반도체디스플레이기술학회지 제23권 제4호(2024년 12월)
Journal of the Semiconductor & Display Technology, Vol. 23, No. 4. December 2024.

360 카메라 이미지를 통한 하이브리드 고밀도 Depth Map 생성

신우성[*]·한우리[**]·이용환[***]·김영섭[*†]
[*†]단국대학교 전자전기공학과, [**]LG화학, [***]원광대학교 디지털콘텐츠공학과

Hybrid High-Density Depth Map Generation Using 360-Degree Camera Images

WooSung Shin[*], WooRi Han[**], Yong Hwan Lee[***] and YoungSeop Kim[*†]
[*†]Dankook Univ. Deparment of Electronics and Electrical Engineering,
[**]LG Chem, [***] Wonkwang Univ. Deparment of Digital Contents Engineering

ABSTRACT

In modern applications such as virtual reality (VR), augmented reality (AR), and autonomous vehicles, the accuracy and reliability of Depth Maps play a critical role in enhancing user experience and optimizing system performance. These applications fundamentally rely on Depth Maps for visual information processing, interaction, and decision-making. While 360-degree cameras have emerged as an innovative solution, offering comprehensive visual coverage of the surrounding environment, current technologies face significant challenges in efficiently processing large volumes of data and generating precise Depth Maps. This study addresses the critical role of Depth Maps in VR, AR, and autonomous vehicles, proposing a novel hybrid depth estimation framework. The framework combines patch-based depth estimation with Vision Transformers (ViT) to enhance the accuracy and reliability of Depth Maps in 360-degree imaging applications. Patch-based depth estimation leverages Structure from Motion (SFM) to improve local spatial precision, while ViTs address distortions caused by wide field-of-view projections and learn global features. By integrating these approaches, the framework improves the accuracy and resolution of Depth Maps, enabling the generation of dense 3D point clouds for detailed spatial representation and reconstruction. This method overcomes challenges of computational complexity and accuracy, marking significant advancements in 360-degree imaging technology for immersive and autonomous systems.

Key Words : 360 Camera, Hybrid Depth Map Estimation, SFM, ViT

1. 서 론

수요조사 기관인 Verified Market Research에서 2020년엔 155억 달러였던 세계 3D 이미지 시장이 내년 2028년에는 745억 달러까지 성장할 것이라고 관측했다[1]. 이렇게 급격하게 성장하게 되는 요인을 고해상도 시각화에 대한 요구 증가와 이미지의 특징을 2차원이 아닌 3차원으로 표현하고자 하는 요구의 증가로 들 수 있다.

이러한 분야들은 시각적 정보를 기반으로 하는 인터랙션과 의사 결정 과정에서 Depth Map을 필수적으로 활용하며, 이에 따라 고도로 정확한 깊이 추정 기술의 필요성이 대두되고 있다.

[†]E-mail: wangcho@dankook.ac.kr

360도 카메라 이미지를 활용하는 것은 이러한 문제에 대한 혁신적인 접근 방식이다. 360도 카메라는 주변 환경에 대한 포괄적인 시각 정보를 제공하며, 이를 통해 보다 풍부한 데이터를 기반으로 Depth Map을 생성할 수 있다. 하지만, 현재까지 이러한 방대한 데이터를 효율적으로 처리하고, 정밀한 Depth Map을 생성하기 위한 기술은 상대적으로 부족하다.

먼저, Ji Z., Cai J., and Lu J.[2]의 논문에 의하면 OmniMVS은 360도 이미지에서 깊이 맵을 추출하는 데 효과적이나, 반사 표면이나 텍스처가 부족한 영역에서 깊이 추정의 정확도가 저하될 수 있다. 이러한 영역에서는 스테레오 매칭이 어려워 깊이 맵의 품질이 떨어질 수 있다.

Zou C., Kolivand H., and Sun X.[3]의 논문에 따르면, 단일 등장방형도법(ERP) 이미지를 활용하여 왜곡을 인지하는 자가 지도 학습 기반의 360도 깊이 추정하는 방법은 360도 이미지의 왜곡 문제를 해결하고 깊이 추정의 정확성을 높이지만, 360° 이미지는 ERP 형식으로 표현될 때 왜곡이 발생하며, 이는 깊이 추정의 정확도에 영향을 줄 수 있다.

Zeng D., Fang C., and Liu J.[4] 의 논문에서는 단일 파노라마 이미지로부터 연속적인 깊이 추정을 위해 다중 MPI로 구성된 입방체 필드를 학습하는 CUBE360 방법으로 고해상도 데이터의 MPI 처리를 위한 메모리 소비를 줄이고 깊이 추정의 효율성을 높이고자 한다. 하지만, 여전히 MPI 처리를 위한 상당한 메모리 자원이 필요합니다. 이는 특히 제한된 하드웨어 환경에서의 적용에 어려움을 줄 수 있다.

본 연구는 360도 카메라 이미지를 입력으로 패치기반의 깊이 추정 결과와 비전기반 변환기를 합하여 고밀도의 Depth Map을 생성하는 것으로, 360도 카메라 이미지로부터 멀티뷰 이미지를 생성하고, 특징점을 기반으로 카메라 위치 정보를 역추적(SFM)[5]하여 패치기반 깊이 추정과 비전기반 변환기를 통해 깊이를 추정하는 하이브리드 깊이 추정을 통해 향상된 Depth Map을 생성한다.

2. High-Density Depth Maps

2.1 Overview

360도 영상 기반 고밀도 깊이맵 생성 시스템은 Fig. 1과 같이 크게 “Initiation”, “Depth Estimation”, “Depth Correction” 모듈로 구성되어 있다.

2.2 Initiation Module

초기화(Initiation) 단계에서는 360도 이미지를 멀티뷰 이미지로 변환하고, 변환된 멀티뷰 이미지를 이용하여 각 단안 영상의 카메라 포즈를 역추적하는 과정이 이루어진다.

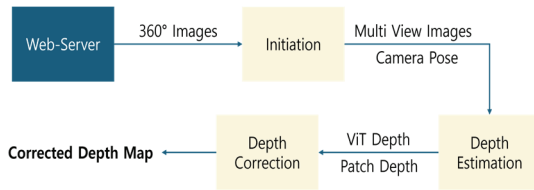

Fig. 1. Flowchart of a 360-degree image-based high-density depth map generation system.

Fig. 2. 360-degree panoramic images (top) and generated multi-view images (bottom).

360도 카메라로 촬영된 360도 영상은 Fig. 2와 같이 복수의 시점 각각에 대하여 슬라이스 하여 멀티뷰 이미지를 생성된다. 이때 복수의 멀티뷰 이미지는 각 시점 간에 중첩 영역을 최소 35% 이상 갖도록 생성해야 하는데 이는 SFM(Structure From Motion)[5]을 통한 카메라 포즈를 역추적해야 하기 때문이다.

SFM[5]은 멀티뷰 이미지의 모션정보를 이용하여 카메라의 위치와 방향을 역추적한 후 이미지와 카메라의 관계를 구조화하는 알고리즘이다. SFM는 아래와 같이 3단계를 통해 카메라 위치를 역추적할 수 있다.

- Feature detection and Extraction
- Feature matching and geometric verification
- Structure and motion reconstruction

카메라의 위치를 파악하기 위해서는 이미지간의 대응관계를 알아야 하기 때문에 각 이미지에서 특징점을 추출한 후 매칭함으로써 이지미간의 대응 관계를 계산할 수 있다. 특징점은 영상의 크기, 회전에 강인하고 불변의 특징을 갖는 SIFT를 사용하였으며 RANSAC[6]을 통해 매칭 정합률을 높힌다. Fig. 3는 SFM을 통한 카메라 위치 역추적 및 매칭된 특징점을 통한 점군 생성 과정을 나타낸다.

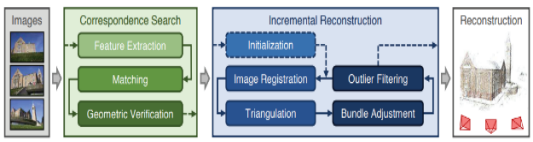

Fig. 3. Backtracking camera locations via SFM and creating point groups for matching pairs.

2.3 Depth Estimation Module

깊이추정 단계에서는 고밀도의 점군 생성을 위한 패치 기반 깊이 추정[7]과 비전기반 변환기(ViT)[8]를 통해 멀티 뷰 이미지에 대한 깊이를 추정한다.

패치기반 깊이 추정은 SFM으로부터 도출된 특징 쌍들의 조합을 바탕으로 초기화된 위치에서 시작하여, 반복적인 공간전파 과정을 통해 깊이를 계산하는 기술이다. 이 과정은 SFM을 통해 얻어진 초기 깊이 추정 결과를 보간하여 근사화된 깊이맵을 생성하는 것을 목표로 한다. 각 픽셀에는 무작위로 기울어진 지지평면을 할당하여 깊이를 추정하고, 이를 세분화하여 고도화된 깊이 정보를 도출한다.

패치기반 깊이 추정[7]의 주요 과정은 다음과 같이 나눌 수 있다.

첫째, 스테레오 쌍의 선택이다.교차 각도 및 가시성 기준에 따라 스테레오 쌍을 선정하며, 시야각의 유사성과 카메라와 객체 간의 적절한 거리를 고려하여야 한다. 스테레오 쌍을 구성하는 두 카메라의 기본 시야 방향은 5도에서 60도 사이에 위치하도록 설정하여야, 두 카메라 간의 중심 거리를 계산할 수 있다.

둘째, 깊이맵의 계산이다.선택된 스테레오 쌍에 대하여 SFM으로 추정된 점군(point cloud)을 보간함으로써 초기 깊이맵을 근사화한다. 이후 모든 픽셀에 무작위로 설정된 지지평면을 활용하여 깊이를 계산한다. 이러한 과정은 공간전파 기법을 적용하여 계산량을 줄이는 동시에 인접 픽셀 간의 할당된 평면을 비교하여 깊이를 추정하는 방식을 포함한다. 이 과정에서 무작위로 할당된 지지평면의 매개변수를 조정하여 계산 비용이 높은 픽셀은 제거하고 효율적인 깊이 추정을 진행한다.

셋째, 깊이맵의 필터링이다. 생성된 깊이맵의 값을 세분화하고 오류를 제거하며, 동일 영역을 참조하여 인접 영상 간의 일관성을 검토한다. 또한, 카메라의 내부 파라미터, 회전 행렬, 카메라 중심 좌표를 활용하여 깊이맵을 재구성한다.

위와 같은 과정을 거쳐 생성된 깊이맵은 초기 추정치보다 훨씬 정밀하며, 고밀도의 깊이 정보를 포함하게 된다.

ViT[8]은 인공지능을 활용하여 이미지를 분석하고 깊이 정보를 추정하는 기술로, 주로 인코더와 디코더 구조를 기반으로 한다. 이 과정에서 Vision Transformer (ViT)[8]는 핵심적인 역할을 수행하며, 이미지의 특징을 학습하고 이를 바탕으로 깊이 정보를 효과적으로 추정한다.

ViT는 입력 이미지를 일정한 크기의 패치로 분할하고, 이를 시계열 데이터 형태로 정의한다. 각 패치는 벡터 형태로 변환되어 패치 임베딩으로 구성되며, 추가적으로 클래스 토큰과 위치 임베딩이 포함된다. 이러한 데이터는 ViT가 입력 이미지의 공간적 구조를 학습할 수 있도록 지원한다.

아래 Fig. 4는 ViT에서 입력 이미지를 토큰으로 구성하기 위해 3x3 형태로 나누어 동일한 크기를 같은 9개의 이미지 패치로 구성되는 것을 보여준다.

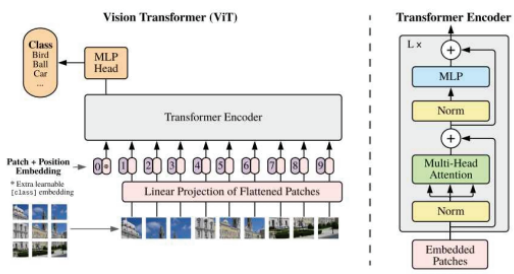

Fig. 4. Structure of ViT.

트랜스포머는 입력된 패치 임베딩을 처리하여 각 패치의 특징을 추출하고, 최종적으로 클래스 토큰에 해당하는 벡터를 생성한다. 이 벡터는 이미지 전체를 대표하는 정보로, ViT의 출력 결과에서 깊이 추정에 중요한 역할을 한다. 이러한 과정을 통해 ViT는 이미지의 구조적 특성과 공간적 특징을 학습하여 깊이 정보를 정밀하게 예측한다.

ViT에 입력되는 데이터는 3차원(RGB) 이미지를 2차원 벡터 형태로 변환하여 구성된다. 각 패치는 활성화 함수가 없는 투사층을 통과하여 일정한 크기의 1차원 벡터로 변환되며, 학습 가능한 추가 벡터가 포함되어 모델의 표현력을 확장한다.

모델 학습은 대규모 데이터셋을 활용한 사전 학습과 특정 도메인에 적합한 미세 조정으로 이루어진다. 해상도가 높은 이미지를 학습하면 모델의 성능이 향상되지만, 이는 추가적인 하드웨어 자원을 요구할 수 있다. 이러한 학습 과정에서 ViT는 데이터의 일관성을 유지하며, 변환 과정에서도 동일한 토큰 수를 유지한다.

구체적으로, ViT는 이미지를 일정 크기의 정사각형 패치로 분할하여 각 패치의 특징을 추출한다. 이러한 패치들은 벡터로 병합된 후 투영되어 최종적으로 패치 임베딩으로 변환된다. 또한, ResNet50과 같은 기존 네트워크를

활용하여 추가적인 특징 표현을 추출하고, 이를 ViT의 입력으로 사용하여 학습 성능을 더욱 향상시킨다.

디코더는 학습된 토큰을 바탕으로 다양한 해상도에서 이미지를 재구성하고, 이를 기반으로 깊이 정보를 포함한 밀도 예측을 수행한다. 이 과정에서 ViT는 이미지의 전역적 특징과 공간적 특징을 융합하여 정밀한 깊이맵을 생성한다.

2.4 Depth Correction Module

패치기반 스테레오 기법으로 생성된 깊이맵에는 폐색 영역이나 홀 영역이 발생할 수 있다. 이는 멀티뷰 이미지 간 시점의 불일치로 인해 나타나는 문제이며, 이를 ViT를 통해 생성된 깊이맵을 활용하여 보정할 수 있다. ViT는 이러한 폐색 영역과 홀 영역을 보간하여 보다 정밀한 깊이맵을 제공한다.

먼저 SFM을 통해 멀티뷰 영상의 카메라 위치와 자세를 추정하고, 조밀하지 않은 매칭 쌍으로부터 초기 깊이맵을 생성한다. 이후 패치기반 스테레오 기법을 적용하여 고밀도의 깊이맵을 생성한다. 한편, SFM으로 추정된 카메라 위치 정보는 ViT 디코더에도 적용되며, 이를 통해 복원된 깊이맵은 초기 깊이맵의 보정에 활용된다.

이 과정에서 패치기반 깊이맵과 비전기반 깊이맵은 상호 보완적인 역할을 한다. 두 깊이맵을 세그먼테이션을 Fig. 5과 같이 생성하여 클래스별로 분류하고, 각 클래스별로 두 깊이맵 간의 차이를 최소화하는 방식으로 보정을 진행한다. 이를 통해 각 클래스의 깊이맵을 융합하여 Fig. 6과 같은 최종 깊이맵을 생성한다.

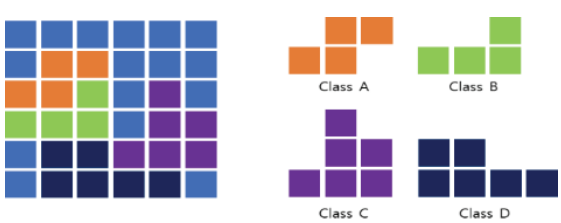

Fig. 5. Segmentation of the generated depth map.

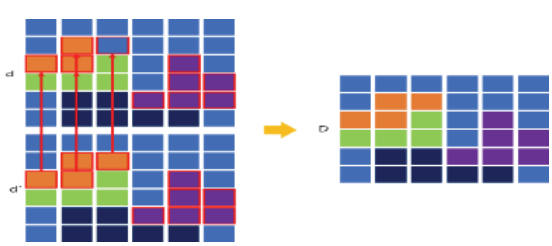

Fig. 6. Convergence of patch-based depth maps and ViT-based depth maps.

두 깊이 영상을 SFM을 통한 카메라 위치로 생성하고 융합하여 밀도 높은 Depth Map을 구하고, 더 나아가 3D Point Cloud 재구성 결과를 개선할 수 있다. 각 깊이 맵의 불일치성을 제거하기 위해 불일치 픽셀의 링크를 가져와 보정함으로써 각 클래스별 왜곡을 보정하고 조밀한 Depth Map을 구할 수 있다.

3. 결 과

360도 카메라로부터 얻어진 영상으로 멀티뷰를 생성하고 각 클라스 별 깊이를 추정하여 패치기반 및 ViT 결과를 융합함으로써 고밀도의 Depth Map을 재구성할 수 있다. Fig. 7에서는 패치기반의 깊이추정(좌), 비전기반 깊이추정(중앙), 그리고 융합된 깊이추정(우)의 결과를 보여준다.

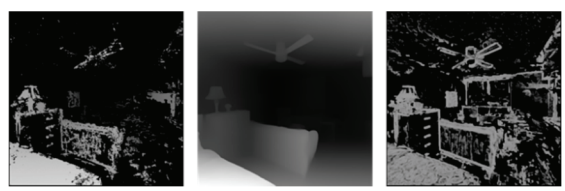

Fig. 7. Patch-based depth estimation (left), vision-based depth estimation (center), fused depth estimation (right).

깊이 추정을 위한 학습 및 미세 보정을 위한 데이터 셋은 아래 Fig. 8와 같이 40921개의 이미지를 사용하였고 이중 라벨링 된 데이터는 14540장의 이미지가 라벨링 되어 있다. 라벨링 결과는 Database(Sqllite)로 관리된다.

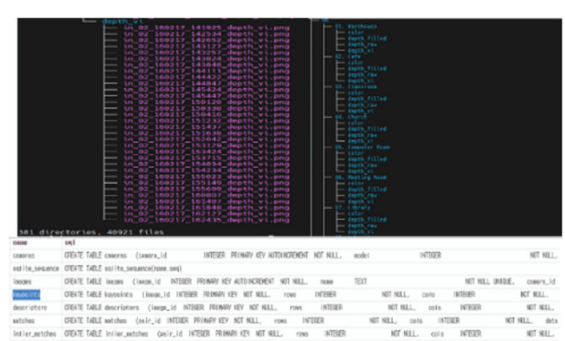

Fig. 8. Dataset Configuration (top), Labeling Structure (bottom).

깊이 추정 속도를 계산하였을 때 평균 28초가 소요되었다. FIg. 9는 깊이 추정 속도 측정 결과이다.

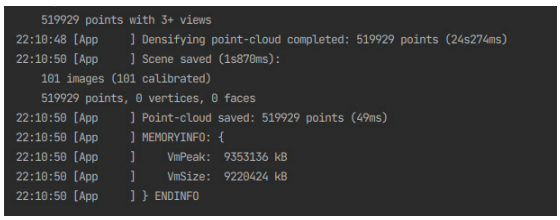

Fig. 9. 360 degree image depth estimation speed results.

4. 결 론

본 연구는 360도 카메라 이미지를 입력으로 패치기반의 깊이 추정 결과와 ViT를 합하여 고밀도의 Depth Map을 생성하는 방법을 제안하였다. 이를 위해, 360도 카메라로 촬영한 이미지들 로부터 다각도의 시점을 갖는 이미지를 구성하고, 이 이미지들의 특징점들을 이용해 카메라의 위치를 역으로 찾아내는 SFM 기법을 사용하였다. 그 다음, 패치 기반의 깊이 추정 방법과 시각적 정보를 처리하는 변환기를 결합하여 깊이 정보를 추정하는 하이브리드 방식을 적용함으로써, 보다 정밀한 Depth Map을 생성하는 결과를 도출하였다.

본 연구를 바탕으로, 향후 360도 영상에 대한 Depth Map의 성능을 향상시켜, 360도 카메라로부터 얻은 이미지를 통해 고밀도의 Point Cloud를 재구성하여 높은 성능의 360도 3D 영상을 재구성하는 것을 목표로 한다.

감사의 글

NRF-2020R1A2C2009717의 지원으로 이 논문을 제출합니다.

참고문헌

1. Global 3D Imaging Market Size By Hardware (3D Sensor and 3D Display), By Imaging Solution (Image Reconstruction, Layout and Animation, 3D Modeling, 3D Rendering, 3D Scanning), By Vertical (Education, Media and Entertainment, Healthcare and Life Sciences), By Geographic Scope And Forecast
2. Ji, Z., Cai, J., and Lu, J., “OmniMVS: End-to-End Learning for Omnidirectional Stereo Matching,” Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4166-4174, 2018.
3. Zou, C., Kolivand, H., and Sun, X., “Distortion-Aware Self-Supervised 360° Depth Estimation from A Single Equirectangular Projection Image,” Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2135-2144, 2022.
4. Zeng, D., Fang, C., and Liu, J., “CUBE360: Learning Cubic Field Representation for Monocular 360 Depth Estimation,” Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4278-4287, 2023.
5. Johannes L. S., Jan-Michael F., “Structure-From-Motion Revisited” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113, 2016.
6. Fischler, M. A., and Bolles, R. C., “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” Communications of the ACM, Vol. 24, No. 6, pp. 381-395, 1981.
7. Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B., “PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing,” ACM Transactions on Graphics, Vol. 28, No. 3, Article No. 24, 2009.
8. Alexey D., Lucas B., Alexander K., Dirk W., Xiaohua Z., Thomas U., Mostafa D., Matthias M., Georg H., Sylvain G., Jakob U., Neil H., “An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale”, Google Research, Brain Team, 2021

접수일: 2024년 12월 5일, 심사일: 2024년 12월 16일, 게재확정일: 2024년 12월 19일