로 다양하게 이루어져 있기 때문에 정확한 비교를 위해 영상의 전체 영역에 대해 4×4 블록 단위를 기준으로 화면 내 예측 모드 비율을 정규화시켜 비교하였다. 제안하는 방법을 수행한 경우, 32.40%로 기존 VTM-12.0의 플라나 모드의 비율이 29.94%인 결과와 비교해보았을 때, 2.46%의 플라나 모드 비율이 증가한 것을 확인할 수 있다.

그림 4-6은 블록 크기를 비교하기 위한 그림으로, Bitstream InSights[21]을 사용하여 CTU의 분할 구조를 시각화하여 표시하였다. 빨간색으로 표시된 블록은 CTU를 의미하고, 검정색으로 표시된 블록은 CU를 의미한다. 그림 4는 그림 3과 같은 프레임으로, 말의 평탄한 몸통 부분에 대해 기존 방법 대비 블록 크기가 64×64로 크게 분할이 된 것을 확인할 수 있다. 또한 64×64 블록의 화면 내 예측 모드가 플라나 모드인 것을 확인할 수 있다. 그림 3과 그림



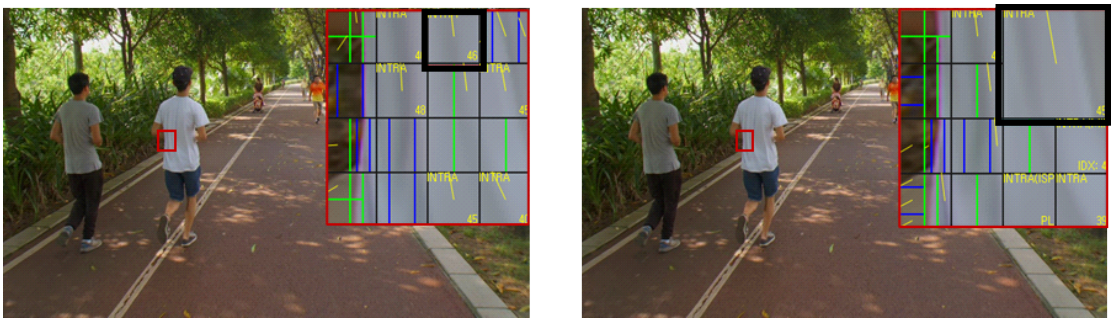
(a) Existing method of intra luma mode (b) Proposed method of intra luma mode

그림 3. AI 환경, QP=37로 부호화된 RaceHorsesC 시퀀스의 13번째 프레임에서 휘도 성분의 화면 내 예측 모드 비교
Fig. 3. Intra luma prediction mode comparison for RaceHorsesC with AI configuration, where the 13th frame is shown (QP=37)



(a) (b)

그림 4. AI 환경, QP=37로 부호화된 RaceHorsesC 시퀀스의 13번째 프레임에서 블록 크기 비교 (a) 기존 방법; (b) 제안하는 방법
Fig. 4. Block size comparison for RaceHorsesC with AI configuration, where the 13th frame is shown (QP=37) (a) Existing method; (b) Proposed method

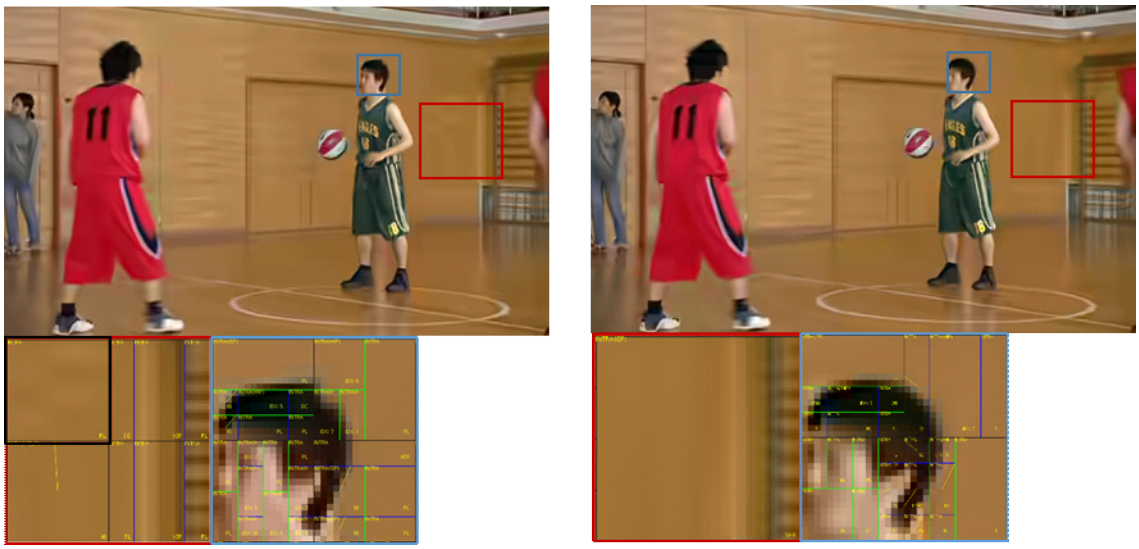


(a) (b)

그림 5. AI 환경, QP=32로 부호화된 ParkRunning3 시퀀스의 0번째 프레임에서 블록 크기 비교 (a) 기존 방법; (b) 제안하는 방법
Fig. 5. Block size comparison for ParkRunning3 with AI configuration, where the 0th frame is shown (QP=32) (a) Existing method; (b) Proposed method

4를 통해 기존 방법 대비 제안하는 매핑 방법이 노이즈가 있는 평탄한 영역에 대해 노이즈를 감소시켜줌으로써 플라나 모드가 더 많이 예측된 것으로 보인다.

그림 5는 AI 환경에서 QP=32로 부호화된 ParkRunning3 시퀀스의 0번째 프레임이고, 그림 6은 AI 환경에서 QP=37로 부호화된 BasketballPass 시퀀스의 0번째 프레임이다.



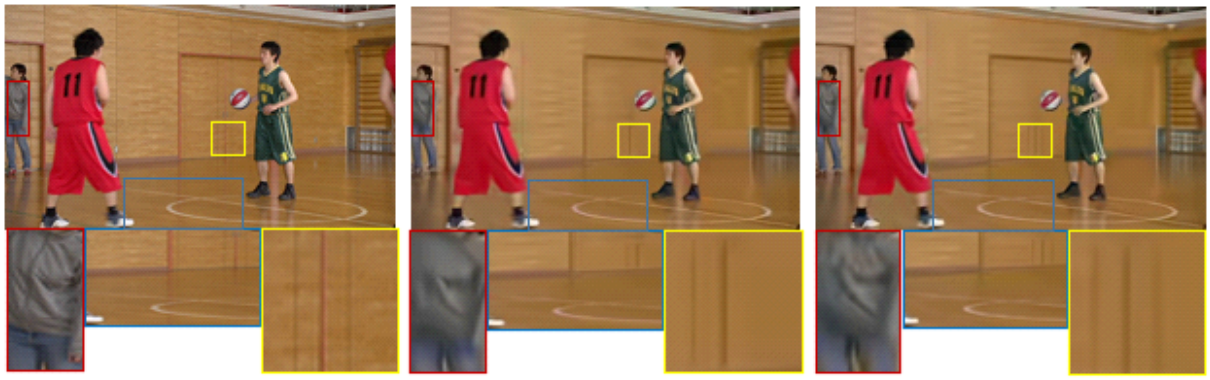
(a) (b)

그림 6. AI 환경, QP=32로 부호화된 ParkRunning3 시퀀스의 0번째 프레임에서 블록 크기 비교 (a) 기존 방법; (b) 제안하는 방법
Fig. 6. Block size comparison for ParkRunning3 with AI configuration, where the 0th frame is shown (QP=32) (a) Existing method; (b) Proposed method

그림 4-6은 평탄한 부분에 대해 기존 방법 대비 블록 크기가 64×64로 블록이 크게 분할된 것을 확인할 수 있다. 이를 통해 노이즈가 있는 영역에 대한 지역적 공간 분산을 계산할 때, 어텐션 맵을 가중치로 제공하여 노이즈가 감소해 평탄한 부분을 강조시킴으로써 블록이 더 크게 분할된 것으로 보인다.

4. 원본 영상 대비 기존 방법 (VTM-12.0)과 제안하는 방법의 복원 영상 비교

그림 7-9는 원본 영상과 복원 영상을 비교하기 위한 그림이다. 그림 7-9는 AI 환경으로 부호화된 영상이다. (a), (b), (c)는 각각 원본 프레임, 기존 방법으로 부호화된 프레임,



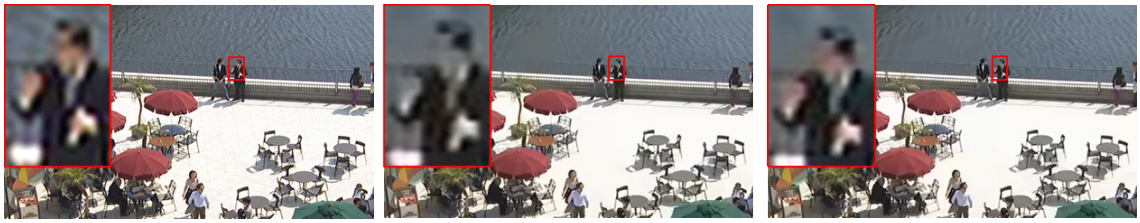
(a) (b) (c)

그림 7. AI 환경, QP=37로 부호화된 BasketballPass 시퀀스의 0번째 프레임에서 원본 영상과 복원 영상 비교 (a) 원본; (b) 기존 방법; (c) 제안하는 방법
Fig. 7. Comparison between original frame and reconstruction frame for BasketballPass with AI configuration, where the 0th frame is shown (QP=37), (a) Original; (b) Existing method; (c) Proposed method

제안하는 방법으로 부호화된 프레임에 대한 그림이다. 그림 7은 QP=37로 부호화된 BasketballPass 시퀀스이고, 0번째 프레임이 사용되었다. 기존 방법 대비 제안하는 방법의 프레임에서 옷의 주름, 농구장 라인, 벽의 선 등의 윤곽선 부분이 원본과 유사하게 표현되었음을 확인할 수 있다. 그림 8에서 비교를 위한 그림은 QP 32로 부호화된 BQSquare 시퀀스이고, 40번째 프레임이 사용되었다. 기존 방법 대비 제안하는 방법으로 부호화된 프레임의 사람 얼굴 색상이 원본과 유사하게 표현되었음을 확인할 수 있다. 이는 인코더에서 제안하는 휘도 순방향 매핑 과정을 수행한 후 이에 대응하는 색차 잔차 신호 사이의 상관관계에 따라 색차 잔차 신호가 보정됨에 따라 오차가 감소하여 결과적으로 원본과 유사하게 얼굴 색상이 표현된 것으로 보인다. 그림 9에서 비교를 위한 그림은 기존 방법 대비 제안하는 방법의 동일 화질에서 비트율 향상 정도가 큰 QP=37로 부호화된 RaceHorsesC 시퀀스의 20번째 프레임이 사용되었다. 기존 방법으로 부호화된 영상같은 경우 주름 부분이 뭉개져 표현된 것에 비해 제안하는 방법으로 부호화된 영상은 사람 무릎 주위의 옷 주름이 원본 영상과 비슷하게 표현되었음을 확인할 수 있다. 말의 다리 부분도 기존 방법으로 부호화된 영상은 근육의 표현이 세밀하게 되지 않은 반면, 제안하는 방법으로 부호화된 영상에서는 말의 근육이 원본 영상과 비슷하게 표현되었음을 확인할 수 있다. 또한 꼬리 부분의 붉은 색상이 제안하는 방법을 수행하였을 때 원본 영상과 더 유사하게 색상 표현이 되었음을 확인할 수 있다. 그림 7-9를 통해 윤곽선 부분과 같이 영상에서 중요한 특징을 가지는 부분의 지역적 공간 분산에 어텐션 맵이 가중치로 곱해져 구간 선형 모델의 기울기가 낮아짐에 따라 결과적으로 에러량이 줄어들어 윤곽선이 강조되는 효과를 얻은 것으로 보인다.

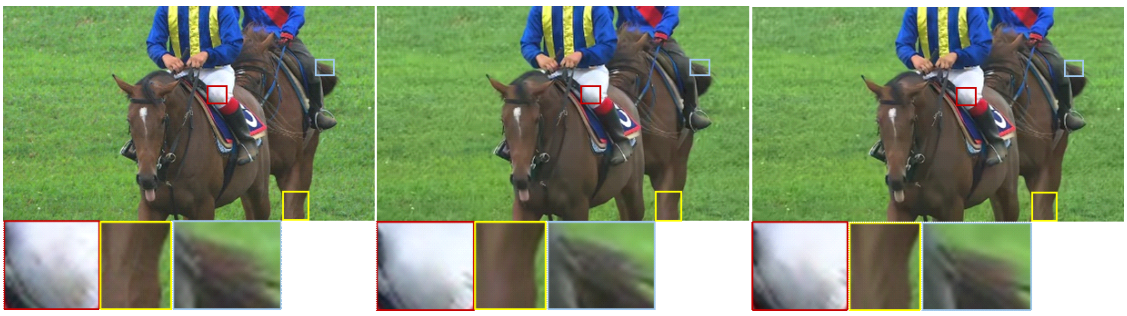
V. 결 론

본 논문에서는 VVC의 LMCS 성능 개선을 위하여 합성



(a) (b) (c)

그림 8. AI 환경, QP=32로 부호화된 BQSquare 시퀀스의 40번째 프레임에서 원본 영상과 복원 영상 비교 (a) 원본; (b) 기존 방법; (c) 제안하는 방법
Fig. 8. Comparison between original frame and reconstruction frame for BQSquare with AI configuration, where the 40th frame is shown (QP=32), (a) Original; (b) Existing method; (c) Proposed method



(a) (b) (c)

그림 9. AI 환경, QP=37로 부호화된 RaceHorsesC 시퀀스의 20번째 프레임에서 원본 영상과 복원 영상 비교 (a) 원본; (b) 기존 방법; (c) 제안하는 방법
Fig. 9. Comparison between original frame and reconstruction frame for RaceHorsesC with AI configuration, where the 20th frame is shown (QP=37), (a) Original; (b) Existing method; (c) Proposed method

곱 신경망의 어텐션 맵을 이용한 휘도 신호 매핑 함수 생성 방법을 제안하였다. 제안하는 방법은 LMCS에서 휘도 성분 코드 워드를 계산할 때, 지역적 공간 분산이 큰 영역에 대해 윤곽선과 같이 중요한 특징 영역인지 노이즈한 영역인지 분별할 수 있도록 합성 곱 신경망의 어텐션 맵을 지역적 공간 분산에 가중치로 사용하여 LMCS 파라미터를 계산하는 방법이다. 본 논문에서 제안하는 방법은 VTM12.0 대비 평균 휘도 성분이 -0.07%의 BD-rate 성능 향상을 보였다. 특히 VVC CTC AI 조건에서 Class A1 영상을 부호화한 결과, 휘도 성분의 BD-rate 성능이 -0.25% 향상이 있었다. 또한 제안하는 방법을 수행할 경우, 기존 방법 대비 평평한 영역에 대해 블록 크기가 크게 분할되고, 플라나 예측 모드가 증가하는 것을 확인하였다. 기존 딥 러닝 기반의 비디오 압축 연구는 부/복호화 시간 복잡도가 증가하는 문제로 실제 적용하기 어렵다는 문제점이 있었지만 제안하는 방법은 부/복호화 시간이 거의 비슷함을 확인할 수 있었다. 하지만 색차 성분의 BD-rate 성능은 VTM12.0 대비 U, V 각각 0.29%, 0.21%의 낮은 성능을 보였다. 이를 개선하기 위해서는 제안하는 방법을 수행함에 있어서 색차 성분의 부호화 효율을 향상시키는 방법에 대한 추가적인 연구가 필요하다.

참 고 문 헌 (References)

[1] G. Sullivan, J. Ohm, W. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” Institute of Electrical and Electronics Engineers (IEEE) Transactions on circuits and systems for video technology, Vol.22, No.12, pp. 1649-1668, Dec. 2012.

[2] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, “Versatile Video Coding (Draft 10),” JVET-S2001, Jul. 2020.

[3] J. Lee, J. Park, H. Choi, J. Byeon, and D. Sim, “Overview of VVC,” Broadcasting and Media Magazine, Vol.24, No.4, pp. 10-25, Oct. 2019.

[4] D. Park, Y. Yun, and J. Kim, "VVC의 In-Loop Filter 기술," Broadcasting and Media Magazine, Vol.24, No.4, pp. 87-101, Oct. 2019.

[5] T. Lu, F. Pu, P. Yin, S. McCarthy, W. Husak, T. Chen, E. Francois, C. Chevance, F. Hiron, J. Chen, R. Liao, Y. Ye, and J. Luo, "Luma Mapping with Chroma Scaling in Versatile Video Coding," Data Compression Conference (DCC), Snowbird, UT, USA, pp. 193-202, 2020.

[6] VTM, https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM

[7] J. Im, U. Im, and D. Sim, "HDR/WCG 영상 압축을 위한 표준 기술 동향," Broadcasting and Media Magazine, Vol.21, No.1, pp. 59-69, 2016.

[8] Rec. ITU-R BT.2100-2, “Image parameter values for high dynamic range television for use in production and international programme exchange”

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neuralnetworks,” In Neural Information Processing Systems (NIPS), 2012.

[10] L. Zhou, X. Song, J. Yao, L. Wang, and F. Chen, “Convolution Neural Network Filter (CNNF) for Intra Frame,” JVET-I0022, Joint Video Exploration Team of ISO/IEC and ITU-T, Gwangju, Korea, Jan 2018.

[11] J. Kang, S. Kim, and K. Lee, “Multi-modal/multi-scale convolutional neural network based in-loop filter design for next generation video codec,” Institute of Electrical and Electronics Engineers (IEEE) International Conference on Image Processing (ICIP), pp. 26-30, 2017.

[12] F. Zhang, C. Feng and D. R. Bull, "Enhancing VVC Through Cnn-Based Post-Processing," Institute of Electrical and Electronics Engineers (IEEE) International Conference on Multimedia and Expo (ICME), pp. 1-6, 2020.

[13] H. Moon, and J. Kim, "CNN Based In-loop Filter in Versatile Video Coding (VVC)," Proceedings of the Korean Society of Broadcast Engineers Conference, The Korean Institute of Broadcast and Media Engineers, pp. 270-271, 2018.

[14] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” In ICLR, 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks." Institute of Electrical and Electronics Engineers (IEEE) transactions on pattern analysis and machine intelligence Vol.39, No.6, pp. 1137-1149, 2017.

[16] E. François, F. Galpin, K. Naser, and P. de Lagrange, “AHG7/AHG15: Signalling of corrective values for chroma residual scaling,” JVET-P0371, Oct. 2019.

[17] J. Deng, W. Dong, R. Socher, L. Li, K. Ki, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database." 2009 Institute of Electrical and Electronics Engineers (IEEE) conference on computer vision and pattern recognition, pp. 248-255, 2009.

[18] F. Bossen, J. Boyce, K. Suehring, X. Li, and V. Seregin, “JVET common test conditions and software reference configurations for SDR video,” JVET-N1010, Mar. 2019.

[19] G. Bjøntegaard, “Calculation of average PSNR differences between RDcurves,” Tech. Rep. VCEGM33, Video Coding Experts Group (VCEG), 2001.

[20] ONNX Runtime, http://github.com/microsoft/onnxruntime, 2019.

[21] Bitstream InSights - VTM, http://www.digitalinsights.co.kr/products/

저 자 소 개

권 나 성

- 2017년 3월 ~ 현재 : 광운대학교 소프트웨어학부 학사과정
- ORCID : https://orcid.org/0000-0002-1796-0564
- 주관심분야 : 영상압축, 컴퓨터비전

이 종 석

- 2016년 2월 : 광운대학교 전자공학과 학사
- 2018년 2월 : 광운대학교 전자공학과 석사
- 2018년 3월 ~ 현재 : 광운대학교 컴퓨터공학과 박사과정
- 2020년 2월 ~ 현재 : 디지털인사이트 선임연구원
- ORCID : https://orcid.org/0000-0001-8045-0244
- 주관심분야 : 영상압축, 스파이킹 심층 신경망, 컴퓨터비전, 고해상도 위성영상 처리

변 주 형

- 2019년 2월 : 광운대학교 컴퓨터공학과 학사
- 2021년 2월 : 광운대학교 컴퓨터공학과 석사
- 2021년 3월 ~ 현재 : 광운대학교 컴퓨터공학과 박사과정
- ORCID : https://orcid.org/0000-0002-6165-9189
- 주관심분야 : 3D 데이터 압축, 영상압축, 컴퓨터비전

심 동 규

- 1993년 2월 : 서강대학교 전자공학과 공학사
- 1995년 2월 : 서강대학교 전자공학과 공학석사
- 1999년 2월 : 서강대학교 전자공학과 공학박사
- 1999년 3월 ~ 2000년 8월 : 현대전자 선임연구원
- 2000년 9월 ~ 2002년 3월 : 바로비젼 선임연구원
- 2002년 4월 ~ 2005년 2월 : University of Washington Senior research engineer
- 2005년 3월 ~ 현재 : 광운대학교 컴퓨터공학과 교수
- ORCID : https://orcid.org/0000-0002-2794-9932
- 주관심분야 : 영상신호처리, 영상압축, 컴퓨터비전