인공지능 기반 3차원 공간 복원 최신 기술 동향

ㅁ 임성훈 / 대구경북과학기술원

요약

최근 스마트폰에서의 증강현실, 미적 효과의 증대(예, 라이브 포커싱) 등의 어플리케이션을 제공하기 위해 모바일 기기에서의 3차원 공간 복원 기술에 대한 관심이 증가하고 있다. 소비자들의 요구에 발 맞춰 최근 스마트폰 제조사는 모든 플래그십 모델에 다중 카메라 및 뎁스 센서(거리 측정 센서)를 탑재하는 추세이다. 본 고에서는 모바일 폰에 탑재되고 있는 대표적인 세 축의 뎁스 추정(공간 복원) 방식에 대해 간단히 살펴보고, 최근 심층학습(Deep learning)의 등장으로 기술 발전의 새로운 국면에 접어 든 다중 시점 매칭(Multi-view stereo) 방법에 대해 소개하고자 한다. 심층 신경망이 재조명 받은 2012년 전까지 주류 연구 방향이었던 전통 기하학 기반의 방법에 대한 소개를 시작으로 심층 신경망기반의 방법론으로의 발전된 형태를 살펴본다. 또한, 신경망기반의 방법론은 크게 3 세대로 나누어 각 세대별 특징에 대해 자세히 살펴보고, 다양한 데이터에 대한 실험 결과를 통해 세대별 공간 복원 결과를 비교 분석한다.

Ⅰ. 서 론

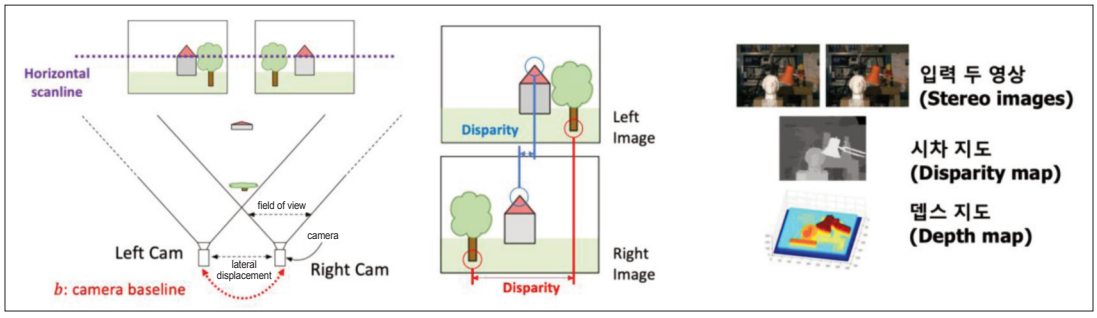
2015년 약 2억여 명 사용자가 Google 포토 앱에 약 240억 셀카를 업로드하였고[1], 현재에도 매일 대량의 미적 효과를 입힌 이미지가 수많은 소셜 미디어에 게시되고 있다. 얼굴의 3차원 정보를 활용한 미학적 효과(라이브 포커싱 등)는 더욱 자연스러운 영상을 생성하는 경향이 있어[2], 3차원 정보 복원에 대한 관심이 높아지고 있다. 뿐만 아니라, 모바일에서 제공되는 얼굴 인식, 증강현실 등 다양한 어플리케이션에서는 가장 기본이 되는 정보로, 해당 정보를 제공하기 위해 다중 카메라(2대 이상) 혹은 뎁스(Depth) 카메라 등이 모바일폰에 탑재되어 있다. 모바일 환경에서의 대표되는 세가지 공간 복원 기술은 (1) 다수의 RGB 카메라 간의 시차(Parallax)를 계산하여 공간을 복원하는 방법(Multi-view Stereo)[3], (2) 구조광 패턴을 장

면에 투영하여 해석하는 방법(SL: Structured Light)[4], (3) 빛의 비행시간을 측정하는 방법(ToF: Time of Flight)[5] 등이 있다. 각 스마트폰 제조사마다 전략을 달리하고 있지만 하드웨어적, 소프트웨어적으로 공간 복원 기술이 꾸준히 발전해 오고 있음은 분명하다.

SL방식은 아이폰 X 이상에서 제공하는 FACE ID 얼굴 인식에 사용되는 방법으로, 특정 패턴의 적외선 빛을 피사체에 방사한 후 피사체 표면 모양에 따라 패턴의 변형된 정도를 분석하여 거리를 측정한다. SL방식은 밀리미터에서 마이크로미터까지 고해상력 뎁스 정보를 추정할 수 있다는 장점이 있지만, 적외선 패턴을 방사함으로 다른 외부의 빛의 간섭에 취약하고 장거리에 취약하다는 단점이 있다. ToF방식은 갤럭시 10(+노트)에 탑재된 3차원 센싱 기술로, 피사체를 향해 레이저를 발사하고 빛이 튕겨져 센서로 돌아오는 시간을 계산해서 거리를 측정하는 방법이다. ToF방식은 레이저를 활용함으로 외부 빛의 간섭에 다소 강인하고 거리 제한에서 비교적 자유롭지만, SL방식에 비해 해상력이 다소 떨어진다. SL방식과 ToF방식 모두 프로젝터를 활용하여 빛을 쏘고 거리를 측정하는 방식으로, 카메라만 활용하는 다중 시점 매칭(Multi-view stereo) 방식과 큰 차이를 보인다.

다중 시점 매칭 〈그림 1〉은 다중 카메라 간의 시차를 계산하여 거리를 측정하는 방법으로 발광체[1]를 사용하는 SL방식이나 ToF방식에 비해 센서 구성이 다소 저렴하고, 카메라 사이의 거리가 충분한 경우 측정 거리 제한에 자유롭다. 하지만, 실내외 촬영 환경(조도 조건)이나 촬영하는 물체 표면의 특성에 따라 고품질 공간 정보 획득에 어려움을 주고 있다. 이를 해결하기 위해 전통적인 기하학기반의 방법론에서부터 최근 심층 신경망기반의 방법론까지 기술 발전이 이루어지고 있으며, 특히 최근에는 심층 신경망의 구조에 초점을 맞춘 연구에 집중되어 있다.

본 고에서는 이와 같이 단일 영상에 대한 화질 개선을 수행하는 최신 기술 동향을 살펴보고자 한다. II장에서는 기하 정보를 이용한 전통적 다중 시점 매칭 방법에 대해 간략히 살펴본 후, 3 세대에 걸쳐 발전된 심층 신경망 기반의 다중 시점 매칭 분야의 최신 연구를 소개한다. III장에서는 다양한 영상에 대해 복원된 결과를 비교 제시하고, 마지막으로 V



<그림 1> 다중 시점 매칭 시각화: 시차(Disparity) 추정 문제

1) 스테레오 비전도 발광체를 함께 사용하기도 한다. 발광체를 활용하는 경우 액티브 스테레오(Active stereo), 카메라만 활용하는 경우 패시브 스테레오(Passive stereo)라고 명명한다. 기본적인 알고리즘은 동일하다.

장에서 결론을 맺는다.

II. 3차원 공간 복원 개선 기술 동향

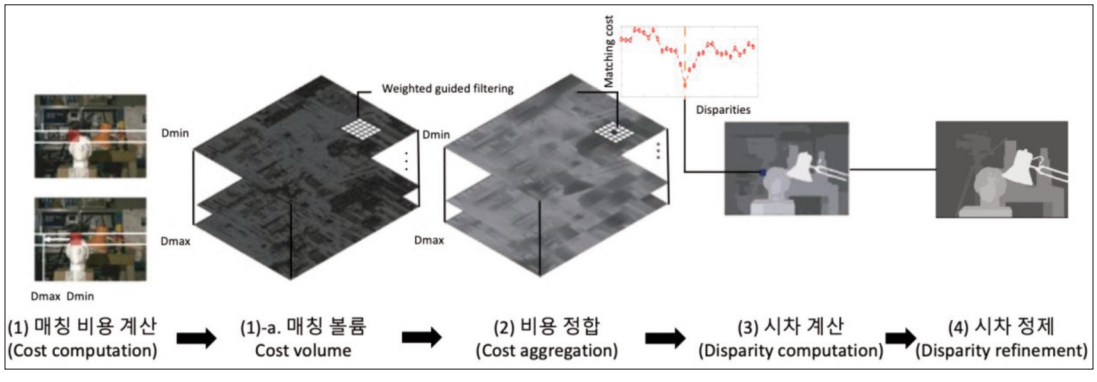
이번 장에서는 3차원 공간 복원 기술(다중 시점 매칭)을 네 가지 방법으로 나누어 자세히 살펴보고자 한다. 먼저 카메라 기하 정보 기반의 다중 시점 매칭 방법에 대해 살펴보고, 3단계로 발전되어 온 심층 학습 기반의 다중 시점 매칭 방법에 대해 살펴본다. 다중 시점 기하학에 대한 전반적인 이해는 Richard와 Zisserman의 저서[6]를 참고하기 바란다.

1. 전통적인(비-학습 기반) 다중 시점 매칭 기술

인간의 스테레오 비전(두 눈)에 대한 계산 원리 정립을 시작으로 지난 수 십 년간 컴퓨터 비전 분야에 다중 시점 매칭 방법이 연구[3]되어 왔다. 본 연구에서는 다중 시점 매칭 알고리즘을 일반적으로 (1) 매칭 비용 계산(matching cost computation), (2) 비용 정합(cost aggregation), (3) 시차 계산/최적화(disparity computation/optimization), (4) 시차 정제(disparity)라는 네 단계를 수행한다고 관찰하였다(〈그림 2〉). 매칭 비용 계산이란 두 영상의 강도(Intensity)의 차이를 픽셀 단위로 계산하여 유사도를 측정하는 것을 말한다. 픽셀 단위로 계산된 매칭 비용을 하나의 볼륨으로 쌓은 것을 비용 볼륨이라고 명명하며, 두 번째 단계인 비용 정합에서 비용 볼륨의 지지 영역(예, 지역 정보)의 정보를 정합하여 계산된 비용 정보의 신뢰도를 높이고자 한다. 이 후 정합된 비용 볼륨에서 시차를 계산하거나 최적화하여 추정하는 시차 계산/최적화 과정을 거치게 된다. 추정된 시차는 카메라의 초점거리와 두 카메라 사이의 거리 정보로 간단히 거리로 변환이 가능하다. 마지막으로, 추정된 시차가 정수 단위의 이산 레벨을 갖고 있어서 연속적 레벨을 갖게 하면서 오류를 줄이는 시차 정제 단계를 거치게 된다.

2. 1세대 학습기반 두 시점 매칭 기술: 비용 계산 심층 신경망 제안

2015년 뉴욕대학교 Zbontar와 LeCun이 처음 심층 신경망을 활용한 뎁스 추정 방법[8]을 발표하였



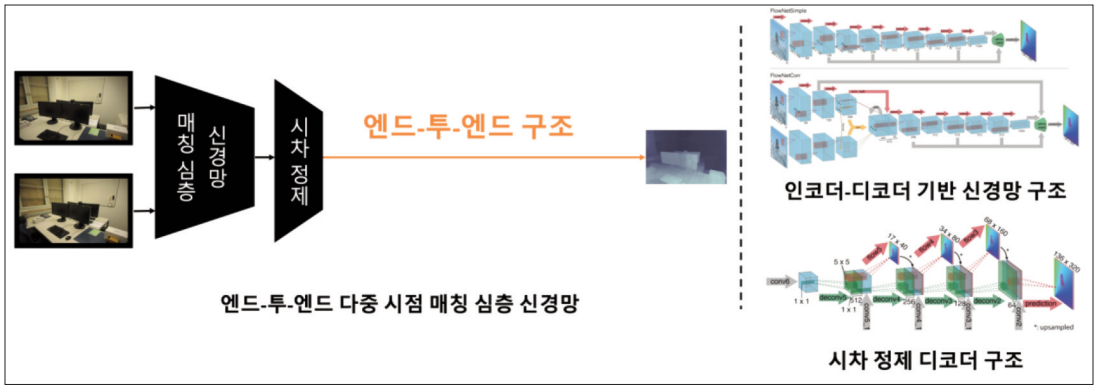
<그림 2> 일반적인 다중 시점 매칭 알고리즘의 4단계 과정

<그림 3> 1세대 학습기반 다중 시점 매칭 기술: 패치기반 비용 계산 심층 신경망

다. 본 연구는 전통 스테레오 방법의 4단계 중 비용 계산 과정만을 심층 신경망으로 대체하였고, 이후 과정은 모두 전통 스테레오 방법과 동일한 과정을 거친다(〈그림 3〉). 비용 계산 신경망은 기준이 되는 영상을 작은 패치 단위로 나누어 다른 영상에서의 패치와 유사도를 계산한다. 영상의 횡축으로 한 픽셀 이동시마다 비용을 계산하고, 동일 과정을 영상 전반에 걸쳐 수행하여 영상 전체 해상도 크기의 비용 볼륨을 만든다. 만들어진 비용 볼륨을 기반으로 전통적 방법에서 사용하던 비용 집합, 깊이 선택, 깊이 정제 알고리즘을 차용하여 뎁스 정보를 추정하였다. 이 결과 논문 제출 당시 KITTI[7]라는 스테레오 매칭 벤치마크 사이트에서 가장 우수한 성능을 보이는 알고리즘이 되었고, 본 연구를 기점으로 스테레오 매칭 연구 분야는 심층 학습기반으로 연구 패러다임이 옮겨가게 되었다.

3. 2세대 학습기반 두 시점 매칭 기술: 엔드-투-엔드 다중 시점 매칭 기술

1세대 학습기반 방법론과 유사한 방식으로 약 1년간 기술 개발이 이뤄지다가 2016년 독일 Freiburg 대학교의 Mayer[9] 등은 전통적 방식의 후처리 과정 없이 오직 심층 신경망을 활용한 깊이 추정 방식인 엔드-투-엔드 다중 시점 매칭 기술을 개발하였다(〈그림 4〉). 본 연구는 심층 신경망 자체에 깊이 정제 역할을 하는 구조를 추가함으로써 추가 후처리 과정 없이 뎁스 정보를 회귀(Regression)하는 신경망을 제안하였다. 2세대 스테레오 매칭 신경망은 패치 단위의 영상을 입력으로 받는 것이 아닌 전체 해상도의 두 영상을 입력으로 받아, 신경망이 깊이 정보를 출력해주는 형태로 구성되어 있다. 신경망은 전형적인 인코더-디코더(Encoder-Decoder) 형태로 구성이 되어 있으며, 디코더 부분에서 이전 레이어에서 추출된 뎁스 정보를 다음 레이어의 입력으로 넣어줌으로써 뎁스를 정제하는 효과를 주었다. 본 연구는 심층 신경망을 제안하였을 뿐 아니라, blender라는 그래픽스 툴로 가상 데이터를 제작하여 가상 데이터를 신경망 학습에 활용하였다. 가상 데이터로 초기 학습을 하고 목표하는 환경에 맞는 데이터셋으로 신경망을 재학습 하는 것이 뎁스 추정 성능을 높일 수 있음을 보였다. 2세대 스테레오 매칭 방법은 1세대 방법에 비해 연산 속도에서 수 백배 이상 빠른 처리 속도를 자랑하면서(1세대:

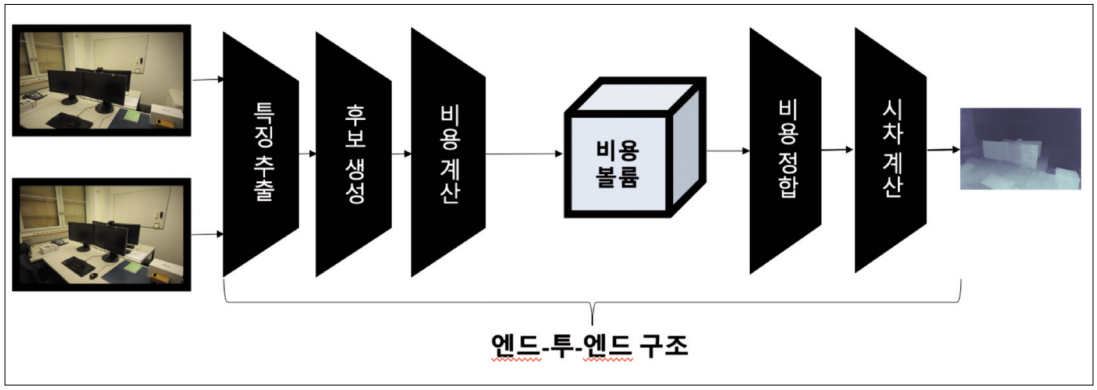


<그림 4> 2세대 학습기반 다중 시점 매칭 기술: 엔드-투-앤드 다중 시점 매칭 심층 신경망

67초 vs 2세대: 0.06초)도, 1세대 방법에 준하는 성능을 보였다.

4. 3세대 학습기반 두 시점 매칭 기술: 멀티클래스 분류 기반 뎁스 추정 심층 신경망

두 세대에 걸친 학습기반 매칭 기술은 유사도 혹은 뎁스 수치를 제공하는 회귀 신경망을 설계함으로써 뎁스를 추정하였다. 2017년 영국 Cambridge 대학교의 Kendall[10] 등은 다중 시점 매칭 문제를 회귀 문제가 아닌 멀티클래스 분류 문제로 접근하였다(〈그림 5〉). 한 픽셀씩 영상을 횡 방향으로 이동(shift) 시키면서 최대 시차까지 비용(유사도)을 계산하여 비용 볼륨을 제작한다. 횡 방향의 각 픽셀을 멀티클래스(후보군)으로 설정이 되고, 이 중 최적의 후보군을 분류하는 문제로 풀게 된다. 두 입력 영상은 동일한 2차원 컨볼루션 레이어를 통과하여 특징점이 추출되고, 추출된 특징점을 픽셀 단위로 이동시켜 모든 특징점을 결부시켜 다중 스케일의 3차원 컨볼루션 레이어를 통과시키는 것으로 비용 볼륨을 제작한다. 제작된 비용 볼륨은 3차

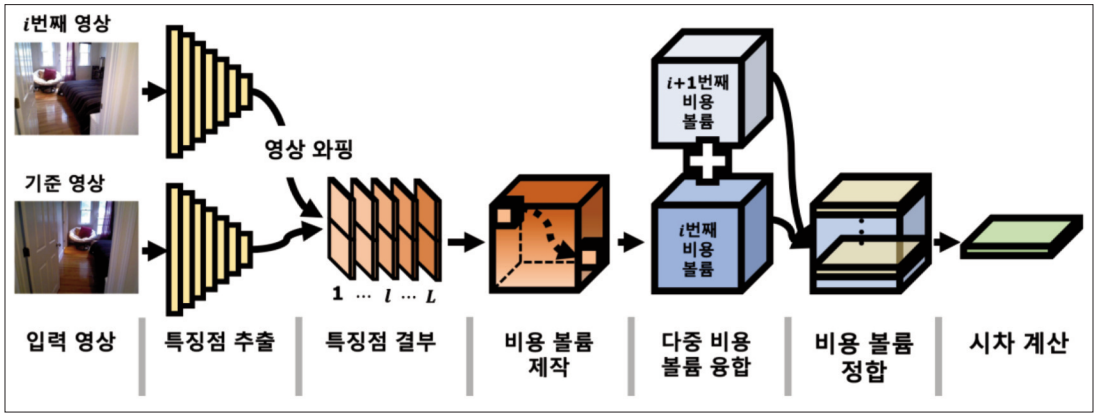


<그림 5> 3세대 학습기반 다중 시점 매칭 기술: 멀티클래스 분류 기반 다중 시점 매칭 심층 신경망

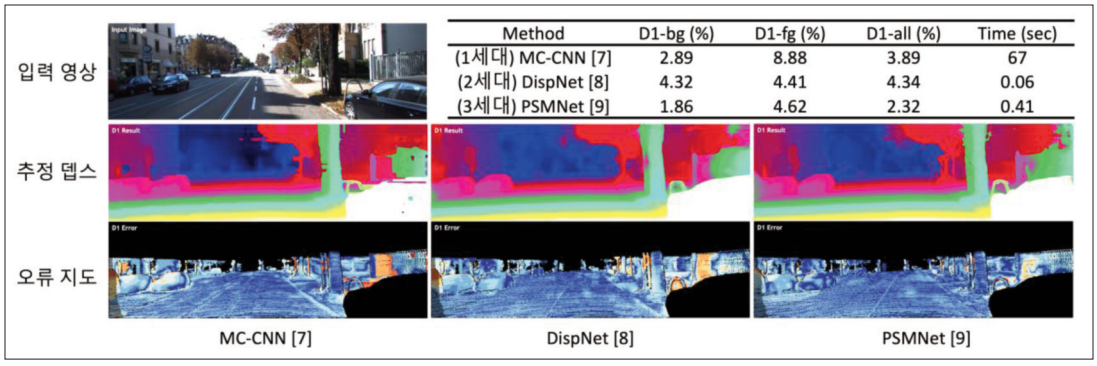
원 디컨볼루션을 통해 비용 볼륨을 집합시키고, SoftMax를 활용하여 깊이를 선택하게 된다. 모든 심층 신경망은 연결되어 엔드-투-엔드로 학습이 가능하고, 정답 가이드는 마지막 레이어를 통과하여 추정된 깊이 정보에만 주어지게 된다. 본 연구의 신경망 구조에 새로운 형태의 특징점 추출 레이어와 비용 볼륨 집합 레이어를 접목한 연구[11]는 2세대 방법에 비해 다소 계산 시간이 오래 걸리지만(2세대 0.06초 vs 3세대 0.41초), 2018년 KITTI 스테레오 벤치마크에서 가장 우수한 성능을 보였다.

5. 학습기반 다중(두 영상 이상) 시점 매칭 기술: 기하 정보 기반 뎁스 추정 심층 신경망

세 세대에 걸친 학습기반 두 시점 매칭 기술은 정류(Rectification)되어 있어서 횡 축으로의 시차 계산을 위한 알고리즘으로 정류가 되지 않은 다중 시점(3대 카메라 이상) 매칭에 직접적 활용이 불가하다. 본 문제를 해결하고자 2019년 본 고의 필자[12]가 다중 시점으로의 확장성을 확보하면서 시점 정보의 효율적 정합 기술을 개발하였다(〈그림 6〉). 특



<그림 6> 학습기반 다중(3개 영상 이상) 시점 매칭 기술: 기하 정보 융합 신경망



<그림 7> 학습기반 두 시점 매칭 기술: 기하 정보 융합 신경망

히, (1) 3세대 두 시점 심층 신경망 구조를 기반으로 카메라간 기하 정보(카메라의 상대적 위치)를 활용한 비용 볼륨 제작 신경망, (2) 입력 영상의 특징점을 가이드로 하여 비용 볼륨의 신뢰도를 높이는 비용 정제 신경망, (3) 다중 영상의 매칭 정보를 간단하면서 효과적으로 집합하는 비용 집합 신경망을 제안하였다. 본 연구는 DeMoN[14]과 ETH3D[17] 등 다양한 데이터셋에 대해 기존의 방법론[13, 14, 15, 16]들과 비교 분석을 수행하였고, 그 결과 기존 방법론이 안고있던 다중 영상으로의 확장성 및 단일 색상 물체에 대한 복원 취약성 등의 문제를 해결한 모습을 보이고 있다(〈그림 8〉).

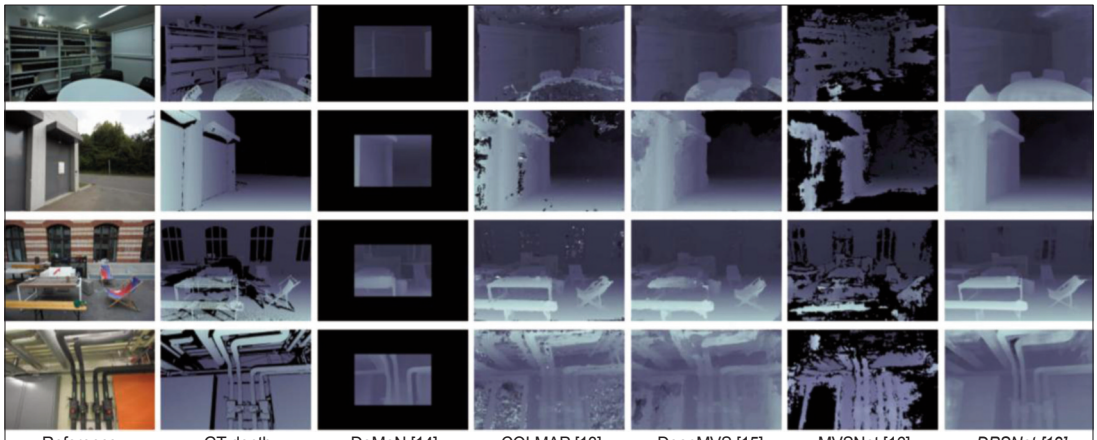
Ⅲ. 뎁스 추정(공간 복원) 성능 비교

이번 장에서는 다중 시점 기반의 뎁스 추정(공간 복원) 방법들의 성능 비교를 해보고자 한다. 먼저 두 시점 매칭에 초점을 둔 1세대에서 3세대 방법(MC-CNN[7], DispNet[8], PSMNet[9])은 주행 환경인 KITTI 데이터셋을 활용하여 비교 분석하였다. 〈그림 7〉에 정성적, 정량적 성능 평가를 수행하였고, 정량적 성능은 D1이라는 오류율을 나타내는 지표로 평가하였다. D1은 '왼쪽 영상에서 3픽셀 이상 시차(disparity) 차이나는 픽셀의 비율'이라는 메트릭으로 낮은 수치일수록 우수한 성능을 보인다. 정량적 수치중 D1-all은 모든 영역, D1-bg는 백그라운드, D1-fg는 포그라운드 영역 내에서 측정한 메트릭을 나타낸다. 1세대 방법론은 패치단위로 매칭점을 찾아 반복적으로 연산을 수행하게 되고, 전통적 스테레오 매칭 방법의 후처리 과정을 적용하여 오랜 연산 시간을 필요로 한다. 2세대 방법론이 계산 시간 측면에서 가장 우수한 성능을 보인 반면, 오류율에서는 1세대보다 높은 수치를 기록했다. 3세대 방법론은 2세대 방법론에 비해 계산 시간은 다소 늘었지만(약 7배) 정성적, 정량적으로 가장 우수한 성능을 보였다.

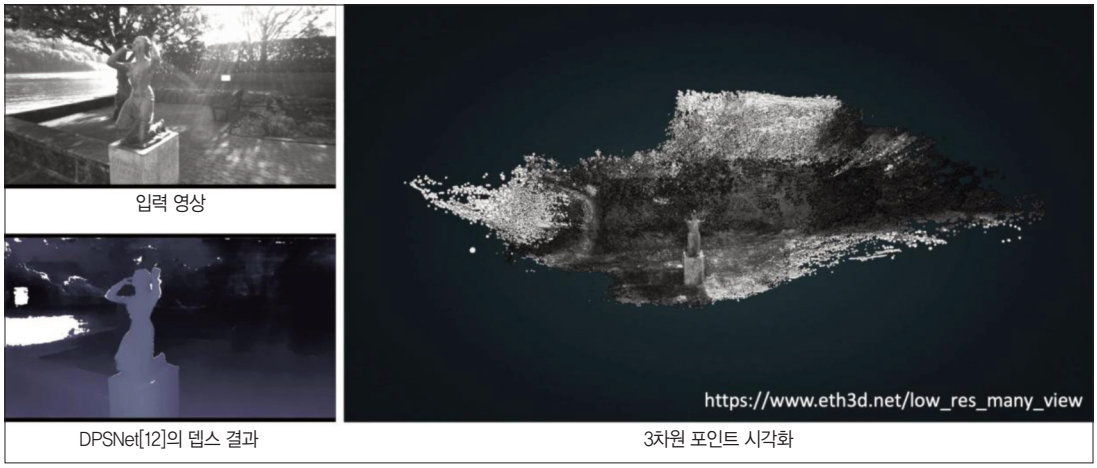
〈표 1〉과 〈그림 8〉은 두 영상 매칭뿐 아니라 두 개 영상 이상을 입력으로 받아 다중 시점 매칭을 하는 연구들을 비교한 정량적, 정성적 결과이다. 정량적 평가에 활용한 지표는 크게 오류 지표(Abs Rel(Absolute Relative), Abs Diff(Absolute Difference), Sq Rel(Square Relative), RMSE (Root-Mean-Square-Error), RMSE log(log 도메인 RMSE)) 및 정확도 지표(오차가 α^t (t=1,2,3, α=1.25) 내에 들어오는 픽셀의 비율)를 사용하였다. 오류 지표 및 정확도 지표 모두 3세대 심층 신경망 기반의 다중 시점 매칭 방법인 DPSNet[12]이 가장 우수한 성능을 보였고, 정성적 비교 〈그림 7〉도 동일한 양상을 보였다.

<표 1> 학습기반 다중(3개 영상 이상) 시점 매칭 기술: 기하 정보 융합 신경망

Method	Error metric					Accuracy metric		
	Abs Rel	Abs Diff	Sq Rel	RMSE	RMSE log	$\delta < \alpha^1$	$\delta < \alpha^2$	$\delta < \alpha^3$
COLMAP [13]	0.324	0.615	36.71	2.370	0.349	**0.865**	0.903	0.927
DeMoN [14]	0.191	0.726	0.365	1.059	0.240	0.733	0.898	0.951
DeepMVS [15]	0.178	0.432	0.973	1.021	0.245	0.858	0.911	0.942
MVNet [16]	1.666	2.165	13.93	3.255	0.824	0.555	0.628	0.686
DPSNet [12]	**0.099**	**0.365**	**0.204**	**0.703**	**0.184**	0.863	**0.938**	**0.963**



<그림 8> 학습기반 다중(3개 영상 이상) 시점 매칭 기술: 기하 정보 융합 신경망



<그림 9> 학습기반 다중(3개 영상 이상) 시점 매칭 기술: 기하 정보 융합 신경망

Ⅳ. 결 론

본 고에서는 다중 시점 매칭 기반의 뎁스 추정(3차원 공간 복원[2]) 방법론을 소개하였고, 각 세대별 대표적인 심층 신경망 기반의 방법론을 비교 분석하였다. DSLR급 성능의 모바일 카메라에 대한 소비자들의 수요가 증가하면서 최근 스마트폰 제조사들은 다중 카메라 및 뎁스 센서를 장착하는 등 하드웨어적으로, 소프트웨어적으로 기술이 빠르게 진보하고 있다. 그 중에서 영상만을 활용한 다중 시점 매칭 분야는 심층 신경망의 등장으로 기술 수준이 한 층 더 성숙한 보습을 보이고 있다. 2015년 심층

2) 추정 된 뎁스 정보는 3차원 공간 상으로 역투영시켜서 공간 복원 결과 획득 가능(〈그림 9〉)

학습기반의 논문이 발표되면서 기하정보만을 활용한 전통적 매칭 방식의 패러다임은 심층학습기반으로 넘어왔고, 3 세대에 걸쳐 기술이 발전되었다. 최신 기술들은 실제 어플리케이션에 활용이 될 정도로 높은 복원 성능을 보이고 있지만, 여전히 분별이 어려운 물체(동일 색상 물체, 반사 물체), 가시성이 떨어지는 환경(야간, 어두운 실내)에서는 복원 성능 저하를 야기한다(〈그림 9〉). 해당 이슈를 해결하기 위해 새로운 방식의 비용 정합이나 깊이 정제 방법[18]들이 제안되고 있으나, 가시광 대역의 영상만을 입력으로 하는 방법론의 근본적인 문제를 해결하는 방향과는 거리가 있다. 본 고의 독자가 심층 신경망의 특성을 깊이 있게 분석하여 현 시점의 방법론들이 품고 있는 근본적인 문제를 극복하는 새로운 방향의 연구를 수행하길 기대한다.

참 고 문 헌

[1] Google photos: One year, 200 million users, and a whole lot of self-ies. https://blog.google/products/photos/google-photos-one-year-200-million/, accessed: 2016-05-27

[2] Augmented faces. https://developers.google.com/ar/develop/java/augmented-faces, accessed: 2019-12-183.

[3] Scharstein, Daniel, and Richard Szeliski. “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms.” International journal of computer vision 47.1-3 (2002): 7-42.

[4] Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light.In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR). vol. 1 (2003)

[5] Foix, Sergi, Guillem Alenya, and Carme Torras. “Lock-in time-of-flight (ToF) cameras: A survey.”?IEEE Sensors Journal 11.9 (2011): 1917-1926.

[6] Hartley, Richard, and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003.

[7] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite.” 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.

[8] Žbontar, Jure, and Yann LeCun. “Stereo matching by training a convolutional neural network to compare image patches.” The journal of machine learning research 17.1 (2016): 2287-2318.

[9] Mayer, Nikolaus, et al. “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[10] Kendall, Alex, et al. “End-to-end learning of geometry and context for deep stereo regression.” Proceedings of the IEEE International Conference on Computer Vision (CVPR). 2017.

[11] Chang, Jia-Ren, and Yong-Sheng Chen. “Pyramid stereo matching network.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[12] Im, Sunghoon, et al. “Dpsnet: End-to-end deep plane sweep stereo.” International Conference on Learning Representations (ICLR) 2019.

[13] Schönberger, Johannes L., et al. “Pixelwise view selection for unstructured multi-view stereo.” European Conference on Computer Vision. Springer, Cham, 2016.

[14] Ummenhofer, Benjamin, et al. “Demon: Depth and motion network for learning monocular stereo.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[15] Huang, Po-Han, et al. “Deepmvs: Learning multi-view stereopsis.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[16] Yao, Yao, et al. “Mvsnet: Depth inference for unstructured multi-view stereo.” Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[17] Schops, Thomas, et al. “A multi-view stereo benchmark with high-resolution images and multi-camera videos.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[18] Cheng, Xinjing, Peng Wang, and Ruigang Yang. “Learning Depth with Convolutional Spatial Propagation Network.” IEEE transactions on pattern analysis and machine intelligence (2019).

필 자 소 개

임 성 훈

– 2018년 2월 ~ 2018년 8월 : Microsoft Research Asia(MSRA) 연구 인턴
– 2019년 6월 ~ 2019년 8월 : Carnegie Mellon University(CMU) 방문 연구
– 2019년 8월 : 한국과학기술원(KAIST) 박사
– 2019년 9월 ~ 현재 : 대구경북과학기술원 정보통신융합전공 조교수
– 주관심분야 : 컴퓨터 비전, 기계학습, 로봇 비전