Journal of Sensor Science and Technology
Vol. 33, No. 2 (2024) pp. 98-104
http://dx.doi.org/10.46670/JSST.2024.33.2.98
pISSN 1225-5475/eISSN 2093-7563

컨볼루션 뉴럴 네트워크와 키포인트 매칭을 이용한 짧은 베이스라인 스테레오 카메라의 거리 센싱 능력 향상

박병재[1,+]

Improving Detection Range for Short Baseline Stereo Cameras Using Convolutional Neural Networks and Keypoint Matching

Byungjae Park[1,+]

Abstract

This study proposes a method to overcome the limited detection range of short-baseline stereo cameras (SBSCs). The proposed method includes two steps: (1) predicting an unscaled initial depth using monocular depth estimation (MDE) and (2) adjusting the unscaled initial depth by a scale factor. The scale factor is computed by triangulating the sparse visual keypoints extracted from the left and right images of the SBSC. The proposed method allows the use of any pre-trained MDE model without the need for additional training or data collection, making it efficient even when considering the computational constraints of small platforms. Using an open dataset, the performance of the proposed method was demonstrated by comparing it with other conventional stereo-based depth estimation methods.

Keywords: Stereo camera, Depth image, Monocular depth estimation

1. 서 론

서비스 로봇이나 드론과 같은 크기가 작은 플랫폼을 이용한 여러 가지 서비스가 등장하고 있다. 작은 플랫폼으로 여러 가지 서비스를 제공하기 위해서는 복잡한 환경 내에서 주변 환경에 대한 지도를 작성하고 장애물 탐지를 할 수 있어야 한다. 서비스 로봇이나 드론과 같은 플랫폼은 크기가 작고 제공할 수 있는 파워가 제한적이기 때문에 라이다와 같이 크기가 큰 액티브(active) 3차원 센서 보다는 스테레오 카메라와 같은 소형의 패시브(passive) 3차원 센서가 주로 사용된다. 스테레오 카메라의 경우 크기가 작고 파워 사용이 적은 반면 풍부한 뎁스(depth)를 제공할 수 있기 때문에 크기가 작은 플랫폼에서 장착되어 지도 작성이나 장애물 탐지에 이용될 수 있다. 그렇지만 스테레오 카메라 만을 사용하여 복잡한 환경에서 지도 작성이나 장애물 탐지를 하는데 있어 몇 가지 이슈가 존재한다. 첫 번째로 스테레오 카메라는 주변 환경의 조명 변화에 영향을 받아 이미지의 퀄리티가 저하될 수 있다. 이를 보완하기 위해서 스테레오 카메라의 노출을 자동으로 제어하는 방법이나 조명 변화에 의해 저하된 이미지를 보정하는 방법이 제안되어 왔다 [1,2]. 두 번째로 소형 플랫폼에 탑재되는 스테레오 카메라는 짧은 베이스라인을 가지고 있기 때문에 탐지 거리가 짧다. 정교한 스테레오 매칭 방법을 사용하는 경우 짧은 베이스라인을 가진 스테레오 카메라의 탐지 범위를 길게 만드는 것이 가능하다.

정교한 스테레오 매칭을 위한 다양한 방법들이 제안되어 왔다. 최근 들어서는 컨볼루셔널 뉴럴 네트워크(CNN)을 이용한 스테레오 매칭 알고리즘 방법이 많이 제안되고 있다 [3,4]. 그렇지만 CNN 기반의 스테레오 매칭 알고리즘의 경우 많은 컴퓨팅 파워가 필요하며 모델을 학습시키기 위해 많은 데이터가 필요하다. 그럼에도 불구하고 모델을 학습시키기 위해 사용한 데이터가 수집된 환경과 다른 환경에서 모델을 인퍼런스 할 때 성능 저하가 심하다는 한계가 있다.

본 논문에서는 소형 플랫폼에 탑재하기 위한 짧은 베이스라인 스테레오 카메라(SBSC)의 짧은 탐지 거리를 해결하기 위해 CNN과 전통적인 스테레오 매칭을 함께 사용하는 방법을 제안한다. 제안하는 방식은 두 단계로 이루어 진다: 1) CNN 기반의 모노 뎁스 추정(MDE) [5]를 이용하여 초기 뎁스를 계산한다; 2) SBSC의 좌우 영상으로부터 sparse visual keypoint들을 추출

[1]한국기술교육대학교 기계공학부 (School of Mechanical Engineering, Korea University of Technology and Education)
1600, Chungjeol-ro, Byeongcheon-myeon, Dongnam-gu, Cheonan-si, Chungcheongnam-do, 31253, Korea
[+]Corresponding author: bjp@koreatech.ac.kr
(Received: Mar. 7, 2024, Revised: Mar. 14, 2024, Accepted: Mar. 21, 2024)

한 다음 이들에 대한 triangulation을 하여 초기 뎁스의 스케일 보정을 수행한다. 제안하는 방법에서 초기 뎁스를 구할 때 CNN 기반의MDE를 사용하는 것은 CNN 기반의 스테레오 매칭을 사용하는 것에 비해서 몇가지 장점이 존재한다: 1) CNN 기반의 MDE는 CNN 기반의 스테레오 매칭 보다 모델의 크기가 작기 때문에 더 적은 컴퓨팅 파워를 요구한다: 2) CNN 기반의 MDE는 학습에 사용한 데이터가 수집된 환경과 다른 환경에서 인퍼런스 할 때 성능 저하가 CNN 기반의 스테레오 매칭에 비해서 적다.

그렇지만 CNN 기반의 MDE를 이용하여 구한 초기 뎁스의 경우 한 가지 중요한 문제가 존재한다. CNN 기반의 MDE를 이용하여 구한 초기 뎁스는 환경의 상대적인 뎁스이기 때문에 스케일이 맞지 않는다. 주변 환경의 지도를 작성하고 장애물 탐지를 하기 위해서는 절대적인 뎁스 정보가 필수적이다. 스케일을 보정하기 위해서 CNN 기반의 MDE 모델을 학습할 때 카메라의 파라미터를 함께 사용하는 방법도 제안되었지만 학습에 사용한 데이터가 수집된 환경과 다른 환경이나 다른 카메라를 사용하여 인퍼런스 하는 경우 스케일이 맞지 않는 문제가 발생한다.

제안하는 방법에서는 CNN 기반의 MDE를 이용하여 초기 뎁스를 구한 다음 스케일을 보정하기 위한 스케일 팩터를 SBSC의 좌우 영상으로부터 sparse visual keypoint들을 이용하여 구한다. 초기 뎁스를 구할 때 제안하는 방법에서는 이미 학습 된 CNN 기반이 MDE를 off-the-shelf 방식으로 사용한다. CNN 기반의 MDE는 학습에 사용된 데이터가 수집된 환경이나 다른 카메라를 사용하여 인퍼런스 하더라도 상대적인 뎁스는 안정적으로 추정하는 것이 가능하다. 추정된 상대적인 뎁스를 절대적인 뎁스로 변환하기 위한 스케일 팩터를 구하기 위해sparse visual keypoint들의 3차원 좌표를 이용한다. 구해진 스케일 팩터를 사용하여 초기 뎁스의 스케일을 보정하여 절대 뎁스를 얻을 수 있다.

제안한 방법을 이용하면 SBSC의 탐지범위를 확장시켜 SBSC만으로 소형 플랫폼이 더 먼 영역까지 지도작성이나 장애물 탐지가 가능할 수 있도록 할 수 있다. 제안하는 방법이 CNN 기반의 스테레오 매칭에 비해 가지고 있는 추가적인 장점으로는 이미 학습된 CNN 기반의 MDE를 off-the-shelf로 사용하기 때문에 소형 플랫폼이 가진 컴퓨팅 파워나 적용되는 환경에 따라 적합한 모델을 추가적인 데이터 수집이나 학습 없이 사용할 수 있다는 것이다.

본 논문은 다음과 같이 구성된다. 2장에서는 CNN 기반의 스테레오 매칭과 MDE 에 대한 관련 연구를 리뷰한다. 3장에서는 제안하는 방법의 전체 파이프라인에 대해 설명한다. 4장에서는 실험 결과를 통해 제한한 방법의 설명을 검증하고 5장에서는 결론 및 향후 과제에 대해 설명한다.

2. 관련 연구

2.1 스테레오 카메라를 이용한 깊이 추정

스테레오 매칭은 스테레오 카메라의 왼쪽과 오른쪽 이미지 사이의 disparity를 구하는 것이다. 뎁스는 구해진 disparity와 스테레오 카메라의 파라미터를 이용하여 구할 수 있다. 전통적인 스테레오 매칭 방법은 스테레오 카메라의 왼쪽과 오른쪽 이미지에서hand-crafted 피쳐를 이용하는 것이다 [6,7]. 이러한 방법은 적은 컴퓨팅 파워만을 사용하면서 실시간으로 disparity를 구하는 것이 가능하다. 그러나 이러한 방법은 복잡한 전처리 및 후처리 과정과 hand-crafted 피쳐를 다루기 위한 세부적인 파라미터 조정이 요구되어 다양한 환경에 적용하기 어려운 한계점이 있다.

CNN 기반의 스테레오 매칭 방법은 전통적인 스테레오 매칭 방법과 달리 hand-crafted 피쳐를 사용하지 않고 모델을 데이터를 기반으로 학습시킨다 [3,4,8]. 이러한 방법은 전통적인 방법에 비해 더 정교한 disparity를 구할 수 있다. 그렇지만 이러한 방법은 전통적인 방법에 비하여 몇 가지 한계점이 있다. 첫 번째로 모델의 학습을 위하여 많은 양의 주의 깊게 수집된 테이터를 필요로 하며 많은 컴퓨팅 파워가 요구된다 [9,10]. 두 번째로 학습된 모델을 이용하여 인퍼런스 할 때 역시 전통적인 방법에 비해 훨씬 많은 컴퓨팅 파워를 요구한다 [3,8]. 세 번째로 학습된 모델을 off-the-shelf로 여러 도메인에서 사용할 수 없다. 모델의 학습을 위해 사용한 데이터가 수집된 환경이나 카메라와 다른 환경이나 카메라에 모델을 적용하여 disparity를 구할 경우 성능의 저하가 심하다. 다른 도메인에서 모델을 적용하기 위해서는 추가적인 데이터를 수집하여 모델을 재학습 시키거나 fine-tunning을 해야만 한다.

2.2 모노 깊이 영상 추정

MDE는 CNN이나 Transformer [11]와 같은 뉴럴 네트워크(NN) 모델을 이용하여 이미지로부터 뎁스를 직접 구하는 것이다. NN 모델을 학습시키기 위해서는 1) supervised, 2) unsupervised, 3) image translation세 가지 접근 방법을 이용할 수 있다. Supervised 접근 방법에서는 이미지와 뎁스 페어로 이루어진 데이터셋을 이용하여 NN 모델을 학습하여 NN 모델이 이미지를 입력 받아 뎁스를 예측할 수 있도록 한다 [12-14]. 그러나 이미지와 뎁스 페어로 이루어진 데이터를 실외에서 수집하기 위해서는 고가의 장비 및 복잡한 후처리 과정이 필요하다는 한계가 있다. Ground truth 뎁스를 얻기 위해서는3D LiDAR가 필수적으로 필요하며 3D LiDAR의 density가 이미지 보다 낮기 때문에 이를 해결하기 위해 여러 위치에서 수집한 3D LiDAR의 point cloud를 registration하는 후처리 작업을 수행해야만 한다 [9].

Unsupervised 접근 방법은 supervised 접근 방법과 달리 ground truth 뎁스가 NN 모델의 학습 과정에서 필요하지 않은 방법이다 [5]. Unsupervised 접근 방법에서는 스테레오 카메라를 탑재한 플랫폼을 이동시키면서 수집한 데이터를 학습에 사용한다. 이를 위해서 left-right consistency라는 개념을 사용한다 [5]. 이는 스테레오 카메라의 캘리브레이션 파라미터를 알고 있고NN

모델이 뎁스를 잘 예측할 수 있다면 한 쪽 카메라 이미지와 뎁스 예측 결과를 가지고 다른 쪽 카메라에서 본 view로 변환한 이미지와 다른 쪽 카메라 이미지와 유사할 것이라는 가정을 바탕으로 한 것이다. 그러나 이 접근 방법 역시 데이터 수집 과정에서 정교하게 캘리브레이션한 카메라를 이용해야 한다는 한계가 있다.

Image-translation 접근 방법 [15,16]은 앞의 두 가지 접근 방법과 달리 generative model을 사용하여 이미지로부터 뎁스를 예측한다. 이 접근 방법에서는 cycle consistency [17]라는 개념을 사용하여 이미지와 뎁스 사이의 명시적인 페어가 존재하지 않는다 해도 이미지와 뎁스 사이의 변환을 할 수 있는 NN 모델을 학습시킬 수 있다. 그러나 이 접근 방법은 앞의 두 가지 접근 방법보다 뎁스 예측의 정확도가 떨어지며 복잡한 loss 함수와 모델의 구조 때문에 안정적으로 학습시키기 어렵다는 한계가 있다.

세 가지 접근 방법 모두 데이터가 수집된 환경과 유사한 환경이나 데이터를 수집할 때 사용한 카메라와 유사한 카메라를 사용해서 얻은 이미지로부터 NN 모델이 절대적인 뎁스를 예측하지만 다른 환경이나 다른 카메라를 사용해서 얻은 이미지로부터 NN 모델이 상대적인 뎁스는 정확하게 예측하나 절대적인 뎁스는 예측하지 못한다는 한계를 가지고 있다.

3. 파이프라인

제안하는 방법은 SBSC의 제한적인 탐지범위를 확장하기 위해 Fig. 1과 같은 파이프라인을 이용한다. Freeze한 MDE 모델을 이용하여 구한 unscaled depth의 scale을 보정하기 위해서 좌우 영상을sparse depth를 이용한다. Sparse depth는 좌우 영상에서 동시에 보이는 keypoint를 추출하여 추출된 keypoint에 해당하는 depth만을 구한다.

3.1 모노 뎁스 추정

좌측 영상으로부터MDE 모델을 사용하여 unscaled depth를 구한다. 이 과정에서 사용하는 MDE 모델인 BTS [12]는 KITTI 데이터셋을 이용하여 학습된 모델의 파라미터를 freeze 하여 재학습 없이 사용하였다. MDE를 재학습 없이 인퍼런스를 하는 경우 인퍼런스 하는 환경과 카메라가 MDE가 학습에 사용한 데이터를 수집한 환경과 카메라의 유사도가 뎁스 추정 결과에 영향을 미친다. 만일 인퍼런스 하는 환경이나 카메라가 학습에 사용한 데이터를 수집한 환경이나 카메라와 유사하지 않다면 재학습 없이 인퍼런스를 통해 추정한 뎁스는 스케일이 맞지 않는다.

3.2 저밀도 뎁스 추정

MDE 모델을 이용하여 구한 unscaled depth의 스케일을 보정

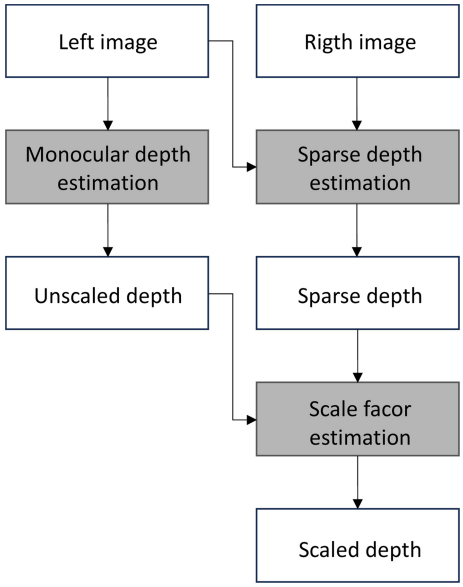

Fig. 1. Pipeline

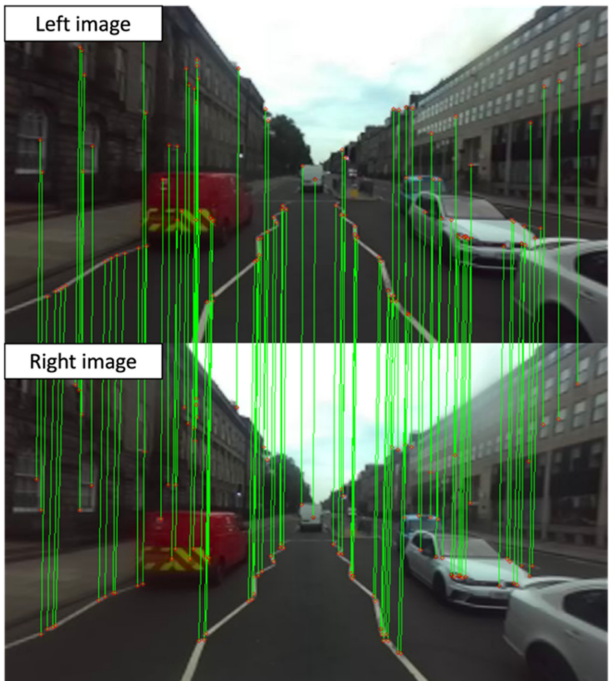

Fig. 2. Sparse stereo matching: upper figure – left camera image, lower figure – right camera image.

해주기 위해 SBSC의 좌우 영상과 파라미터를 이용하여 구한 sparse depth를 이용한다.

Sparse depth를 구하기 위해서 먼저 좌우 두 영상으로부터 keypoint detector를 이용하여 좌우 영상 각각keypoint와 keypoint의 descriptor를 추출한다. 이 과정에서 keypoint detector로 FAST [17]를 이용하였고descriptor로 AKAZE [18]를 이용하였다. Keypoint는 영상 내에서 골고루 분포하는 것이 scale feactor를 더 정확하게 추정할 수 있도록 할 수 있기 때문에 bucketing을

사용한다 [19]. Bucketing은 영상을 일정한 크기의 grid로 분할한 다음 한 개의 grid cell에서 추출될 수 있는 keypoint의 최대 숫자를 제한하여 영상의 특정 영역에서 keypoint가 과도하게 추출되지 않도록 억제한다.

Bucketing을 사용하여 추출한 좌측 영상의keypoint와 우측 영상의 keypoint 사이의 matching을 수행한다. Keypoint matching은 keypoint의 descriptor의 유사도를 이용하여 왼쪽 영상의 keypoint들과 우측 영상의 keypoint들의 tentative match를 찾은 다음 epipolar geometry를 적용하여 신뢰할만한 match만을 남겨둔다.

왼쪽 영상의 keypoint들과 오른쪽 영상의 keypoint들의 신뢰할만한 match를 구한 다음 match 되는 왼쪽 영상과 오른쪽 영상의 keypoint들과 SBSC의 파라미터를 이용하여 3차원 triangulation을 수행한다.

이 과정을 통해 왼쪽 영상의 sparse depth를 구할 수 있다. 이는 Stereo matching을 통하여 얻은 depth 보다 훨씬 sparse 하다. 그러나 Sparse한 keypoint의 위치에 해당하는 depth는 stereo matcing을 한 것보다 더 정확할 수 있다.

3.3 스케일 팩터 추정

MED 모델을 이용하여 구한 unscaled depth의 스케일을 보정하기 위한 scale factor를 sparse depth를 이용하여 구한다. 이를 위해서 우선 sparse depth 데이터 포인트 $p_i^{sd} = [r_i, c_i, d_i^{sd}]^T$ 에서 d_i^{sd} 에 대응하는 unscaled depth $p_j^{ud} = [r_j, c_j, d_j^{ud}]^T$ 의 d_j^{ud} 의 매칭 페어를 찾는 것이다.

Sparse depth와 unscaled depth의 매칭 페어를 찾는 과정에서 두 가지 이슈가 있다. 첫 번째 이슈는sparse depth데이터 포인트의 r_i 와 c_i 의 값은 sup-pixel 레벨의 부동소수점으로 표현되지만 unscaled depth 데이터 포인트의 r_j 와 c_j 의 값은 pixel 레벨의 정수로 표현된다는 것이다.

두 번째 이슈는 MDE에 의해 구해진 unscaled depth 데이터 포인트의 p_j^{ud} 의 값이 환경 내의 오브젝트 형태에 따라 오차를 포함할 수 있다는 것이다. MDE 모델의 uncertainty를 추정하기 위하여 MC-Dropupt을 적용한 결과를 볼 때 [20] p_j^{ud} 의 값은 오브젝트의 경계나 표면이 불규칙한 오브젝트의 포면에서 오차가 커지는 것을 확인할 수 있다.

두 가지 이슈를 해결하기 위하여 p_i^{sd} 와 p_j^{ud} 의 매칭 페어를 찾을 때 (1) sub-pixel 레벨로 표현되는 r_i 와 c_i 주변의 일정 범위의 윈도우를 linear interpolation을 이용하여 remapping한 패치를 추출한 다음 (2) 추출한 패치에서 가장 평면에 가까운 픽셀 좌표에 해당하는 d 값을 p_j^{ud} 의 값으로 사용한다.

$$p_j^{ud*} = \arg\min_{r_j, c_j} \nabla d^{ud} = \arg\min_{r_j, c_j} \left\{ \left(\frac{\partial d^{ud}}{\partial r_j} \right)^2 + \left(\frac{\partial d^{ud}}{\partial c_j} \right)^2 \right\}^{\frac{1}{2}} \quad (1)$$

위 식에서 r_j, c_j 의 범위는 윈도우 사이즈에 의해 결정된다. 이 과정을 통해 p_i^{sd} 와 p_j^{ud} 의 매칭 페어를 찾을 수 있다.

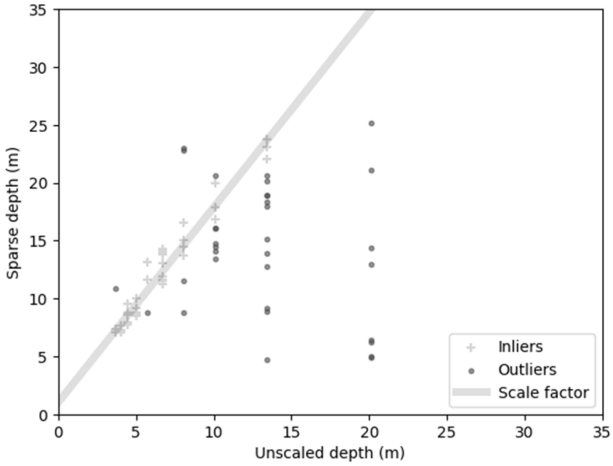

Fig. 3. Unscaled depth and sparse depth matching.

Sparse depth 데이터 포인트와 unscaled depth데이터 포인트의 매칭 페어를 찾은 다음 적절한 스케일 팩터를 찾는 과정을 수행한다. Sparse depth와 unscaled depth 사이에는 바이어스가 있는 선형 관계가 있다고 가정한다.

$$d_i^{ud} = \theta_0 + \theta_1 \cdot d_j^{sd} \quad (2)$$

스케일 팩터에 대한 파라미터 $\theta = [\theta_0, \theta_1]^T$ 는 RANSAC과 least squares를 이용해 구한다. Fig. 3은 unscaled depth의 데이터 포인트와 scaled depth데이터 포인트의 매칭 페어를 바탕으로 스케일 팩터의 파라미터를 추정한 결과이다.

3.4 스케일 보정

스케일 팩터에 대한 파라미터 θ를 매 프레임마다 구한 다음 이를 이용하여 unscaled depth에 적용하여 MDE의 depth의 스케일을 보정할 수 있다.

스케일 팩터를 추정하는 과정에서 영상 내에 텍스쳐가 부족하여 keypoint가 충분히 추출되지 않거나 keypoint가 균일하게 추출되지 않는 경우 스케일 팩터에 대한 파라미터가 제대로 추정되지 않을 수 있다. 이에 대응하기 위하여 스케일 팩터에 median 필터를 적용하여 사용한다.

4. 실험 결과

제안하는 방법을 이용하여 SBSC의 제한적인 탐지거리를 확장할 수 있다는 것을 정성적 또는 정량적으로 보여주기 위하여 실제 환경에서 수집된 오픈 데이터셋을 이용한 실험을 수행하였다. 실험을 수행하기 위해 제안된 방법은 Python과 PyTorch를 이용하여 구현 되었으며 Nvidia RTX 3070 (8Gb)를 탑재한 노트북에서 구동되었다. 모노 뎁스 추정을 위해서는 KITTI [9]

데이터셋을 이용해 학습된 ResNet50 backbone 기반의 BTS 모델을 사용하였으며 재학습을 하지 않았다.

4.1 거리 센싱 능력 향상 결과

SBSC를 장착한 플랫폼으로 수집된 오픈 데이터셋 중 하나인 RADIATE 데이터셋 [21]을 이용하여 제안하는 방법의 성능을 정성적 또는 정량적으로 검증하였다.

RADIATE 데이터셋은 승합차에 SBSC인 ZED 카메라와 차원 라이다와 레이다를 함께 탑재하여 도심 및 교외 지역을 다양한 날씨 조건에서 주행하면서 수집한 데이터이다. ZED 카메라는 약 12 cm의 베이스라인을 가지고 있으며 rectification을 다음 입력으로 사용하였으며 입력에 사용한 좌우 영상의 크기는 672×372 이다.

제안하는 방법의 성능을 정성적으로 비교하기 위하여 SBSC에서 뎁스를 추정할 수 있는 전통적인 스테레오 매칭 방법인 SGBM [6]과 CNN기반의 스테레오 매칭 방법인 PSMNet [3]을 함께 사용하였다. PSMNet은 KITTI 데이터셋을 이용해 학습된 모델을 사용하였다.

Fig. 4는 RDIATE 데이터셋 중 하나인 city_1_1의 한 프레임을 이용하여 SGBM, PSMNet 그리고 제안하는 방법을 이용하여 깊이 추정을 수행한 결과를 보여준다. Fig. 4 (b)는 SGBM을 이용하여 뎁스 추정을 한 결과를 보여준다. 뎁스 추정 결과를 볼 때 전체적인 구조를 어느 정도 나타내고 있지만 상당히 많은 artifact가 존재하며 유사한 뎁스를 가지고 있는 평면에서도 뎁스가 불균일하게 추청되는 것을 확인할 수 있다. 또한 뎁스 추정 결과 그림에서 왼쪽 부분은 뎁스를 제대로 추정하지 못하는데 이는 SGBM이 매칭 과정에서 양쪽 영상의 오버랩이 충분하지 못한 경우 뎁스 추정을 수행하지 못하기 때문이다.

Fig. 4 (c)는 PSMNet을 이용하여 뎁스 추정을 한 결과를 보여준다. KITTI 데이터셋과 RADIATE 데이터셋은 둘 다 비슷한 도시 환경에서 수집되었음에도 불구하고 KITTI 데이터셋으로 학습된PSMNet은 RADIATE 데이터셋에서 깊이 추정을 제대로 하지 못하고 있는 것을 확인할 수 있다. 뎁스 추정 결과를 볼 때 전체적인 구조만을 어느 정도 나타내고 있으며 근거리의 물체에 대한 뎁스를 제대로 추정하지 못하고 있다는 것을 확인할 수 있다.

Fig. 4 (d)는 제안하는 방법을 이용하여 뎁스 추정을 한 결과를 보여준다. BTS는 PSMNet과 동일하게 KITTI 데이텃셋으로 학습되었지만 PSMNet과 달리 환경의 전체적인 구조 뿐 아니라 환경 내에 존재하는 여러 가지 물체의 구체적인 윤곽도 추정할 수 있다는 것을 확인할 수 있다. 또한 같은 유사한 뎁스를 가지고 있는 평면의 뎁스도 균일하게 추정할 수 있다는 것을 확인할 수 있다. 그러나 BTS는 상대적인 뎁스만을 추정할 수 있기 때문에 제안한 방법을 이용하여 스케일 보정을 하면 절대적인

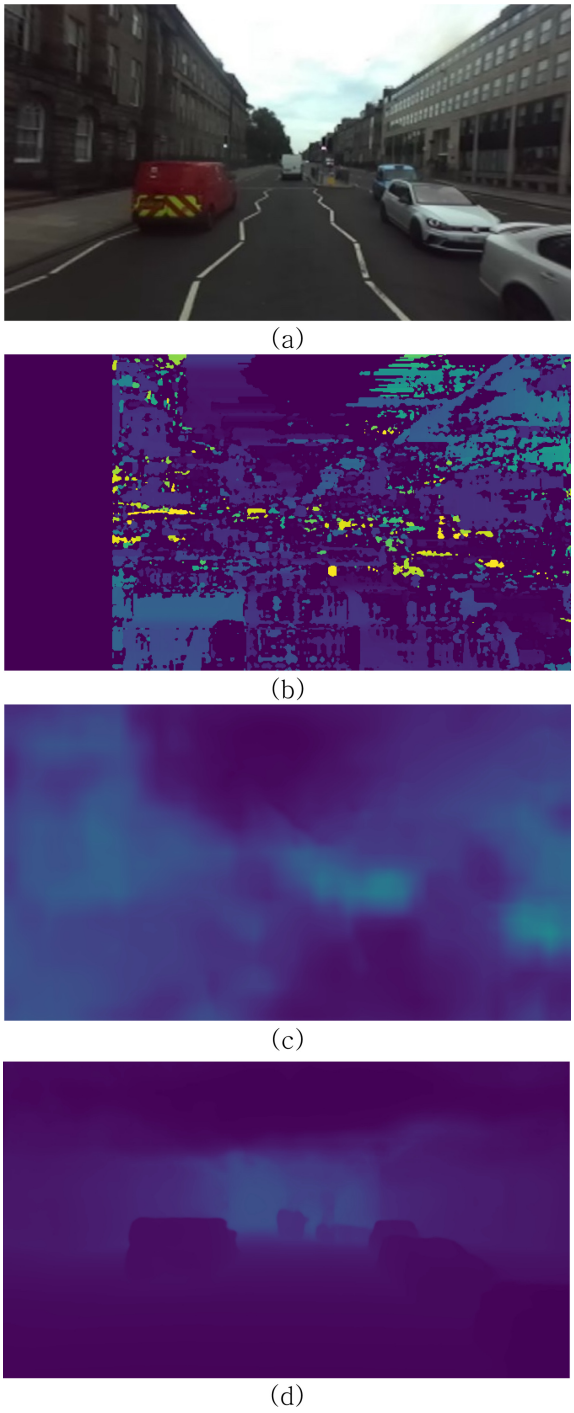

Fig. 4. Depth estimation result (a) input image, (b) SGBM, (c) PSMNet, (d) proposed method.

뎁스를 추정할 수 있다.

제안하는 방법을 이용하여 절대적인 뎁스를 추정한 결과를 확인하기 위하여 스케일 보정이 되지 않은 BTS와 제안한 방법으로 스케일이 보정된 BTS의 평균 뎁스 오차를 거리에 따른 구간 별로 확인하였다.

Fig. 5 (a)와 5 (b)는 각각 RADAIATE city_1_1 데이터와

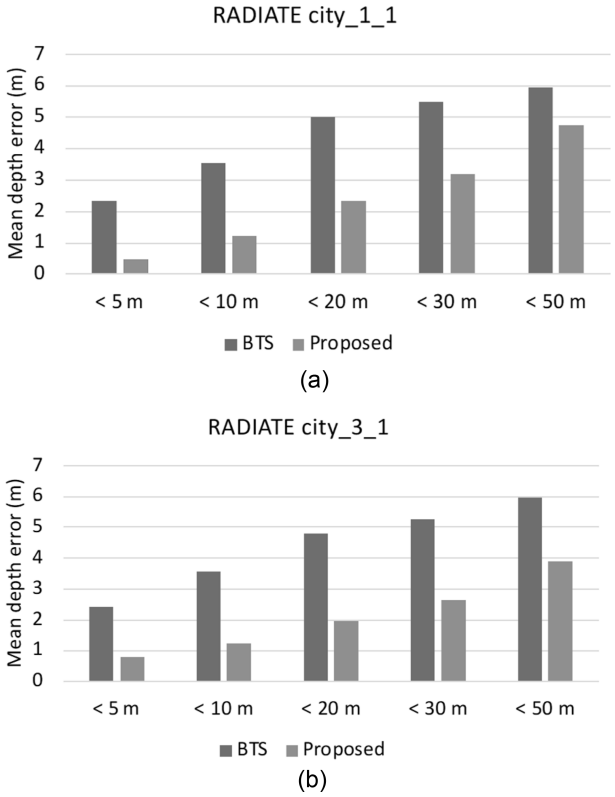

Fig. 5. Mean depth estimation errors of BTS [BTS] and the proposed method. (a) city_1_1 dataset, (b) city_3_1 dataset

Table 1. Computing time.

Sparse depth estimation	Unscaled depth estimation	Scale factor estimation
6.6 ms	18 ms	28 ms

RADIATE city_3_1 데이터에서의 두 방법의 구간 별 평균 뎁스 오차를 나타낸다. BTS는 KITTI 데이터셋을 이용하여 학습 되었기 때문에 RADIATE 데이터셋을 사용하여 뎁스 추정을 하는 경우 상대적인 뎁스는 잘 추정할 수 있지만 스케일 보정을 수행하지 않으면 상당히 큰 뎁스 추정 오차가 발생하는 것을 확인할 수 있다. 제안하는 방법을 사용하여 스케일 보정을 수행하면 약 30m 범위 내에서는 뎁스 추정 오차를 상당히 감소시킬 수 있다는 것을 확인할 수 있다.

Table 1은 제안하는 방법의 연산 시간을 나타낸다. 모노 뎁스 추정의 경우 MDE 모델의 복잡도에 따라 연산 시간이 결정된다. 더 빠른 연산 속도를 필요로 하는 경우 BTS 보다 더 경량화된 MDE 모델 [14]을 사용한다면 연산 속도를 더 향상 시키는 것이 가능하다. 저밀도 뎁스 추정과 스케일 보정은 keypoint detector와 descriptor의 특성 및 추출하는 최대 keypoint 개수 그리고 bucketing 파라미터에 의해서 영향을 받는다.

5. 결론 및 향후 과제

본 논문에서는 소형 플랫폼에 탑재할 수 있는 SBSC의 짧은 탐지거리를 해결하기 위한 방법을 제안하였다.

제안한 방법은 스케일 팩터를 추정하기 위해 초기 뎁스와 저밀도 뎁스 사이의 관계를 선형으로 표현하는 모델을 사용한다. 그러나 두 개의 뎁스 사이의 관계는 영상 중심점으로부터 위치나 환경의 형태에 의해 영향을 받는 비선형성을 가지고 있는 것으로 추측된다. 따라서 향후 연구에서는 이러한 비선형성을 고려할 수 있는 모델을 사용하여 스케일 팩터를 추정할 계획이다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. 2021R1F1A1057949).

본 과제(결과물)는 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다 (2021RIS-004).

REFERENCES

[1] J. Kim, Y. Cho, and A. Kim, “Proactive Camera Attribute Control Using Bayesian Optimization for Illumination- Resilient Visual Navigation”, *IEEE Trans. Robot.*, Vol. 36, No. 4, pp. 1256-1271, 2020.

[2] J. Kim, M.-H. Jeon, Y. Cho, and A. Kim, “Dark Synthetic Vision: Lightweight Active Vision to Navigate in the Dark”, *IEEE Robot. Autom Lett.*, Vol. 6, No. 1, pp. 143-150, 2020.

[3] J.-R. Chang and Y.-S. Chen, “Pyramid Stereo Matching Network”, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5410-5418, Salt Lake City, USA, 2018.

[4] J.-R. Chang, P.-C. Chen, and Y.-S. Chen, “Attention- Aware Feature Aggregation for Real-time Stereo Matching on Edge Devices”, *Proc. of Asian Conference on Computer Vision (ACCV)*, pp. 1-16, 2020.

[5] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostrowm, “Digging into Self-Supervised Monocular Depth Prediction”, *Proc. of International Conference Computer Vision (ICCV)*, pp. 3828-2838, Seoul, Korea, 2019.

[6] H. Hirschmuller, “Stereo Processing by Semiglobal Matching and Mutual Information”, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 30, No. 2, pp. 328-341, 2008.

[7] A. Geiger, M. Roser, and R. Urtasun, “Efficient Large-Scale Stereo Matching”, *Proc. of Asian Conference on Computer Vision (ACCV)*, pp. 25-38, Queenstown, New Zealand, 2010.

[8] D. Sun, X. Yang. M.-Y. Liu, and J. Kautz, “PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume”, *Proc. of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pp. 8934-8943, Salt Lake City, USA, 2018.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset", *Int. J. Rob. Res.*, Vol. 32, No. 11, pp. 1231-1237, 2013.

[10] Z. Li and N. Snavely, "MegaDepth: Learning Single- View Depth Prediction from Internet Photos", *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2041-2050, Salt Lake City, USA, 2018.

[11] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision Transformers for Dense Prediction", *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179-12188, 2021.

[12] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation", arXiv preprint arXiv:1907.10326, pp. 1-11, 2021.

[13] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth Estimation using Adaptive Bins", arXiv preprint arXiv: 2011.14141, pp. 1-13, 2020.

[14] D. Wofk, F. Ma, T.-J, Yang, S. Karaman, and V. Sze, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems", *Proc. of International Conference on Robotics and Automation (ICRA)*, pp. 6101-6108, Montreal, Canada, 2019.

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2223-2232, Venice, Italy, 2017.

[16] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal Unsupervised Image-to-Image Translation", *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 172-189, Munich, Germany, 2018.

[17] E. Rosten, R. Porter, and T. Drummond, "Faster and better: a machine learning approach to corner detection", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 32, No. 1, pp. 105-119, 2008.

[18] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces", *Proc. of the 24th British Machine Vision Conference (BMVC)*, pp. 1281-1298, Bristol, UK, 2013.

[19] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D Reconstruction in Real-time", *Proc. of IEEE Intelligent Vehicles Symposium (IV)*, pp. 963-968, Baden-Baden, Germany 2011.

[20] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding", arXiv preprint arXiv:1511.02680, pp. 1-11, 2015.

[21] M. Sheeny, E. D. Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "RADIATE: A Radar Dataset for Automotive Perception in Bad Weather", arXiv preprint arXiv:2010.09076, pp. 1-15, 2020.