Journal of Korea Society of Industrial Information Systems Vol.29 No.2, Apr. 2024 :35-45
http://dx.doi.org/10.9723/jksiis.2024.29.2.035 ISSN:1229-3741

CNN과 Attention을 통한 깊이 화면 내 예측 방법.+

(Intra Prediction Method for Depth Picture Using CNN and Attention Mechanism)

윤 재 혁[1], 이 동 석[2], 윤 병 주[3], 권 순 각[4]*
(Jae-hyuk Yoon, Dong-seok Lee, Byoung-ju Yun, and Soon-kak Kwon)

요 약 본 논문에서는 CNN과 Attention 기법을 통한 깊이 영상의 화면 내 예측 방법을 제안한다. 제안하는 방법을 통해 예측하고자 하는 블록 내 화소마다 참조 화소를 선택할 수 있도록 한다. CNN을 통해 예측 블록의 상단과 좌단에서 각각 수직방향과 수평 방향의 공간적 특징을 검출한다. 두 공간적 특징은 예측블록과 참조 화소들에 대한 특징을 예측하기 위해 각각 특징차원과 공간적 차원으로 병합된다. Attention을 통해 예측 블록과 참조 화소간의 상관성을 입력된 공간적 특징을 통해 예측한다. Attention을 통해 예측된 상관성은 CNN 레이어를 통해 화소 도메인으로 복원되어 블록 내 화소 값이 예측된다. 제안된 방법이 VVC의 인트라 모드에 추가되었을 때 화면 예측 오차가 평균 5.8% 감소하였다.

핵심주제어: 영상부호화, 화면 내 예측, 깊이 영상, Attention Mechanism

Abstract In this paper, we propose an intra prediction method for depth picture using CNN and Attention mechanism. The proposed method allows each pixel in a block to predict to select pixels among reference area. Spatial features in the vertical and horizontal directions for reference pixels are extracted from the top and left areas adjacent to the block, respectively, through a CNN layer. The two spatial features are merged into the feature direction and the spatial direction to predict features for the prediction block and reference pixels, respectively. the correlation between the prediction block and the reference pixel is predicted through attention mechanism. The predicted correlations are restored to the pixel domain through CNN layers to predict the pixels in the block. The average prediction error of intra prediction is reduced by 5.8% when the proposed method is added to VVC intra modes.

Keywords: video coding, intra prediction, depth picture, attention mechanism

* Corresponding Author: skkwon@deu.ac.kr
+ 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평과원의 지원을 받아 수행된 지역지능화혁신인재양성사업(IITP-2024-2020-0-01791, 100%)과 부산광역시 및 (재)부산테크노파크의 BB21plus 사업임.
Manuscript received March 18, 2024 / revised April 14, 2024 / accepted April 15, 2024
1) 동의대학교 컴퓨터소프트웨어공학과, 제1저자
2) 동의대학교 인공지능그랜드ICT연구센터, 제2저자
3) 경북대학교 전자공학부, 제3저자
4) 동의대학교 컴퓨터소프트웨어공학과, 교신저자

1. 서 론

깊이 영상은 카메라로부터의 실제 거리를 화소 단위로 저장한다. 깊이 영상은 거리 정보를 통해 3차원 정보를 획득할 수 있다. 깊이 영상을 통해 획득된 3차원 정보를 통해 특정 형태를 가지는 객체를 검출하고 추적하거나 (Kwon et al., 2017; Ren et al., 2017; Zhao et al., 2017; Jiang et al., 2019; Lee et al., 2019), 특정 행동을 인식할 수 있다 (Li et al., 2012; Ren et al., 2013; Li et al., 2016; Oh et al., 2018). 또한 깊이 영상을 통해 지표면이나 벽 등을 검출하여 동시적 위치추정 및 지도작성(SLAM: Simultaneous Localization and Mapping)에 적용할 수 있다 (Aguilar et al., 2017; Sun et al., 2017). 이들 응용의 광범위한 적용을 위해서 깊이 영상의 전송, 저장이 필요하다. 이를 위해서는 깊이 영상의 효율적인 부호화 방법이 필요하다.

영상 부호화 표준에서는 화면 내 예측과 화면 간 예측을 통해 영상 내 유사성을 제거하고 영상을 압축할 수 있다. 화면 내 예측에서는 공간 내 인접한 화소 간 유사한 화소 값이 분포할 확률이 크다는 가정을 토대로 다수의 인트라 모드를 제공한다. 해당 가정은 깊이 영상에서도 유효할 수 있다. 하지만 깊이 영상의 공간적으로 이웃한 화소 간의 상관관계는 색상 영상과 다르다. 특히 화면 내 예측하고자 하는 블록 내에 서로 다른 객체 또는 객체와 전경의 혼합 등의 요인으로 인해 화면 내 예측을 하고자 하는 블록 내에 두 계층 이상의 화소 값 분포를 가지는 영역들로 분할될 수 있다. 특히 깊이 영상은 화소 값의 범위가 색상 영상보다 매우 크기 때문에 이러한 요소는 깊이 영상의 화면 내 부호화의 효율을 떨어뜨린다. Lee and Kwon(2022)의 연구에서는 이러한 문제를 해결하기 위해 CNN 계층을 통해 블록 내 클러스터 영역을 예측하는 방법을 제안했다. 하지만 이전 연구에서는 블록의 클러스터 영역의 화소 값을 일괄적으로 해당 클러스터에 속하는 참조 화소의 평균으로 예측하였다. 깊이 영상의 화면 내 예측의 어려움을 해결하기 위해서는 정해진 한 방향의 화소들만을 참조하는 것이 아닌, 예측 블록 내 각각의 화소마다 참조 화소를 결정할 필요가 있다.

본 논문에서는 깊이 영상에 대해 Attention 기법을 통해 참조 화소들의 중요도를 예측하여 화면 내 예측을 하는 방법을 제안한다. 제안하는 방법은 공간적 특징 추출 단계, Attention 기법을 통한 참조 화소와 예측 블록 내 화소 간 상관성 예측 단계, 그리고 상관성을 통한 블록 내 화소 예측 단계로 구성된다. 공간적 특징 추출 단계에서는 예측 블록의 참조 화소들로 활용되는 상단과 좌단 영역에서 각각 수직 방향과 수평 방향의 공간적 특징을 CNN 계층을 통해 추출한다. 참조 화소와 예측 블록 간 상관성 예측 단계에서는 참조 화소에서 얻은 공간적 특징에 Attention 기법을 적용하여 예측 블록의 공간적 특징을 예측한다. 예측 블록 내 각 화소마다 상단에 있는 참조 화소의 공간적 특징과 좌단에 있는 참조 화소의 공간적 특징을 병합한다. 그 후, 병합된 공간적 특징은 Attention 기법을 통해 모든 참조 화소의 공간적 특징들과의 상관성이 계산된다. 이를 통해 예측 블록의 공간적 특징을 예측한다. 예측 블록 내 화소의 공간적 특징은 상관성이 높게 예측된 참조 화소의 특징이 크게 반영된다. 이를 통해 예측 블록 내 화소마다 다른 화소를 참조할 수 있고, 한 화소에서 다수의 참조 화소를 활용할 수 있다. 블록 내 화소 예측 단계에서는 예측 블록의 공간적 특징을 CNN 계층을 통해 화소 값으로 변환한다.

기존 연구들은 화면 내 예측을 위해 화소를 참조할 때 블록 내 화소의 평균 등의 통계적인 방법을 사용하거나 특정한 방향에 위치한 화소를 참조한다는 한계가 있다. 깊이 화면 내 객체 모서리 영역이나 두 선이 교차하는 패턴 등에서는 화소별로 다른 방향의 화소 참조가 필요하지만, 기존 방법에서는 이러한 문제를 해결하기 어렵다. 제안된 방법을 통한 화면 내 예측은 예측 블록 내 화소마다 공간적 상관성이 큰 참조 화소를 Attention 기법을 통해 검출한다. 이를 통해 예측할 화소는 서로 다른 화소를 참조할 수 있다. 이를 통해 기존 화면 내 예측 방법이 참조 화소를 한 방향으로만 예측할 수 있다는 한계를 극복할 수 있다.

2. 기존 깊이 영상의 화면 내 예측 방법

영상의 중복성은 시간적으로 인접된 프레임에서 반복해서 나타나는 화소 배열의 유사성과 한 화면 내 인접한 화소 값의 유사성으로 나타난다. 영상은 이러한 시간적, 공간적 방향의 중복성을 제거하여 압축될 수 있다. H.264/AVC (Kwon et al., 2005), H.265/HEVC (Sullivan et al., 2012), VVC (Bross et al., 2021) 등의 색상 영상 부호화 표준들은 공간적 중복성을 제거하기 위해 다양한 인트라 모드를 지원하고, 시간적 중복성을 제거하기 위해 블록 단위의 움직임 추정 알고리즘을 제공한다.

색상 영상에서 공간적으로 인접한 화소는 높은 빈도로 유사한 값을 가진다. 하지만 깊이 영상에서는 색상 영상과 달리 표면의 성질에 따라 인접한 화소 간 비선형적으로 증가 또는 감소할 수도 있다. 따라서 효율적인 깊이 영상에 대한 부호화를 하기 위해서는 깊이 영상이 가지는 인접한 화소가 가지는 성질, 즉 공간적 유사성을 고려할 필요가 있다.

8비트 화소만을 지원하는 H.264/AVC를 깊이 영상에 적용하기 위해, Nenci et al.(2014)는 하나의 채널을 가지는 깊이 영상을 8비트 화소를 가지는 다채널로 분할하는 방법을 제안하였고, Stankiewicz et al.(2013)는 깊이 데이터의 비선형 변환을 제안하였다. 하지만 이러한 방법들은 단순한 화소 값의 변환으로, 깊이 영상의 특성을 고려하지 않는다는 한계가 있다. 이전 연구(Lee et al., 2021)는 평면 추정을 통해 깊이 영상을 화면 내 예측하는 방법을 적용한 인트라 모드를 제안했다. 참조 화소를 통해 블록 내 화소를 이루는 평면이 모델링되고, 해당 평면을 통해 예측 블록 내 화소가 예측된다. 하지만 이 방법은 평면이 아닌 표면이 속한 영역에 대해 예측 성능이 거의 개선되지 않는다는 문제가 있다. Lee and Kwon(2022)은 참조 화소들로부터 예측 블록의 영역을 예측하고, 각 영역에 속하는 참조 화소의 평균을 통해 예측 블록 내 해당 영역에 속하는 화소들을 예측하였다. 하지만 이 연구는 각 영역에 대해 참조 화소의 평균으로 예측하기 때문에 영상의 공간적 유사성이 무시된다는 한계를 가진다.

3. CNN과 Attention을 통한 깊이 화면 내 예측

본 논문은 Attention 기법을 통해 예측 블록 내 화소마다 참조 화소와의 상관성을 계산하여 상관성이 높은 참조 화소로 예측하는 방법을 제안한다. 제안하는 네트워크의 입력은 깊이 화면 내 $m \times m$ 크기의 예측 블록에 대해 $t \times m$ 상단 영역의 블록과 $m \times t$ 좌단 영역의 블록이다. 상단 및 좌단 블록에서 CNN 계층을 통해 공간적 특징이 추출된다. 상단 및 좌단 블록에서 추출된 공간적 특징들은 병합된 후, Attention 기법 적용을 위해 위치 정보를 추가하는 Positional encoding을 적용한다. 그 후 Attention 기법을 적용하여 예측 블록과 참조 화소 간 상관성을 검출한다. 그 후 CNN 계층을 통해 공간적 특징들을 화소로 변환한다. Fig. 1은 제안하는 방법의 흐름을 보인다.

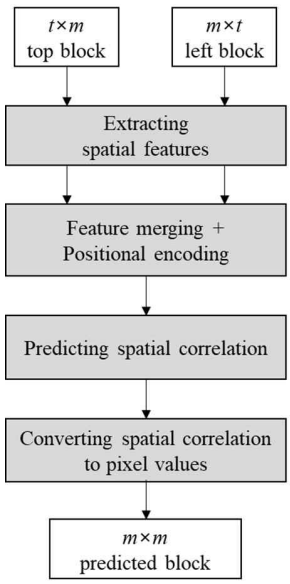


Fig. 1 Flow of the proposed method

3.1 참조 화소에서 공간적 특징 추출

예측 블록의 상단 좌단의 영역의 화소들에서

공간적 특징을 추출한다. 최대 10000 이상의 값을 가질 수 있는 깊이 영상 내 화소 값은 네트워크에 입력되기 전 아래 식을 통해 [-1, 1] 범위 내 값으로 변환하여 네트워크 학습이 원활하게 한다.

$$input_{pos}(i,j) = (p_k(i,j) - \beta)/(\alpha - \beta)) \quad (pos = top, left) \tag{1}$$

여기서 (i, j)는 상단 영역과 좌단 영역의 좌표, α와 β는 각각 입력 블록들의 화소의 최댓값과 최솟값, p_{top}와 p_{left}는 각각 상단 영역과 좌단 영역의 원래 화소 값, $input_{top}$와 $input_{left}$ 는 각각 네트워크에 입력되는 상단 영역과 좌단 영역에 대한 블록들이다.

상단 영역 내 참조 화소들에서 공간적 특징 추출을 수행한다. f 개의 3x3 커널을 가지는 CNN 계층을 통해 해당 영역들 내 국소 영역에서 이웃한 화소와의 관계를 검출한다. 그 후 커널의 크기는 높이가 입력 블록과 동일한 $3\times m$인 CNN 계층을 통해 해당 영역의 수직 성분에 대한 공간 특징들이 $1\times m\times f$ 형태로 출력된다. 이 CNN 계층은 수직 방향에 대해 패딩을 적용하지 않도록 하여 출력의 높이가 1이 되도록 한다. Fig. 2는 공간적 특징 추출을 위한 모듈 구성을 보인다.

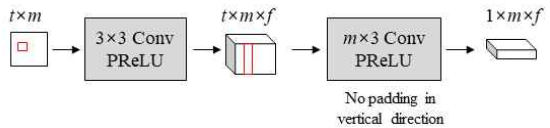


Fig. 2 Module structure for spatial feature extraction

좌단 영역 내 화소들에 대해서도 같은 모듈을 통해 공간적 특징을 추출한다. 이를 위해 공간적 특징 추출 모듈에 입력하기 전 수평 차원과 수직 차원에 대해 전치를 수행한다. 이는 같은 모듈을 통해 상단 및 좌단 영역에서 공간적 특징들을 추출할 수 있도록 하여 네트워크의 가중치 수를 줄이고, 학습 효율성을 개선한다.

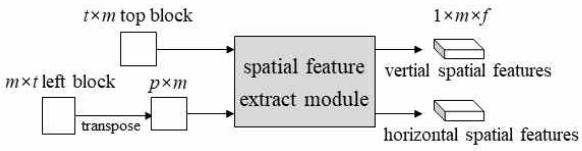


Fig. 3 Spatial feature extraction for top and left blocks

3.2 공간적 특징 병합 및 Positional Encoding을 통한 위치 정보 추가

Attention 기법을 적용하여 예측 블록과 참조 화소 간 공간 특징의 상관성을 예측한다. 이를 위해 먼저 공간적 특징 추출 모듈에서 얻은 수직, 수평 공간적 특징들을 병합한다. 상단과 좌단 영역들에서 얻은 $1\times m\times f$ 형태의 공간적 특징 내 값들을 수직 방향으로 복제하여 $m\times m\times f$ 형태가 되도록 한다. 그 후 좌단과 상단 블록들에 대한 공간 특징들을 특징 차원으로 연결한다. 이 때, 좌단 영역의 공간적 특징은 병합되기 전 전치되어 수평 방향의 공간적 특징으로 표현되도록, 즉 수평 성분의 특징들이 모두 같도록 한다. 이를 식으로 나타내면 아래와 같다.

$$F(i,j,k) = \begin{cases} S_V(1,j,k) & \text{if } k < f \\ S_H(1,i,k) & otherwise \end{cases} \quad (0 < i,j \le m, 0 < k \le 2f) \tag{2}$$

여기서 S_V와 S_H는 각각 상단 영역과 좌단 영역에서 얻은 $1\times m\times f$ 형태의 공간적 특징이다. 그 후 $m\times m\times 2f$ 형태의 공간적 특징의 특징 차원, 즉 맨 마지막 차원은 커널 크기가 1×1인 CNN 계층을 통해 f로 축소한다.

Attention 기법에서 예측 블록 내 화소들의 공간적 특징의 상관성을 비교하는 대상으로 활용하기 위해, 참조 화소들의 공간적 특징들을 병합한다. 이를 위해 S_V와 S_H를 아래와 같이 수평 차원으로 병합하여 참조 화소에 대한 $1\times 2m\times f$ 형태의 공간적 특징 R을 얻는다.

$$R(1,i,k) = \begin{cases} S_V(1,i,k) & \text{if } 0 < i \le m \\ S_H(1,i-m,k) & otherwise \end{cases} \quad (0 < i \le 2m, 0 < k \le f) \tag{3}$$

Fig. 4는 Attention 기법 적용을 위한 공간적 특징의 병합을 보인다.

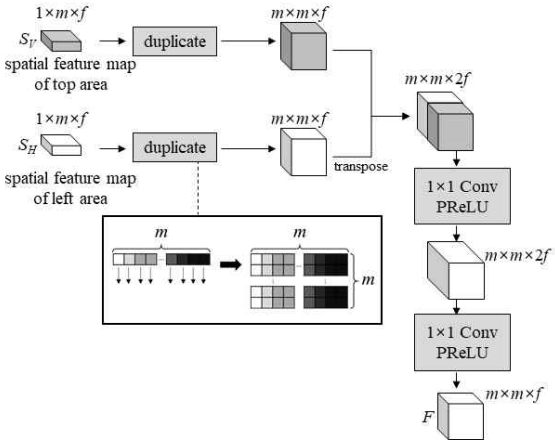


Fig. 4 Merging spatial feature maps for target block and reference pixels

기존 영상 부호화 표준에서 화면 내 예측에 쓰이는 인트라 모드에서는 블록 내 좌표에 따라 참조할 블록이 결정된다. 제안된 네트워크에서 각 화소의 좌표에 대한 정보를 추가하기 위해 별도의 차원에 좌표 인덱스를 추가할 수도 있지만, Positional encoding을 통해 입력 크기를 축소할 수 있다. Positional encoding은 특징들에 대해 해당 특징에 대한 위치를 임베딩하여 이를 더하는 방법이다. 이를 통해 위치에 대한 차원을 추가할 필요 없이 위치 정보를 추가할 수 있다. 제안된 네트워크에서는 아래 식을 통해 Positional encoding을 적용한다.

$$PE(i,j,k)=\begin{cases}P_H(i,k), \text{if } k<f/2\\ P_V(j,k), otherwise\end{cases} \quad (4)$$
$$(0<i,j\le m,\ 0<k\le f)$$

여기서 P_H와 P_V는 각각 위치를 나타내는 i과 j에 대한 위치 특징들이 저장된 학습 가능한 테이블이다. 즉 (i, j) 좌표에 대해 PE는 i값에 대응되는 P_H의 특징들과 j값에 대응되는 P_V의 특징들이 병합된 것이다. $m\times m\times f$ 형태의 공간적 특징 F는 아래와 같이 PE와 합쳐진다.

$$F_{PE}(i,j,k)=F(i,j,k)+PE(i,j,k) \quad (5)$$

3.3 Attention 기법을 통한 참조 화소와 예측 블록 내 화소 간의 상관성 예측

Attention 기법 (Vaswani et al., 2017)은 입력 데이터의 다양한 부분에 선택적으로 집중하여 특정 요소에 더 많은 중요성을 부여하여 특징을 검출하는 딥러닝 구조이다. Attention 기법은 Query, Key, Value의 3종류 데이터를 입력받는다. Query와 Value는 각각 입력 데이터에 대한 특징과 참조할 데이터의 특징을 의미하고, Key는 Value에 대응되는 값으로써, Query를 통해 알맞은 Value를 찾기 위한 것이다. 즉, Attention 기법은 Query와 유사한 Key가 가지는 Value를 찾는 것이다. Query, Key, Value에 대한 입력을 각각 X_Q, X_K, X_V라고 했을 때, 이들은 각각 학습 가능한 테이블 W_Q, W_V, W_K를 참조하여 다음과 같이 변환된다.

$$Q=X_Q\otimes W_Q \quad (6)$$
$$K=X_K\otimes W_K$$
$$V=X_V\otimes W_V$$

여기서 $\otimes$는 행렬 곱이다. Attention 기법에서 학습은 최적의 W_Q, W_V, W_K를 찾는 것으로 정의될 수 있다. Q와 K의 유사도는 Q와 K간 수평 성분들의 내적들을 통해 구한다. 해당 연산은 다음과 같이 Q와 전치된 K 간의 행렬 곱으로 계산될 수 있다.

$$score=\frac{Q\otimes K^T}{\sqrt{f}} \quad (7)$$

그 후 softmax 함수를 적용하여 모든 요소에 대한 합이 1이 되는 확률 형태로 변환한 후, 다음과 같이 Query와 Key 간 유사도에 따른 Value의 가중합을 계산한다.

$$Z=softmax(score)\otimes V \quad (8)$$

그 후 Z는 다음 식으로 정의되는 FFNN (Feed- Forward Neural Network)를 거친다.

$$FFNN(Z) = a_2 \odot (\sigma(a_1 \odot Z + b_1)) + b_2 \quad (9)$$

여기서 a_1, a_2, b_1, b_2는 각각 학습 가능한 행렬이며, σ(.)는 활성화 함수, ⊙는 행렬 요소 간의 곱셈(element-wise product)이다. FFNN는 전연결층을 두 번 적용한 것과 동일하지만, 마지막 단계에서 활성화 함수를 적용하지 않는다. 이를 통해 Attention 기법을 위한 연산이 비선형적이 되도록 한다. 활성화 함수는 PReLU(Parametric Rectified Linear Unit)를 적용한다. PReLU는 입력에 대해 양수일 경우 해당 값을 그대로 출력하고, 음수일 경우 학습 가능한 변수를 곱한 값이 출력된다. PReLU는 입력 x에 대해 다음과 같이 출력한다.

$$PReLU(x) = \begin{cases} x, & \text{if } x > 0 \\ ax, & \text{if } x \le 0 \end{cases} \quad (10)$$

여기서 a는 학습 가능한 변수이다. 본 논문에서는 일련의 Attention 기법 적용을 $ATTN(X_Q, X_K, X_V)$로 표기한다.

예측 블록의 공간적 특징에 Attention을 적용하기 위해 F_{PE}의 수직 차원과 수평 차원을 병합하여 $m^2 \times f$ 형태로 변환한다. 그 후 예측 블록과 참조 화소 간 공간적 특징의 상관성을 찾기 위해 Query를 F_{PE}로 하고, Key와 Value를 R로 하여 Attention 기법을 다음과 같이 적용한다.

$$A = ATTN(F_{PE}, R, R) \quad (11)$$

그 후 $m^2 \times f$ 형태의 A를 변환하여 특징의 크기를 $m \times m \times f$가 되게 한다. 식 (11)의 Attention 기법 적용을 통해 예측 블록 내 각 화소는 참조 화소의 공간적 특징들에 대해 공간적 특징의 상관성으로 가중하여 더해진다. 즉 큰 상관성을 가지는 참조 화소의 공간적 특징이 예측 블록 내 화소의 공간적 특징에 많이 반영된다.

3.4 공간적 특징의 화소 값 변환을 통한 화면 내 예측

1×1 커널의 4개의 CNN 계층을 통해 특징 차원을 축소하고, 특징들에서 화소 값으로 복원되도록 하여 최종적으로 $m \times m \times 1$의 블록을 출력하도록 한다. 활성화 함수는 마지막 CNN layer를 제외한 나머지는 PReLU이며, 마지막 layer는 sigmoid이다. Fig. 5는 공간적 특징을 화소 값으로 변환하는 흐름을 보인다.

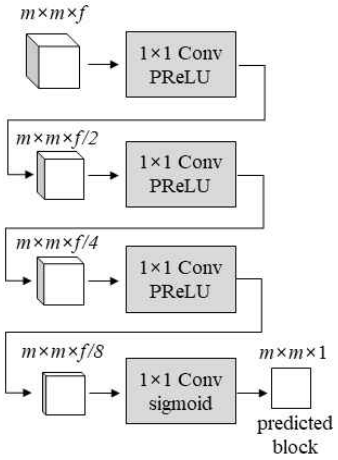


Fig. 5 Conversion of spatial features to pixel values

공간적 특징에서 변환된 화소 값은 [0, 1]의 범위 내에 있다. 이를 원래 화소 값 범위로 변환하기 위해 아래 식을 적용한다.

$$\hat{y}(i,j) = \lfloor R(i,j,1) * (\alpha - \beta) + \beta \rfloor \quad (12)$$

여기서 ⌊ ⌋는 내림 함수로, 입력된 실수에 대해 정수부만 출력한다.

3.5 데이터셋과 학습

제안하는 네트워크를 학습하기 위한 데이터셋으로 NYU Depth Dataset (Silberman et al., 2012)을 활용한다. 해당 데이터셋은 1,446장의 다른 장면을 촬영한 화면이다. Fig. 6은 해당 데이터 셋의 일부를 보인다. 각각의 화면에 대해 $(m+t) \times (m+t)$ 크기의 블록들로 분할하여 데이터셋을 생성한다. 네트워크 학습은 전체 화면의 80%인 1,157장에서 추출된 약 151,000개의 블록

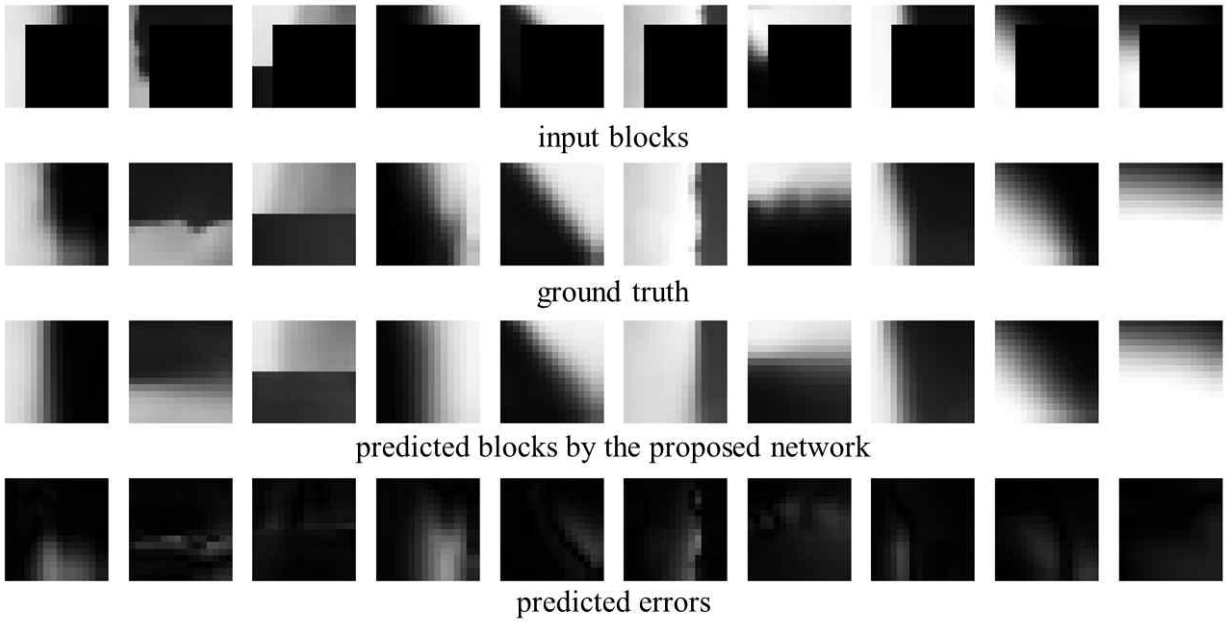


Fig. 7 Block prediction results through the proposed method

을 통해 이루어지고, 네트워크 학습 검증은 나머지 화면에서 추출된 약 51,000개의 블록을 통해 이루어진다.

학습을 위한 손실 함수는 다음의 MSE(Mean Square Error)를 사용한다.

$$loss = \frac{1}{m^2}\left(\sum_{i}^{m}\sum_{j}^{m}(\hat{y}(i,j)-y(i,j))^2\right) \quad (13)$$

Fig. 6 Samples of pictures for simulation

4. 실험결과

본 논문에서 제안된 네트워크의 정확도를 측정한다. 네트워크 학습의 하이퍼 파라미터로 배치크기는 128, 에포크는 128, 학습률은 1×10^{-3}으로 한다. 이는 다른 하이퍼 파라미터로 여러 번 학습을 하여 이 중 최고의 성능이 나오는 값으로 결정한 것이다. 제안된 네트워크의 파라미터로 네트워크 내 특징 수 f와 예측 블록의 크기 m, 그리고 상단 영역의 높이와 좌단 영역의 너비인 t는 각각 16, 16, 4로 한다.

Fig. 7은 제안된 방법의 예측 결과를 보인다. 이 때 제안된 방법은 인트라 모드의 방향성 모드와 유사한 결과를 보인다. 하지만 인트라 모드와 달리 제안된 방법은 곡선 등의 비선형적인 방향에 대해 예측을 하는 것을 보인다. 이는 제안된 방법이 영상 부호화 표준의 방향성 모드를 대체할 수 있음을 보인다.

제안된 네트워크의 파라미터인 f, m, t 에 따른 예측 정확도를 검증셋을 통해 비교한다. Table 1은 특징 개수 f에 따른 MSE를 보인다. f가 커질수록 예측 정확도가 높아진다. 하지만 f가 24 이상일 경우 예측 정확도는 거의 증가하지 않는다. 이는 최적의 f는 16에서 24 범위 내의 값이라는 것을 의미한다. f가 너무 클 경우 네트워크의 큰 복잡성으로 인해 비효율적인 학습이 이루어진다.

Table 1 Prediction errors for f

f	8	16	24	32
MSE	197.16	192.58	169.61	169.35

Table 2는 예측 블록의 크기 m에 따른 MSE를 보인다. m이 클수록 예측 정확도가 급격하게 떨어지는 것을 보인다. 이는 블록의 크기가 클수록 참조 화소와의 거리가 멀어져 공간적 상관성이 떨어지기 때문이다.

Table 2 Prediction errors for m

m	8	16	32
MSE	67.57	192.58	445.22

Table 3은 입력된 블록의 크기인 t 에 따른 예측 정확도를 보인다. t가 4 이상일 경우 예측 정확도가 거의 변하지 않는다. 오히려 t가 너무 클 경우 예측 오차가 다소 커지는 것을 보인다. 이는 지나치게 큰 참조 영역에 대해서는 오히려 공간 특징 검출이 부정확해져서 예측이 부정확해지기 때문이다.

Table 3 Prediction errors for t

t	2	4	6	8
MSE	313.54	192.58	193.14	200.70

제안한 방법의 화면 내 예측 성능을 비교하기 위해 기존 영상 부호화 표준인 VVC의 인트라 모드와 비교한다. 실험 영상으로 NYU Dataset의 4개의 깊이 영상을 Fig. 8과 같이 사용한다. 이 깊이 영상들은 Kinect로 촬영되었으며, 해상도는 640x480이다. 각 영상들의 첫 프레임 화면을 실험에 사용한다.

제안하는 방법의 부호화 성능의 개선을 평가하기 위해 VVC의 화면 내 예측 방법과 성능을 비교한다. VVC의 화면 내 예측은 DC모드, planar 모드와 방향성 모드(Angular mode)를 포함한 총 67개의 화면 내 부호화 모드가 있다. 제안한 방법을 해당 VVC의 화면 내 예측에서의 새로운 인트라 모드로 추가하여 원래의 화면 내 예측 및 이를 통한 부호화와 비교하여 평가

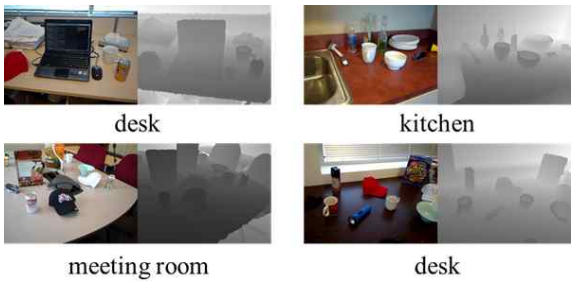


Fig. 8 Depth videos for comparison simulation with VVC

한다. Table 4는 VVC의 인트라 모드들과 제안한 방법의 실제 영상에 대해 MSE를 측정한 결과로, 평균 5.8%의 MSE 개선을 보인다. 이는 제안된 방법이 기존 화면 내 예측을 위한 인트라 모드를 개선할 수 있다는 것을 보인다.

Table 4 Comparison of intra prediction between VVC and the proposed method

Video	MSE		improved rate
	VVC Intra modes	including proposed method	
desk	198.82	186.32	6.3%
kitchen	131.33	124.46	5.2%
meeting room	194.53	184.46	5.1%
desk	136.32	127.58	6.4%

5. 결 론

본 논문에서는 깊이 영상의 부호화에서 화면 내 예측을 위한 네트워크를 제안했다. 예측하고자 하는 블록의 상단과 좌단 영역에 있는 참조 화소로부터 수직 방향과 수평 방향의 공간적 특징들을 동일한 CNN 계층을 통해 검출되고, 특징 방향과 공간 방향으로 병합되었다. Attention 기법을 통해 참조 화소와 예측 블록 내 화소 간 상관성을 예측하고, 이를 통해 예측 블록의 공간적 특징을 예측하였다. 예측 블록 내 화소는

CNN 계층을 통해 공간적 특징을 화소 값으로 변환함으로써 예측된다. 본 논문에서 제안된 방법의 화면 내 예측 성능을 VVC 인트라 모드와 비교한 결과 기존 VVC 인트라 모드만 적용했을 때보다 평균 5.8%의 정확도 향상을 보였다. 제안된 방법은 기존 인트라 모드들이 직선 방향으로만 참조 화소를 예측할 수 있다는 한계에서 벗어나 비선형 방향의 참조를 통한 예측이 가능하다. 또한 제안된 방법은 기존 영상 부호화 표준 내 다수의 방향성 인트라 모드를 Attention 기법이 적용된 네트워크로 대체할 수 있다는 것을 보였다. 하지만 현재 연구에서는 동영상 부호화에 대한 연구는 부족하다는 한계가 있다. 화면 내 예측의 개선과 깊이 동영상의 부호화간의 관계에 대한 후속연구를 수행할 예정이다. 제안된 방법을 통해 깊이 영상의 주된 응용인 라이더 영상에서의 객체 검출이나 VR/AR 영역에서의 활용 등에서 고해상도의 깊이 영상의 저장 및 전송을 개선시킬 수 있다.

References

Aguilar, W. G., Rodríguez, G. A., Álvarez, L., Sandoval, S., Quisaguano, F. and Limaico, A. (2017). Visual SLAM with a RGB-D Camera on A Quadrotor UAV Using On-board Processing, *Proceedings of the Advances in Computational Intelligence: 14th International Work-Conference on Artificial Neural Networks*, June 14-16, Cadiz, Spain., pp. 596-606.

Bross, B., Wang, Y., Ye, Y., Liu, S., Chen, J., Sullivan, G. J. and Ohm, J. (2021). Overview of The Versatile Video Coding (VVC) Standard and Its Applications, *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10), 3736 - 3764.

Jiang, M. X., Luo, X. X., Hai, T., Wang, H. Y., Yang, S. and Abdalla, A. N. (2019). Visual Object Tracking in RGB-D Data via Genetic Feature Learning, *Complexity*, 4539410.

Kwon, S. K., Kim, H. J. and Lee, D. S. (2017). Face Recognition Method Based on Local Binary Pattern using Depth Images, *Journal of Korea Society of Industrial Information Systems*, 22(6), 39 - 45.

Kwon, S. K., Tamhankar, A. and Rao, K. R. (2006). Overview of H.264/MPEG-4 Part 10, *Journal of Visual Communication and Image Representation*, 17(2), 186 - 216.

Lee, D. S. and Kwon, S. K. (2022). Intra Prediction Method for Depth Video Coding by Block Clustering through Deep Learning, *Sensors*, 22(24), 9656.

Lee, D. S., Kim, B. G. and Kwon, S. K. (2021). Efficient Depth Data Coding Method Based on Plane Modeling for Intra Prediction, *IEEE Access*, 9, 29153 - 29164.

Lee, D. S. and Kwon, S. K. (2019). Vehicle Plate Detection Method by Measuring Plane Similarity Using Depth Information, *Journal of Korea Society of Industrial Information Systems*, 24(2), 47 - 55.

Li, Y. (2012). Hand Gesture Recognition Using Kinect, *Proceedings of the 2012 IEEE International Conference on Computer Science and Automation Engineering*, June 22-24, Beijing, Chian, pp. 196 - 199.

Li, Y., Miao, Q., Tian, K., Fan, Y., Xu, X., Li, R. and Song, J. (2016). Large-scale Gesture Recognition with A Fusion of RGB-D Data Based on The C3D Model, *Proceedings of the 23rd international conference on pattern recognition*, Dec. 4-8, Cancun, Mexico, pp. 25 - 30.

Nenci, F., Spinello, L. and Stachniss, C. (2014). Effective Compression of Range Data Streams for Remote Robot Operations Using H.264, *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 14-18, Chicago, IL, USA, pp. 3794 - 3799.

Ren, C. Y., Prisacariu, V. A., Kähler, O., Reid, I. D. and Murray, D. W. (2017). Real-time Tracking of Single and Multiple Objects from Depth-colour Imagery Using 3D Signed Distance Functions, *International Journal of Computer Vision*, 124, 80 - 95.

Ren, Z., Yuan, J., Meng, J. and Zhang, Z. (2013). Robust Part-based Hand Gesture Recognition Using Kinect Sensor, *IEEE transactions on multimedia*, 15(5), 1110 - 1120.

Oh, K. J., Han, D. H. and Kwon, S. K. (2018). Character Floating Hologram Using Detection of User's Height and Motion by Depth Image, *Journal of Korea Society of Industrial Information Systems,* 23(4), 33 - 40.

Silberman, N., Hoiem, D., Kohli, P. and Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images, *Proceedings of the 12th European Conference on Computer Vision*, Oct. 7-13, Florence, Italy, pp. 746 - 760.

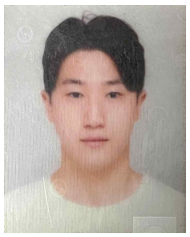
Stankiewicz, O., Wegner, K. and Domański, M. (2013). Nonlinear Depth Representation for 3D Video Coding, *Proceedings of the IEEE International Conference on Image Processing*, Sep. 15-18, Melbourne, Australia, pp. 1752 - 1756.

Sullivan, G. J., Ohm, J. R., Han, W. J. and Wiegand, T. (2012). Overview of The High Efficiency Video Coding (HEVC) Standard, *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12), 1649 - 1668.

Sun, Y., Liu, M. and Meng, M. Q. H. (2017). Improving RGB-D SLAM in Dynamic Environments: A Motion Removal Approach, *Robotics and Autonomous Systems*, 89, 110 - 122.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Ł. Kaiser. and Polosukhin, I. (2017). Attention is All You Need, *Proceedings of the Neural Information Processing Systems*, Dec. 4-9, Long Beach, CA, USA, pp. 5998-6008.

Zhao, Y., Carraro, M., Munaro, M. and Menegatti, E. (2017). Robust Multiple Object Tracking in RGB-D Camera Networks, *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sep. 24-28, Vancouver, Canada, pp. 6625-6632.

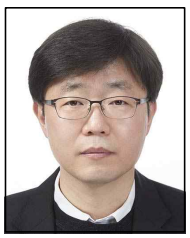


윤 재 혁 (Jae-hyuk Yoon)

- 동의대학교 컴퓨터소프트웨어공학과 공학사
- 동의대학교 컴퓨터소프트웨어공학과 석사과정
- 관심분야 : 영상딥러닝, 행동인식

이 동 석 (Dong-seok Lee)

- 정회원
- 동의대학교 컴퓨터소프트웨어공학과 공학석사
- 동의대학교 컴퓨터소프트웨어공학과 공학박사
- 동의대학교 인공지능그랜드ICT연구센터 연구교수
- 관심분야 : 멀티미디어 신호처리, 영상딥러닝



윤 병 주 (Byoung-ju Yun)

- 정회원
- 경북대학교 전자공학과 학사
- 한국과학기술원 전기및전자공학과 석사
- 한국과학기술원 전자전산학과 박사
- 경북대학교 IT대학 전자공학부 초빙교수(교수)
- 관심분야: 영상신호처리, 컴퓨터 비전, HDR 컬러영상향상, HCI 등

권 순 각 (Soon-kak Kwon)

- 정회원
- 경북대학교 전자공학과 공학사
- KAIST 전기및전자공학과 공학석사
- KAIST 전기및전자공학과 공학박사
- 동의대학교 컴퓨터소프트웨어공학과 교수
- 관심분야 : 영상딥러닝, IoT