OSMNet: 3D Octave Convolution을 통한 효율적 비용 정합*

전영민[0], 신동군
성균관대학교 전자전기컴퓨터공학과
mkjsym@g.skku.edu, dongkun@skku.edu

OSMNet: Efficient Cost Aggregation with 3D Octave Convolution

Youngmin Jeon[0], Dongkun Shin
Department of Electrical and Computer Engineering, Sungkyunkwan University

요 약

딥러닝 기반 스테레오 매칭 분야에서 PSMNet(Pyramid Stereo Matching Network)은 비용 정합(Cost Aggregation) 단계에서 Stacked-hourglass 구조를 사용하여 높은 정확도를 달성하였다. 이 구조는 반복적인 3D 컨볼루션 및 3D 디컨볼루션을 통해 다중 스케일(multi-scale) 특징을 효과적으로 학습하지만, 연산 비용이 매우 크다는 한계점을 가진다. 본 논문에서는 이러한 한계를 극복하기 위해, PSMNet을 기반으로 연산 효율성을 크게 개선한 OSMNet(Octave Stereo Matching Network)을 제안한다. OSMNet은 비용 정합 단계에서 표준 3D 컨볼루션을 3D 옥타브 컨볼루션(3D Octave Convolution)으로 대체한 옥타브-모래시계(Octave-hourglass) 모듈을 도입하여, 연산량을 줄이면서도 효과적인 다중 스케일 특징 학습을 가능하게 하였다. KITTI 2012 데이터셋을 이용한 실험에서, 제안하는 OSMNet은 3-pixel error 0.847, 추론 시간(Latency) 84ms를 달성하였다. 이는 베이스라인 모델인 PSMNet 대비 정확도는 약 10% 향상되었고, 연산량(FLOPs)은 38% 감소했으며, 추론 시간은 29% 단축된 결과로, 제안 모델의 우수한 효율성과 성능을 입증한다.

1. 서 론

스테레오 매칭은 인간의 양안 시각 원리를 모방하여, 서로 다른 두 위치에서 촬영된 스테레오 영상(좌, 우 영상)으로부터 깊이 정보를 추출하는 컴퓨터 비전 기술이다. 스테레오 매칭을 통해 3차원 공간 정보를 담고 있는 깊이 맵(depth map)을 생성할 수 있으며, 이는 로보틱스, 자율주행, 3D 모델링 등 다양한 분야에서 중요하게 활용된다.

최근 딥러닝 기술의 발전은 스테레오 매칭 분야의 성능을 획기적으로 향상시켰다. 초기에는 컨볼루션 신경망(CNN)을 이용하여 스테레오 영상으로부터 특징을 추출하고 깊이 맵을 추정하였으며, 이후 비용 볼륨(cost volume)을 구성하고 이를 3D 컨볼루션 네트워크로 처리하는 방식들이 등장하며 높은 정확도를 달성하였다.

그러나 높은 정확도에도 불구하고, 3D 컨볼루션 기반의 딥러닝 스테레오 매칭 방식들은 본질적인 한계를 지닌다. 첫째, 3D 컨볼루션 연산 자체의 막대한 계산 복잡도와 메모리 요구량은 실시간 처리나 자원이 제한된 임베디드 환경에서의 배포를 어렵게 만든다. 표준 3D 컨볼루션은 특징 볼륨 내에 존재하는 다양한 공간 주파수(spatial frequency) 정보를 구분 없이 동일하게 처리함으로써, 특히 변화가 적은 저주파 성분 정보 처리에 있어 연산 및 메모리 자원을 비효율적으로 사용하는 문제를 안고 있다.

이러한 비효율성 문제를 해결하기 위해 본 연구는 기존에 2D CNN을 위해 제안된 옥타브 컨볼루션(Octave Convolution) 기법에 주목한다. 하지만 기존 옥타브 컨볼루션은 2D 컨볼루션 연산에 초점이 맞추어져 있어, 3D 컨볼루션이 핵심인 스테레오 매칭의 비용 정합 단계에 직접 적용하기 어렵다. 본 논문에서는 이러한 한계를 극복하고 효율적인 3D 비용 정합을 통해 스테레오 매칭 성능을 개선하는 새로운 네트워크인 OSMNet(Octave Stereo Matching Network)을 제안한다. 이를 위해 기존 2D 옥타브 컨볼루션 개념을 3D 컨볼루션에 맞게 확장한 3D 옥타브 컨볼루션을 설계하였으며, 이를 기반으로 연산 효율성을 극대화한 새로운 비용 정합 모듈인 '옥타브-모래시계(Octave-hourglass)' 모듈을 제안한다.

2. 배 경

2.1 딥 러닝 기반 스테레오 매칭

최근 딥러닝 기반 스테레오 매칭 연구는 Vision Transformer(ViT)와 같은 최신 아키텍처를 도입하여 벤치마크 성능을 지속적으로 경신하고 있다. 그러나 이러한 ViT 아키텍처 기반의 SOTA(State-of-the-Art) 모델[1]은 1억 개 이상의 파라미터를 가지며 막대한 계산량과 메모리를 요구한다. 이는 고성능 GPU 환경이 필수적일 뿐만 아니라, 실시간 처리가 중요한 로보틱스나 모바일 기기와 같이 자원이 제한된 환경에 배포하기 어렵다는 실질적인 한계를 가진다.

PSMNet(Pyramid Stereo Matching Network)[2]은 3D

* 이 논문은 삼성전자 미래기술육성센터의 지원을 받아 수행된 연구임(과제번호 SRFC-IT2402-15)

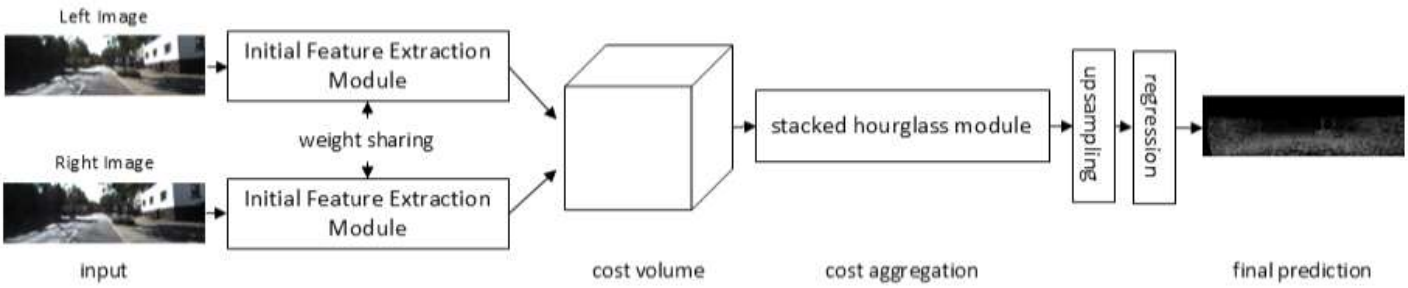

그림 1. PSMNet 구조

컨볼루션 기반의 비용 정합(Cost Aggregation) 방식을 사용하여 높은 정확도를 달성한 대표적인 모델로, 이후 많은 연구에서 비교 기준으로 활용될 만큼 그 구조와 성능이 널리 검증되었다. PSMNet은 여전히 경쟁력 있는 성능을 제공하지만 3D 컨볼루션 기반 비용 정합 과정의 높은 계산 복잡성 및 메모리 비효율성 문제가 존재한다.

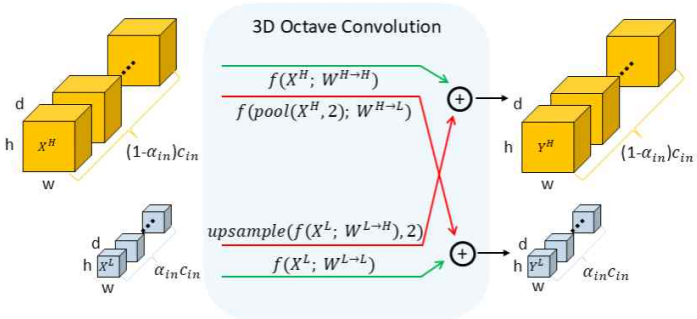

그림 2. 3D 옥타브 컨볼루션 구조

2.2 다중 스케일 비용 정합

PSMNet과 같은 3D 컨볼루션 기반 비용 정합(Cost Aggregation) 방식의 주요 병목은 깊게 쌓인 3D 컨볼루션 레이어 자체에서 발생한다. 이 표준적인 3D 컨볼루션은 비용 볼륨(Cost Volume) 내의 모든 위치(공간 및 시차축) 정보를 동일한 해상도에서 균일하게 처리한다. 하지만 비용 볼륨 내에는 상대적으로 변화가 적은 저주파(low-frequency) 성분과 객체 경계나 텍스처와 같이 복잡한 고주파(high-frequency) 성분이 혼재하는데, 저주파 성분까지 고해상도로 처리하는 것은 불필요한 계산 중복성(computational redundancy)을 야기할 수 있다.

이러한 문제점을 해결하기 위해, 본 연구는 2D 컨볼루션 신경망의 연산 효율성을 높이기 위해 제안되었던 옥타브 컨볼루션(Octave Convolution)[3]의 원리를 3D 비용 정합 과정에 적용하고자 한다. 특징 정보를 주파수 대역별로 분리하여 저주파 성분을 효율적으로 처리하는 방식을 3D 공간으로 확장하여 비용 정합 과정에 적용한다면, 정확도에 필수적인 고주파 정보는 유지하면서도 전체 계산량과 메모리 사용량을 크게 최적화할 수 있다.

3. OSMNet

3.1 Octave-hourglass 모듈

옥타브-모래시계(Octave-hourglass) 모듈은 기본적인 인코더-디코더 형태의 모래시계 골격을 유지하되, 내부의 모든 표준 3D 컨볼루션 및 3D 디컨볼루션(Deconvolution) 레이어를 각각 새롭게 설계된 3D 옥타브 컨볼루션 및 3D 옥타브 디컨볼루션 레이어로 대체하였다.

이 모듈 내부에서 3D 특징 볼륨(feature volume)은 명시적으로 고주파(High-frequency, H) 성분과 저주파(Low-frequency, L) 성분으로 분리되어 처리된다. 핵심적인 차이는 3D 옥타브 컨볼루션/디컨볼루션 연산을 통해, 저주파 성분(L)은 해당 모듈 단계의 고주파 성분(H)보다 각 차원(높이 H, 너비 W, 깊이/시차 D)에서 절반의 해상도, 즉 전체적으로 1/8의 볼륨(volume)을 가지는 낮은 3차원 해상도(1/2H×1/2W×1/2D)에서 처리된다는 점이다. 이는 상대적으로 변화가 적은 저주파 정보 처리에 필요한 연산량과 메모리 사용량을 크게 절감하는 효과를 가져온다. 이러한 옥타브-모래시계 모듈 설계를 통해, 기존 모래시계 구조의 다중 스케일(multi-scale) 비용 정합 능력은 효과적으로 보존하면서도 비용 정합(Cost Aggregation) 단계의 계산 복잡도(FLOPs)를 효과적으로 감소시킬 수 있다.

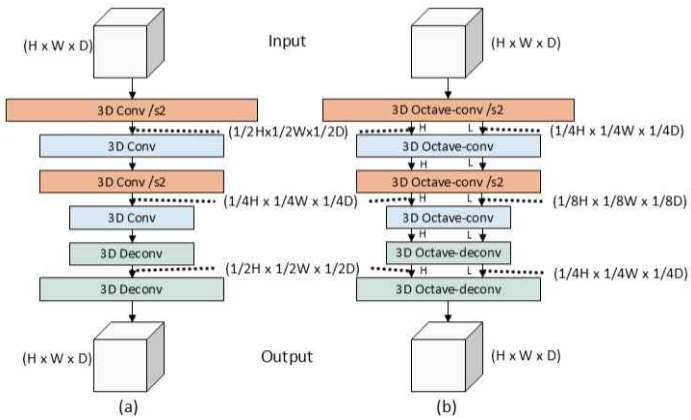

그림 3. (a) Hourglass 구조, (b) Octave-hourglass 구조

3.2 OSMNet 구조

본 논문의 핵심 기여는 계산 비용이 가장 많이 소요되는 3D 비용 정합(Cost Aggregation) 단계를 최적화한 것이다. 이를 위해 3D 옥타브 컨볼루션(3D Octave Convolution)을 기반으로 하는 새로운 '3D 옥타브-모래시계(3D Octave-hourglass)' 모듈을 설계하고, 이를 비용 정합 단계에 적용하였다.

초기 특징 추출 단계의 효율성을 높이기 위해, PSMNet에서 사용된 깊은 잔차(residual) 2D 컨볼루션 블록의 일부를 파라미터 수와 연산량(FLOPs) 측면에서 더 효율적인 포인트와이즈 컨볼루션(pointwise convolution)과 그룹 컨볼루션(group convolution)의 조합[4][5]으로 대체하였다. 포인트와이즈 컨볼루션은 채널 간의 상호작용을 담당하고, 그룹 컨볼루션은 채널 그룹 내에서 공간적 특징을 추출함으로써, 적은 비용으로 효과적인 특징 표현을 학습하도록 설계하여 초기 특징 추출 단계의 계산 부담을 줄였다.

결과적으로 OSMNet은 베이스라인 모델(PSMNet)의 강력한 다중 스케일 비용 정합 능력은 효과적으로 보존하면서도, 전체 모델의 파라미터 수와 연산량을 현저히 감소시켜 경량화 및 추론 속도 향상을 달성하였다.

4. 실험 및 결과

4.1 실험 환경

본 논문에서 제안하는 OSMNet의 성능을 평가하기 위해 주로 KITTI-2012 데이터 세트를 사용하였다. 모든 훈련 및 평가에 사용된 하드웨어 사양은 AMD Ryzen Threadripper 2950X 16-Core Processor CPU, NVIDIA GeForce RTX 3090 GPU이다. 모델 구현 및 실험에는 PyTorch 딥러닝 프레임워크를 사용하였으며, 모델의 추론 시간을 측정할 때 입력 이미지의 크기는 1216x352, 배치 크기는 16으로 설정하였다.

표 1. 기존 PSMNet 및 OSMNet의 성능 비교

Method	Parameters (M)	FLOPs (T)	3-pixel error (%)	Latency (ms)
PSMNet	5.22	1.68	0.942	118
OSMNet	4.97	1.04	0.847	84

4.2 실험 결과

먼저 모델 복잡도 측면에서, OSMNet의 파라미터 수는 약 497만개로 PSMNet의 약 522만개 대비 약 4.8% 감소하였으며, 연산량(FLOPs)은 1.68G에서 1.04G로 약 38% 감소하였다. 이는 본 논문에서 제안한 초기 특징 추출 단계의 효율화와 특히 3D 옥타브 컨볼루션을 활용한 옥타브-모래시계 모듈이 연산량을 효과적으로 절감했음을 보여준다. 정확도 측면에서도 OSMNet은 3-pixel error 0.847을 기록하여 PSMNet의 0.942 대비 더 낮은 에러율, 즉 더 높은 정확도를 달성하였다. 이는 연산량 감소를 위한 모델 구조 변경에도 불구하고, 옥타브-모래시계 모듈이 다중 스케일 정보를 효과적으로 정합하여 성능 저하 없이 오히려 정확도가 향상되었음을 시사한다. 마지막으로 추론 속도 측면에서 OSMNet의 Latency는 84ms로 측정되어 PSMNet의 118ms 대비 약 29% 감소하였다. 이는 감소된 연산량(FLOPs)이 실제 추론 시간 단축으로 이어진 결과로 해석할 수 있다.

5. 결 론

본 논문에서는 PSMNet을 기반으로, 3D 옥타브 컨볼루션(3D Octave Convolution)을 활용하여 비용 정합(Cost Aggregation) 과정의 효율성을 개선한 새로운 스테레오 매칭 네트워크인 OSMNet을 제안하였다. OSMNet의 핵심인 옥타브-모래시계(Octave-hourglass) 모듈은 기존의 표준 3D 컨볼루션 기반 모래시계 모듈보다 적은 연산량(FLOPs)으로도 효과적인 다중 스케일(multi-scale) 특징 정합을 가능하게 한다. 실험 결과를 통해 제안하는 OSMNet이 베이스라인 모델인 PSMNet 대비 훨씬 적은 연산량과 파라미터 수를 가지며, 이를 통해 더 빠른 추론 속도를 달성함을 확인하였다. 동시에, 이러한 효율성 증대에도 불구하고 경쟁력 있는 수준의 정확도를 달성함을 입증하였다.

참 고 문 헌

[1] Junda, C., et al. Monster: Marry monodepth to stereo unleashes power. arXiv preprint (arXiv:2501.08643). (2025).

[2] Chang, J., et al. Pyramid stereo matching network. CVPR (pp. 5410-5418). (2018).

[3] Chen, Y., et al. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3435-3444). (2019).

[4] Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. CVPR (pp. 14600-14609). (2020).

[5] Andrew, G, H., et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint (arXiv:1704.04861). (2017).

neural networks (pp. 92-101). Berlin, Heidelberg: Springer Berlin Heidelberg. (2010, September).