비지도 학습 깊이 예측 모델을 이용한 가상시점 합성

*송민기 **양지희 ***황동호 **박구만

서울과학기술대학교 *일반대학원 미디어IT공학과, **나노IT디자인융합대학원
정보통신미디어공학전공, ***전자IT미디어공학과

*thdalsrl10@gmail.com

Virtual view synthesis using unsupervised learning depth estimation model

*Song, Min-Ki **Yang, Ji-Hee ***Hwang, Dong-Ho ***Park, Goo-Man

Seoul National University of Science & Technology *Graduate School, Dept. of Media IT Engineering, **Graduate School of Nano IT Design Fusion, Dept. of Information Technology and Media Engineering ***Dept. of Electronics and IT Media Engineering

요약

본 논문에서는 기존의 DERS, VSRS를 이용한 가상시점 합성이 가지고 있는 문제점을 해결하기 위해 비지도 학습 방식의 학습 모델을 이용하여 가상시점 합성에 적용하는 방식을 제안한다. 제안한 방식에서는 기존의 DERS와 달리 Disparity의 탐색 범위를 지정하지 않고 Depth의 예측이 가능하며 단안의 영상에서 Depth를 예측하기 때문에 가상시점 합성 시 더 넓은 시점을 합성 할 수 있다. 또한 기존 방식은 Depth와 합성 영상을 각각 처리해야하지만 제안하는 방식은 한 번에 작업이 이루어지며, GPU를 기반으로 구현하였기 때문에 기존의 합성 방식 보다 처리 속도가 우수하다.

1. 서론

3-DTV(3-D Television), FTV(Free view Television) 에서는 가상시점 합성을 이용한 이미지의 재구성으로 임의의 시점에서 영상을 제공함으로써 사용자와 상호작용이 가능한 실감영상 표출을 제공한다[1,2,3]. 이 외에도 이미지의 합성은 그래픽스, 자율주행 자동차 등 다양한 분야에서 중요한 문제점으로 꼽히고 있다[2, 3, 4, 5].

MPEG-FTV 에서는 이미지의 재구성을 위해 Depth 기반의 가상시점 합성 방식을 사용하는 DERS(Depth Estimation Reference Software), VSRS(View Synthesis Reference Software)를 제안하였다[6, 7]. 하지만 DERS와 VSRS를 이용하기 위해서는 데이터마다 결과 값을 도출하기 위해 실측 기반으로 Disparity range를 역산하는 불필요한 과정이 소요되고 3개의 시점을 이용하여 가운데 시점의 Depth를 예측하기 때문에 데이터 시퀀스 내에 양끝 2개의 시점에서는 Depth를 예측할 수 없다는 문제가 있다.

또한 딥러닝을 이용한 지도학습 기반의 Depth 예측 연구가 진행되었다[8, 9]. 하지만 기존 지도학습 방식에는 LIDAR와 같은 고가의 장비를 통해서 Ground Truth를 취득하기 때문에 비용적인 문제가 발생한다. 이 문제를 해결하기 위해 추가적인 장비 없이 이미지만을 이용하여 Depth를 예측하기 위한 비지도 학습 방식의 모델들이 제안되었다[10, 11, 12]. 지도학습은 Depth와 직접적으로 오차를 계산하지만 비지도 학습에서는 모델을 통해 나온 결과물을 이용하여 이미지를 재구성해 원본 이미지와 비교하는 방식으로 Depth 생성을 유도한다. 본 논문에서는 비지도 학습 방식을 적용한 Depth 기반의 이미지 합성 방식을 제안한다.

2. 본론

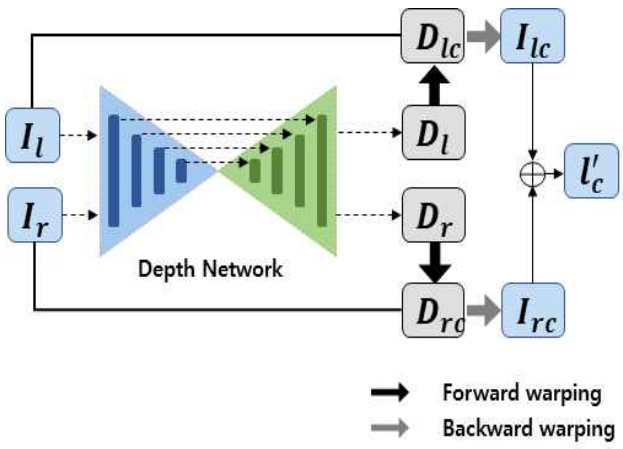

그림 1. View synthesis Architecture

그림 1은 본 논문에서 제안하는 모델의 구성도이다. I_l과 I_r은 각각 합성하기 위한 영상의 좌, 우 이미지를 나타내며, 네트워크를 지나면 각각의 Depth 영상인 D_l과 D_r이 생성된다.

$$P_c R_l^{-1}\left(K_l^{-1}\begin{bmatrix} x \\ y \\ D_{l_{xy}} \end{bmatrix} - t_l\right) = \begin{bmatrix} iD_{lc_{ij}} \\ jD_{lc_{ij}} \\ D_{lc_{ij}} \end{bmatrix} \quad \cdots (1)$$

여기서 P, R, K, t는 카메라 파라미터를 나타내며, 표현식은 다음과 같다.

$$P = K[R|t] \quad \cdots (2)$$

D_l과 D_r은 Forward warping을 통해 가운데 시점에서의 Depth (D_{lc}, D_{rc})로 변형된다. 식 (1)에서 x, y와 i, j는 각각 D_l과 D_{lc}의 이미지 좌표를 나타낸다. 이때, 생성되지 않은 픽셀에 대해서 Depth 이미지 크기의 hole 이미지(h_l, h_r)를 만든다.

I_{lc}와 I_{rc}은 각각 I_l, D_{lc}와 I_r, D_{rc}를 이용해 Backward warping으로 가운데 시점을 재구성한 이미지를 나타낸다[10, 11].

$$D_{mask_l} = \begin{cases} 1, \text{if } D_{lc_{ij}} \le D_{rc_{ij}} \\ 0, else \end{cases} \quad \cdots (3)$$

$$M = D_{lc_{ij}} \wedge D_{rc_{ij}} \quad \cdots (4)$$

$$m_l = h_r \vee D_{mask_l} \quad \cdots (5)$$

$$I_c' = I_{lc} \otimes (m_l - M) + I_{rc} \otimes (m_r - M) + M \otimes ((I_{lc} + I_{rc}) \times 0.5) \quad \cdots (6)$$

식 (3)은 Forward warping을 통해 생성된 Depth를 비교해 더 작은 값을 가지는 영역을 각각 D_{mask_l}와 D_{mask_r}으로, 같은 영역은 식 (4)의 M으로 나타내며, 반대편 이미지의 hole에 해당하는 부분을 합쳐 식 (5)의 m_l, m_r로 나타낸다. 최종적으로 I_c'는 식 (6)에 의해 생성된다.

$$L_{hole} = \frac{1}{N}\sum_{1}^{N}(h_l \wedge h_r) \quad \cdots (7)$$

$$pe(I_a, I_b) = 0.85(1 - SSIM(I_a, I_b)) + 0.15 \| I_a - I_b \|_1 \quad \cdots (8)$$

$$L_{smooth} = \frac{1}{N}\sum_{ij} |\partial_x D_{ij}| e^{-\|\partial_x I_{ij}\|} + |\partial_y D_{ij}| e^{-\|\partial_y I_{ij}\|} \quad \cdots (9)$$

hole loss, reconstruction loss는 식(7, 8)과 같이 정의 하였으며, 식 (9)는 기존에 제안된 disparity smoothness loss이다[10, 11]. 최종 loss function은 다음과 같다.

$$loss = L_{hole} + pe(I_c', I_c) + (pe(I_{lc} \otimes m_l, I_c \otimes m_l) + pe(I_{rc} \otimes m_r, I_c \otimes m_r)) + L_{smooth} + (\alpha - PSNR(I_c', I_c)) \quad \cdots (10)$$

Depth 예측 네트워크는 ResNet-18 Encoder를 갖는 U-Net 구조를 사용하였다[13].

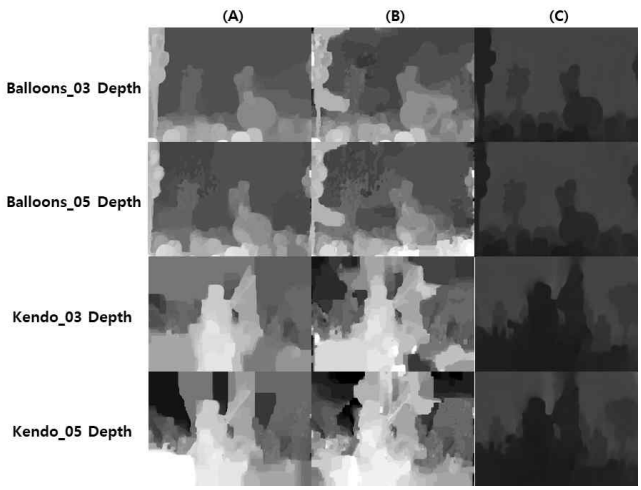

그림 2. Depth 생성 결과

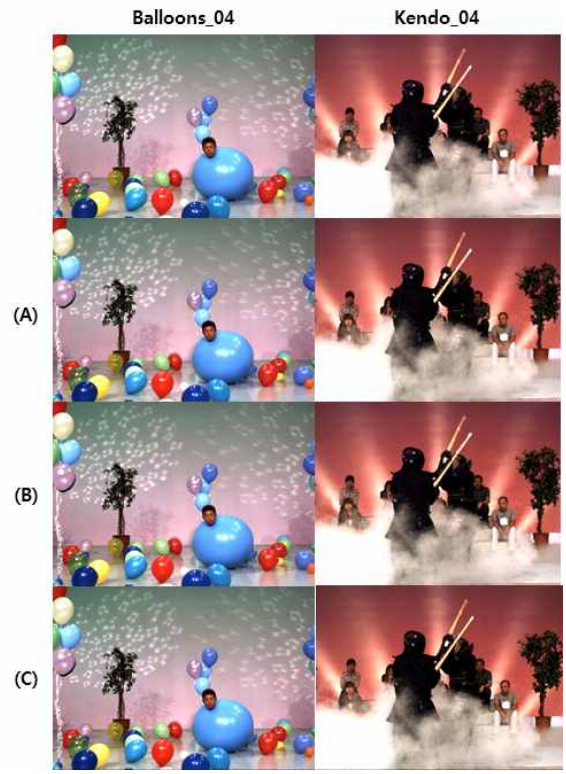

그림 3. 가상시점 합성 결과

3. 실험 결과

본 논문에서 제안하는 모델은 GTX 1080Ti 환경에서 Python, Tensorflow를 사용하여 구현하였으며, 학습 및 평가 비교를 위해 Ground truth data가 있는 Nagoya 대학의 Balloons, Kendo Sequence[14]를 이용하였다.

그림 2와 그림 3은 Nagoya에서 제공하는 이미지 시퀀스와 DERS, 및 VSRS, 제안하는 모델로 Depth와 시점을 각각 생성한 결과이며 표 1과 표 2 는 합성된 이미지와 원본을 비교하여 SSIM 과 PSNR을 측정한 표이다. (A)와 (B)는 각각 Nagoya에서 제공한 Depth와 DERS를

통해 생성된 Depth를 이용해 합성한 결과이며, (C)는 제안하는 모델로 합성한 시점에 대한 측정 결과이다. PSNR 다소 떨어지지만 SSIM은 비슷한 결과를 보이며 3개의 시점을 사용해 Depth를 생성하는 DERS와 달리 제안하는 모델은 단안의 영상에서 Depth를 예측하기 때문에 더 넓은 범위를 합성할 수 있다.

표 1. Balloons Sequence 실험 결과

	Frame	Method	Balloons01	Balloons02	Balloons03	Balloons04	Balloons05
(A)	300	SSIM	-	0.87517	-	0.92241	-
		PSNR	-	27.3809	-	33.3975	-
(B)	500	SSIM	-	0.92098	0.92667	0.92070	-
		PSNR	-	32.0898	33.1580	31.0029	-
(C)	500	SSIM	0.91331	0.91723	0.91727	0.92063	0.87472
		PSNR	29.7753	29.5269	29.3445	29.5227	27.4855

표 2. Kendo Sequence 실험 결과

	Frame	Method	Kendo01	Kendo02	Kendo03	Kendo04	Kendo05
(A)	300	SSIM		0.91754		0.93380	
		PSNR		32.4894		34.7713	
(B)	400	SSIM		0.92384	0.93713	0.93343	
		PSNR		31.9632	33.6659	31.7955	
(C)	400	SSIM	0.91841	0.91903	0.92184	0.92732	0.91338
		PSNR	28.1449	28.1798	28.4604	28.2739	27.96512

표 3은 DERS, VSRS와 제안된 모델의 생성 시간을 측정한 결과이다. (B)의 Depth 생성 시간은 1개의 시점에 대해 나타냈으며 하나의 가상시점을 합성하기 까지 12.380 sec가 소요 된다. 반면 제안된 방식에서는 GPU 기반으로 모델이 동작하기 때문에 2개의 시점에 대한 Depth예측과 가상시점 합성 까지 0.09533 sec가 소요 된다.

표 3. 생성 시간 측정 결과

	Depth 생성 시간	가상 시점 생성 시간	총 생성 시간
(B)	5.7493 sec	0.881202 sec	12.380 sec
(C)	-	-	0.09533 sec

4. 결론

본 논문에서는 비지도 학습의 Depth 예측 모델을 이용해 가상시점을 합성하는 모델을 구현하였다. 비지도 학습을 이용함으로써 기존의 Depth ground truth 데이터 수집에 대한 문제점을 극복하였다. 또한, 기존의 DERS가 데이터 마다 수동적인 작업을 거쳤던 부분을 줄였으며 양끝의 시점에 대해서도 depth 예측이 가능하다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2019년도 문화기술 연구개발 지원 사업으로 수행되었음. (R2017030041, 자유선택 시점에서의 문화 콘텐츠 감상 체험 극대화 기술)

참고 문헌

[1] 호요성, 윤승욱, 김성열, “실감 방송과 차세대 실감형 미디어”, TTA 저널, 제 100호

[2] S. Lu, T. Mu, and S. Zhang. “A survey on multiview video synthesis and editing”. Tsinghua Science and Technology, 21(6):678 - 695, 2016.

[3] C.-C. Lee, A. Tabatabai, and K. Tashiro, “Free viewpoint video (FVV) survey and future research direction,” APSIPA Trans. Signal Inf. Process.,vol. 4, Oct. 2015, Art. no. e15.

[4] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. In ECCV, 2016.

[5] R. Hartley, A. Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2000.

[6] M. Tanimoto, T. Fujii and K. Suzuki, “Improvement of Depth Map Estimation and View Synthesis”, ISO/IEC JTC1/SC29/WG11, M15090, January 2008.

[7] M. Tanimoto, T. Fujii and K. Suzuki, “Reference Software of Depth Estimation and View Synthesis for FTV/3DV”, ISO/IEC JTC1/SC29/WG11, M15836, October 2008.

[8] D. Eigen, C. Puhrsch, and R. Fergus. “Depth map prediction from a single image using a multi-scale deep network.” In NIPS, 2014.

[9] F. Liu, C. Shen, G. Lin, and I. Reid. “Learning depth from single monocular images using deep convolutional neural fields.” PAMI, 2015.

[10] Clement Godard, Oisin Mac Aodha, and Gabriel J Bros- ´ tow. “Unsupervised monocular depth estimation with left-right consistency.“ In CVPR, 2017.

[11] C. Godard, O. Mac Aodha, G. Brostow, "Digging into self-supervised monocular depth estimation", arXiv preprint arXiv:1806.01260, 2018.

[12] Ravi Garg, Vijay Kumar BG, and Ian Reid. “Unsupervised CNN for single view depth estimation: Geometry to the rescue.” In ECCV, 2016

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “UNet: Convolutional networks for biomedical image segmentation.” In MICCAI, 2015

[14] Nagoya University Multi-view Sequences, http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/