Smart Media Journal / Vol.14, No.11 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2025.14.11.166

AI 기반 비정형 물류 팔레타이징 시스템 비교 실증

(Experimental Comparison of AI-Based Palletizing Systems for Unstructured Logistics Environments)

윤봉식*, 백상윤**
(Bong Shik Yun, Sang Yun Baek)

요 약

본 연구는 비정형 물류 환경에서 자동 적재 시스템 고도화를 위해 RGB-Depth 융합 기반 3D 비전 기술과 AI 객체 인식 모델을 결합한 프로토타입을 실증하는 데 목적이 있다. 이를 위해 CNN 기반 YOLOv11, Transformer 기반 Swin Transformer, 그리고 Point Cloud 기반 PointNet++ 모델을 동일 조건하에 실험하여 객체 인식정확도, 상단인식률, 자세추정 오차(RPY), 연산 속도(FPS) 등 주요 성능지표를 비교·분석하였다.

실험 결과, YOLOv11은 mAP50 99.5%, 상단 인식률 96.4%, RPY 오차 ±4.2°, 52.1 FPS로 전체적으로 가장 우수한 성능을 보여 실시간 산업 환경 적용 가능성이 입증되었다. 이 밖에도 Swin Transformer는 Occlusion 환경에서의 견고성, PointNet++는 자세 인식 정밀도에서 강점을 보였으며, 본 연구의 결과는 협동로봇 기반 팔레타이징 시스템에 적용 가능한 통합 인식-제어 구조 개발의 기초 자료로 활용될 수 있다.

■ 중심어 : 3D 비전 ; 비정형 객체 인식 ; RGB-Depth 융합 ; AI 적용기술 실증

Abstract

This study aims to prototype an automated palletizing system for unstructured logistics environments by integrating RGB-Depth 3D vision with AI-based object recognition. The CNN-based YOLOv11, Transformer-based Swin Transformer, and Point Cloud-based PointNet++ models were experimentally evaluated under controlled conditions using 300 labeled samples.

The results indicate that YOLOv11 achieved the highest performance (mAP50 99.5%, top recognition 96.4%, ±4.2° RPY error, 52.1 FPS), demonstrating its feasibility for real-time industrial deployment. These findings provide foundational evidence for developing integrated vision-control architectures applicable to collaborative robot palletizing systems

■ keywords : 3D Vision ; Unstructured Object Recognition ; RGB-Depth Fusion ; Experimental Verification of AI Technology

Ⅰ. 서 론

1. 연구 배경 및 목적

최근 물류 및 제조 산업에서는 생산성 증대 및 작업자 효율성 개선을 목적으로 자동화 시스템의 도입이 활발히 진행되고 있다. 다양한 형태의 포대, 사료, 곡물 포장 등 비정형 객체는 크기, 형태, 기울기, 재질 등 특성이 불균질하기 때문에, 기존의 정형화된 자동화 장비로는 정확한 개체의 인식과 적재 정보에 대한 정확성 확보에 한계가 있다. 이러한 환경에서 3D 비전 기반 인식 기술과 AI 객체 검출 알고리즘과 3D 비전 기술의 융합은 자동화 수준을 향상시키는 핵심 대안으로 주목받고 있다.

YOLO 계열의 CNN 기반 객체 검출기는 실시간

* 정회원, 남부대학교 자동차기계공학과
** 정회원, 다소솔루션 대표
본 과제(결과물)는 2025년도 교육부 및 광주광역시의 재원으로 광주RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-05-007)
접수일자 : 2025년 10월 17일
수정일자 : 2025년 11월 10일
게재확정일 : 2025년 11월 13일
교신저자 : 백상윤 e-mail : daso_solution@naver.com

처리성과 구조가 단순하다는 장점에 반해, 가림(Occlusion), 반사 재질, 기울기 변화 등 복잡한 조건에서는 검출 정확도가 급격히 하락하는 문제점이 보고 되어 왔다[1].

반면, Transformer 기반 비전 모델과 포인트 클라우드 기반 점군 네트워크는 이러한 변형 및 복잡 조건에도 더 높은 강인성을 보이는 사례가 다수 발표되고 있다[2-3].

또한, 물류자동화 기술이 고도화될수록 비정형 객체 인식 정확성과 처리속도가 전 공정효율에 미치는 영향은 더욱 커지고 있고, 복잡한 현장 조건을 견딜 수 있는 인공지능 기반 적응형 인식시스템 개발이 시급한 과제로 대두되고 있다

2. 연구 목적

본 연구의 목적은 RGB 이미지와 Depth 정보를 융합한 모델 구조 하에서, 비정형 포대 객체 검출 성능을 YOLO 기반 CNN, Swin Transformer 기반 모델, PointNet++ 기반 점군 네트워크 간에 비교·분석하는데 있다[4].

이를 통해 실제 물류 환경에서 발생이 가능한 가림, 색상 변화, 반사, 기울기 등 다양한 조건에서도 안정적으로 동작할 수 있는 인식 구조 조합을 탐색하고, 향후 팔레타이징 자동화 시스템 설계의 기반 기술로 활용하고자 한다[5].

3. 기존 연구 동향 분석

maechMind Robotics는 물류 현장 비정형 포대 및 박스 대응을 위한 AI 3D 비전 시스템을 개발했으며, 포인트 클라우드 기반 인식 및 충돌 회피, 상단 우선 적재 전략 등을 제시한 바 있다[6].

Fizyr는 Zivid의 고정밀 3D 카메라와 딥러닝 기반 객체 인식 알고리즘을 결합하여 불규칙한 형태의 객체에 대한 파지 및 분류 성능을 향상한 연구를 발표하였다[7].

Swin Transformer은 "Hierarchical Vision Transformer using Shifted Windows" 논문에서 계층적 구조와 윈도우 기반 self-attention 기법을 제안하며, 다양한 크기의 객체 탐지와 Occlusion 대응 성능이 기존 CNN 대비 우수하다는 것을 입증하였다[2].

PointNet++ 논문은 포인트 클라우드 입력을 계층적으로 처리하면서 점 집합의 지역 구조를 반영하는 네트워크 구조를 제시하였고, 복잡한 3D 객체 인식에서 뛰어난 성능을 보임을 실험적으로 증명하였다[3].

Ⅱ. 이론적 배경 및 관련 기술

1. 3D 비전 시스템

3D 비전 시스템은 객체의 입체 정보를 인식하여 로봇 제어, 검사, 팔레타이징 등 다양한 산업 분야에서 활용되고 있다. 2D 비전이 평면 상의 색상·형상 정보에 국한되는 반면, 3D 비전은 깊이(Depth) 정보와 위치, 기울기 등의 정량적 좌표 데이터를 제공한다[5].

대표적인 3D 센서 기술에는 스테레오 비전, ToF(Time-of-Flight), 구조광 방식(Structured Light), 레이저 스캐닝 등이 있으며, 물류·제조 분야에서는 높은 프레임 속도와 비교적 저렴한 구조광 및 ToF 센서가 널리 사용된다[8].

3D 비전 시스템은 대개 RGB 이미지와 Depth 맵을 동시에 수집하는 방식으로 구성되며, 이때 생성된 포인트 클라우드(Point Cloud)는 후속 AI 기반 인식 및 로봇 제어의 핵심 입력 데이터가 된다. 최근에는 RGB-Depth 융합 네트워크가 성능 향상에 유효함이 입증되며, 연구가 활발히 진행되고 있다[9].

2. YOLO 계열 CNN 객체 검출기

YOLO(You Only Look Once)는 단일 CNN 네트워크가 객체 위치와 클래스 예측을 동시에 수행하는 대표적인 실시간 객체 검출 프레임워크로 v1~v5까지는 backbone 구조의 개선 및 anchor box 전략 강화에 초점을 맞췄고, 최근 발표된 v7과 v8은 transformer 블록 도입, 모델 경량화,

segmentation/pose estimation 통합 기능을 통해 복합 시각 작업에 대응할 수 있도록 발전하였다[10-11].

표 1. mAP: Mean Average Precision / Occlusion 인식률

모델	mAP (%)	Occlusion 인식률 (%)	연산 속도 (FPS)
YOLOv8	87.3	78.5	58.0
Swin Transformer	92.1	91.8	35.0
PointNet++	90.4	89.2	42.0

YOLO 계열 모델은 빠른 속도와 mAP 기준 비교적 높은 정확도를 제공하지만, 객체의 부분 가림(Occlusion)이나 기울기 변화에 민감하고 동일한 물체가 겹쳐진 경우(Bounding box 중첩) 처리 오류가 발생할 수 있거나 3D 형태 정보가 부재하여 공간 정렬 및 회전 추정이 불가능하는 등의 한계를 보이고 있다. 이렇듯 YOLO 기반 모델은 RGB 영상 기반의 전면 검출에는 적합하지만, 비정형 물류에서의 3D 자세 추정 및 정확한 파지 위치 인식에는 한계를 고려해야 한다[12].

이에 본 연구에서는 YOLO 계열 모델 중 최신 버전인 YOLOv11을 실험 대상으로 선정하였다. YOLOv11은 기존 YOLOv8 대비 C2f 기반 백본 구조의 개선, Deformable Convolution Layer의 도입, 그리고 EIoU(Extended Intersection over Union) Loss 함수의 향상된 정합성을 통해 비정형 객체의 형태 왜곡(Deformation) 및 부분 가림(Occlusion) 상황에서도 높은 검출 안정성을 확보하였다. 이러한 구조적 개선은 모델의 공간 적응성(Spatial Adaptability)과 수렴 안정성(Convergence Stability)을 동시에 향상시킬 수 있어 비정형 물류 포대와 같이 불규칙한 형태를 갖는 객체 인식 환경에서 실시간 검출 성능을 극대화할 수 있다[15].

3. Swin Transformer 구조

본 연구에서는 가림(Occlusion)이나 복잡한 배경 조건에서의 강인성을 검증하기 위해 기존 CNN 구조의 한계를 극복하고자 제안된 Swin Transformer 모델을 채택하였다. 이 모델은 Shifted Window 기반 Self-Attention 구조를 통해 지역적 세부 특징과 전역적 문맥 정보를 동시에 학습할 수 있고, 연산을 통해 다양한 크기의 객체를 안정적으로 인식할 수 있다[2]. 또한 RGB 영상에서 발생하는 반사·음영·기울기 변화에도 비교적 안정적 인식 성능을 제공한다[16].

기존 CNN이 지역 패턴에 민감한 반면, Swin Transformer는 전역적 문맥 정보를 기반으로 물체를 인식함으로써 Occlusion 상황에서 인식률이 상대적으로 높고, 윈도우 기반 attention 구조 덕분에 연산 효율도 보장되어 산업용 Edge- AI 시스템에도 적용이 용이한 장점이 있다. 다만, Swin Transformer는 3D 인식 구조가 내재 되어 있지 않아, Depth map 또는 Point Cloud의 깊이 정보와 융합이 필수적이며, 해당 융합 전략에 따라 인식 성능의 차이가 발생할 수 있다[13].

4. PointNet++ 기반 점군 네트워크

PointNet++는 불규칙한 포인트 클라우드 입력을 처리하기 위한 딥러닝 모델로, 공간상의 점(Points) 집합에서 지역적 특징(Local Feature)과 전역적 특징(Global Feature)을 함께 추출할 수 있는 구조를 가진다[2]. 또한 정형화되지 않은 물체, 부분 가림, 기울어진 상태 등 다양한 3차원 조건에서도 객체를 안정적으로 인식할 수 있다. Point Set의 순서 불변성, 적응적 neighborhood sampling, 자세 인식에 활용 Roll/Pitch/Yaw 값 추정이 가능하는 등의 차별적 특징이 있다.

비정형 포대 인식에는 포대의 윗면, 옆면, 경계점 정보 등 불규칙한 형태 분석이 필요한데, PointNet++은 이러한 비정형 정보를 정밀하게 분석할 수 있어 팔레타이징 로봇의 제어 정확도를 높이는 데 매우 유용하다[14].

5. RGB-Depth 융합 전략

최근 연구에서는 다음의 표2와 같이 RGB와 Depth 정보를 동시에 활용하는 융합 인식 구조(Fusion Architecture)가 활발히 연구되고 있다

[15,17,18].

표 2. RGB-Depth 융합 전략

융합방식	주요 개념	특징 및 적용 예시
Early Fusion	RGB/Depth 데이터를 채널 단위 결합, 입력 단계에서 통합	단순한 구조, 연산 효율은 높으나, RGB·Depth 간의 특성 차이 반영 어려움
Mid-level Fusion	RGB와 Depth 각각 Encoder에서 특징 추출 후, 중간계층 병합	두 모달리티의 독립적 특징을 유지 상호 보완, 균형 잡힌 인식 성능 확보
Late Fusion	RGB 및 Depth의 독립적 추론 결과를 후단에서 결합하여 최종 판단 수행	각 모듈의 최적화를 유지할 수 있으나, 실시간 응용에는 연산 부담이 증가

RGB-Depth 융합은 서로 다른 modality의 시각 정보를 활용함으로써 반사/색상/기울기/가림 등 다양한 조건에서도 안정적인 인식이 가능하도록 한다[13]. 특히 산업용 팔레타이징 시스템에서는 3D 좌표 추정과 위치 정렬을 동시에 수행해야 하므로, RGB-Depth 융합 구조의 실효성이 매우 높다[14].

Ⅲ. 연구 방법

1. 실험 목적 및 환경 구성

본 실험은 RGB-Depth 융합 기반 비정형 포대 인식 시스템을 설계하고, 세 가지 대표적 인식 구조인 CNN 기반의 YOLO v11, Transformer 기반의 Swin Transformer, Point Cloud 기반 PointNet++의 성능을 동일 조건 하에 비교·분석하는 것을 포함한다. 특히 가림(Occlusion), 기울기(Rotation), 색상/재질의 변화 등 현실적인 물류 환경 변수를 반영하여 인식 정확도, 기울기 추정 오차, 연산 속도, 상단 포대 인식률을 실증 실험 과정을 통해 정량적으로 측정하였다.

표 3. 실험환경 구성

항목	내용
실험 장소	광주광역시 소재 N대학교, 기업연구소
조명 환경	확산광 조명 (5,600K), 자연광 차단 실내 조건
테스트베드	비정형 포대류 3종 (색상·크기·재질 다양), 시험용 팔레트, 회전기구
센서 장비	Intel RealSense D455 (RGB + Depth 동시 촬영)
연산 플랫폼	NVIDIA RTX 4080 GPU / CUDA 12.1 / Ubuntu 22.04 / PyTorch 2.0
비교 모델	YOLOv8 (Ultralytics), Swin Transformer (Timm), PointNet++ (Open3D)
평가 횟수	조건당 반복 측정, 총 300장 실측 수행

2. 실험 조건 설계

가. 데이터셋 구성

그림 1 샘플 촬영 및 객체 정보화

총 300장의 RGB+Depth 동기화 샘플을 촬영하고, 라벨링 도구(RoboFlow 및 LabelMe)를 활용해 바운딩 박스와 상단 객체 정보를 부여하였다.

나. 실험 데이터 표준화

실험 데이터의 훈련과 검증 및 테스트 비율 6:2:2로 분할하였다.

그림 2 실험 데이터 표준화

다. 전처리 및 증강

Depth 이미지에는 Gaussian 및 Median 필터를 적용해 표4와 같이 노이즈를 제거하였다.

표 4. 필터링 알고리즘 적용

Filter Type	Kernel Size	σ (Std. Dev.)	Purpose
Gaussian	5×5	1.0	Noise smoothing
Median	3×3	—	Edge-preserving denoise

라. 데이터 증강

데이터 증강을 위해 Random Flip, Rotation (±30°), Brightness 값을 ±20% 조정하고, Occlusion 조건은 반쯤 가려진 포대 배치로 인위적으로 생성하였다.

그림 3 데이터 증강

마. 모델 학습 설정

YOLOv8은 Epoch 200, batch size 16, lr = 0.001로 정하고, Swin Transformer는 Epoch 100, pretrained weight 사용하였다. PointNet++는 1024 points 샘플링, 3-layer SA 모듈, batch norm 적용하여 각 모델이 동일한 연산 환경에서 재현성 확보를 위해 seed를 고정하였다.

바. 측정 항목별 평가

mAP는 COCO 평가 기준 IOU ≥ 0.5에서 측정하고, 상단 인식률은 Occlusion 조건에서 수동 정답 일치율로 측정하였다. FPS는 평균 연산 속도를 반복 산출하고, 기울기 오차는 Roll/Pitch /Yaw 값과 인식값의 차이로 계산하였다 [11]

표 5. 실험 변수 설계

변수명	분류	수준 (예시)	설명
포대 상태	독립변수	정형 / 찌그러짐 / 겹침	형태 변화에 따른 인식 영향 확인
조명 조건	독립변수	일반광 / 반사광 / 그림자 포함	센서 반응 차이 측정 목적
입력 방식	독립변수	RGB / Depth / RGB+Depth 융합	정보 조합별 성능 비교
인식 모델 구조	독립변수	YOLOv8 / Swin Transformer / PointNet++	구조별 성능 비교
인식 정확도 (mAP)	종속변수	IOU ≥ 0.5 기준 평균정밀도 (%)	모델 성능 핵심 지표
상단 인식률	종속변수	Occlusion 상황에서 상단 포대 검출 성공률	상단 우선 인식 능력
기울기 추정 오차	종속변수	Roll/Pitch/Yaw 오차 ±도 단위 (%)	자세 추정 능력

3. 성능 비교 지표

mAP, FPS, 기울기 오차, 인식률 등의 성능 비교 지표는 각 별 성능 비교를 위해 동일한 실험 조건에서 YOLOv8, Swin Transformer, PointNet++ 모델에 대해 실험이 진행될 수 있도록 다음과 같이 설정하였다.

표 6. 실험환경 구성

지표명	정의
mAP50	IOU ≥ 0.5에서의 평균 정밀도
mAP50~95	IOU ≥ 0.5에서의 평균 정밀도
상단 인식률(%)	겹침/가림 조건에서 최상단 객체 검출 성공 비율
기울기 오차 (°)	RPY(Roll, Pitch, Yaw) 추정값–실제값 평균 편차
FPS	초당 프레임 처리 속도 (Frames Per Second)

국내에서는 1990년대 중반부터 3차원 영상 및 3DTV 시스템에 대한 연구가 시작되어 현재까지 활발하게 진행되고 있다. 에지 검출을 수행하여 수직 및 수평 성분을 분석하여 텍스트 영역을 검출한다. 본 연구에서는 간판의 텍스트 영역이 영상의 중심선에 수평으로 존재한다고 가정하였으며[2,3] 하나의 간판에 대하여 고려한다. 일반적으로 텍스트 영역에서는 텍스트 구성 획의 특성이 수직 및 수평의 패턴으로 문자를 형성하기 때문에 수직/수평의 에지 히스토그램의 분포에 의하여 간단하게 검출을 수행할 수 있다[1-3]. 일반적으로 텍스트 검출을 위한 방법은 영상에서 텍스트 영역을 검출하기 위하여 low-level의 특징들을 사용한다.

Ⅳ. 실험 및 분석 결과

1. 실험모델 설계 결과

실험용 모델은 객체 인식과 상단 후보 분류과정을 통해 객체들을 인식한 뒤, Bounding box 상단 좌표를 기준으로 가장 위쪽 객체 후보군을 추출하고, 깊이 정보 적용 및 상단 객체 선택을 위해 RealSense D405 카메라의 Depth Map을 기반으로

후보 객체들의 Z값을 비교하여 실제 상단에 존재하는 객체 1개를 최종 선정하도록 설계하였다.

무게중심 추정과 파지 경로 전달은 선정된 객체의 Bounding box로부터 중심좌표를 계산하고, 해당 중심점을 기반으로 로봇 파지 경로를 자동 산출하여 TCP/IP 프로토콜 기반 UR3 로봇 제어기를 통해 그리퍼 파지 동작 수행하는 공정을 부여하였다.

표 7. 실험 환경 구성 세부 사양

구성 요소	세부 사양
산업용 로봇	UR3, 6자유도 협동로봇
그리퍼	OnRobotics VGC10 공압식 그리퍼
Depth 카메라	Intel RealSense D405 (RGB + Depth 동시 취득)
AI 인식 모델	YOLO v11 (Ultralytics 기반 최신 CNN 객체 검출기)
입력 데이터	자체 구축 비정형 포대 이미지 300장 (RGB + Depth), 다양한 기울기·변형 포함
출력 항목	Bounding box, class label, 중심좌표, 무게중심 추정

본 실험과정에서는 다음의 그림과 같은 실험로봇을 구성하여 인식과 상단 선정 및 파지 정보를 송출하는 단계에 대한 실증으로 진행되었다.

그림 4. 팔레타이징 실험로봇 구성

2. 객체 인식 성능의 실증 결과

YOLO v11 모델은 비정형 포대 300장의 학습 데이터로 훈련되었으며, 이후 테스트용 신규 이미지에 IOU ≥ 0.5 기준 평균 정밀도 mAP50 0.995를 보였다.

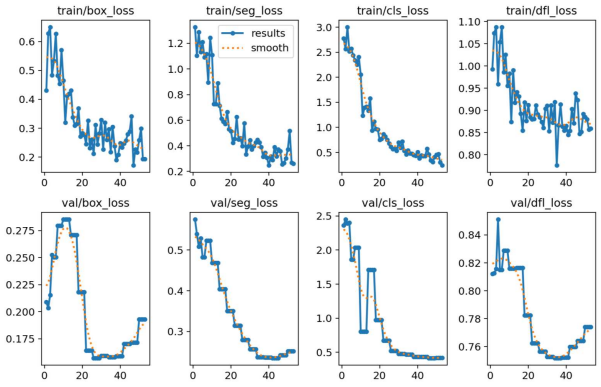

그림 5. mAP50 실증 데이터

계단형 IOU 평균 정밀도 실험에서는 mAP50 -95에 0.98503를 보여 높은 성능을 나타냈다.

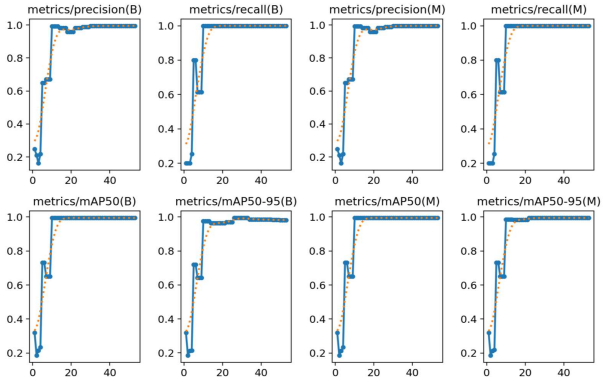

그림 6. mAP50-95 실증 데이터

이는 기존 YOLOv8 대비 세밀한 객체 윤곽 검출 능력과 가림 상태에서도 높은 정확도 유지가 가능함을 의미한다. 특히 불규칙하게 찌그러지거나 회전된 포대의 상단 위치를 정밀하게 인식하는 데 강점을 보였다.

본 연구에서는 동일한 환경에서 YOLOv11, Swin Transformer, 그리고 PointNet++ 세 가지 모델을 대상으로 인식 정확도, 상단 인식률, 자세(RPY) 추정 오차, 연산 속도(FPS)를 비교·분석하였다.

실험 결과, YOLOv11은 mAP50 기준 99.5%, 상단 인식률 96.4%, RPY 추정 오차 ±4.2°로 측정되었으며, 초당 52.1 FPS의 빠른 처리 속도를 보여 상대적으로 실시간 인식 및 로봇 연동에 적합한 모델로 평가되었다. 특히 비정형 포대의 찌그러짐, 기울기, 부분 가림 등의 환경에서도 높은 인식 안정성을 유지하였다.

표 8. 적용 모델별 성능 비교 결과

모델	mAP50 (%)	상단 인식률 (%)	RPY 추정 오차 (°)	연산 속도 (FPS)
YOLOv11	99.5	96.4	±4.2	52.1
Swin Transformer	92.1	91.8	±6.8	35.0
PointNet++	90.4	89.2	±3.5	42.0

한편, Swin Transformer는 mAP50 92.1%, 상단 인식률 91.8%로 YOLOv11보다는 다소 낮은 수치를 보였으나, Occlusion 상황에서의 인식 강인성이 가장 우수한 모델로 확인되었으며, 이는 전역적 문맥 정보를 학습하는 self-attention 구조의 특성에 기인한 것으로 판단된다.

PointNet++의 경우, 전체 인식 정확도는 mAP 50 90.4%, 상단 인식률 89.2%로 나타났고, RPY 추정 오차가 ±3.5°로 가장 낮게 측정되어 3차원 자세 인식 정밀도 측면에서 강점을 보였다.

연산 속도는 42.0 FPS로 중간 수준이었으나, 점군(Point Cloud) 기반의 공간 구조 이해 능력 덕분에 불규칙한 형태의 포대 식별에는 높은 신뢰성을 확보하였다.

일반적으로 산업용 팔레타이징 로봇의 파지 공정에서는 ±5° 이내의 자세 오차(Roll, Pitch, Yaw)가 허용 가능한 정렬 오차 범위로 제시된다. 본 연구에서 측정된 PointNet++의 평균 RPY 오차(±3.5°)와 YOLOv11의 평균 오차(±4.2°)는 모두 이 허용 한계 내에 해당하므로, 두 모델 모두 실제 로봇 파지(grasping) 및 적재 작업에 적용 가능한 수준의 자세 추정 정밀도를 확보한 것으로 판단된다[18].

이러한 결과는 3D 비전 기반 인식 모델이 단순한 객체 검출을 넘어, 실시간 로봇 제어 및 파지 정렬 단계에서도 충분한 신뢰성을 제공할 수 있음을 실증적으로 보여준 사례가 되며, 종합적으로 YOLOv11은 속도와 인식 정확도 모두에서 우수한 성능을 보여 실시간 산업용 비전 시스템에 가장 적합한 구조로 평가되었다.

반면 Swin Transformer는 가림·복잡 배경 상황에서의 인식 신뢰성, PointNet++는 자세 인식 및 3D 구조 해석 능력 측면에서 경쟁력을 보여, 향후 실증 환경에서는 YOLOv11을 중심으로 Swin Transformer 및 PointNet++의 장점을 결합한 하이브리드 융합 구조가 실질적 성능 향상 방안으로 제시될 수 있을 것이다.

V. 결론 및 향후 과제

본 연구에서는 비정형 물류 환경에서의 자동화 팔레타이징 시스템 구현을 목표로 RGB-Depth 융합 기반 객체 인식 구조를 비교·분석하였다. 실험 결과, YOLO v11 모델이 전체 평가 항목에서 가장 우수한 성능(mAP50 99.5%)과 빠른 연산 속도를 보였으며, 복잡한 형상의 비정형 포대에 대해서도 안정적인 인식이 가능함을 확인하였다. 특히 상단 인식률에서 높은 정확도를 유지하여 실시간 산업 응용 측면의 실효성을 입증하였다. Swin Transformer는 가림(Occlusion) 환경에서의 인식 강인성, PointNet++는 RPY 기반 자세 인식의 정밀도 측면에서 경쟁력을 보였다.

이러한 비교 결과는 물류 환경의 요구 조건에 따라 속도 지향형 또는 정밀도 지향형 모델을 선택하기 위한 합리적 기준을 제시한다. 또한 RGB-Depth 융합을 통한 상단 객체 선별 및 중심 추정 기능의 실험적 검증을 통해, 협동 로봇(UR3)과 공압 그리퍼(VGC10)을 이용한 자동 적재 공정 실현 가능성을 확인하였다.

다만, 본 연구는 300개의 비정형 포대 이미지와 실내 환경에서 수행되었으므로 조도 변화, 반사체 특성, 먼지 및 진동 등의 실제 산업 조건에 따른 성능 저하 가능성이 존재한다.

이에 후속 연구에서는 도메인 랜덤화(Domain Randomization), 다중 조도 환경 촬영, 교차 환경 검증(Cross-Environment Validation) 등을 도입하여 모델 일반화 성능을 강화 중이다[18].

또한 다양한 물류 품목을 포함한 범용 데이터셋 구축, 객체 인식 - 로봇 제어 통합 실증, 3차원 무게 중심 추정 기반 실시간 보정 알고리즘 개발, AI 모델 경량화 및 엣지 디바이스 연동 평가, 강화학습

기반 적재 순서 최적화 기법 개발 등을 통해 실증적 자동화 팔레타이징 시스템의 기술적 완성도를 제고하고자 한다.

REFERENCES

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 779 - 788, 2016.

[2] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992 - 10002, 2021.

[3] Charles Ruizhongtai Qi, Li Yi, Hao Su, Leonidas Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space,” Advances in Neural Information Processing Systems 30 (NeurIPS 2017), pp.5099 - 5108, 2017

[4] A. M. Lavoie, “3D Vision Systems for Industrial Robotics,” IEEE Transactions on Industrial Electronics, vol. 67, no. 8, pp. 6702 - 6712, Aug. 2020.

[5] H. Zhang et al., “RGB-D object recognition using fusion of deep convolutional neural networks,” Computer Vision and Image Understanding, vol. 173, pp. 21 - 29, 2018.

[6] Mech Mind Robotics, “3D Vision Guided Case Depalletizing,” Mech Mind Solutions, Mech Mind Robotics Technologies Co., Ltd.,

[7] Fizyr, “Picking the Unknown: Flexible Automation in Logistics,” Fizyr News, 15 Feb. 2021, Fizyr.

[8] C.-Y. Wang et al., “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” arXiv preprint, arXiv:2207.02696, 2022.

[9] A. Bochkovskiy et al., “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv preprint, arXiv:2004.10934, 2020.

[10] Wu, Y., et al., “3D Object Detection with Transformer and Graph Neural Networks,” IEEE Access, vol. 10, pp. 22465 - 22475, 2022.

[11] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” International Conference on Learning Representations (ICLR), 2021.

[12] K. Hu et al., “Depth-aware CNN for RGB-D segmentation,” Pattern Recognition Letters, vol. 130, pp. 306 - 313, 2020.

[13] Jung, S. et al., “Performance analysis of Intel RealSense depth cameras for industrial robotics”. Sensors, vol. 21, no. 4, 1234, 2022.

[14] Kim, D. & Park, J. “Palletizing optimization using AI-based packing algorithms”. International Journal of Advanced Logistics, vol. 9, no. 2, 113 - 124, 2020.

[15] A. Glenn Jocher et al., “YOLOv11: Next-Generation Real-Time Object Detection,” Ultralytics Paper, arXiv preprint, 2024.

[16] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992 - 10002, 2021.

[17] M. Li et al., “Domain Randomization and Cross-Scene Generalization for Robust Object Detection in Industrial Environments,” IEEE Transactions on Industrial Informatics, vol. 20, no. 3, pp. 5124 - 5136, 2024.

[18] H. Tanaka, K. Fujimoto, “Pose Accuracy Evaluation for Industrial Palletizing Robots Using 3D Vision Systems,” IEEE Transactions on Robotics and Automation, vol. 40, no. 2, pp. 1248 - 1256, 2024.

저 자 소 개

윤봉식(정회원)

1998년 전북대학교 산업디자인학과 학사 졸업.

2000년 전북대학교 일반대학원 산업디자인학과 석사 졸업.

2018년 전북대학교 일반대학원 디자인제조공학과 박사 졸업.

<주관심분야 : 데이터기반 프로세스, UI/UX디자인, 성과 분석, 산학융합연구>

백상윤(정회원)

2010년 전남대학교 기계시스템공학과 학사 졸업.

2012년 광주과학기술원 기전공학부 석사 졸업.

2018년 광주과학기술원 기전공학부 박사 졸업.

<주관심분야 : 머신비전, 로봇제어, 자동화 시스템>