레터논문 (Letter Paper)
방송공학회논문지 제26권 제5호, 2021년 9월 (JBE Vol.26, No.5, September 2021)
https://doi.org/10.5909/JBE.2021.26.5.652
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

고속 VVC 부호화를 위한 신경망 기반 움직임 벡터 해상도 결정 알고리즘

백 한 결[a)], 박 상 효[a)‡]

Motion Vector Resolution Decision Algorithm based on Neural Network for Fast VVC Encoding

Han-gyul Baek[a)] and Sang-hyo Park[a)‡]

요 약

Versatile Video Coding(VVC)의 압축 효율을 끌어올리기 위하여 다양한 화면 간 예측(inter prediction)기법 중 적응적 움직임 벡터 해상도(Adaptive motion vector resolution, 이하 AMVR)기술이 채택되어 왔다. 다만, AMVR을 적용하여 최적의 해상도를 결정하기 위해서는 매 부호화 유닛마다 다양한 테스트를 진행해야 하며, 이는 율-왜곡 비용의 계산 복잡도 증가를 야기한다. 따라서 VVC의 부호화 복잡도의 감소를 위해 효과적으로 최적의 AMVR 모드를 찾아야 한다. 본 논문에서는 보다 다양한 데이터셋 기반 하에 경량화된 신경망 기반의 AMVR 결정 알고리즘을 제안한다.

Abstract

Among various inter prediction techniques of Versatile Video Coding (VVC), adaptive motion vector resolution (AMVR) technology has been adopted. However, for AMVR, various MVs should be tested per each coding unit, which needs a computation of rate-distortion cost and results in an increase in encoding complexity. Therefore, in order to reduce the encoding complexity of AMVR, it is necessary to effectively find an optimal AMVR mode. In this paper, we propose a lightweight neural network-based AMVR decision algorithm based on more diverse datasets.

Keyword : VVC, inter prediction, motion vector resolution, encoding complexity, Multi-layer perceptron

a) 경북대학교 컴퓨터학부(School of Computer Science and Engineering, Kyungpook National University)
‡ Corresponding Author : 박상효(Sang-hyo Park)
E-mail: s.park@knu.ac.kr
Tel: +82-53-950-6373
ORCID: https://orcid.org/0000-0002-7282-7686

※This study was supported in part by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394) and was supported in part by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2020R1I1A3072227)).
· Manuscript received August 24, 2021; Revised September 24, 2021; Accepted September 24, 2021.

Ⅰ. 서 론

Versatile Video Coding(VVC)[1]의 화면 간 예측(inter prediction)의 적응적 움직임 벡터 해상도(Adaptive motion vector resolution, 이하 AMVR)는 움직임이 큰 동영상에서 움직임 벡터 차이(Motion vector difference, 이하 MVD) 값이 크게 나타나면 비트 수의 증가가 초래되기 때문에 압축 효율의 저하의 원인이 되는 비효율적 MVD를 부호화하는 방법이다. 현재 VVC의 복잡도는 HEVC/H.265[2] 대비 3000%가 넘는 것으로 알려져 있다[3]. AMVR을 사용하여 비효율적인 MVD를 효과적으로 부호화하여 비트율을 줄이기 위해서는 매번 다양한 MVD에 대한 테스트를 진행하게 되는데 이는 율-왜곡 비용(rate-distortion cost)의 계산이 필수적으로 필요하다. 이러한 필수적인 계산으로 인해 부호화 복잡도의 증가는 필연적일 수밖에 없다. 따라서 AMVR의 부호화 복잡도를 줄이는 것이 하나의 중요한 연구라고 할 수 있다.

관련 연구 중 AMVR의 부호화 복잡도를 통계적인 기법으로 경량 하려는 연구가 있었다[4]. 또한, AMVR의 모드 결정을 위하여 Multi-layer perceptron(MLP)로 조기에 다른 정보를 가지고 판단할 수 있는 신경망 모델을 사용한 연구가 있다[5]. 그러나 신경망 모델의 성능과 일반화를 위한 고려가 부족하여, 다양한 영상에서도 효과적으로 최적의 모드를 찾을 수 있을지는 미비한 실정이다. 따라서 본 논문에서는 보다 다양한 데이터셋 기반하에 훨씬 경량화된 신경망 기반의 AMVR 결정 알고리즘을 제안한다.

Ⅱ. 제안기법

AMVR 모드는 최대 4가지가 나올 수 있다. 또한 일반적인 움직임 벡터 예측(motion vector prediction, MVP)인지 Affine을 위한 MVP 인지, 화면 내 블록 카피(intra block copy)인지에 따라 MV 해상도가 달라질 수 있다. 일반적인 MVP일 경우, MV의 해상도는 1/4, 1/2, 1(정수 기본단위), 4의 4가지 모드이며, Affine을 위한 MVP일 경우, MV의 해상도는 1/4, 1/16, 1(정수 기본단위)의 3가지이다. 기본값은 HEVC에서 흔히 쓰이는 1/4이다. 따라서 만약 AMVR을 쓰지 않기로 한다면(즉, AMVR off 모드), MV의 해상도는 1/4로 해석하지만, 그 외에는 AMVR을 위한 flag가 켜지면서 율-왜곡 비용 기준으로 최적의 해상도를 같이 전달해야 한다. 제안기법에서는 AMVR이 필요 없는 경우, 즉 AMVR off 모드와 AMVR이 필요한 경우의 2가지로 문제를 단순화하여 이진 분류 문제로 변경한다. 여기서는 AMVR off일 경우는 True, AMVR on일 경우에는 False로 분류하는 문제로 정의한다.

이진분류 문제를 해결하기 위한 부호화 문맥 정보를 추출하는 과정은 아래와 같다. 먼저 현재 부호화 유닛(coding unit, CU)의 부호화 과정 중에 추출할 수 있는 특징 벡터를 x라고 할 때, 다음에서 나오는 15개의 문맥 정보를 feature로 정의한다. $x0$과 $x1$은 CU의 너비와 높이를 각각 128로 나눈 값이며, $x2$는 현 프레임의 QP 값을 QP 최댓값인 63으로 나눈 값이다. $x3$과 $x4$는 Quadtree(QT)와 Multi-type tree(MTT)의 깊이를 각각 3과 4로 나눈 값이다. 이때, $x4$는 1보다 커지는 모든 경우를 1로 간주한다. $x5$, $x6$, $x7$, $x8$은 각각 부모 CU의 최종 AMVR 모드 4가지를 의미한다. 예컨대 부모 CU가 AMVR off를 최종으로 선택했다면, $x5$는 1을 갖고, $x6$부터 $x8$은 0을 갖는다. 이와 비슷하게 $x9$, $x10$, $x11$, $x12$, $x13$은 각각 부모 CU의 트리 구조를 의미하며, 차례대로 QT, binary tree(BT) 가로 분할, BT 세로 분할, ternary tree(TT) 가로 분할, TT 세로 분할을 의미한다. 만약 부모 CU가 QT라면 $x9$만 1을 갖고, $x10$부터 $x13$은 0을 갖는다. 만약 어떠한 분할도 아닌 CU일 경우에는 $x9$부터 $x13$까지는 모두 0을 갖는다. 마지막으로, 기존 연구[5]와 다른 새로운 input feature를 추가하였다. 새로 추가한 $x14$는 현재 CU의 다양한 모드를 율-왜곡 비용 기준으로 평가할 때, Merge 모드의 율-왜곡 비용이 AMVR off 상태의 기본적인 움직임 예측을 수행한 율-왜곡 비용보다 작을 경우 1을 갖도록 하고, 반대로 Merge 모드의 율-왜곡 비용이 더 적어서 나머지 AMVR 모드를 테스트하기 전까지는 최적의 모드라고 간주하고 있는 경우, $x14$는 0을 갖도록 한다.

대부분의 경우 AMVR off의 경우가 아닌 경우보다 훨씬 많이 나온다. 이러한 경우, 기계학습을 통해 학습을 진행하려 하여도 데이터셋에서의 클래스 간 불균형이 심하여, 기존 연구[5]에서도 모델의 학습 성능을 크게 저하하거나 Underfitting을 유발하기 쉽다. 따라서 본 연구에서는 클래스간 균형을 맞추기 위하여, feature 데이터셋을 생성 시,

AMVR off의 클래스는 AMVR on의 클래스보다 10배 희소하게 feature를 출력하도록 하여 데이터셋 불균형 문제를 해결하고자 하였다.

AMVR 데이터셋은 HEVC 표준과 Ultra Video Group (UVG)[7] 영상에서 추출하였으며, 이 중 train set을 위한 영상은 Bosphorus, HoneyBee, Jockey, Ready- SteadyGo, Traffic, Kimono1, ParkScene이다. 중복되지 않는 validation set을 사용하여 Overfitting을 감지하였으며 전체 학습 데이터셋의 Sample 수는 492,968개이다. 이번 실험에서 사용된 데이터셋은 기존 연구[5]에서 사용하였던 Feature의 종류와는 거의 유사하다. 하지만 기존 연구[5] 데이터셋은 VVC의 공통시험조건(common test condition, CTC)[8] 영상에 대한 Feature를 추출하였기 때문에 Feature를 추출하는 영상이 다르다는 점과 하나의 Feature 수 증가, loss를 이진 교차 엔트로피와 optimizer를 Adam으로 변경하여 최적화하였다는 차이점이 존재한다.

본 논문에서 최종적으로 제안된 신경망 모델의 구조는 15x5x1의 MLP이며 이는 아래 수식 (1) 과 같이 표현할 수 있다. 인풋 특징 벡터인 x의 차원이 다르다는 점 외에 구조와 식의 전개는 [6]와 같다. 먼저 첫 번째 레이어(l = 1)에 15개의 인풋 특징 벡터 x가 있고 두 번째 레이어(l = 2)의 j번째 노드에 있는 뉴런에 x를 비선형 가중치 합을 공급한다. 계산은 (1)에서 정의된다:

$$y_j = f(\sum_i w_{ji} x_i + b_j), \text{ for } j = 1, ..., \varphi, \quad (1)$$

여기서 x_i는 i번째 입력의 값을 나타낸다. $w_{j,i}$는 i번째 입력에 대한 j번째 뉴런에 해당하는 가중치 값이고 ϕ는 레이어당 뉴런의 수이다. 또한 b_j는 j번째 뉴런의 바이어스 값이다. $f(\cdot)$는 활성화 함수로서의 로지스틱 시그모이드 함수를 나타내며 y_j는 j번째 뉴런의 $f(\cdot)$의 결과를 나타낸다. 본 모델은 기존 연구[5]보다 노드의 수를 감소시켜 경량화하였다.

제안하는 신경망 기반의 고속 AMVR 결정기법은 아래 그림 1과 같다. 먼저 매 CU마다 AMVR을 테스트할 때를 기점으로 본문에서 제안하는 기 언급한 feature들을 모두 가지고 있다고 가정한다. 다만, 언급한 feature들을 모두 가지고 있지 않다면 신경망을 사용하지 않고 기존기법[4]을 그대로 사용한다. 그림 1에 대한 수식 (1)의 표기법으로 첫

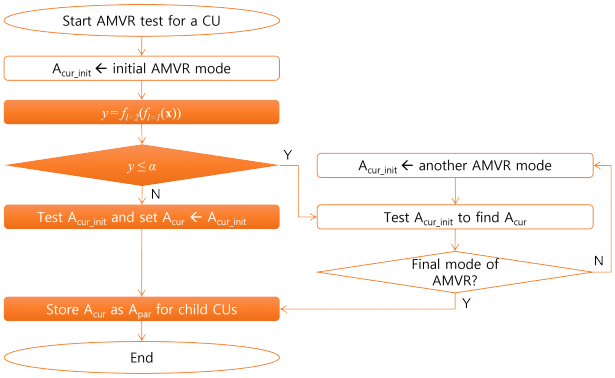

그림 1. 신경망 기반 고속 AMVR 결정 기법
Fig. 1. Fast AMVR decision method based on neural network

번째 레이어(l = 1)는 f_{l-1}로 표기하고 두 번째 레이어(l = 2)는 f_{l-2}로 표기한다. 또한 A_{cur_init}은 초기의 AMVR값이며 A_{cur}은 현재 CU의 최적의 AMVR 정보를 의미한다. 신경망 출력값을 y라고 하면, 이 출력값은 최종적으로 0에서 1 사이의 값을 갖는다. 이때 임의의 상숫값 a보다 작거나 같으면, AMVR의 모든 모드들을 테스트해보아야 함을 의미하고, a보다 크면 AMVR이 필요 없는 경우로 판단한다.

a의 값에 따라 압축 성능과 부호화 복잡도가 좌우될 수 있으며, a 값을 키울수록 AMVR 테스트를 생략하는 경우가 작아질 가능성이 높으므로, a 값은 0.5에서 1사이 값을 적절히 정해야 한다. 우리는 0.5, 0.9, 0.93, 0.96 등의 값과 압축성능을 비교해보며 압축손실이 가장 적은 0.96으로 결정하였다. 또한, y 값의 여부에 상관없이 관련된 모든 AMVR 과정을 마치게 되면, 최종적으로 가장 좋은 AMVR 모드 값을 메모리에 저장하여 추후 자녀 CU에게 전달해주어 신경망의 입력 벡터에 활용한다.

Ⅲ. 실험결과

본문에서 제안하는 경량화된 신경망의 성능을 평가하기 위하여, VVC Test Model(VTM) 6.0에 구현하여 성능 평가를 진행하였다. 실험 환경은 Random Access 시나리오로 진행하였으며, VTM의 방대한 자원 소모를 단순화하기 위하여 CTC에서 부호화 프레임 숫자는 일괄적으로 FPS * 2 (즉, 2초 동안) 만큼으로 감소 시켜 진행하였다. 또한, 기존 기법[4]과 함께 비교를 진행하였다. 기존 기법[4]은 확률에 기

반하여 AMVR을 on/off의 여부를 결정하는 알고리즘이었던 반면에 제안기법은 신경망 모듈을 VVC에 추가하여 결정하는 기법이다. 부호화 복잡도의 정밀한 평가를 위하여 하이퍼쓰레딩과 터보모드는 끈 상태로 진행하였으며, 부호화 과정은 Intel i7-10700 CPU @ 2.90GHz와 64비트 Windows 10에서 실행을 진행하였다.

제안기법과 기존기법[4]의 VVC 대비 성능 변화를 압축효율인 BD-rate(BDBR)과 부호화 시간(Encoding time, 이하 T)을 측정한 결과는 표 1과 같다. 부호화 시간 T를 수식으로 정의하면 $T = T_{proposed} / T_{anchor}$로 정의하며 여기서 T_{anchor}는 VTM 6.0의 전체 부호화 시간을, $T_{proposed}$는 제안기법의 전체 부호화 시간을 의미한다.

표 1. VVC 대비 제안기법과 기존 기법[4]의 성능 비교
Table 1. Performance comparison with the proposed method and the existing method [4] in comparison with VVC

Class	Sequence	[4]		Proposed	
		Y_BDBR	T	Y_BDBR	T
Class A1	Tango2	2.57%	85%	2.97%	83%
	FoodMarket4	1.34%	87%	1.72%	87%
	Campfire	0.61%	83%	0.68%	83%
Class A2	CatRobot	1.33%	84%	1.29%	82%
	DaylightRoad2	1.77%	83%	2.28%	81%
	ParkRunning3	0.89%	84%	0.96%	83%
Class B	MarketPlace	1.46%	83%	1.57%	80%
	RitualDance	1.45%	83%	1.54%	81%
	Cactus	0.73%	86%	0.85%	85%
	BasketballDrive	1.81%	84%	2.02%	82%
	BQTerrace	0.48%	92%	0.63%	76%
Class C	BasketballDrill	1.36%	89%	1.52%	75%
	BQMall	1.03%	92%	1.13%	76%
	PartyScene	0.58%	85%	0.60%	75%
	RaceHorsesC	0.96%	86%	0.99%	70%
Class D	BasketballPass	1.33%	84%	1.41%	76%
	BQSquare	0.44%	83%	0.50%	74%
	BlowingBubbles	0.61%	86%	0.75%	81%
	RaceHorses	0.76%	85%	0.91%	73%
Average		1.13%	85%	1.28%	79%

기존기법은 휘도성분 기준으로 평균 1.13%의 BDBR 손실을 일으키면서 전체 부호화 시간을 85%로 감소시킨 반면, 제안기법은 1.28%의 BDBR 손실 대비 전체 부호화시간을 79%로 감소시켰다. 제안기법은 기존기법[4] 대비 약간의 BDBR 손실을 일으켰지만 부호화 시간을 6% 더 감소시켰음을 알 수 있다. 또한 기존기법[4]과 달리 제안기법은 a값 조정을 통해 수요자의 필요에 맞게 성능을 조절할 수 있다는 장점이 있다. 표 1을 해상도 별로 보았을 때, Class C, D와 같이 저해상도 영상의 경우에는 압축률 손실이 적으면서 부호화 시간이 많이 늘어났다. 또한 제안기법의 최대 부호화시간 감소 영상인 RaceHorsesC의 경우 최대 70%까지 감소한 반면 기존기법[4]은 최대가 83%라는 한계가 있다.

Ⅳ. 결 론

신경망은 일반적으로 복잡하다고 알려져있어서 부호화기의 복잡도를 줄이는데 사용하기에 도전적인 측면이 있으나, 본문에서 제안하는 매우 가벼운 MLP를 채택할 경우, AMVR의 최적의 모드를 비교적 정확하게 분류하면서 동시에 효과적으로 부호화 복잡도를 줄일 수 있음을 알 수 있었다. 이러한 신경망 기반의 분류 기법을 활용한다면, AMVR 외의 다양한 화면 간 예측 기법의 최적의 모드를 분류하는 작업에도 적용할 수 있을 것으로 기대한다.

참 고 문 헌 (References)

[1] B. Bross, J. Chen, S. Liu, Y. -K. Wang, “Versatile Video Coding (Draft 7),” Joint Video Experts Team (JVET) of ITU-T and ISO/IEC, Document JVET-P2001, 2019.

[2] High Efficiency Video Coding (HEVC), Rec. ITU-T H.265 and ISO/IEC 23008-2, ITU-T and ISO/IEC JTC 1, 2013 (and subsequent editions).

[3] H. Han, J. Choe, D. Gwon, and H. Choi, “VVC intra prediction and encoding key technology”, Broadcasting and Media Magazine, Vol.24, No.4, pp.39-59, October 2019.

[4] S. -h. Park, "Fast Decision Method of Adaptive Motion Vector Resolution", The Korean Institute of Broadcast and Media Engineers, Vol.25, No.3, pp.305-312, May 2020, http://dx.doi.org/10.5909/JBE.2020.25.3.305

[5] H. Baek and S. -h. Park, "Neural Network-Based Adaptive Motion Vector Resolution Discrimination Technique", Proceedings of the Korean Society of Broadcast Engineers Conference, pp. 49-51, 2021, in press

[6] S. -h. Park and J. Kang, "Fast Multi-type Tree Partitioning for Versatile Video Coding Using a Lightweight Neural Network, " in IEEE Transactions on Multimedia, pp.1-1, Dec 2020, doi:10.1109/TMM.2020.3042062.

[7] Ultra video group (UVG) dataset, http://ultravideo.cs.tut.fi/#testsequences (Accessed Oct. 30, 2020).

[8] F. Bossen, J. Boyce, K. Suehring, X. li, and V. Seregin, “JVET common test conditions and software reference configurations for SDR video,” Joint Video Experts Team (JVET) of ITU-T and ISO/IEC, Document JVET-N1010-v1, 2019.