딥뉴럴 네트워크 기반 가상 축구 데이터셋을 이용한 단일 이미지 깊이 예측

신은섭 김유민 배성호
경희대학교
kairos9603@khu.ac.kr, rladbals0733@gmail.com, shbae@khu.ac.kr

Monocular Image Depth Estimation using Synthetic Soccer Data based on Deep Neural Networks

Eunseop.Shin Youmin Kim Sung-Ho Bae
Kyung Hee University

요 약

학습 기반 단일 이미지 깊이 예측은 최근 몇년간 활발하게 진행되어왔다. 효과적인 단일 이미지 생성 모델을 학습하기 위해서는 고화질 영상 및 이에 상응하는 깊이 맵에 대한 데이터 확보가 필수적이다. 본 논문에서는 축구 영상에 대한 단일 이미지 깊이 예측을 위한 딥 뉴럴 네트워크 모델을 제시한다. 현존하는 축구 영상 및 깊이 영상 데이터가 거의 없기 때문에, 이를 극복하기 위해 FIFA 게임 영상으로부터 획득한 합성 축구 영상 및 깊이 맵을 도메인 전이 기술을 통해 실제 영상과 같이 만들어 학습하는 방법을 제안한다. 즉, 제안 모델은 실제 축구 영상을 입력으로 받아 내부적으로 이를 합성 영상(FIFA 게임영상)으로 변환하고, 합성된 영상에 깊이 맵을 생성함으로써 데이터 부족을 효과적으로 해결한다. 실험 결과, 제안 방법은 실제 축구 영상으로부터 효과적으로 깊이 맵을 생성했다.

1. 서 론

만약 우리가 월드컵 결승을 눈앞에서 3D로 본다면 어떨까? 화면 속에서만 움직이던 선수들이 화면 밖으로 튀어나와 눈앞에서 돌아다닌다면 생생한 현장 분위기를 느끼며 더욱 축구에 빠져들어 열광할 수 있을 것이다.

단일 2D 영상 또는 다수개의 2D영상들로부터 3D 영상을 생성하는 기술은 이미 다양한 비전 알고리즘으로 구현이 되어 있다. Multi-View geometry[4] 알고리즘은 특정 공간 주위에 수십 개의 카메라를 설치한 뒤 카메라 영상을 동기화 하여 3D로 복구해 내는 방법이다. 실제로 이 기술은 Intel사의 True View[1]라는 이름으로 상용화 되어 있다.

3D 콘텐츠는 VR시장의 확대, 3D 디스플레이의 발전과 멀티미디어 방송 기술의 발전으로 수요가 꾸준히 증가하고 있다. 기존의 3D 콘텐츠를 만들기 위해서는 3D 카메라나, True View에서 사용하는 것과 같은 전문 장비가 필요했다. 그러나 이러한 전문장비들은 매우 비싸기 때문에 3D 영상을 제작, 보급하는데 큰 걸림돌이 되고 있다. 이를 해결하기 위해 전문 장비를 사용하지 않고 기존의 2D 콘텐츠를 3D 콘텐츠로 변환하려는 다양한 시도가 이루어 지고 있는데, 이 변환에서 가장 중요한 것이 바로 단일 영상 깊이 예측이다.

*본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음.(2017-0-00093)

기존의 깊이 예측 방법은 매우 다양하다. 대표적으로 스테레오 대응[5], 움직임 구조 파악[6], 빛과 그림자의 확산으로부터 깊이 측정[7] 등의 방법들이 있다. 그러나 위와 같은 대부분의 방법들은 몇가지 중요한 문제점을 가지고 있는데 대표적으로 깊이 불균형, 깊이 측정 불가(구멍, holes), 계산 복잡성, 지식기반의 특정 후처리 의존 등이 있다[8].

위 방법의 해결책으로서 단일 이미지를 사용한 딥 뉴럴 네트워크 방법이 최근에 많이 연구되었다. 가장 먼저 등장한 방법이 지도학습으로 이미지의 특징들과 깊이를 매핑하여 예측하는 방법이다[9]. 이러한 방법들은 실제 이미지에 해당하는 깊이 맵이 존재하여야 학습이 가능하다. 그러나 이미지와 1대1로 대응되는 깊이 맵을 갖는 대용량 데이터 셋을 만드는 것은 시간과 비용이 많이 소모되는 단점이 있다.

이와 다른 방법으로 비지도 학습법이 위의 단점을 해결할 방안으로 제시 되었다. 비지도 학습법은 이미지와 직접적으로 매핑 되는 깊이 맵 없이, 깊이 맵을 추정 가능한 다른 데이터를 이용하여 학습시키는 방법이다[10]. 이 방법은 실제 깊이 맵 없이, 쉽게 얻을 수 있는 2차 데이터를 이용하여 학습할 수 있는 장점이 있으나, 이미지 화질 저하, 깊이 불균형 등의 문제가 있다.

이후 실제 깊이 맵 데이터 셋이, 얻기 어렵고 매우 적은 것을 보완하고자 제안된 방법이 가상 데이터 셋을 이용하여

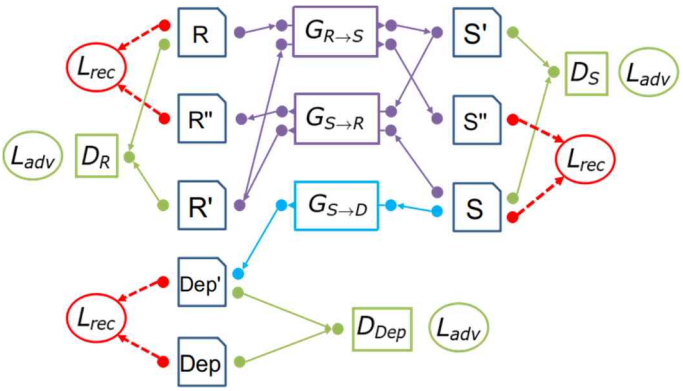

그림1. 제시하는 딥러닝 모델. 실제 이미지(r)에서 가상 이미지(s)로 도메인 변환을 한 후 깊이(dEP)를 예측한다. R, S, Dep는 원본 데이터, R`, S`, Dep`은 GAN을 통해 만들어 낸 데이터, R``, S``은 순환 재구성을 통해 재구성 한 데이터를 의미한다. 보라색 네트워크는 실제 이미지(R)과 가상 이미지(S)를 상호 변환하는 네트워크로 2.2에서 설명하고 있다. 하늘색 네트워크는 가상 데이터를 이용하여 깊이 맵을 예측하는 네트워크이고, 2.3에서 설명하고 있다.

학습 시키는 방법이다[11]. 게임이나 시뮬레이션을 통해 얻어진 가상 데이터는 수집 과정이 쉽고 다량의 데이터를 빨리 모을 수 있을 뿐 아니라 깊이 맵 측정 측면에서 직접 측정하는 것에 비해 매우 정확하다는 장점이 있다.

가상 데이터를 이용하여 딥 뉴럴 네트워크를 학습시키는 것은 기존에도 시도된 방법이지만[12], 깊이 예측 부분에서 도메인 적응은 한 도메인과 다른 도메인 사이 즉, 실제 이미지와 가상 이미지 간의 확연한 차이로 인해 학습이 잘 안되는 문제점을 가지고 있다. 이 문제를 극복하려는 다양한 시도가 있었는데 그 방법 중 하나로 이미지 스타일 변환을 통해 두 도메인 간의 차이를 줄이는 방법이 있다[13].

축구 데이터를 이용하여 깊이를 예측하려는 시도는 기존에도 있었다[14]. 기존의 연구들은 전경 및 후경 분리, 물체 인식, 물체에 따른 깊이 예측 하는 방법이 있었으며, 주로 전후경 분리 및 물체를 인식하는 다양한 비전 알고리즘을 사용하였다. 그러나 기존의 알고리즘과 마찬가지로 계산량이 매우 많은 것과 깊이 불균형 등의 문제점을 가지고 있다.

본 연구에서는 위의 문제점들을 극복하기 위한 새로운 게임 축구 데이터 셋을 이용하여 단일 사진의 깊이를 예측 및 3D 영상을 만들어주는 딥 뉴럴 네트워크 모델을 제시한다.

2. 제시하는 모델

본 논문에서 제시하는 모델은 그림1과 같다. 먼저 남색 귀퉁이가 잘린 사각형 부분이 2.1에서 설명 할 가상 데이터 셋을 이용한 부분, 보라색 부분이 2.2에서 설명할 이미지 스타일 변환, 하늘색 부분이 2.3에서 설명할 깊이 예측이다. 연두색부분은 각 생성 모델과 함께 학습되는 분별 모델이고, 빨간 타원은 Loss를 의미한다.

2.1 가상 축구 데이터 셋

이전의 몇몇 축구 영상 깊이 예측 연구에서 이미 가상 축구 데이터 셋을 사용하고 있었으며, 주로 Electric Art사의 FIFA 게임을 이용한 가상 데이터를 사용하였다.

Calagari et al.[15]에서는 Microsoft의 DirectX 디버깅 툴인 PIX[3]를 사용하여 FIFA 게임으로부터 화면과 깊이 데이터를 얻어 왔고, Rematas et al.[14]에서는 RenderDoc[2] 프로그램을 사용하여 FIFA 게임 엔진과 GPU 사이의 GPU Call을 캡쳐 하였고, GPU memory에서 깊이와 색상 버퍼를 프레임 별로 추출하였다. 우리는 위 논문에서 영감을 받아 FIFA 게임에서 추출한 화면 및 깊이 쌍 데이터 셋을 이용하여 딥 뉴럴 네트워크 모델을 학습시켰다.

2.2 이미지 스타일 변환

우리는 깊이 예측에 앞서, 가상데이터를 사용하여 깊이 예측을 하면 도메인 간의 차이로 인해 학습이 잘 되지 않는 문제점을 해결하기 위해 이미지 스타일 변환을 사용하였다. 스타일 변환 모델로 JY Zue et al.[16]의 Cycle GAN을 사용하여, 실제 데이터 셋을 가상 데이터 셋으로 변환하였다. 학습 모델은 ResNet 을 변형한 모델을 사용하였다.

모델의 목적함수로는 *Adversarial Loss, Cycle Loss, Identity Loss*를 사용하였다. 두개의 Generator가 학습되어야 하기 때문에 두개의 *GAN Loss*를 사용하였다. 먼저 실제 데이터를 가상 데이터로 바꾸는 G_{XtoY}, D_Y의 Loss는 다음과 같다:

$$L_{adv}(X \to Y) = \min_{G_{XtoY}} \max_{D_Y} \mathbb{E}_{y\sim\mathbb{P}_d(y)}[logD_Y(y)] + \mathbb{E}_{x\sim\mathbb{P}_d(x)}[log(1 - D_Y(G_{XtoY}(x)))] \quad (1)$$

$\mathbb{P}_d$는 데이터의 분포, X는 source domain, Y는 target domain을 나타낸다. x와 y는 각각 X의 샘플, Y의 샘플을 의미한다. G_{XtoY}, D_Y와 마찬가지로 G_{YtoX}, D_X의 *Loss*는 다음과 같다:

$$L_{adv}(Y \to X) = \min_{G_{YtoX}} \max_{D_X} \mathbb{E}_{x\sim\mathbb{P}_d(x)}[logD_X(x)] + \mathbb{E}_{y\sim\mathbb{P}_d(y)}[log(1 - D_X(G_{YtoX}(y)))] \quad (2)$$

Cycle GAN에서는 위에서 학습시킨 두개의 GAN을 모두 통과한 뒤 원래의 이미지와 같은지를 측정하는 *Cycle-consistency Loss*가 존재한다. 이 Loss는 Generator가 영상을 왜곡하지 않고 원하는 스타일로만 변환 하도록 해준다. 이 loss는 원본 샘플 x를 G_{XtoY}에 통과시켜 y' 을 만들고 다시 y' 을 G_{YtoX}에 통과시킨 뒤 나온 x'와 x를 비교한다. 수식은 다음과 같다:

$$L_{cyc} = ||G_{YtoX}(G_{XtoY}(x)) - x||_1 + ||G_{XtoY}(G_{YtoX}(y)) - y||_1 \quad (3)$$

또한, 원래 논문에서는 그림을 사진으로 변환하는 실험에서 *Identity loss*를 도입하였는데 이 것은 원본 X를 G_{YtoX}에 넣었을 때 외곡 없이 원본 X그대로 나오게 하기 위함이다. 이를 통해 의도치 않은 색의 변화를 줄일 수 있다. *Identity Loss*의 수식은 다음과 같다:

$$L_{identity} = ||G_{XtoY}(y) - y||_1 + ||G_{YtoX}(x) - x||_1 \quad (4)$$

따라서 최종적인 *Loss*는 다음과 같다:

$$L = L_{adv}(X \to Y) + L_{adv}(Y \to X) + \lambda_1 L_{cyc} + \lambda_2 L_{idntity} \quad (5)$$

<Real Image>	<Synthetic Image>	<Depth Map>

그림 2. 축구 영상 데이터 셋 모델 테스트 결과이다. 왼쪽부터 실제 이미지, 실제 이미지를 축구 게임 이미지로 변환한 가상 이미지, 변환된 가상 이미지로 부터 예측한 깊이 맵이다.

λ_1과 λ_2는 실험적으로 얻어진다. 모델의 Optimizer는 ADAM을 사용하였으며, Learning Rate = 0.001, Momentom =0.5, β_1 = 0.99를 사용하였다.

2.3 깊이 예측

깊이 예측 문제는 2.2절과 마찬가지로 Image to Image 변환 문제로 볼 수 있다. 2.2절과 같이 CycleGAN으로 문제를 해결할 수도 있지만, 일반적으로 일대일변환은 PIX2PIX와 같이 직접적으로 Input과 label을 비교하는 방법이 정확도가 더 높다.

우리는 실시간으로 깊이를 예측하기 위해 U-Net 모델을 사용하여 깊이 예측을 학습하였다. U-Net은 weight가 적은 가벼운 모델로 Image to Image변환에 자주 사용되어왔다. 모델의 *Loss*로는 2.2절의 (1)과 같은 *Adversal loss*와 생성된 깊이와 실제 깊이를 1:1로 비교하는 *Reconstruction loss*를 사용하였다.

$$L_{rec} = ||G_{XtoY}(x) - y||_1 \qquad (6)$$

최종 *Loss*는 다음과 같다:

$$L = L_{adv}(X \to Y) + \lambda_3 L_{rec} \qquad (7)$$

모델의 Down Convolution에서는 Leaky ReLU(slope=0.2)를, Up Convolution에서는 ReLU를 사용하였다. Optimizer는 ADAM을 사용하였으며, Learning Rate = 0.001, Momentom = 0.5, β_1 = 0.99를 사용하였다.

3. 실험 결과

모델에 실제 축구 영상 이미지(real)를 입력하면 실제 이미지를 가상 이미지(synthetic)로 바꿔준 후, 가상 이미지를 통해 깊이 맵을 구하는 결과이다. **그림 2**는 실제 수행 결과를 보이며, 제안 방법이 도메인 전이를 통해 효과적으로 실제 축구 영상에 대한 깊이 맵을 생성하는 것을 확인할 수 있다.

4. 결론

본 연구는 축구 영상 깊이 맵 데이터 부족 문제를 해결하기 위해 실제 이미지와 축구 게임 영상의 가상 이미지를 Cycle GAN을 통해 가상 이미지로 변환하도록 학습한 후, 축구 게임 영상의 가상 이미지와 해당 이미지의 깊이 맵을 통해 가상 이미지를 깊이 맵으로 변환시켜주는 모델을 학습하였다. 기존의 연구는 본 모델을 KITTI Dataset으로 학습하여 좋은 성과를 내었다. 본 연구는 본 모델에 축구 영상 데이터를 사용하여 어느 정도의 결과는 도출하였지만, 아직 완벽한 결과를 도출하지는 못하였다. 이러한 원인으로는 다음과 같이 추측할 수 있다.

- KITTI Dataset은 데이터의 View Point가 하나지만, 축구 데이터 셋은 View Point가 다양하기 때문에 학습이 더 어렵다.
- 실제 축구 영상 데이터는 축구 영상을 매 프레임 이미지들로 잘라낸 것들이기 때문에 경기 대진표나 점수 판, 동영상 마크 등 학습에 지장을 줄 수 있는 노이즈가 추가되어 학습이 어렵다.

본 연구는 앞으로 데이터를 정제하고 다양한 view point로 인해 발생할 수 있는 문제점들을 분석하여 더 좋은 결과를 낼 수 있도록 할 것이다.

참 고 문 헌

[1] Intel True View. https://www.intel.com/content/www/us/en/sports/technology/true-view.html.

[2] RenderDoc. https://renderdoc.org/.

[3] Microsoft PIX. https://blogs.msdn.microsoft.com/pix/.

[4] M. Germann, T. Popa, R. Keiser, R. Ziegler, and M. Gross. Novel-view synthesis of outdoor sport events using an adaptive view-dependent geometry. Eurographics, 2012.

[5] O. Hamilton and T. Breckon. Generalized dynamic object removal for dense stereo vision based scene mapping using synthesised optical flow. In Proc. Int. Conf. Image Processing, pages 3439-3443, 2016.

[6] P. Cavestany, A. Rodriguez, H. Martinez-Barbera, and T. Breckon. Improved 3d sparse maps for high-performance structure from motion with low-cost omnidirectional robots. In Proc. Int. Conf. Image Processing, pages 4927-4931, 2015.

[7] A. Abrams, C. Hawley, and R. Pless. Heliometric stereo: Shape from sun position. Proc. Euro. Conf. ComputerVision, pages 357-370, 2012.

[8] A. Atapour-Abarghouei and T. Breckon. Depthcomp: Real- time depth image completion based on prior semantic scene segmentation. In Proc. British Machine Vision Conference, 2017.

[9] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proc. Int. Conf. Computer Vision, pages 2650-2658, 2015.

[10] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In Advances in Neural Information Processing Systems, pages 730-738, 2016.

[11] A. Gaidon, Q.Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In Proc. Conf. Com- puter Vision and Pattern Recognition, pages 4340-4349, 2016.

[12] A. Ruano Miralles. An open-source development environ- ment for self-driving vehicles. 2017.

[13] A. Atapour-Abarghouei and T. P. Breckon. Real-Time Monocular Depth Estimation using Synthetic Data with Domain Adaptation via Image Style Transfer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[14] K. Rematas et al. Soccer on Your Tabletop. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[15] K. Calagari et al. Data Driven 2-D-to-3-D Video Conversion for Soccer. IEEE Transactions on Multimedia. 2018.

[16] J. Zhu et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ICCV. 2017.