KICSP
한국 융합신호처리학회
융합신호처리학회논문지 Vol. 22, No. 1 : 30-37 March. 2021

딥러닝 기술을 이용한 3차원 객체 추적 기술 리뷰

박한훈*
부경대학교 전자공학과

A Review of 3D Object Tracking Methods Using Deep Learning

Hanhoon Park*
Department of Electronic Engineering, Pukyong National University

요 약 카메라 영상을 이용한 3차원 객체 추적 기술은 증강현실 응용 분야를 위한 핵심 기술이다. 영상 분류, 객체 검출, 영상 분할과 같은 컴퓨터 비전 작업에서 CNN(Convolutional Neural Network)의 인상적인 성공에 자극 받아, 3D 객체 추적을 위한 최근의 연구는 딥러닝(deep learning)을 활용하는 데 초점을 맞추고 있다. 본 논문은 이러한 딥러닝을 활용한 3차원 객체 추적 방법들을 살펴본다. 딥러닝을 활용한 3차원 객체 추적을 위한 주요 방법들을 설명하고, 향후 연구 방향에 대해 논의한다.

• 주제어 : 3차원 객체 추적, 카메라 포즈 추정, 비전 기반, 딥러닝, 증강현실

Abstract Accurate 3D object tracking with camera images is a key enabling technology for augmented reality applications. Motivated by the impressive success of convolutional neural networks (CNNs) in computer vision tasks such as image classification, object detection, image segmentation, recent studies for 3D object tracking have focused on leveraging deep learning. In this paper, we review deep learning approaches for 3D object tracking. We describe key methods in this field and discuss potential future research directions.

• Key Words : 3D object tracking, Camera pose estimation, Vision-based, Deep learning, Augmented Reality

Received 17 February 2021, **Revised** 28 March 2021, **Accepted** 29 March 2021

* **Corresponding Author** Hanhoon Park, Department of Electronic Engineering, Pukyong National University, 45, Yongso-ro, Nam-gu, Busan, Korea. E-mail: hanhoon.park@pukyong.ac.kr

Ⅰ. 서론

증강현실(augmented reality)은 현실(real) 세계에 가상(virtual) 콘텐츠를 병치하여 인간의 감각과 인식 범위를 확장시켜 주는 실감미디어 기술로서, 최근 교육, 훈련, 방송, 게임을 포함한 다양한 분야에서 활용되고 있다[1, 2].

증강현실의 가장 기본적이고 핵심적인 기술은 관심 객체(object)나 장면(scene)의 3차원 움직임 또는 포즈(pose)를 추적(tracking)하는 것이다. 여기서, 3차원 움직임은 객체 좌표계와 카메라 좌표계 사이의 변환(transformation)을 말하며, 3 자유도(degree of freedom)의 회전(rotation)과 3 자유도의 위치 이동(translation)으로 구성된다[9]. 본 논문에서는 컴퓨터 비전 기반 3차원 객체 추적 기술, 즉, 카메라 영상 정보를 이용하여 객체의 3차원 움직임을 추적하는 기술, 을 중심으로 개념 및 대략적인 방법을 설명하고, 딥러닝(deep learning) 기술을 활용한 최근 방법들을 소개한 후, 향후 발전 방향에 대해 논의한다.

1.1 리뷰 범위

3차원 객체 추적 기술은 카메라와 대상 객체 사이의 상대적인 3차원 포즈 변환을 추정하는 것이기 때문에, 3차원 객체 추적은 카메라 추적으로 볼 수도 있다. 카메라 추적은 크게 ego-centric 카메라 추적과, exo-centric 카메라 추적으로 나뉠 수 있는데, 본 논문에서 논의되는 3차원 객체 추적 기술은 객체를 중심으로 움직이는 카메라의 포즈를 추정하는 exo-centric 카메라 추적 기술과 동일한 의미를 가진다. 카메라를 중심으로 주변 장면을 바라보면서 움직이는 ego-centric 카메라 추적은 visual SLAM[3] 분야의 주요 기술로, 본 논문에서는 거의 논의되지 않는다. 또한, exo-centric 카메라 추적은 일반적으로 실내 환경을 대상으로 하기 때문에 실외 환경은 고려하지 않는다.

Exo-centric 카메라 추적을 위해 딥러닝 기술을 활용한 연구 사례는 많지 않다. 이는 ego-centric 카메라 추적에 비해 시점 변화에 따른 영상 정보 사이의 연관성(correlation)이 부족하여 학습을 통한 추적이 용이하지 않기 때문이다. 따라서, 관련 기술에 대한 이해를 돕기 위해 일부 exo-centric 카메라 추적이 아닌 방법들에 대한 설명이 포함된다.

대부분의 3차원 객체 추적 기술은 단일 RGB 카메라를 사용하여 단일 강체(rigid object)를 추적하는 방법과 관련된다. 그러므로, 본 논문에서는 이와 관련된 연구를 중점적으로 논하고자 한다. 적외선 카메라를 사용한 3차원 객체 추적 기술은 어두운 조명 환경이나 관심 객체가 적외선 영역에서 특징적인 정보를 가질 경우 유용할 수 있으나, 여전히 적외선 영상의 낮은 대비(contrast)와 부족한 영상 특징 정보로 인해 제약된 성능을 가진다[4]. 또한, 모바일 환경에서 적외선 영상을 획득하는 것은 여전히 보편적이지 않다.

1.2 기존 리뷰 논문과의 차별성

객체 또는 카메라 추적 기술에 대한 여러 리뷰 논문들이 있다[5-8]. 그러나, 대부분의 리뷰 논문들은 넓은 범위의 기술을 포괄적으로 다루거나, 최근 기술(주로 딥러닝 기술을 활용)에 대한 논의가 부족하다. 따라서, 딥러닝 기술을 활용한 3차원 객체 추적(즉, exo-centric 카메라 추적) 기술에 대한 심층적인 리뷰 논문은 보고된 적이 없다.

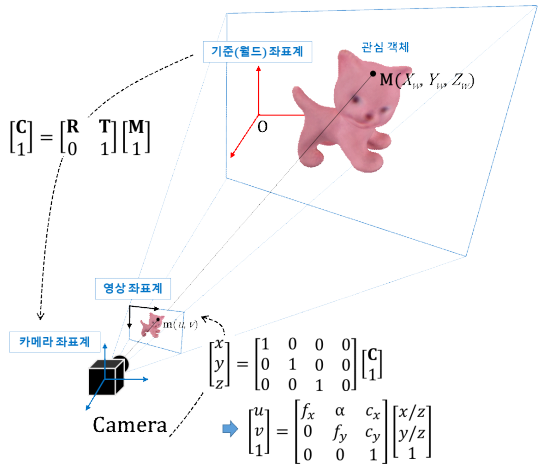


Fig. 1. Pinhole camera model that represents the relationship between 3D object coordinates and their corresponding 2D image coordinates

Ⅱ. 비전 기반 3차원 객체 추적 기술

비전 기반 객체 추적 기술은 객체를 연속적으로 촬영한 카메라 영상으로부터 영상을 촬영한 카메라 또는 촬영된 객체의 3차원 움직임을 추정하는 기술을 말하며, 카메라와 객체 사이의 상대적인 3차원 포즈 변환(그림 1에서 R과 T 행렬)을 추정하는 기술이기도 하다. 이

를 위해서는 객체의 3차원 좌표가 카메라 영상 좌표로 투영(projection)되는 과정에 대해 알아야 하는데, 객체의 3차원 좌표(**M**)와 대응하는(corresponding) 2차원 영상 좌표(**m**)는 카메라 투영 행렬(**P**)로 표현될 수 있다(그림 1 참조). 카메라 투영 행렬은 다음과 같이 카메라 내부(intrinsic) 행렬 **K**와 외부(extrinsic) 행렬 **E**로 구성된다[9].

$$[\mathbf{m} \quad 1]^{\mathrm{T}} = \lambda \mathbf{P}[\mathbf{M} \quad 1]^{\mathrm{T}},$$
$$\text{where } \mathbf{P} = \mathbf{KE} = \begin{bmatrix} f_x & \alpha & c_x \\ 0 & f_x & c_y \\ 0 & 0 & 1 \end{bmatrix} [\mathbf{R} \quad \mathbf{t}]. \tag{1}$$

내부 행렬은 카메라의 고유한 광학 특성과 관련된 파라미터를 포함하며, 카메라 보정(calibration) 방법[5]들에 의해 미리 계산될 수 있다. 외부 행렬은 객체 좌표계(또는 월드 좌표계)와 카메라 좌표계 사이의 상대적인 3차원 포즈 변환과 관련된 파라미터를 포함하며, 카메라 내부 행렬이 주어졌을 때, 일정 수 이상의 (**M**, **m**) 쌍이 주어지면 다양한 최적화 기법을 사용하여 **PM**과 **m** 사이의 거리를 최소화하거나, 밝기, 색상, 텍스처, 그래디언트(gradient)와 같은 특징 정보의 차를 최소화함으로써 추정할 수 있다[5, 7, 9]. 일반적으로 이러한 방법을 최적화 기반 방법 또는 3차원 구조(structure) 기반 방법이라고 한다.

Ⅲ. 딥러닝 기술 활용

최근 딥러닝 기술이 주목받으면서 3차원 객체 추적 분야에도 적극적으로 활용되고 있다. 앞 장에서 설명한 대로 비전 기반 3차원 객체 추적 기술은 주어진 카메라 영상으로부터 특징 정보를 찾고, 특징 정보의 3차원-2차원 대응 관계를 수식화한 후, 최적화 과정을 통해 카메라 외부 행렬을 추정하는 과정을 포함한다. 이러한 각 과정은 CNN(Convolutional Neural Network)과 FCN(Fully Connected Network)을 사용하여 대체될 수 있다. 특히, 딥러닝 기술을 이용하여 광대한 양의 영상 데이터로부터 3차원 객체 추적에 필요한 정보를 추출, 학습함으로써 기존의 3차원 객체 추적 기술이 직면해 왔던 다양한 문제들(복잡한 배경, 조명 변화, 모션 블러, 가려짐, 객체 모양 변형 등)에 효과적으로 대처할 수 있다.

3차원 객체 추적과 관련하여 딥러닝 기술을 적용하는 방법은 적용 목적이나 형태에 따라 크게 세 가지로 분류될 수 있다: 템플릿(template) 매칭, E2E(End-to-End) 접근법, 특징 추출/매칭.

3.1 템플릿 매칭

입력 영상으로부터 객체의 전체 혹은 부분 영역을 표현하는 템플릿이나 로컬 패치(local patch)를 잘 구별되도록 표현하는 특징 벡터를 CNN나 CAE(Convolutional Auto-Encoder)을 이용하여 획득할 수 있다[21, 22](그림 2 참조). 특징 벡터를 데이터베이스에 저장된 ground truth 포즈 정보를 가지고 있는 템플릿이나 로컬 패치와 비교하여 객체의 포즈를 얻을 수 있다. 그러나, 객체 전체 영역을 포함하는 템플릿으로부터 특징 벡터를 획득하도록 학습하는 것은 가려짐(occlusion)에 취약하거나 객체의 부분적인 모양 변화(shape variation)를 가지는 객체에 대응할 수 없고, 로컬 패치로부터 특징 벡터를 획득하도록 학습하는 것은 잡음에 취약하기 때문에 두 방법론을 결합함으로써 보다 강건한 객체 추적이 가능하다[23, 24]. 예를 들어, 객체 전체 영역을 포함하는 템플릿을 이용한 CNN 학습을 통해 객체의 후보 포즈를 획득하고, 로컬 패치를 이용한 CAE 학습을 통해 각 로컬 패치의 후보 포즈와 3차원 좌표를 획득한 후, ICP(Iterative Closest Point) 알고리즘을 통해 템플릿의 후보 포즈와 로컬 패치의 후보 포즈에 근접한 객체 포즈를 최적화할 수 있다[23]. 유사한 예로, 객체 전체 영역을 포함하는 템플릿을 이용한 CNN 학습을 통해 객체의 영역을 검출하고, 검출된 객체 영역 내 로컬 패치를 이용한 CAE 학습 및 보팅(voting)을 통해 대략적인 객체 포즈를 얻은 후, PSO(particle swarm optimization)를 이용하여 정확한 객체 포즈를 계산할 수 있다[24].

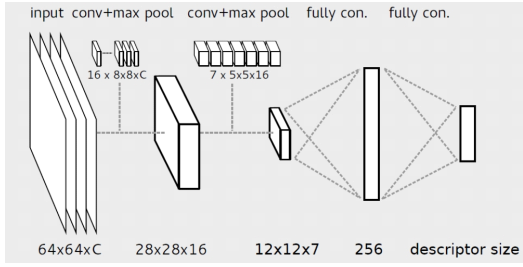


Fig. 2. Description of a template using a CNN made of two convolutional layers and two fully connected layers [21]

템플릿 매칭과 유사한 방법으로, 객체의 부분(parts) 내 미리 정의된 7개 제어 점(control points)의 2차원 투영 좌표를 CNN을 이용하여 학습(그림 3에서 보는 것처럼 각 부분의 중심점을 중심으로 하는 패치를 입력으로 사용)함으로써, 입력 영상에서 가려지지 않은 부분의 제어점의 투영 좌표를 검출하고, PnP(Perspective-N-Point) 알고리즘을 적용하여 객체의 포즈를 얻을 수 있다[25]. 또는 객체를 감싸는 바운딩 박스(bounding box)의 3차원 코너(corner) 점과 객체의 무게중심 점의 투영 좌표를 학습하고, 입력 영상으로부터 각 점의 투영 좌표를 검출하여 객체의 포즈를 얻을 수도 있으며[26], 반대로 SURF[27] 특징점을 중심으로 하는 RGB 로컬 패치를 입력으로 하여 로컬 패치 중심의 3차원 좌표를 얻는 CNN(RGBD 영상을 이용하여 학습)을 이용함으로써, 객체 포즈 추정을 위한 3차원-2차원 대응 좌표 쌍을 얻을 수도 있다[28].

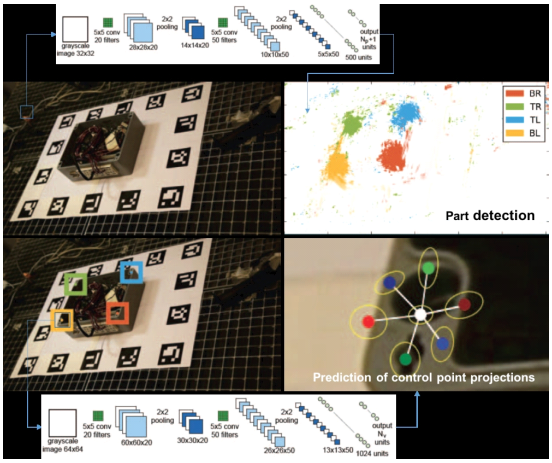


Fig. 3. Detection of object parts and prediction of control point projections in each part using CNNs [25]

3.2 E2E 접근법

E2E 접근법은 주어진 입력 영상으로부터 객체의 3차원 포즈를 직접적으로 추정하는 방법을 말한다. 일반적으로 CNN을 이용하여 입력 영상으로부터 특징 정보를 추출한 후, FCN을 이용하여 이산적으로 나누어진 포즈로 분류(classification)하거나[11, 24, 32], 회귀(regression)[14, 23, 29, 40]를 통해 객체의 포즈를 추정한다(그림 4 참조). 예를 들어, 입력 영상이 주어지면 객체 검출을 위해 미리 학습된 VGG-M[33]에서 마지막 완전 연결 계층(fully connected layers) 두 계층을 삭제한 네트워크로부터 특징 정보를 추출하고, 추출된 특징 정보를 입력으로 하여 그림 5의 FCN을 이용한 회귀를 통해 객체의 3차원 포즈를 추정할 수 있다[29]. 유사한 방법으로, RPN(Region Proposal Network)[37]을 이용하여 입력 영상으로부터 객체 후보 영역을 검출한 후, 각 객체 영역별로 CNN을 이용하여 특징 정보를 추출한 후, 그림 5와 유사한 FCN을 이용하여 객체 포즈를 추정함으로써, 입력 영상 내 여러 객체의 3차원 포즈를 동시에 추정할 수도 있다[38]. 그러나 입력 영상으로부터 특징 정보 추출과 회귀를 통해 직접적으로 추정된 포즈는 정확성이 떨어지기 때문에, 추정된 포즈가 실제(real)인지 위조(fake)인지를 판별하는 네트워크(discriminator)를 추가한 적대적 학습(adversarial learning)을 통해 추정된 포즈의 정확성을 향상시킬 수 있다[40].

연속된 프레임 사이에서의 포즈 변화를 직접적으로 추정할 수도 있다. 주어진 객체 포즈를 이용하여 객체를 투영한 영상($\mathbf{I}$)과 주어진 객체 포즈에 무작위로 포즈 변화량을 추가하고 투영된 객체를 임의의 RGBD 배경 영상(SUN3D 데이터 셋[34])과 합성한 후, 광원 위치 변경, 잡음 추가, 블러 합성, 가상 객체에 의한 가려짐 생성 등의 영상 생성 과정에 영향을 주는 다양한 효과를 모방하는 영상($\mathbf{J}$)을 생성한다. 생성된 두 영상($\mathbf{I}$, $\mathbf{J}$)을 CNN의 입력으로 하여 두 영상 사이의 상대적인 포즈 변화를 학습함으로써, 연속된 프레임 사이의 상대적인 포즈를 추정할 수 있다[30]. 다만, 입력으로 깊이 정보를 포함한 RGBD 영상이 사용되었다.

Fig. 4. E2E approach for object pose estimation

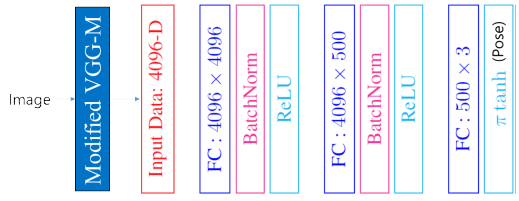


Fig. 5. Direct object pose regression [25]

3.3 특징 추출/매칭

입력 영상으로부터 반복성(repeatability)이 높은 특징점을 추출하고 특징점을 표현하는 서술자(descriptor)를

얻는 과정은 3차원 객체 추적을 위한 요소 기술이다. 딥러닝 기술을 활용하여 특징점을 검출하고 서술자를 획득하는 과정을 학습함으로써[16-19], SIFT[20], SURF[27]와 같은 기존 실험적, 경험적으로 고안된 필터나 계산식을 사용하는 방법보다 강건한 특징점 검출 및 매칭이 가능하다. 대표적인 예로, LIFT[16]는 SIFT의 주요 단계인 특징점 추출, 회전(orientation) 추정, 서술자 획득 과정을 CNN을 이용하여 학습하였다. 다른 예로, 특징점의 위치를 알고 있는 영상에 대해 각 특징점을 중심으로 가우시안(Gaussian) 함수를 적용하여 생성된 히트맵(heatmap)을 이용하여 영상으로부터 특징점 위치를 표현하는 히트맵을 생성하도록 stacked hourglass 구조의 네트워크를 학습시킬 수 있다[17]. 이를 통해 의미론적(semantic) 변환을 포함한 다양한 영상 변환에 강건한 특징점을 추출할 수 있다.

단일 영상으로부터 특징점을 추출하거나 특징점의 서술자를 도출하는 방식과 달리, 영상 쌍으로부터 특징점을 검출하고, 특징점 사이의 기하학적(geometric), 의미론적 대응 관계(correspondence)를 보다 정확하게 도출하는 데 초점을 둔 연구들도 있다. UCN[18]은 특징점의 위치를 알고 있는 입력 영상 쌍이 주어지면 CNN을 이용하여 각 특징점에서의 서술자를 도출하고, 서술자 사이의 거리를 최소화하도록 CNN을 학습시켰다. 따라서, 학습된 CNN은 대응 관계에 있는 특징점들의 서술자가 매우 유사하도록 서술자를 생성한다. 또한, 학습 시 대조 손실(contrastive loss) 함수를 사용하여 잘못된 대응 관계를 가지는 특징점 쌍도 학습에 이용함으로써 학습 능력이 크게 향상되었다. SuperPoint[19]는 학습 데이터셋을 얻기 위해 특징점의 위치를 알고 있는 합성 영상을 이용하여 지도 학습(supervised learning)을 통해 영상으로부터 특징점을 검출하는 CNN을 학습한 후, homographic adaptation을 통해 합성 영상을 이용하여 학습된 CNN이 다양한 영상 변환을 포함하는 실제 영상에서 특징점을 검출할 수 있도록 하였다. 특징점의 위치 및 특징점 사이의 대응 관계를 알고 있는 입력 영상 쌍(학습 데이터셋의 임의의 영상으로부터 특징점을 검출하고, 무작위로 주어진 homography를 이용하여 영상을 변형(warp))이 주어지면, UCN과 마찬가지로 샴(siamese) 네트워크를 이용하여 각 영상으로부터 특징점을 검출하고 각 특징점의 서술자를 생성한 후, 각 특징점의 위치는 ground truth에 가깝도록 하고 서로 대응되는 특징점의 서술자는 유사해지도록 샴 네트워크를 학습시켰다.

특징점 매칭과 관련된 연구로, 참조(reference) 영상에서 검출된 각 특징점을 중심으로 하는 패치에 다양한 영상 변환을 적용한 후, 영상 변환이 적용된 패치를 입력으로 받아 각 특징점의 레이블(label)을 출력하도록 CNN을 학습시킴으로써, 입력 영상에서 특징점이 검출되면 특징점의 서술자를 얻지 않고, CNN을 이용한 분류를 통해 참조 영상에서 검출된 특징점과 매칭할 수 있다[31]. 또는 영상 쌍으로부터 SIFT와 같은 방법을 이용하여 특징점을 검출하고 특징점의 서술자 사이에 최근린법(nearest neighbor)을 적용하여 대응 쌍을 구한 후, CNN 입력으로 특징점 대응 쌍의 좌표가 주어지면 inlier인지 outlier인지 판별할 수 있다[36].

에지(edge) 기반 객체 추적 기술의 경우[10], 복잡한 배경에서 정확성이 떨어지는 문제가 있는데, 최근 GAN(Genarative Adversarial Network)을 활용하여 입력 영상으로부터 객체의 경계선(boundary)을 정확하게 추출하고(그림 6 참조) 대략적인 객체(카메라)의 초기 포즈를 함께 추정(auto-encoder로부터 추출된 특징 정보를 분류)함으로써, 복잡한 배경에 의한 영향을 크게 줄일 수 있는 방법이 제안되었다[11].

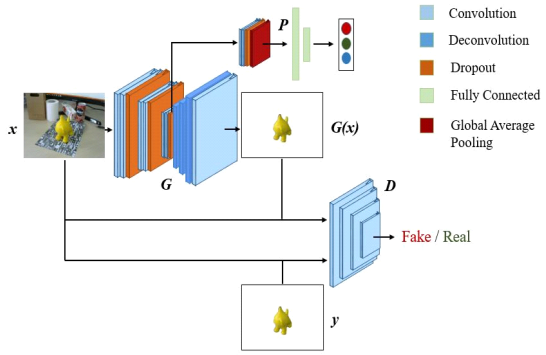


Fig. 6. Object boundary extraction using adversarial learning [11]

Ⅳ. 딥러닝 기술 활용의 한계

딥러닝 기술을 사용하기 위해서는 다양한 조건이나 환경에서 획득된, ground truth 포즈를 가진 영상 데이터를 필요로 하는데, 이러한 영상 데이터를 실제로 획득하는 것은 매우 어렵다. 그러므로, 많은 연구들에서 다양한 조건이나 환경을 시뮬레이션해서 생성된 합성(synthetic) 영상을 활용하고 있다. 그러나, 렌더링 시스

템의 성능이 크게 향상되었고 합성 영상을 생성하기 위한 효과적인 툴[12]도 제안되고 있음에도 불구하고, 여전히 다양한 환경적인 정보를 포함한 실제 영상과 유사한 합성 영상을 생성하는 것은 매우 어려우며, 이로 인해 객체 추적의 성능이 저하될 수 있다. 이를 해결하기 위해 최근 실제 영상과 합성 영상 사이의 차이를 줄이기 위한 연구가 진행되고 있다. 한 예로, auto-encoder를 이용하여 실제 영상에서 추출된 특징점 벡터와 다양한 조명 조건을 가진 합성 영상에서 추출된 특징점 벡터 사이의 관계를 학습함으로써, 실제 영상에서 추출된 특징점 좌표를 합성 영상에서 추출된 특징점 좌표로 변환할 수 있다[13]. 또는 색상을 포함한 appearance 정보를 없애고 pencil 필터를 사용하여 에지 정보를 강조한 합성 영상을 사용함으로써 실제 영상과 합성 영상의 차이를 줄일 수도 있다[14]. 보다 일반적인 방법으로, 충분히 다양한 환경 조건에서 무작위로 랜더링된 합성 영상을 생성함으로써(domain randomization 방법[15]), 실제 영상에 포함된 환경 조건의 영향을 최소화할 수 있다.

앞서 설명한 것처럼 딥러닝 기술만을 사용하여 추정된 3차원 객체의 포즈는 정확성이 떨어진다[35]. 그러므로, 기존 최적화 기반의 3차원 객체 추적 방법과 결합하는 것이 필요하며, 일반적으로 최적화 기반의 3차원 객체 추적을 시작하거나 복원(recovery)하기 위한 대략적인 포즈 정보를 제공하는 데 활용되고 있다.

Ⅴ. 향후 전망

딥러닝 기술은 앞으로도 3차원 객체 추적 분야에서 다양한 형태로 활용될 것이다. 그러나, 기존 연구들은 단일 영상으로부터 객체의 3차원 포즈를 추정하기 위한 다양한 형태의 솔루션을 제공하고 있지만, 연속된 프레임에서 객체를 정확하고 안정적으로 추적하기 위한 연구는 많지 않다. 그러므로, 향후 연구는 연속된 프레임 사이의 상관성을 분석, 활용하기 위해 딥러닝 기술을 적용하는 방향으로 확대될 필요가 있다.

대표적인 예로, 기존 연구 성과를 활용하면 각 프레임 내에서 객체의 절대적인(absolute) 3차원 포즈는 추정할 수 있지만, 프레임 사이의 상관성를 고려하지 않기 때문에 객체 추적 시 지터링(jittering) 문제가 발생할 수 있다. 안정적인 객체 추적을 위해서는 연속된 프레임 사이의 상대적인 포즈 변화를 추정하는 것이 적절하며, 이를 위해 연속된 프레임 사이의 정보 변화를 CNN을 이용하여 추출하고 학습[30]하거나 RNN(Recurrent Neural Network)을 이용하여 여러 프레임에서 추출된 정보를 동시에 학습[39]하는 방법 등을 개선, 발전시킬 필요가 있다. 특히 RGBD 영상이 아닌 RGB 영상을 입력으로 하는 방법이 필요하다.

영상은 촬영 조건이나 환경에 따라 정보가 크게 달라지고, 특정 객체(예, 반짝이거나 투명한 객체)의 경우 영상(특히 RGB 영상)으로부터 추적에 필요한 충분한 정보를 추출할 수가 없으며, 학습에 노출된 적이 없는 객체를 추적하는 것은 매우 어렵다[35]. 그러므로, 안정적인 3차원 객체 추적을 위해서는 단순히 영상 정보만을 활용하지 않고, IMU와 같은 센서로부터 획득된 정보를 함께 학습하기 위해 딥러닝 기술을 활용하는 연구[41]가 보다 활성화되어야 한다. 이러한 센서 융합은 IMU를 포함한 각종 센서를 탑재한 스마트폰, 태블릿과 같은 모바일 기기를 이용한 3차원 객체 추적 응용 분야(예, 증강현실)에서는 필수적인 연구 분야이다.

대부분의 3차원 객체 추적 기술들은 PC 환경에서 개발, 검증되고 있으며, 모바일 환경에서 구현할 경우 정확성이나 강건성이 크게 떨어질 수 있다. 이는 컴퓨팅 리소스가 부족하기 때문에 PC 환경에 비해 간소화(simplification)되거나 근사화된(approximated) 알고리즘이나 모델을 사용해서 구현하는 데서 비롯된다[42]. 그러므로, 모바일 환경에서도 성능을 유지할 수 있도록 가볍고(light) 얕은(shallow) 네트워크 구조를 개발할 필요가 있다. 센서 융합은 계산 효율성 측면에서도 모바일 환경에서 좋은 방안이 될 수 있다.

ACKNOWLEDGMENTS

이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2018R1D1A1B07045650).

REFERENCES

[1] A. Dey, M. Billinghurst, R. W. Lindeman, and J. E. Swan, “A systematic review of 10 years of augmented reality usability studies: 2005 to 2014,” Front. Robot. AI, vol. 5, article 37, 2018.

[2] K.-M. Lee and J.-I. Kim, “Design and implementation of hybrid VR lock system by Arduino control,” The Journal of Korea Institute of Signal Processing and Systems, vol. 15, no. 3, pp. 97-103, 2014.

[3] Y. Wu, F. Tang, and H. Li, “Image-based camera localization: an overview,” Visual Computing for Industry, Biometric, and Art, vol. 1, article number: 8, 2018.

[4] V. A. Knyaz, O. Vygolov, V. V. Kniaz, Y. Vizilter, and V. Gorbatsevich, “Deep learning of convolutional auto-encoder for image matching and 3D object reconstruction in the infrared range,” Proc. of ICCVW, pp. 2155-2164, 2017.

[5] E. Marchand, H. Uchiyama, and F. Spindler, “Pose estimation for augmented reality: a hands-on survey,” IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 12, pp. 2633-2651, 2016.

[6] H. Park and J.-I. Park, “Recent trends and analysis on AR technology - focused on 3D object tracking methods,” Proc. of The Korean Institute of Broadcast and Media Engineers Summer Conference, pp. 299-300, 2018.

[7] P. Han and G. Zhao, “A review of edge-based 3D tracking of rigid objects,” Virtual Reality & Intelligent Hardware, vol. 1, no. 6, pp. 580-596, 2019.

[8] Y. Shavit and R. Ferens, “Introduction to camera pose estimation with deep learning,” arXiv preprint arXiv:1907.05272, 2019.

[9] R. Hartley and A. Zisserman, Multiple View Geometry, 2nd Ed., Cambridge University Press, 2003.

[10] B. Wang, F. Zhong, and X. Qin, “Pose optimization in edge distance field for textureless 3D object tracking,” Proc. of the Computer Graphics International Conference, article no. 32, 2017.

[11] X. Liu, J. Zhang, X. He, X. Song, and X. Qin, “6DoF pose estimation with object cutout based on a deep autoencoder,” Proc. of ISMAR-Adjunct, 2019.

[12] S. Zhang, C. Song, and R. Radkowski, “Setforge - synthetic RGB-D training data generation to support CNN-based pose estimation for augmented reality,” Proc. of ISMAR-Adjunct, pp. 227-232, 2019.

[13] S. Shoman, T. Mashita, A. Plopski, P. Ratsamee, Y. Uranishi, and H. Takemura, “Illumination invariant camera localization using synthetic images,” Proc. of ISMAR-Adjunct, pp. 143-144, 2018.

[14] J. Rambach, C. Deng, A. Pagani, and D. Stricker, “Learning 6DoF object poses from synthetic single channel images,” Proc. of ISMAR-Adjunct, pp. 164-169, 2018.

[15] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” Proc. of IROS, pp. 23-30, 2017.

[16] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “LIFT: learned invariant feature transform,” Proc. of ECCV, pp. 467-483, 2016.

[17] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, “6-DoF object pose from semantic keypoints,” Proc. of ICRA, pp. 2011-2018, 2017.

[18] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, “Universal correspondence network,” Proc. of NIPS, pp. 2414-2422, 2016.

[19] D. DeTone, T. Malisiewicz, and A. Rabinovich, “SuperPoint: self-supervised interest point detection and description,” Proc. of CVPRW, 2018.

[20] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” IJCV, vol. 60, no. 2, pp. 91-110, 2004.

[21] P. Wohlhart and V. Lepetit, “Learning descriptors for object recognition and 3D pose estimation,” Proc. of CVPR, pp. 3109-3118, 2015.

[22] W. Kehl, F. Milletari, F. Tombari, S. Ilic, and N. Navab, “Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation,” Proc. of ECCV, vol. 3, pp. 205-220, 2016.

[23] K. Park, J. Prankl, and M. Vincze, “Mutual hypothesis verification for 6D pose estimation of natural objects,” Proc. of ICCVW, pp. 2192-2199, 2017.

[24] H. Zhang and Q. Cao, “Combined holistic and local patches for recovering 6D object pose,” Proc. of ICCVW, pp. 2219-2227, 2017.

[25] A. Crivellaro, M. Rad, Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit, “A novel representation of parts for accurate 3D object detection and tracking in monocular

images," Proc. of ICCV, pp. 4391-4399, 2015.

[26] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," Proc. of CVPR, pp. 292-301, 2018.

[27] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359, 2008.

[28] N.-D. Duong, A. Kacete, C. Sodalie, P.-Y. Richard, and J. Royan, "xyzNet: towards machine learning camera relocalization by using a scene coordinate prediction network," Proc. of ISMAR-Adjunct, pp. 258-263, 2018.

[29] S. Mahendran, H. Ali, and R. Vidal, "3D pose regression using convolutional neural networks," Proc. of ICCVW, pp. 2174-2182, 2017.

[30] M. Garon and J.-F. Lalonde, "Deep 6-DOF tracking," IEEE Trans. on Vis. and Comp. Grap., vol. 23, no. 11, pp. 2410-2418, 2017.

[31] O. Akgul, H. I. Penekli, and Y. Genc, "Applying deep learning in augmented reality tracking," Proc. of SITIS, pp. 47-54, 2016.

[32] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views," Proc. of ICCV, pp. 2686-2694, 2015.

[33] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convolutional nets," Proc. of BMVC, 2014.

[34] J. Xiao, A. Owens, and A. Torralba, "SUN3D: a database of big spaces reconstructed using SfM and object labels," Proc. of ICCV, pp. 1625-1632, 2013.

[35] T. Sattler, Q. Zhou, M. Pollefeys, Laura Leal-Taixe, "Understanding the limitations of CNN-based absolute camera pose regression," Proc. of CVPR, pp. 3297-3307, 2019.

[36] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," Proc. of CVPR, pp. 2666-2674, 2018.

[37] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[38] T.-T. Do, M. Cai, T. Pham, and I. Reid, "Deep-6DPose: recovering 6D object pose from a single RGB image," arXiv preprint arXiv:1802.10367, 2018.

[39] T. X. Qing, W. Fan, and Z. Y. Tao, "Camera pose estimation method based on deep neural network," Proc. of ICDLT, pp. 85-90, 2019.

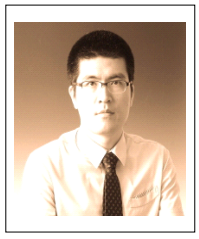
[40] M Bui, C. Baur, N. Navab, S. Ilic, and S. Albarqouni, "Adversarial networks for camera pose regression and refinement," Proc. of ICCVW, pp. 3778-3787, 2019.

[41] J. R. Rambach, A. Tewari, A. Pagani, and D. Stricker, "Learning to fuse: a deep learning approach to visual-inertial camera pose estimation," Proc. of ISMAR, pp. 71-76, 2016.

[42] V. A. Prisacariu, O. Kahler, D. W. Murray, and I. D. Reid, "Real-time 3D tracking and reconstruction on mobile phones," IEEE Trans. on Vis. and Comp. Grap., vol. 21, no. 5, pp. 557-570, 2015.

저자 소개

박 한 훈 (Hanhoon Park)



2000년 2월 : 한양대학교 전자통신전파공학과(공학사)
2002년 2월 : 한양대학교 전자통신전파공학과(공학석사)
2002년 2월 : 한양대학교 전자통신전파공학과(공학박사)
2012년 3월~현재 : 부경대학교 전자공학과 교수
관심분야 : 증강현실, 인간컴퓨터상호작용, 컴퓨터비전/그래픽스