컴퓨터비전 깊이추정 기술 동향 분석

1. 서론

컴퓨팅 기술 및 환경이 발달함에 따라 컴퓨터가 처리할 수 있는 범위와 한계가 높아졌다. 특히 자율주행 자동차나 사람의 얼굴을 인식하여 발열 체크하는 모습은 일상에서 쉽게 볼수 있다. 이렇듯 사람의 시각정보를 대신 할 수 있는 컴퓨터비전 분야는 딥러닝 기술이 발달함에 따라 함께 접목되어 계산 구조 및 데이터가 한결 빠르고 간단하게 개발이 되고 있다. 딥러닝 모델은 대게는 객체의 검출과 분류하는 모델을 평가하며 평가모델의 성능은 ImageNet Large Scale Visual Recognition Challenge (ILSVRC)등의 대회를 개최하여 평가된다. 사람의 인지율과 비슷한 성능을 낸 GoogLeNet[1]등의 모델이 합성곱 신경망(CNN)의 발전을 가속화 시켰으며 객체위치 검출에 있어서 R-CNN[2], Fast R-CNN[3], YOLO[4]로 발전되었다. 그러나 딥러닝을 접목한 연구의 경우 데이터의 질과 양이 모델 성능에 중요한 요인이 된다는 것은 오랜 시간동안 많은 연구를 통해 입증이 되었으며, 라벨이 필요한 지도학습의 경우 전처리에 소요되는 인력과 시간이 많이 걸린다는 단점이 있고 장비가 비싸고 가공하기 힘든 데이터 셋을 필요로 하는 경우 대량의 자료를 확보하기 어렵다. 따라서 양안 영상을 이용한 영상깊이추정의 단점을 보완해가며 단일영상을 이용한 깊이 추정으로 발전한 과정을 비지도 학습 논문을 기준으로 동향을 분석한다.

2. 영상 깊이 추정 기술 동향

가. 양안영상 깊이 추정(Stereo Depth Estimation)

깊이 Z값의 추정은 그림1에서 볼수있듯이 왼쪽영상 x-left와 오른쪽영상 x-right 사이의 disparity를 통해 구해진다는 것이 기본이 된다. 대표적으로 MC-CNN[5]은 컨볼루션과 ReLU의 여러 레이어를 쌓아 Feature를 학습하고 양쪽 영상을 묶어 여러층의 Fully Connected Layer들과 시그모이드로 similarity score를 구하는 방식으로 학습하였다[6]. 대표적인 학습 데이터로는 KITTI, Make3D, Cityscapes, NYU Depth v2 등이 있다. 데이터들은 레이저 및 광센서를 통하여 Ground Truth를 얻으며 특히 KITTI는 LiDAR 센서로 도로 주행에서 촬영된 데이터 셋이다.

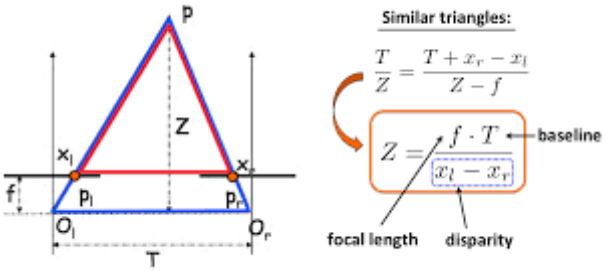

그림 1 양안 거리를 이용한 깊이 추정 공식

나. 단일영상 깊이 추정 (Monocular Depth Estimation)

양안 영상이 아닌 하나의 영상에서는 Camera pose, Segmentation 등 입체감과 관련한 주변 환경 요소를 고려하여 한계를 극복한다. Godard[7]는 라벨이 필요하지 않은 비지도학습 방식으로 깊이추정 연구를 진행했다. 학습 방식은 왼쪽영상을 입력값으로 하여 CNN을 통해 양쪽의 Disparity를 구한다. 구해진 Disparity를 통해 오른쪽 이미지를 출력값으로 결과를 얻을 수 있다. 사용한 Loss는 left-right disparity consistency loss, Appearance Matching Loss, Disparity Smoothness Loss 세 가지가 있다. 이는 depth의 라벨이 필요하지 않다는 장점이 있으나 반대쪽 영상으로부터의 픽셀을 샘플링 하므로 양쪽의 영상이 모두 필요하다는 단점이 있다.

본 수업에서 소개된 monodepth2[8]는 비지도학습을 유지한 채로 시간차 특성으로 Strero 영상처럼 사용한다는 특징을 가진다. 일반적인 Unet구졸르 따르며 Depth와 Pose 네트워크에서 standard ResNet18 인코더를 사용하였고 프레임쌍이 입력으로 사용되었기 때문에 첫 번째 컴볼루션 레이어에서 여섯 개의 채널로 수정되었다는 점에서 차이가 있다. 하나의 영상만을 사용했을 때 나타날 수 있는 문제들은 크게 세 가지로 해결하고자 했는데, Occluded 픽셀은 학습의 loss를 방해하는 요소가 되므로 Appearance loss를 사용하였으며 관계없는 카메라의 모션을 제거하는 Auto-Masking을 젱한하여 Depth-hole문제를 해결하고 예측된 작은영상을 키워서 Ground Truth와 비교하는 Multi-Scale을 사용한 loss를 제안하였다.

한편 이미지 기반의 비지도 학습 모델인 GAN을 활용한 연구도 등장하였다. Pix2Pix[9]는 GAN 기법의 구조를 갖는 모델로서 가상의 이미지를 생성해내는 생성자(Generator)와 진짜 이미지와 가상의 이미지를 비교하여 구별해내는 판별자(Discriminator)로 이루어졌다. Pix2Pix는 오코인코더와 같은 압축 기능을 하면서 엣지의 스머징효과를 줄이면서 객체 추출이 가능하다는 특징을 가진다. Pix2pix 모델을 활용한 단일 영상의 깊이맵 추출(Gang, Su Myung et al, 2019)[10] 논문에서는 계산 속도를 향상시키기 위하여 생성자의 컨볼루션 형태인 Unet 구조를 Depthwise컨볼루션으로 변경하여 파라미터수를 줄였다. Depthwise 컨볼루션은 일반적인 컨볼루션과 다르게 입력채널의 전체 정보가 아닌 단일 채널에 대해서만 수행하여 특정 채널만이 공간적 특징을 추출한다. 모식도는 그림2와 같다. 결과적으로 기존 구조의 성능을 유사하게 따르면서 계산 속도가 64%감소한 결과를 도출한다.

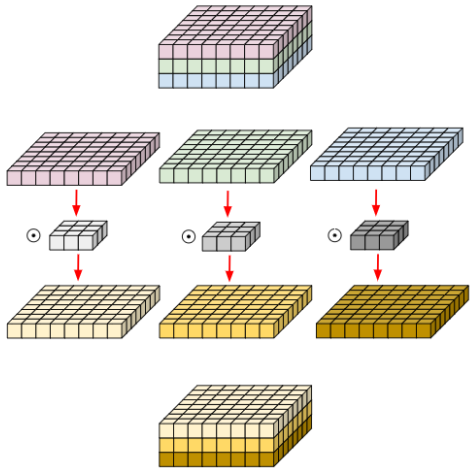

그림 2 Depthwise Convolution의 원리

3. 결론

컴퓨터비전의 발전은 딥러닝 기법이 적용됨에 따라 더욱 빠르고 정확하게 진행되었으며 영상깊이 추정연구에서 거리를 측정하기위하여 요구되었던 양안의 영상 대신에 단일영상만으로도 같은 성능을 낼 수 있는 모델 또한 개발되었다. 대부분의 딥러닝 모델의 개발이 CNN을 기반으로 발전했다는 특징이 있지만 최근에는 GAN을 이용한 단일 영상깊이 추정 연구가 있다는 점을 알게 되었다. KITTI데이터는 도로 위 주행영상으로 LiDAR를 통하여 Ground Truth영상을 얻은 데이터이다. 지도학습(Supervised)은 깊이 Depth에 대한 라벨이 필요로 하는 반면에 비지도 학습(Unsupervised, Self-Supervised)은 정답이 없어도 스스로 학습한 자는 장점이 있어서 효율적이다. 살펴본 논문을 통해 추후 적용 가능한 데이터를 탐구해보면 각 모델의 성능비교 및 장단점을 심도있게 체감 할 수 있을 것으로 기대된다.

4. 참고논문

[1] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[2] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[3] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[4] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[5] Zbontar, Jure, and Yann LeCun. "Stereo matching by training a convolutional neural network to compare image patches." J. Mach. Learn. Res. 17.1 (2016): 2287-2318.

[6] 김혜진, 지수영 "Depth Estimation 기술의 원리 및 동향" 정보기술융합공학논문지 9.1 pp.37-43 (2019) : 37.

[7] Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monocular depth estimation with left-right consistency." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[8] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[9] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[10] Gang, Su Myung, and Joon Jae Lee. "Depth map extraction from the single image using pix2pix model." Journal of Korea Multimedia Society 22.5 (2019): 547-557.