Journal of Korea Robotics Society (2025) 20(3):546-553
https://doi.org/10.7746/jkros.2025.20.3.546 ISSN: 1975-6291 / eISSN: 2287-3961

# LLM 기반 시맨틱 스타일 분석을 활용한 사용자 추적 시스템

# User Identification and Tracking System via LLM-Based Semantic Style Analysis

경 도 현[1] · 양 견 모[2] · 오 장 석[2] · 서 갑 호[†]
Dohyun Kyoung[1], Kyon-Mo Yang[2], Jangseok Oh[2], Kap-Ho Seo[†]

**Abstract:** This paper proposes a user tracking system based on semantic style recognition using a Large Language Model (LLM). To overcome the limitations of traditional face recognition in non-frontal or occluded scenarios, the proposed system extracts clothing styles and colors using GPT-4o, a multimodal LLM, and combines this information with YOLOv5-based segmentation and DeepSORT tracking. The extracted style features serve as robust identity cues that enable consistent tracking even when a user's face is not visible. Experiments conducted on RGB-D image sequences with three subjects demonstrate the system's high recognition accuracy and improved ID stability compared to conventional tracking methods. The integration of LLM-based style analysis enables ID restoration after tracking failures and enhances tracking performance in complex environments. The proposed framework shows strong potential for use in indoor robotics, smart surveillance, and human-robot interaction applications.

**Keywords:** Style-based User Recognition, LLM-based Style Analysis, User Tracking, Semantic Recognition, Non-face-based Identification

## 1. 서 론

인공지능 기술의 발전에 따라 최근에는 농업, 물류, 실내 서비스 등 다양한 도메인에서 특정 대상자를 인식하고 추적하는 기술이 활발히 활용되고 있다[1,2]. 특히, 사람과의 협업이 중요한 로봇 시스템이나 스마트 감시 환경에서는 특정 대상을 정확하게 인식하고, 그 변화를 실시간으로 추적하는 능력이 필수적으로 요구된다. 이러한 인식 및 추적 능력은 시스템이 사람의 위치와 상태를 파악하여, 필요한 시점에 적절한 서비스를 제공하거나 상호작용을 유도하는 기반이 된다.

Received : May. 22. 2025; Revised : Jun. 17. 2025; Accepted : Jun. 20. 2025
※ This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture and Forestry (IPET) through High Value-added Food Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (RS-2022-IP322054).
1. Researcher, Korea Institute of Robotics and Technology Convergence (KIRO), Seoul, Korea (kyoungdh@kiro.re.kr)
2. Senior Researcher, Korea Institute of Robotics and Technology Convergence (KIRO), Seoul, Korea (kmyang, dueleldi@kiro.re.kr)
† Chief Researcher, Adjunct Professor, Corresponding author: KIRO, Seoul, and Robot and Smart System Engineering, Kyungpook National University, Daegu, Korea (neoworld@kiro.re.kr)



기존의 사용자 인식 기술은 주로 RGB 카메라 기반의 얼굴 인식, 형상 기반의 추적 알고리즘, 또는 LiDAR 센서를 통한 위치 기반 식별 기법들이 활용되어 왔다[3-5]. 그러나 얼굴 인식 방식은 원거리에서 정확도가 급격히 저하되며, 마스크 착용이나 후면 촬영 등 얼굴이 부분적으로 가려진 상황에서는 신뢰성이 크게 떨어진다[6]. 형상 기반 추적의 경우, 다수의 인원이 겹치거나 부분 가림이 발생하면 ID 전환 또는 추적 실패가 빈번히 발생한다[7]. 또한 LiDAR 기반 인식은 개별 대상의 특징이 부족하여 유사한 실루엣을 가진 인물 간 오인식이 발생하기 쉽다[8]. 이는 기존 인식 방식이 형태 중심의 시각 또는 구조 정보에 집중되어 있고, 시맨틱(semantic) 정보는 활용되지 않았기 때문에 발생하는 한계다. 특히, 의상의 색상이나 형태, 헤어스타일과 같은 스타일 정보는 각 개인을 구별할 수 있는 중요한 시맨틱 단서임에도 불구하고, 기존 인식 시스템에서는 거의 사용되지 않았다.

이러한 한계를 극복하고자 최근에는 대규모 언어 모델(LLM)을 활용하여 이미지 내 객체의 시맨틱 정보를 해석하고, 이를 기반으로 인식 및 추적을 수행하는 연구들이 주목받고 있다. ChatTracker[9]는 멀티모달 LLM의 대화형 응답을 통해 객체의 시각적 특징을 정교하게 보완하며, DTLLM-VLT[10]는 다양한

텍스트 표현을 통해 스타일 기반 추적의 유연성을 확보하고자 하였다. 또한, MemVLT[11]는 시간 정보를 LLM에 반영함으로써 시계열 기반의 시맨틱 일관성을 유지하는 방식을 제안하였다. 이처럼 LLM 기반 방식은 기존의 형태 중심 인식 기법이 가지는 한계를 극복하고, 의미 기반의 사용자 인식을 가능하게 하는 방향으로 진화하고 있다.

본 논문에서는 이러한 연구 흐름을 반영하여, 사람의 스타일 정보를 중심으로 인식을 수행하는 새로운 사용자 인식 프레임워크를 제안한다. 제안하는 시스템은 RGB 영상으로부터 사람을 검출하고, 의상 색상, 형태, 헤어스타일 등 다양한 시맨틱 스타일 정보를 LLM (Large Language Model) 기반으로 도출한다. 이후 생성된 스타일 정보는 트리(Tree) 구조 기반 데이터베이스에 저장되어 유사도 기반 사용자 식별에 활용되며, 시계열 추적기와 결합함으로써 시간적 연속성과 시맨틱 정보가 통합된다. 이로써, 추적 중 ID가 변경되어 타겟을 놓치는 상황에서도 스타일 정보를 기반으로 다시 타겟의 ID를 재지정할 수 있어, 연속적인 사용자 추적이 가능하다. 본 논문의 주요 기여는 다음과 같다.

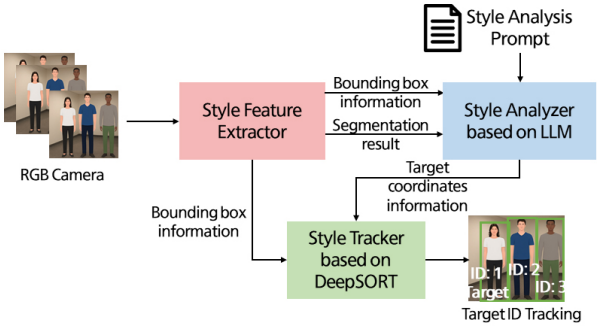
- LLM 기반 시맨틱 스타일 인식을 활용한 사용자 식별 프레임워크 제안: 대규모 언어 모델(LLM)을 활용하여 의상 색상, 형태, 헤어스타일 등의 시맨틱 스타일 정보를 도출하고, 이를 기반으로 다양한 각도 및 조건에서 실시간 사용자 인식을 수행하는 시스템 개발
- 트리 구조 기반의 스타일 정보 관리 및 유사도 평가 기법 제안: 사용자 스타일 정보를 트리(Tree) 구조로 저장하고 유사도 기반으로 관리함으로써, 실시간 데이터 갱신과 효율적인 사용자 구분이 가능한 관리 체계를 구축
- 시맨틱 정보와 시계열 추적기의 결합을 통한 안정적인 사용자 ID 유지 기법 구현: 스타일 인식 정보를 시계열 추적기와 결합하여, 겹침이나 가림 등 복잡한 환경에서도 사용자 ID 전환 문제를 완화 및 추적 안정성 향상

본 논문의 구성은 다음과 같다. 2장에서는 전체 시스템 구조와 스타일 정보의 LLM 기반 도출 과정 및 트리 구조 관리 방식에 대해 설명한다. 3장에서는 실험을 통해 제안 시스템의 유효성을 기존 방식과 비교하여 검증한다. 마지막으로 4장에서는 결론 및 향후 연구 방향을 제시한다.

## 2. LLM 기반 스타일 인식 시스템

본 장에서는 시맨틱한 스타일 정보와 시계열 기반 추적 정보를 통합하여, 시계열 기반 추적기가 타겟의 ID를 놓치는 경우 인식된 스타일 정보를 이용하여 타겟의 추적을 유지할 수 있는 LLM기반의 스타일 인식 시스템 구조를 설명한다.

### 2.1 제안하는 시스템 구성



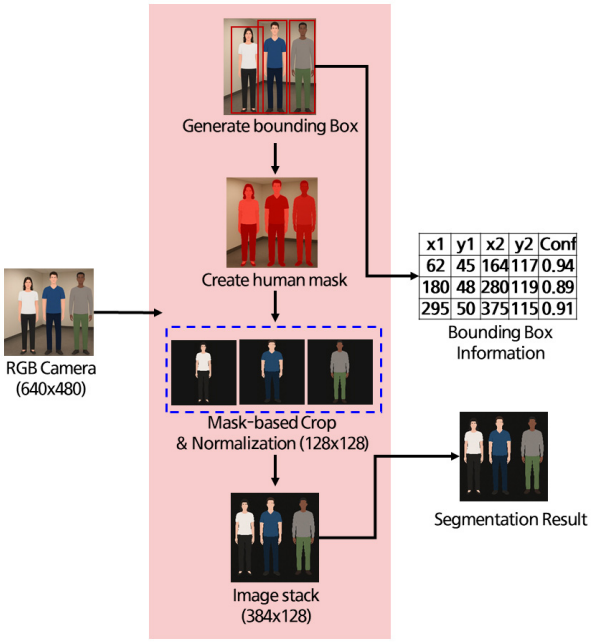

[Fig. 1] Architecture of LLM-based style recognition system

[Fig. 1]은 제안하는 시스템의 구조이다. 시스템은 스타일 정보추출기(Style Feature Extractor), LLM 스타일 분석기(LLM-based Style Analyzer), 그리고 스타일 추적기(Style Tracker) 모듈로 구성된다. 첫째, 스타일 정보 추출기(Style Feature Extractor)는 RGB 카메라로부터 입력된 이미지를 세그멘테이션(Segmentation) 모델을 이용하여 사람 객체를 분리하고, 각 객체의 바운딩 박스(Bounding Box)를 생성한다. 다음으로, LLM 스타일 분석기에서는 세그멘테이션 이미지, 바운딩 박스 정보, 스타일 분석 프롬프트를 입력받아 각 사람의 상의 및 하의 스타일 정보를 자연어로 도출하고, 이를 사전 정의된 타겟의 스타일과 비교하여 타겟의 좌표를 추출한다. 마지막으로, 스타일 추적기는 스타일 정보추출기와 LLM 스타일 분석기의 결과를 기반으로 타겟의 ID를 맵핑하고, 추적시 타겟을 놓치는 경우 스타일 정보를 기반으로 타겟의 ID를 재지정하여 연속적인 추적을 수행한다.

본 논문에서 제안된 시스템은 입력 영상으로부터 추적 결과를 도출하기까지의 모든 처리가 연속적으로 연결된 End-to-End (E2E) 기반의 온라인(Online) 구조로 이루어진다. 각 모듈은 프레임 단위의 입력을 받아 순차적으로 처리되며, 중간 결과는 직접 전달된다. 특히, LLM 스타일 분석기는 모든 프레임에서 수행되는 것이 아니라 타겟이 새로 초기화 되거나 특정 플래그 조건이 활성화된 시점에서만 실행되어 시스템의 처리 성능을 효율적으로 유지한다.

### 2.2 스타일 정보 추출기

[Fig. 2]은 스타일 정보 추출기 구조도를 보여준다. 타일 정보추출기는 시스템의 첫 번째 모듈로, 입력된 RGB 이미지에서 사람 객체를 검출하여 개별 이미지를 생성한다. 본 연구에서는 YOLOv5 기반 세그멘테이션 모델인 Yolov5n-seg을 활용하여 바운딩 박스와 마스크를 생성하고, 마스크 영역만을 잘라내어 개별 사람 이미지를 분리한다[12]. 이후 각 이미지는 분석 효율 향상을 위해 128×128 해상도로 정규화되며, 모든 이미지가 좌우로 스택된 하나의 통합 이미지로 구성된다. 이 통합 이미지는




[Fig. 2] RGB camera-based style feature extractor

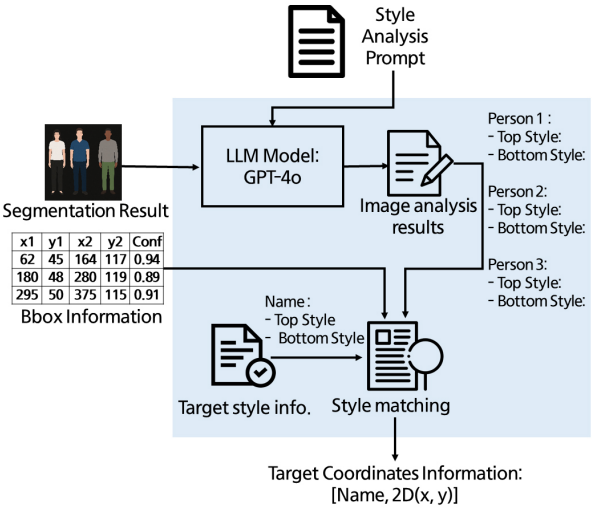
사람 순서가 바운딩 박스 탐지 순서와 일치하도록 정렬되어, 이후 스타일 분석 결과와의 인덱스 기반 매칭이 가능하도록 한다. 추출된 바운딩 박스 중심 좌표는 타겟 위치 산출 및 추적기 입력에 사용된다.

## 2.3 LLM 스타일 분석기

LLM 스타일 분석기는 스타일 정보추출기에서 생성된 통합 이미지, 바운딩 박스 정보, 그리고 스타일 분석 프롬프트를 기반으로 시맨틱한 스타일 정보를 생성한다. [Fig. 3]는 제안하는 LLM 스타일 분석기의 구조이다. 본 논문에서는 OpenAI의 GPT-4o 모델을 활용하며, 이 모델은 시각 정보와 언어 정보를 함께 처리할 수 있어 자연어 기반 스타일 인식이 가능하다[13]. 아래는 입력 정보를 분석하기 위한 분석프롬프트와 결과프롬프트의 정의 및 예이다.

- 분석 프롬프트: 입력 정보에 대하여 사용자의 스타일을원하는 형태로 인식하기 위한 명령 프롬프트, “해당 이미지에서 각 사람마다 옷 스타일을 간단히 설명해줘. 색상, 옷 , 그리고 상·하의 구성을 중점적으로 알려줘.”
- 결과 프롬프트: 분석 프롬프트와 이미지를 입력받아 GPT-4o 모델에서 도출되는 결과 프롬프트, “사람 1: 상의 스타일: 반팔+흰색, 하의 스타일: 긴바지+파란색“

스타일 매칭에서는 LLM에서 출력된 문장을 사전 정의된 타겟 스타일 정보와 비교한다. LLM 응답에서 각 사람에 대한 스



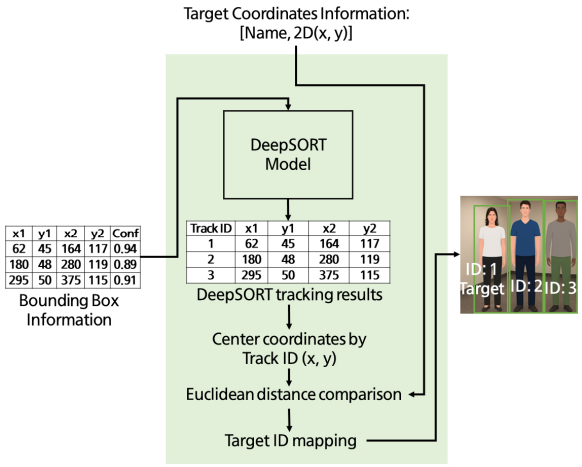

[Fig. 3] Architecture of style analyzer based on LLM

타일 문장을 순차적으로 분석하며, 해당 문장에 타겟 스타일 정보에 포함된 키워드가 존재하는지를 문자열 기반으로 확인한다. 즉, 특정 인물의 스타일 설명에 “상의 스타일”과 “하의 스타일”에 정의된 키워드가 모두 포함될 경우, 해당 사람은 타겟으로 식별된다. 예를 들어, 타겟 스타일 정보가 “상의 스타일: [흰색]”, “하의 스타일: [파란색]”인 경우, 키워드가 모두 포함되어 있기 때문에 해당 사람은 타겟으로 판단된다. 인식된 타겟이 있으면, 스타일 추적기의 입력으로 사용하기 위해 바운딩 박스의 중심 좌표를 타겟 위치로 전달한다.

또한, 본 논문에서는 이러한 스타일 정보를 트리 구조로 분해하여 저장한다. 트리의 상위 노드는 “상의 스타일”, “하의 스타일” 등의 항목이며, 하위 노드는 해당 항목의 세부 속성으로 구성된다. 새로운 LLM 응답도 같은 방식으로 파싱되어 스타일 속성 단위로 비교된다. 이 구조는 각 항목별 키워드를 기준으로 일치 여부를 평가하므로 표현의 다양성과 문장 구조의 차이에 영향을 받지 않고 일관된 타겟 식별이 가능하다. 추가적으로 항목별 비교 결과를 기반으로 유사도를 정량화하며, 각 항목에 포함된 키워드의 일치 정도에 따라 완전 일치 시 1점, 일부 색상만 일치하는 경우 0.7점, 전혀 일치하지 않는 경우 0점, 문법 오류나 응답 누락, 판단 불가 등은 -1점으로 처리한다. 이 유사도 평가 방식은 트리 구조 기반 매칭 방식과 결합되어 정확한 타겟 매칭이 가능하다.

## 2.4 스타일 추적기

세 번째 모듈은 대상자의 이동 경로를 추적하는 객체추적을 위한 스타일 추적기이다. 본 연구에서는 객체추적 알고리즘으로 DeepSORT[14]를 활용하며, 이 모듈은 바운딩 박스 정보와 타겟 좌표를 입력으로 받아 해당 타겟의 ID를 지속적으로 추적한다.




[Fig. 4] Architecture of style tracker based on DeepSORT

[Fig. 4]는 DeepSORT 기반 스타일 추적기의 구조를 나타낸다. 바운딩 박스 정보와 스타일 분석기에서 도출된 타겟의 중점 좌표 정보를 입력받고 DeepSORT의 추적 결과와 비교하여 타겟을 설정한다. 이를 위해 각 프레임에서 DeepSORT가 도출한 바운딩 박스 중심 좌표와 타겟 중점 좌표 간의 유클리드 거리를 계산하고, 설정된 거리 임계 값 이하로 가장 근접한 객체를 타겟으로 간주한다. 이 때 추적된 ID와 타겟 이름 간의 매핑이 생성되며, 이후 동일 ID에 대해 타겟 식별 상태가 유지된다.

[Table 1]은 DeepSORT 추적 결과와 타겟 좌표 정보 매칭하기 위한 의사코드(pseudo-code)이다. 스타일 분석기를 통해 도출된 타겟의 중심 좌표 (tx, ty)를 기반으로, 현재 프레임에서

[Table 1] Pseudo-code for target ID matching

| Line | Pseudocode |
|---|---|
| 1 | for each bbox in tracked_results do |
| 2 | (x1, y1, x2, y2, track_id) ← bbox |
| 3 | cx ← $\frac{(x1+x2)}{2}$ |
| 4 | cy ← $\frac{(y1+y2)}{2}$ |
| 5 | dist ← $\sqrt{(cx-tx)^2+(cy-ty)^2}$ |
| 6 | if dist < threshold and dist < min_dist then |
| 7 | min_dist ← dist |
| 8 | matched_id ← track_id |
| 9 | end if |
| 10 | end for |
| 11 | if matched_id != None then |
| 12 | name_by_id[matched_id] ← target_name |
| 13 | end if |

DeepSORT가 추적한 객체들 중에서 타겟에 가장 가까운 객체를 식별한다. 입력으로는 타겟 좌표 정보와 함께, 각 객체에 대한 바운딩 박스 좌표와 추적 ID로 구성된 리스트(tracked_results)가 주어진다. 먼저, 각 객체에 대해, 바운딩 박스의 좌상단 좌표 (x1, y1)와 우하단 좌표 (x2, y2)를 이용하여 중심 좌표 (cx, cy)를 이용하여 해당 객체가 2D 화면상 어디에 위치하는지를 인식한다. 다음으로, 계산된 중심 좌표와 타겟의 중심 좌표 사이의 유클리드 거리(dist)를 계산한다. 다음으로, 유클리드 거리를 기반으로 임계값(threshold)보다 작고, 현재까지 측정된 최소 거리(min_dist)보다도 작은 경우에만 해당 객체가 타겟에 일치하는 것으로 판단한다. 이때, min_dist 값을 현재 거리로 갱신하고, 해당 객체의 추적 ID를 matched_id에 저장한다. 이 과정을 모든 객체에 대해 반복한 후, 가장 가까운 객체의 추적 ID가 결정된다. 마지막으로, 유효한 매칭 결과(matched_id)가 존재한다면, 해당 ID와 타겟 이름을 name_by_id 딕셔너리에 저장한다. 이것은 추적 ID와 타겟 이름 간의 매핑을 유지하며, 이후 프레임에서도 같은 ID가 나타날 경우 해당 객체가 동일한 타겟으로 식별한다. 타겟과 다른 사람의 바운딩 박스의 겹침, 화면 밖으로 사라졌다 들어온 상황등의 발생에 의해 이후 프레임에서 ID가 변경될 경우 변경된 시점의 스타일 정보 추출기와 스타일 인식기 결과를 이용하여 객체 추적기에서 변경된 ID와 타겟좌표를 다시 매핑시켜 타겟 추적을 유지한다.

### 2.5 모듈 간 통합 및 효과

이와 같은 구조를 통해, 시맨틱한 스타일 정보와 시계열 기반 추적 정보를 통합하여, 단순한 위치 기반 추적이 아닌 ID 복원이 가능한 의미 기반의 지능적 사용자 추적이 가능하다. 특히, 일시적으로 가려지거나 복수 인원이 겹치는 상황에서도 DeepSORT의 외형 기반 ID 유지 기능과 스타일 기반 타겟 식별이 상호 보완적으로 작용하여 높은 추적 안정성을 확보할 수 있다.

본 시스템은 전체 파이프라인이 실시간으로 연동되는 End-to-End 기반의 온라인 구조로 구성되어 있으며, 스타일 분석모듈은 특정 시점에서만 조건부로 실행됨으로써 처리 지연을 최소한하고 실시간 응답성을 유지한다. 이러한 구조는 즉각적인 반응성이 요구되는 응용 환경에서의 적용 가능성을 높인다.

## 3. 실 험

본 장에서는 제안한 시스템의 성능을 검증하기 위한 하드웨어 구성과 제안하는 스타일 인식기의 성능, 그리고 기존 방법인 DeepSort를 단일로 사용하였을 때와 제안하는 방법과 함께 사용하였을 때의 추적 성능을 비교하였다.

### 3.1 실험 환경 및 구성

본 실험을 위해 RGB-D 카메라인 Intel RealSense D435를 사용하여 3명이 한 공간에서 자유롭게 움직이며, 일부 프레임에서는 사용자 간의 위치가 겹치거나 일시적으로 가려지는 상황을 포함한 RGB 및 Depth 이미지 영상 1,244프레임의 Rosbag 데이터를 수집하였다. 실험은 Intel i7-14700 CPU와 NVIDIA RTX 4070 Super GPU가 탑재된 환경에서 수행되었다. Yolov5n-seg 모델의 하이퍼파라미터는 신뢰도(confidence) 임계값은 0.7, IoU (Intersection over Union) 임계값은 0.45로 설정하였다. 검출 클래스는 사람(classes=0)만을 대상으로 필터링하였다.

스타일 분석을 위한 LLM 입력 프롬프트는 상의, 하의 스타일, 그리고 색상 목록에 기반하여 구성되며, 각 사람에 대해 사전 정의된 선택지에서 하나씩 선택하여 서술하도록 유도하였다. 사람 구분은 “사람 1”, “사람 2”와 같이 이미지의 좌측부터 순차적으로 부여되며, 응답 형식은 다음 예시와 같이 지정된다: “상의 스타일: 티셔츠 + 색상: 흰색, 하의 스타일: 바지 + 색상: 파란색”. 분석 프롬프트는 다음과 같은 형태로 구성되었다. 사용된 스타일 목록은 [Table 2]와 같다.

- 이미지에서 사람이 2명 이상일 경우, 사람 1, 사람 2, …형태로 구분해야 하며, 이미지에서 왼쪽부터 사람 1로 명명된다.
- 다음 스타일 목록 중에서 사람에 맞는 항목을 하나씩 고른 후, 각 스타일 항목에 대해 색상도 함께 지정하여 응답해야 한다.
- 반드시 제공된 스타일 목록과 색상 목록 중에서만 선택하여 응답하도록 제한한다.

[Table 2] Style list

| | |
|---|---|
| Top Style | T-shirts, Dress, Blouse, Hoodie |
| Bottom Style | Pants, Skirt |
| Color List | Black, White, Blue, Red, Gray, Brown, Green, Yellow |

[Table 3] Color mapping list

| Original Color | Normalization Color |
|---|---|
| Dark blue, Navy | Blue |
| Wine | Red |
| Off-white, Ivory | White |
| Coal-black | Black |
| Charcoal | Gray |
| Beige | Brown |
| Light green, Lime | Green |
| Flesh | Yellow |

프롬프트에 제시되지 않은 색상 표현이 LLM 응답에 포함될 수 있는 점을 고려하여, 사전 정의된 색상 정규화(Color Mapping)를 적용하였다. 이 매핑은 색상의 표현 다양성을 통일하기 위한 전처리 과정으로 사용되었으며, 주요 매핑은 [Table 3]과 같다.

### 3.2 LLM 인식 정확도 실험

인식 정확도 실험에서는 유사도 점수 체계를 활용해 정확도를 측정했다. 3명의 사용자를 대상으로 수행되었으며, 각 사용자의 상의 및 하의 색상 정보를 기반으로 LLM 응답 결과의 정확도를 프레임 단위로 분석하였다. LLM의 응답 속도를 고려하여 1,244프레임의 Rosbag 데이터를 20프레임당 한번씩 각 사용자에 대한 인식을 수행하여 총 62프레임에 대해 실험을 진행하였다.

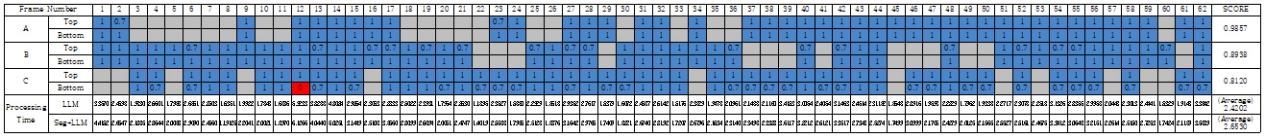
사전 정의된 타겟 스타일 정보는 상·하의 색상 키워드를 포함하고 있으며, LLM 응답에서 해당 키워드가 정확히 포함될 경우 정답(1), 일부 색상 일치 시 부분 정답(0.7), 불일치 시 오답(0), 판독불가(-1)으로 평가하였다. LLM 응답 결과가 문법적 오류를 포함하거나, 응답이 누락되거나 “사람 없음”, “판단 불가” 등의 표현이 포함된 경우는 해당 프레임을 판독불가로 간주하고 -1 점수로 처리하였다. [Fig. 5]는 LLM 스타일 분석기의 인식 정확도를 확인하는 실험 결과이다. 이것은 각 사용자에 대해 프레임 단위로 측정된 LLM 응답 정확도를 나타낸다. [Table 4]는 [Fig. 5]을 정리한 표로써, 전체 프레임에 대해 평균 점수를 계산하여 인식 정확도를 산출하였다. 점수 계산식은 식 (1)과 같다 :

$$\mathrm{SCORE}_X = \frac{1}{N_X}\sum_{i=1}^{N_X}\left(X_{up}^{(i)} \times X_{down}^{(i)}\right), X \in \{A, B, C\} \qquad (1)$$

$N_X : X_{up} \neq -1$인 유효 항목의 값

$X_{up}^{(i)}, X_{down}^{(i)}$ : i번째 유효 항목의 값

실험 결과에서는 상의 89.38, 하의 100.00, C 연구원은 상의 100.00, 하의 81.20의 정확도를 기록하였다. 전반적으로 상의보다는 하의 스타일에서 더 높은 인식 정확도가 나타났으며, 이



[Fig. 5] Recognition Accuracy of LLM-based Style Analyzer

[Table 4] Average recognition accuracy of LLM response

| Target | Average recognition accuracy (SCORE) |
|---|---|
| A | 98.57 |
| B | 89.38 |
| C | 81.20 |

는 하의의 시각적 특징이 더 명확하거나 덜 가려졌기 때문으로 판단된다. 또한, 일부 오차는 LLM 응답의 색상 표현 차이나 스타일 용어 누락으로 인한 것으로 보이며, 전반적으로 제안된 방식은 다양한 표현에 대해 일정 수준 이상의 인식 성능을 유지함을 확인하였다.

추가로, 측정된 스타일 인식기의 프레임 단위 처리 속도 평균 응답 시간은 약 2.42초였다. 최소 응답 시간은 1.5초, 최대 응답 시간은 5.9초로 확인되었으며, 이것은 LLM 서버 처리 지연 및 입력 이미지 복잡도에 따라 다소 차이를 보이는 것으로 분석된다. 스타일 분석기 모듈은 모든 프레임에서 실행되는 구조가 아니며, 타겟 식별이 필요한 시점에만 수행되도록 설계되었다. 따라서 처리 시간은 스타일 초기화 지연으로 간주되며, 전체 추적 루프의 프레임 속도에는 직접적인 영향을 주지 않는다.

### 3.3 추적 유지 성능 실험

제안하는 방법의 사용자 추적 안정성을 평가하기 위해, 기존 DeepSORT 추적 방식과 본 논문에서 제안한 DeepSORT 기반 스타일 추적 방식 간의 유지 성능을 1,244 프레임의 Rosbag데이터를 이용하여 비교하였다. 수행되는 DeepSORT 추적기의 주요 파라미터설정은 [Table 5]와 같다.

외형 특징(Appearance feature) 간 유사도 거리(MAX_DIST)와 Kalman 예측에 기반한 조건(MAX_IOU_DISTANCE)은 오탐 방지를 고려해 설정하였으며, MAX_AGE는 일시적인 가림 상황에서도 동일 ID를 유지할 수 있도록 90으로 설정하였다. 또한, 빠른 ID 초기화를 위해 N_INIT는 1로 설정하고, 임베딩 메모리 사용량 제한을 위해 NN_BUDGET은 100으로 설정하였다.

[Table 5] Parameters of DeepSORT

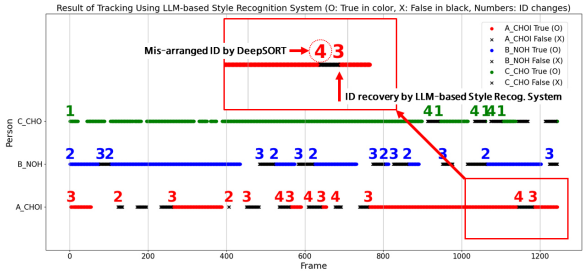
| Parameters | Value |
|---|---|
| MAX_DIST | 0.4 |
| MIN_CONFIDENCE | 0.4 |
| NMS_MAX_OVERLAP | 0.5 |
| MAX_IOU_DISTANCE | 0.7 |
| MAX_AGE | 90 |
| N_INIT | 1 |
| NN_BUDGET | 100 |

DeepSORT 기반 스타일 추적기는 초기 프레임에서 LLM 기반 스타일 분석을 통해 타겟 사용자에 대한 스타일 정보를 획득하고, 해당 사람의 중심 좌표와 DeepSORT 추적 결과를 비교하여 타겟 ID를 지정한다. 이후 동일 ID가 프레임 간 얼마나 일관되게 유지되는지를 측정하였다.

[Fig. 6]는 기존 DeepSORT 단독 방식과 LLM 기반 스타일 인식 시스템을 결합한 방식에서의 객체추적결과를 프레임별로 표현한 그림이다. 점선이 끊긴 부분은 DeepSORT 자체에 프서 탐지하지 못한 상황을 표현한 것이다. DeepSORT 단독 방식을 사용하였을 경우, 처음 할당된 ID가 변경이 되는 경우에 더 이상의 추적이 불가능하다. 반면에, LLM 기반 스타일 인식 시스템의 경우에는 다시 ID를 복원하는 기능이 있으므로, 지속적인 추적이 가능한 것을 확인했다.

[Table 6]은 기존 DeepSORT 단독 방식과 LLM 기반 스타일 인식 시스템을 결합한 방식 간의 사용자 추적 유지 성능을 비교한 결과를 보여준다. DeepSORT만을 사용한 경우, 일부 사용자의 추적 정확도가 낮고 정답 ID 유지가 불안정한 경향을 보였다. 반면, LLM 기반 방식은 초기 프레임에서 스타일 정보로 타겟을 식별하고 추적기와 결합함으로써 ID가 변경된 이후에도 스타일 정보를 기반으로 동일 ID를 복원할 수 있었으며, 전반적으로 추적유지 성능을 크게 높였다. 이는 시맨틱 정보가 외형 변화나 가림 상황에서의 보완 정보로 작용함을 의미하며, 제안한 시스템의 추적신뢰성을 실험적으로 입증하였다.


(a) Tracking results using DeepSORT only




(b) Tracking results using the LLM-based style recognition system

[Fig. 6] Frame-by-frame results of human tracking

[Table 6] Tracking performance of LLM-based style recognition system

(a) Summary of tracking results using DeepSORT only

| Target | Number of total frames (frames) | Number of correctly tracked frames (frames) | Tracking success rate (%) |
|---|---|---|---|
| A | 908 | 51 | 5.62 |
| B | 1,018 | 77 | 7.56 |
| C | 1,066 | 812 | 76.17 |

(b) Summary of tracking results using the LLM-based style recognition system

| Target | Number of total frames (frames) | Number of correctly tracked frames (frames) | Tracking success rate (%) |
|---|---|---|---|
| A | 908 | 671 | 73.90 |
| B | 1,018 | 762 | 74.85 |
| C | 1,066 | 930 | 87.24 |

## 4. 결론 및 향후 연구

본 논문에서는 LLM 기반 시맨틱 스타일 인식과 객체 추적 알고리즘을 결합하여, 얼굴 인식이 어려운 상황에서도 사용자를 안정적으로 식별하고 추적할 수 있는 시스템을 제안하였다. 제안 시스템은 RGB 이미지로부터 YOLOv5를 활용해 사람 객체를 검출하고, OpenAI GPT-4o 모델을 통해 상의 및 하의 스타일 정보를 자연어 형태로 분석하였다. 이후 DeepSORT 기반의 시계열 추적기와 연계함으로써, 시간적 연속성과 시맨틱 정보를 동시에 고려한 사용자 추적이 가능함을 실험을 통해 입증하였다.

LLM 인식 정확도 실험에서는 프레임 단위 평균 정확도가 대부분 사용자에서 90% 이상으로 나타났으며, 일부 스타일 표현의 다양성을 고려한 색상 정규화를 통해 추가적인 인식 정확도 향상을 확인하였다. 또한, DeepSORT와의 결합은 ID 전환 빈도를 감소시키며, 복잡한 환경에서도 안정적인 추적 성능을 보였다. 향후, 제안된 시스템을 표준 벤치마크 환경에 적용 가능한 구조로 확장하고, Ground Truth ID와의 프레임 단위 비교를 통해 MOTA, MOTP 등 정량 지표를 포함한 성능 평가를 진행할 계획이다.

본 시스템은 실내 서비스 로봇, 스마트 감시 시스템, 사람-로봇 협업 환경 등 다양한 분야에 활용될 수 있으며, 향후 연구로는 다양한 환경에서의 일반화 성능 검증, 사용자 행동 패턴 기반의 동적 추론 기능 확장, 그리고 다중 타겟 추적 환경으로의 확장뿐 아니라, 실시간 대응성을 높이기 위한 경량화와 스타일 기반 인식의 정밀도 향상을 위한 고도화 연구도 수행할 계획이다.

## References


[1] J. Gwak, K. Yang, J. Lee, J. Koo, and K. Seo, “Development of Autonomous Mobile Robot Control System for Changeable Target-of-interest Object Tracking using Specific Vest,” *Journal of Institute of Control, Robotics and Systems*, vol. 29, no. 5, pp. 452-457, May, 2023, DOI: 10.5302/J.ICROS.2023.23.0006.

[2] L. Saraceni, I. M. Motoi, D. Nardi, and T. A. Ciarfuglia, “AgriSORT: A Simple Online Real-time Tracking-by-Detection framework for robotics in precision agriculture,” *2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, pp. 2675-2682, 2024, DOI: 10.1109/ICRA 57147.2024.10610231.

[3] H. L. Gururaj, B. C. Soundarya, S. Priya, J. Shreyas, and F. Flammini, “A Comprehensive Review of Face Recognition Techniques, Trends, and Challenges,” *IEEE Access*, vol. 12, pp. 107903-107926, 2024, DOI: 10.1109/ACCESS.2024.3424933.

[4] L. Shaikewitz, S. Ubellacker, and L. Carlone, “A Certifiable Algorithm for Simultaneous Shape Estimation and Object Tracking,” *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 11873-11880, Dec., 2024, DOI: 10.1109/LRA.2024.3501684.

[5] C. Zhang, C. Zhang, Y. Guo, L. Chen, and M. Happold, “Motion Track: End-to-End Transformer-based Multi-Object Tracking with LiDAR-Camera Fusion,” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vancouver, BC, Canada, pp. 151-160, 17-24, 2023, DOI: 10.1109/CVPRW59228.2023.00020.

[6] F. Pleško, T. Goldmann, and K. Malinka, “Reconstruction and enhancement techniques for overcoming occlusion in facial recognition,” *EURASIP Journal on Image and Video Processing*, May, 2025, DOI: 10.1186/s13640-025-00670-7.

[7] Y. H. Wang, J. W. Hsieh, P. Y. Chen, M. C. Chang, H. H. So, and X. Li, “SMILEtrack: similarity learning for occlusion-aware multiple object tracking,” *AAAI Conference on Artificial Intelligence*, vol. 38, no. 638, pp. 20-27, 2024, DOI: 10.1609/aaai.v38i6.28386.

[8] W. Guo, Z. Pan, Y. Liang, Z. Xi, Z. Zhong, and J. Feng, “LiDAR-Based Person Re-Identification,” *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 17437-17447, 2024, DOI: 10.1109/CVPR52733.2024.01651.

[9] Y. Sun, F. Yu, S. Chen, Y. Zhang, J. Huang, C. Li, Y. Li, and, C. Wang, “Chattracker: Enhancing visual tracking performance via chatting with multimodal large language model,” *arXiv:2411.01756*, 2024, DOI: 10.48550/arXiv.2411.01756.

[10] X. Li, X. Feng, S. Hu, M. Wu, D. Zhang, J. Zhang, and K. Huang, “DTLLM-VLT: Diverse text generation for visual language tracking based on LLM,” *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, pp. 7283-7292, 2024, DOI: 10.1109/CVPRW63382.2024.00724.

[11] X. Feng, X. Li, S. Hu, D. Zhang, M. Wu, J. Zhang, X. Chen, and

K. Huang, “MemVLT: Vision-language tracking with adaptive memory-based prompts,” *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, Vancouver, BC, Canada, pp. 14903-14933, 2024, [Online], https://neurips.cc/virtual/2024/poster/94643.

[12] *v7.0 - YOLOv5 SOTA realtime instance segmentation*, [Online], https://github.com/ultralytics/yolov5/releases, Accessed: Feb. 14, 2025.

[13] OpenAI, “GPT-4o system card,” *OpenAI*, Aug., 2024, [Online], https://openai.com/index/gpt-4o-system-card/, Accessed: Sept. 26, 2024.

[14] N. Wojke, A. Bewley and D. Paulus, “Simple online and realtime tracking with a deep association metric,” *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, pp. 3645-3649, 2017, DOI: 10.1109/ICIP.2017.8296962.

**경 도 현**

2021 국립강릉원주대학교 컴퓨터공학과 (공학사)

2023 동국대학교 자율사물지능학과 (공학석사)

2024~현재 한국로봇융합연구원 지역연구본부 주임연구원

관심분야: 인공지능, 딥러닝, 사물 지능 제어

**양 견 모**

2011 세종대학교 디지털콘텐츠학과(공학사)

2014 연세대학교 컴퓨터과학(공학석사)

2018~현재 한국로봇융합연구원 지역연구본부 선임연구원

관심분야: 인공지능, 지식추론, 상황인식

**오 장 석**

2004 고려대학교 전자및정보공학과(공학사)

2006 동대학(공학석사)

2016 동대학(공학박사)

2018~현재 한국로봇융합연구원 지역연구본부 선임연구원

관심분야: 3차원복원, 영상처리, 모바일로봇

**서 갑 호**

1999 고려대학교 전기공학과(공학사)

2001 KAIST 전기및전자공학(공학석사)

2009 동대학(공학박사)

2009~현재 한국로봇융합연구원 지역연구본부 수석연구원

2021~현재 경북대학교 겸임교수

2024~현재 대동로보틱스 상무

관심분야: 시스템 제어, 농업로봇, 웨어러블로봇, 모바일로봇