

Frontal Following 을 위한 ToF Depth 카메라 기반 사용자 보행 의도 인식 및 위치 추정

User Walking Intention Recognition and Position Estimation Using a ToF Depth Camera for Frontal Following

최 원 석¹, 최 준 혁², 최 문 택^{1*}

(Wonseok Choi¹, Joonhyuk Choi², and Mun-Taek Choi^{1*})

¹Department of Intelligent Robotics, Sungkyunkwan University

²Department of Electrical and Computer Engineering, Sungkyunkwan University

Abstract: Smart walkers that provide walking assistance for the elderly or disabled increasingly require the ability to recognize the condition of the user and realize real-time interactions. Among the core components of such systems, recognizing the walking intention and estimating the position of the user are essential. This study proposes a technique that simultaneously performs walking-intention recognition and position estimation using depth image sequences acquired from a single time-of-flight depth camera. For user-walking-intention recognition, the depth image sequence is used as the input to a transformer model. This enables the network to learn spatial depth features and temporal walking patterns in an end-to-end manner, allowing it to predict walking intention. For user-position estimation, pixel-level depth variation analysis based on depth differences is applied to quantitatively classify user movements. A Kalman filter is then employed to enhance the stability of distance estimation between humans and walking devices. The experimental results achieved an F1-score of approximately 74% for walking-intention prediction and an average position error of 0.723 ± 0.027 m for distance estimation. The proposed method can serve as a core perception and intention recognition technology for smart assistive-walker control. Future research will aim to further enhance the prediction performance of the developed system by integrating motion trajectory data from smart walkers.

Keywords: ToF Depth Camera, Transformer, Gait Intention Prediction, Frontal Following, Human–Walker Localization

1. 서론

고령화 사회가 가속화됨에 따라, 보행 약자를 지원하기 위한 스마트 워커(smart walker) 기술의 중요성이 점점 커지고 있다[1,2]. 이러한 스마트 워커는 단순히 사용자의 신체적 부담을 줄여주는 이동 보조 기구를 넘어, 사용자의 상태를 인식하고 상황 변화에 능동적으로 반응하는 지능형 인간-로봇 협업 시스템으로 발전하고 있다[3]. 이에 따라 최근에는 스마트 워커의 자율성과 안전성을 향상시키기 위해 다양한 센서를 활용한 보행 의도 파악 및 사용자 위치 추정 연구가 활발히 진행되고 있다 [4-7].

특히, 스마트 워커가 사용자의 이동을 능동적으로 지원하기 위해서는 사람의 움직임을 실시간으로 인식하고 이에 맞게 반응하는 동작 방식이 필요하다. 이때 스마트 워커가 사용자의 전방에서 걸음 움직임을 인식하며 함께 이동하는 frontal following 방식이 요구된다[8,9]. 이 방식은 사용자의 시야를 가리지 않으면서도 스마트 워커가 전방의 장애물 정보를 선제적으로 인식하고, 적절한 거리와 방향을 유지하며 이

동을 지원할 수 있다는 장점이 있다. 따라서, 보행자의 위치를 안정적으로 추정하고 보행 의도를 실시간으로 파악할 수 있는 인식 기술이 frontal following 제어의 핵심이 된다.

이를 구현하기 위한 다양한 센서 기반 접근이 제안되어 왔으나, 보행자의 움직임을 정밀하게 인식하기 위해서는 다수의 센서를 신체에 부착하거나 스마트 워커 본체에 여러 개 설치해야 하는 경우가 많아 복잡한 구성, 높은 비용, 착용 불편함 등의 한계가 존재한다. 이에 비해 카메라 기반 인식 기술은 단일 장치만으로도 보행자의 걸음으로부터 보행 의도를 정밀하게 분석할 수 있다는 점에서 유용하다[10]. 일반적으로 고성능 깊이(depth) 카메라는 넓은 화각과 높은 해상도를 제공하지만, 비용이 높다는 한계가 있다. 반면, 저가형 ToF (Time-of-Flight) 깊이 카메라는 상대적으로 비용 부담이 적으나, 화각이 좁고 해상도가 낮아 보행자의 움직임을 정밀하게 인식하기 어렵다. 따라서, 저가형 ToF 깊이 카메라로 사용자의 보행 의도를 강건하게 파악하고 사람의 위치를 안정적으로 인식할 수 있는 기술이 필요하다.

* Corresponding Author

Manuscript received November 5, 2025; revised December 1, 2025; accepted January 5, 2026

최원석: 성균관대학교 지능형로봇학과 대학원생(bbmon650@skku.edu, ORCID[®] 0009-0006-2468-703X)

최준혁: 성균관대학교 전자전기컴퓨터공학과 대학원생(emin38@skku.edu, ORCID[®] 0009-0001-1998-0227)

최문택: 성균관대학교 지능형로봇학과 교수(mtchoi@skku.edu, ORCID[®] 0000-0002-5025-8785)

※ 본 연구는 보건복지부의 재원으로 장애인·노인 자립생활을 위한 보조기기 실용화 연구개발사업 지원에 의하여 이루어진 것임(RS-2025-03373011).

※ 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임 (RS-2025-25463654, 실세계 조작과 보행을 위한 멀티모달 휴머노이드 파운데이션 AI 모델 개발).

최근 보행 의도 분석 연구에서는 보행 데이터를 자동으로 classification하는 접근이 활발히 이루어지고 있으며, 스마트 워커가 사용자를 일정 거리에서 추종할 수 있도록 자동 제어의 기반이 되는 사용자 위치 추정 연구도 활발히 진행되고 있다[11]. 보행 의도 분석에는 머신러닝 기반의 통계적 방법이나 딥러닝 기반의 학습 모델이 주로 활용되고 있다. 그중 딥러닝은 복잡한 보행 패턴의 시공간적 특성을 데이터로부터 직접 학습할 수 있다는 점에서 높은 표현력과 일반화 성능을 제공한다. 보행자의 움직임은 하지의 형태적 변화와 공간적 패턴이 뚜렷하게 나타나므로, 이러한 국소적 시각 정보를 효과적으로 학습하기 위해 CNN (Convolutional Neural Network)을 적용하였다[12-14]. CNN은 깊이 영상 내 발과 다리의 움직임과 같은 공간적 특징을 추출하고, 복잡한 깊이 영상 데이터를 구조적으로 압축함으로써 학습 효율을 향상시킨다. 이를 통해 보행자의 걸음 움직임을 세밀하게 표현하면서도 깊이가 데이터를 효율적으로 처리할 수 있다. 이후, Transformer를 적용하여 시퀀스 형태의 입력 데이터를 처리하고, 보행 패턴의 연속성과 장기 의존성을 학습하였다[15-17]. Transformer의 self-attention 구조를 통해 모델은 시간에 따른 움직임 변화의 중요도를 동적으로 반영할 수 있으며[18], 이는 전통적인 Recurrent Neural Network 계열 모델보다 더 높은 표현력과 안정적인 학습 성능을 제공한다.

사용자 위치 추정에는 LiDAR, 초음파 센서 등 다양한 센서를 활용하여 사람-스마트 워커 간 거리를 측정하는 방법이 제안되어 왔다[19,20]. 그러나 이러한 방식은 센서가 사람의 옆면을 기준으로 거리를 측정하기 때문에, 보행 중 신체나 스마트 워커 프레임에 의해 시야가 부분적으로 가려지는 가림(occlusion) 현상이 발생할 수 있다. 또한, 사용자의 상대적인 이동 위치가 변하거나 회전 동작이 발생할 경우, 센서의 관측 각도가 달라지면서 거리 측정 오차가 커지는 한계가 존재한다. 이에 따라, ToF 기반 깊이가 카메라를 활용하여 보행자의 전면 움직임을 인식하는 구조를 적용하면, 보행 중 신체나 스마트 워커 프레임에 의한 가림을 줄이고, 관측 방향이 일정하게 유지되어 거리 추정의 일관성을 확보할 수 있다. 이러한 접근은 여러 센서 조합 없이 단일 센서만으로 보행 의도 인식과 사용자 위치 추정을 동시에 수행할 수 있어, 시스템 구성을 단순화하고 스마트 워커의 자동 제어에 필요한 핵심 정보를 효율적으로 통합할 수 있다는 실용적 장점을 가진다.

본 논문에서는 저가형 ToF 깊이가 카메라를 활용하여 정상인을 대상으로 8가지 보행 의도를 포함한 보행 데이터를 수집하였다. 데이터는 사전에 정의된 시나리오에 따라 다양한 보행 동작을 연속적으로 수행하는 방식으로 획득되었으며, 각 동작은 시간 구간별로 구분하여 라벨링하였다. 수집된 깊이 영상은 거리 구간을 5가지 영역으로 masking한 후, CNN을 적용하여 공간적 특징을 추출하고 Transformer를 이용해 시공간적 정보를 통합적으로 학습하였다. 또한, 동일한 분류된 영역별 깊이 영상을 기반으로 깊이 차이(depth difference) 방법을 통해 움직임과 정지 상태를 구분하고, 각 영역의 중심점을 이용해 ROI (Region of Interest)를 생성하였다 ROI 중심 좌표를 기반으로 스마트 워커에 부착된 카메라 기준의 사람-카

메라 간 상대 거리값을 계산하여, 사용자의 위치를 추정하였다. 이러한 구조는 스마트 워커 자동 제어의 토대를 마련하기 위한 인식 기술로서, 보행 의도를 실시간으로 제공하는 종단간(end-to-end) 학습 구조로 구현되었으며, 단일 저가형 ToF 센서만으로도 효율적인 인식이 가능하다. 향후 모터 궤적 데이터를 추가하여 인식 성능을 개선하고, 스마트 워커 제어 시스템과의 연계 가능성을 검증할 예정이다.

II. 방법

본 연구에서는 ToF 깊이가 카메라를 이용한 보행 의도 예측 기법을 제안한다. 제안된 방법은 수집된 깊이 영상을 전처리하여 학습용 시퀀스를 구성하고, Transformer 기반 모델을 통해 사용자의 보행 의도를 예측하는 구조로 이루어진다. 그림 1은 제안된 시스템의 전체 구조를 나타낸다.

1. 데이터 수집 및 전처리

본 논문에서는 ToF 깊이가 카메라(Tau Lidar 3D Depth Camera, Onion Corporation, Hong Kong)를 이용하여 정상인 3명을 대상으로 보행 데이터를 수집하였다. 데이터 수집에 사용된 스마트 워커는 그림 2에 제시되어 있다. 피험자는 동일한 실내 환경에서 총 4회에 걸쳐 각각 2분간 측정되었으며, 각 측정 세션은 사전에 정의된 8가지 보행 의도(전진, 전진하며 좌회전, 전진하며 우회전, 후진, 제자리에서 좌회전, 제자리에서 우회전, 정지, 사람 없음)를 순차적으로 15초 마다 수행하는 방식으로 진행되었다. 실험은 외부 요인의 영향을 최소화하기 위해 넓은 실내 개방형 공간에서 수행되었다. 공간은 약 20-30 m 길이의 직선형 구조를 가지고 있으며, 바닥은 반사율이 일정한 폴리싱 처리된 평면으로 구성되어 센서 입력의 변동을 최소화하였다. 실내 조명은 천장 LED를 통해 균일하게 확산되는 형태로, 그림자나 급격한 밝기 변화가 거의 발생하지 않아 깊이 영상 안정성이 확보되었다. 실험 공간 내에는 장애물이나 이동 물체가 존재하지 않도록 사전 정비되었으며, 벽면은 단색과 단순 구조로 이루어져 깊이 기반 보행 인식 과정에 불필요한 간섭이 발생하지 않도록 하였다. 카메라는 스마트 워커 전면 하단부에 부착되어 피험자의 하체(발~다리 부위)를 향하도록 설치되었으며, 보행 중 사람-카메라 간 거리 변화를 포함한 연속적인 깊이 영상을 기록하였다.

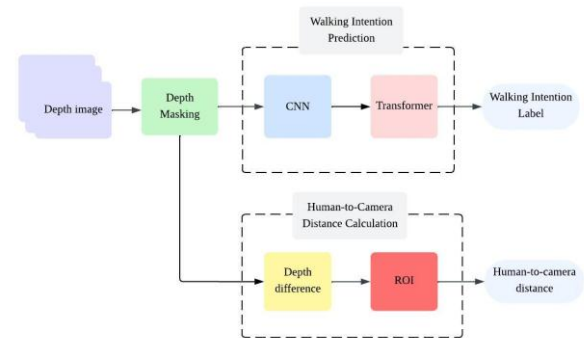


그림 1. ToF 기반 보행 의도 예측 및 사람-카메라 거리 추정 시스템의 전체 구조.

Fig. 1. Overall architecture of ToF-based walking intention prediction and human-camera distance estimation system.

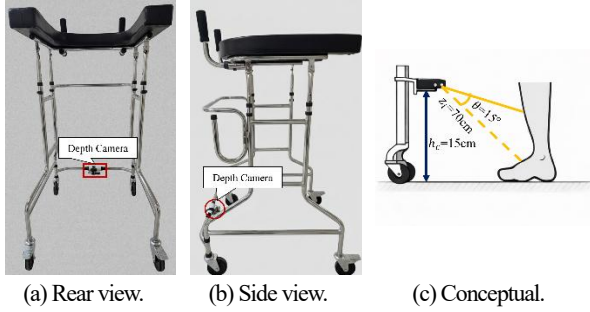


그림 2. 스마트 워커에 설치된 ToF 깊이 카메라의 설치 위치 및 기하학적 구성.

Fig. 2. Installation and geometric configuration of the ToF depth camera on the smart walker.

수집된 원시 깊이 영상마다 median filter를 적용하여 센서 노이즈를 제거한 후, 5가지 영역으로 masking을 수행하였다. 거리 값은 [20]에서 제시된 기준을 그대로 적용하여 close: 0~0.1 m, normal: 0.1~0.5 m, far: 0.5~0.8m, out-of-zone: 0.8~0.9 m, background: 0.9m~ ∞ 의 다섯 구간으로 구분하였다. 이러한 구간 설정은 배경을 제거하고 보행 움직임이 발생하는 영역을 명확히 파악하기 위함이다. 이후, 각 깊이 영상을 사람-카메라 간 거리 구간에 따라 다중 서브 영상으로 분할하고, 원본 영상을 포함하여 총 6장의 시퀀스로 구성하였다. 이 과정을 통해 모델이 깊이 구간별 공간적 주의(attention)를 학습할 수 있도록 하여 배경 노이즈를 억제하고, 사람의 보행 패턴 인식 성능을 향상시킨다. 그림 3는 masking 과정을 통해 생성된 깊이 영상을 보여준다.

라벨링은 사전에 정의된 8가지 보행 의도를 약 15초 간격으로 구분하여, 각 프레임에 해당 의도 라벨을 부여하는 방식으로 수행하였다.

이후, 정제된 깊이 영상 시퀀스는 슬라이딩 윈도우 방식으로 분할하여 학습용 입력 시퀀스를 구성하였다[21,22]. 각

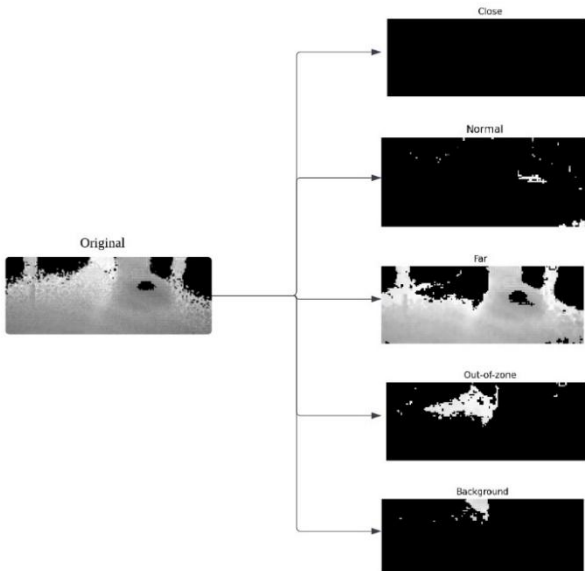


그림 3. 깊이 영상의 masking 과정 및 구간 분할.

Fig. 3. Masking and depth range segmentation of depth images.

윈도우는 15프레임으로 설정하였으며, 이는 실험에서 사용된 기본 설정 값으로, 모델 성능이나 데이터 특성에 따라 조정 가능한 하이퍼 파라미터이다. 본 설정을 통해 시간적 연속성을 보존하면서도 학습 데이터의 다양성과 균형성을 확보하였다.

모델 학습을 위해 전체 데이터 약 43,200프레임을 학습 세트와 테스트 세트로 분할하였다. 비록 피험자 수는 제한적이지만, 각 피험자당 충분한 길이의 반복 측정 데이터를 포함하고 있어 개인 내 변동성을 반영한 학습이 가능하도록 구성되었다. 데이터 분할은 측정 세션 기준으로 수행되었으며, 각 피험자별로 무작위로 1회 세션을 테스트 데이터로 분리하고, 나머지 3회 세션을 학습 데이터로 사용하여 약 8:2 비율의 train-test split을 구성하였다. 이와 같은 분할 방식에 따라 동일 세션의 데이터가 학습과 테스트 세트에 중복되지 않도록 분리되었으며, 이를 통해 반복 측정 간 일반화 성능을 평가하였다.

입력 특징 값은 Scikit-Learn의 Standard Scaler 함수를 사용하여 평균 0, 표준편차 1로 정규화하였다[23]. 또한 클래스 간 데이터 불균형을 완화하기 위해 resampling 기법을 적용하였으며, 이는 학습 데이터에 대해서만 수행되었다. 테스트 데이터에는 어떠한 resampling도 적용하지 않았다. 이를 통해 모델이 특정 클래스에 편향되지 않도록 하였다.

2. 보행 의도 예측

본 절에서는 ToF 깊이 카메라로부터 획득한 깊이 영상 시퀀스를 입력으로 하여 사용자의 보행 의도를 분류하기 위한 딥러닝 기반 예측 모델을 제안한다. 제안된 모델은 CNN과 Transformer Encoder를 결합한 하이브리드 구조로, 깊이 영상의 공간적 특징(spatial feature)과 시계열적 의존성(temporal dependency)을 통합적으로 학습하도록 설계하였다. 제안된 모델의 구조는 그림 4와 같다.

입력 깊이 영상은 2차원 합성곱 연산과 활성화 함수를 통해 지역적 특징을 추출한 후, GAP (Global Average Pooling)을 통해 압축된 형태로 변환된다. CNN 블록의 연산은 다음과 같이 정의된다.

$$F_{cnn} = GAP(\sigma(W_2 * (\sigma(W_1 * X + b_1)) + b_2)) \quad (1)$$

여기서 $*$ 는 convolution 연산, $\sigma(\cdot)$ 는 ReLU activation function, $GAP(\cdot)$ 은 Global Average Pooling을 의미한다. W_1 과 W_2 는 각각 첫 번째 및 두 번째 convolution layer의 가중치를 나타내고, b_1 과 b_2 는 각 레이어에 대응하는 bias term을 나타낸다. 이를

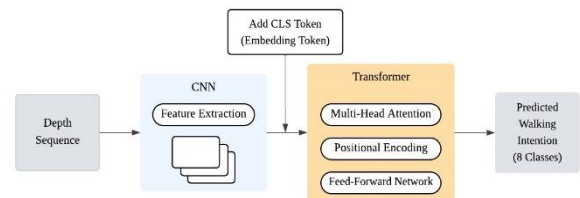


그림 4. CNN-Transformer 기반 보행 의도 예측 모델의 구조.

Fig. 4. Architecture of the proposed CNN-transformer-based walking intention prediction model.

통해 CNN은 깊이 영상 내의 공간적 구조 정보를 압축된 벡터 형태로 표현한다. 이후 CNN에서 추출된 feature 벡터는 Transformer Encoder의 입력으로 전달된다. Transformer는 self-attention 메커니즘을 통해 프레임 간 전역적 상호 관계를 학습하며, attention 계산은 다음과 같이 정의된다.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

여기서 Q, K, V는 각각 Query, Key, Value 행렬을 나타내며, d_k 는 Key의 차원이다. Transformer Encoder는 self-attention, Feed-Forward Network, Layer Normalization으로 구성된 표준 구조를 따르며, 시퀀스 전 구간의 상관관계를 반영하여 temporal feature를 학습한다.

최종적으로 CLS token 출력을 Fully Connected Layer에 통과시켜 8가지 보행 의도 클래스에 대한 Softmax 확률을 산출한다. 제안된 구조는 깊이 영상의 공간적 패턴과 시간적 변화를 통합적으로 고려함으로써, 보행자의 의도를 보다 정교하게 인식할 수 있다는 장점을 가진다. 이를 구현하기 위해, 본 연구에서는 CNN는 기존연구[20]의 모델 파라미터를 참조하여 총 2개의 convolution layer로 구성하였으며, 각 레이어는 커널 크기 3을 사용하고 필터 수는 64로 설정하였다. Transformer는 $d_{model} = 64$, multi-head attention의 head 수는 8로 고정하였으며, encoder layer 수와 feed-forward network의 차원은 Optuna [24] 기반 탐색을 통해 각각 3과 128로 선정하였다.

3. 사용자 위치 추정

본 절에서는 ToF 깊이 카메라로부터 획득한 깊이 영상을 기반으로 사용자-스마트 워커 간 거리를 추정하는 방법을 제안한다. 깊이 영상은 거리 범위에 따라 Close 영역과 Far 영역으로 마스킹되며, 연속 프레임 간 깊이 변화량을 이용하여 사용자의 움직임 상태를 판별한다. 정지 및 움직임 상태에서의 전체 처리 과정은 그림 5에 제시된 바와 같다.

움직임 검출은 깊이 차이 방식으로 수행된다. 현재 프레임 $D_t(x, y)$ 와 이전 프레임 $D_{t-1}(x, y)$ 간의 깊이 차이를 계산하고, 그 절대값이 임계값 T_{diff} 를 초과할 경우 해당 픽셀을 움직임이 있는 영역으로 판단한다. 이렇게 계산된 차이값을 기반으로 이전 마스크 $M_t(x, y)$ 를 생성하여 움직임 영역을 구분한다.

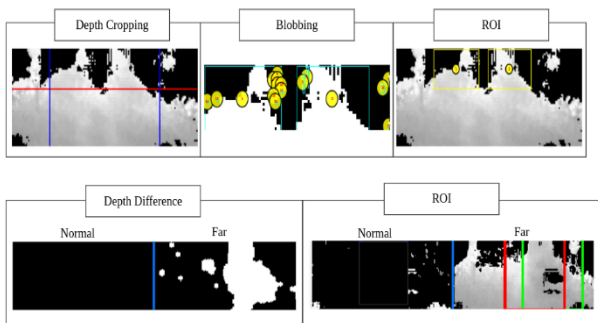


그림 5. 정지(상단) 및 움직임(하단) 상태에서의 사용자 위치 추정 과정.

Fig. 5. User position estimation process in static (top) and dynamic (bottom) conditions.

$$M_t(x, y) = \begin{cases} 1, & |D_t(x, y) - D_{t-1}(x, y)| > T_{diff} \\ 0, & otherwise \end{cases} \quad (3)$$

판단된 모션 상태에 따라 ROI 검출 및 거리 계산 방식이 달라진다. 이를 위해 본 연구에서는 실험적 경험에 의해서 정규화된 깊이 영상 간의 차이를 기준으로 $T_{diff} = 3$ 을 설정하였으며, 노이즈에 의해 발생하는 미소 영역을 제거하기 위해 최소 contour 면적을 500 픽셀로 제한하였다.

정지 상태에서는 Far 마스크만을 사용하여 blob 검출을 수행한다. 여기서 blob은 깊이 영상에서 인접한 픽셀들로 형성된 연결된 영역(connected component)을 의미하며, 이는 사용자의 다리나 발과 같은 독립된 물체 영역에 해당한다. 깊이 영상의 상단 영역에서 연결 성분 분석을 통해 독립된 blob 후보를 탐색하고, 가장 큰 blob을 ROI로 설정한다. ROI 내 픽셀을 카메라 좌표계로 변환되며, 변환된 좌표의 수직 방향(y-axis) 성분 y_i^{cam} 과 전방 방향(z-axis) 성분 z_i^{cam} 을 사용하여 θ 와 h_c 를 고려한 z_i 를 계산한다.

$$z_i = y_i^{cam} \sin \theta + z_i^{cam} \cos \theta \quad (4)$$

정지 상태에서는 각 blob의 중앙값(median)을 계산하고, 모든 blob의 평균(mean)을 최종 거리로 산출한다.

$$d_t = mean(median(z_i)) \text{ for each blob} \quad (5)$$

움직임 상태에서는 Close 및 Far 마스크를 모두 사용하여 깊이 차이 결과로부터 ROI를 구성하고, ROI 내부의 깊이값 z_i 의 단순 평균을 통해 거리를 계산한다. 여기서 R_t 는 t번째 프레임에서 깊이 차이로 추출된 ROI를 구성하는 유효 픽셀들의 집합이며, $N = |R_t|$ 는 해당 ROI 내 픽셀의 총 개수이다. 이때 계산된 거리는 이전 정지 상태에서 산출된 기준 거리 값을 참조하여 보정되며, 이를 통해 프레임 간 거리의 일관성을 유지한다.

$$d_t = \frac{1}{N} \sum_{i \in R_t} z_i \quad (6)$$

움직임 과정에서 급격한 거리 변동이 발생할 수 있으므로, 예측 기반 칼만 필터를 적용하여 거리의 연속성과 안정성을 확보하였다. 거리 변화량이 설정된 임계 값을 초과할 경우에는 칼만 필터의 예측값을 사용하여 이상치나 급격한 깊이 변화를 완화하였다.

III. 결과 및 논의

Transformer 모델의 하이퍼파라미터 튜닝은 Optuna [24]를 이용하여 수행하였다. 탐색 과정에서 accuracy를 목적함수로 설정하여 최적의 파라미터 조합을 도출하였으며, 이를 기반으로 Transformer 모델을 학습시켰다. 이후 테스트 데이터셋을 대상으로 accuracy, precision, recall, F1-score를 사용하여 성능을 평가하였다.

Transformer 모델의 학습 및 검증 과정에서의 손실 변화는 그림 6에 나타내었다. 학습 초기에는 Train Loss가 급격히 감소하며, 모델이 보행 패턴의 기본적인 특징을 빠르게 학습하는

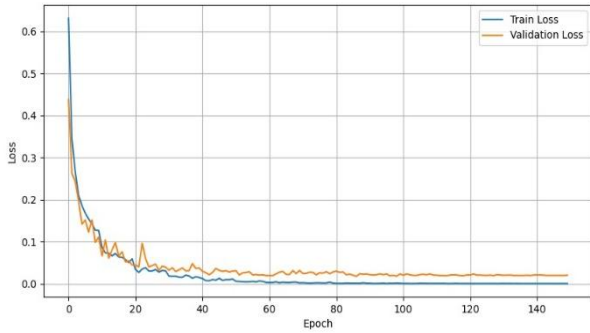


그림 6. 최적 Transformer 모델의 학습 곡선.

Fig. 6. Learning curve of the best transformer model.

양상을 보였다. 이후 약 20 epoch 이후부터는 손실 값이 완만하게 감소하며 안정적으로 수렴하였다. Validation Loss 역시 학습 초반의 진동 구간을 지나 빠르게 안정화되었으며, 최종 수렴 구간에서 Train Loss와 Validation Loss는 각각 약 0.02와 0.03 수준으로 유지되었다. 전체적으로 손실 값이 매우 낮게 수렴하였고, 두 곡선이 거의 일치하는 형태를 보여 과적합(overfitting) 현상은 발생하지 않았다. 이러한 안정적인 학습 결과는 Optuna 기반의 하이퍼파라미터 튜닝과 교차검증(cross validation) 적용을 통해 모델의 일반화 성능이 향상된 결과로 해석된다.

그림 7은 테스트 데이터셋에 대한 혼동 행렬(confusion matrix)을 나타낸 것이다. 각 클래스별 결과를 살펴보면, 전진(forward)은 비교적 안정적으로 인식되었다. 전진하며 좌회전(forward left)은 낮은 인식률을 보였으며, 주로 전진으로 오분류되었다. 전진하면서 회전하는 경우의 낮은 인식률은 본 연구에서 사용된 데이터셋은 표본 수가 제한적이므로, 이러한 경향이 일반적인 보행 특성에 기인한 것인지 또는 데이터 분포 및 수집 환경에 따른 결과인지는 추가적인 데이터 수집과 분석을 통해 검증할 필요가 있다. 후진(backward)은 전진 계열 동작과 구분되는 깊이 패턴을 가지고 있어 중간 수준의 인식률을 보였다. 제자리 좌회전(in-place left)과 제자리 우회전(in-place right)은 두 클래스의 동작 특성이 매우 유사하여 서로 간의 오분류가 많이 발생하였다. 정지(stop)는 배경(background)으로 오분류되는 비율이 특히 높았다. 이는 정지 상태에서는 사람의 실루엣 변화가 거의 없어 깊이 영상에서 사람이 없는 장면과 유사하게 나타나기 때문으로, stop과 background 간 구분이 구조적으로 어려운 특성이 반영된 결과로 볼 수 있다. 마지막으로 배경 클래스는 가장 명확하게 구분되었으며, 대부분 올바르게 분류되었다.

이러한 결과는 Transformer 모델이 전진과 같이 이동 패턴과 깊이 변화가 뚜렷한 동작에서는 상대적으로 안정적인 성능을 보이지만, 정지나 회전처럼 움직임이 작고 깊이 변화가 미세한 동작에서는 배경과의 구분이 상대적으로 어려워 오분류가 발생할 수 있음을 시사한다. 이를 보완하기 위해서는 사람 존재 여부를 명시적으로 표현할 수 있는 occupancy/실루엣 기반 특징을 추가하거나[25], 정지·배경과 유사한 프레임 하드 네거티브[26]로 활용하는 학습 전략을 통해 클래스 간 경계를 강화할 필요가 있다. 또한 짧은 시간 구간에서 발생하는 미세 움직임을 안정적으로 포착할 수 있도록 temporal

window를 확장하거나, 정지 및 회전 클래스에 대한 손실 가중치를 조정하여 모델이 해당 패턴을 더 명확히 학습하도록 유도하는 방법도 효과적일 것으로 보인다[27]. 이러한 개선은 클래스 간의 혼동을 줄이고, 보행 의도 인식의 전반적인 신뢰성을 향상시키는 데 기여할 것으로 기대된다.

Transformer 기반 보행 의도 예측 모델의 성능은 표 1에 요약하였다. 전체 8개 클래스에 대해 평균 accuracy 0.7536, precision 0.7699, recall 0.7480, F1-score 0.7461을 기록하였다. 전반적으로 양호한 성능을 보였으며, 동일하게 깊이 기반 정보를 활용하여 보행 의도를 예측한 연구 [20]에서 Ablation Study Table 1에 DTA-Mot 성능이 약 0.66 수준임을 고려할 때, 본 연구의 Transformer 모델은 단일 모달리티 환경에서도 충분히 경쟁력 있는 결과를 확보한 것으로 해석될 수 있다. 이는 다양한 보행 의도 간 깊이 패턴의 유사성을 고려할 때 Transformer Encoder가 보행 시퀀스의 시계열적 의존성을 효과적으로 학습하였음을 의미한다. 또한 주요 평가 지표 간 편차가 크지 않아 모델이 전반적으로 안정적인 분류 경향을 보였으며, 이는 제한된 데이터셋 조건에서도 Transformer 구조가 보행 의도의 시간적 변화를 일관되게 반영할 수 있음을 시사한다.

의도 인식 모듈의 실시간성에 대한 추정은 다음과 같다. Nvidia Jetson Nano 환경에서 비전 모듈(FP32)의 단독 추론 성능은 평균 약 67 fps (프레임당 약 15 ms)로 확인하였다. 그러나 실제 시스템 환경에서는 카메라 입력 처리, 전·후처리,

Forward	85.88	4.98	2.37	5.58	0.24	0.12	0.83	0.00
Forward Left	37.72	52.19	3.91	2.61	0.00	0.00	0.24	3.32
Forward Right	4.50	4.99	85.89	1.09	0.12	1.58	0.00	1.82
Backward	2.84	5.06	12.82	72.01	0.99	0.00	0.00	6.29
In-place Left	1.11	0.00	6.78	1.60	77.31	11.47	1.73	0.00
In-place Right	0.00	0.12	1.48	2.34	31.20	63.01	1.85	0.00
Stop	0.00	0.49	0.00	0.00	2.34	3.82	62.89	30.46
Background	0.00	0.00	0.00	0.00	0.00	0.00	0.82	99.18
	Forward	Forward Left	Forward Right	Backward	In-place Left	In-place Right	Stop	Background

그림 7. 테스트 데이터셋에 대한 보행 의도 분류의 혼동 행렬.
Fig. 7. Confusion matrix of walking intention classification on the test dataset.

표 1. 보행 의도 예측 모델의 성능 평가 지표.

Table 1. Performance evaluation metrics of the walking intention prediction model.

	Accuracy	Precision	Recall	F1-Score
Macro Average	0.7536	0.7699	0.7480	0.7461

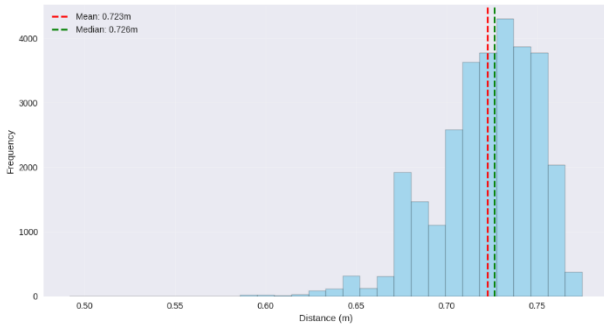


그림 8. 추정된 사람-카메라 거리의 분포.

Fig. 8. Distribution of the estimated human-camera distance.

CPU-GPU 데이터 전송, 시스템 스케줄링 오버헤드 등 다양한 요소가 추가로 포함되므로, 최종 추론 속도는 약 17~20 fps 수준으로 감소할 것으로 예상된다. 이에 따라, 실시간 처리 성능 확보 및 시스템 자원 활용 효율 개선을 위해 향후 양자화(INT8/FP16) 적용 및 TensorRT 기반 최적화와 같은 추가적인 성능 개선 기법의 도입이 필요하다.

깊이 차이 기반의 ROI 추정과 칼만 필터 기반 거리 보정 알고리즘의 성능을 검증하였다. 정지 상태와 움직임 상태를 통합하여 거리 추정 오차를 계산하였으며, 실제 거리(스마트 워커와 사람 사이의 기준 거리)는 줄자를 이용하여 0.70 m로 설정하였다. 이 기준값과 추정 거리의 평균 및 표준편차는 $0.723 \text{ m} \pm 0.027 \text{ m}$ 로 측정되었다. 평균 오차 약 0.023 m는 절대적인 깊이 편차를 포함한 값이지만, 발의 전-후방 움직임과 같은 상대적 거리 변화는 안정적으로 포착할 수 있을 만큼 충분한 분해능을 제공한다. 이는 스마트 워커가 향후 제어 단계에서 사용자의 이동에 따라 일정 거리 간격을 유지하도록 하는 알고리즘을 구현할 때 필요한 거리 변화를 정확히 감지할 수 있음을 의미한다. 또한 표준편차 0.027 m는 프레임 간 변동성이 매우 작음을 보여주며, 거리 추정 결과가 시간적으로 일관되게 유지됨을 의미한다. 거리 분포 또한 평균 주변에 밀집되어 있어 전반적으로 안정적인 추정 성능을 그림 8에서 확인할 수 있다. 이러한 결과는 제안된 방식이 단일 ToF 깊이 카메라만으로도 사용자의 정적 및 동적 위치 변화를 안정적이고 신뢰성 있게 인식·추적할 수 있음을 실험적으로 검증한 것이다.

IV. 결론

본 연구는 ToF 기반 깊이 영상과 Transformer 모델을 결합하여, 사용자의 보행 의도와 위치를 동시에 인식할 수 있는 스마트 워커 핵심 기술을 구현하였다. 제안한 방법은 단일 저가형 깊이 카메라만으로도 보행자의 공간적·시계열적 특징을 정밀하게 분석하여 보행 의도를 인식하고, 사용자의 위치를 안정적으로 추정할 수 있음을 입증하였다. 특히, 하이퍼파라미터 최적화를 통해 보행 의도 인식의 성능을 향상시켰으며, 움직임 구간의 거리 계산을 정지 상태에서 측정된 기준 거리값을 기반으로 보정함으로써, 거리 추정의 일관성과 안정성을 확보하였다. 이를 통해 고가의 복합 센서나 장비 없이도 보행 의도 인식과 거리 추정을 단일 센서로 동시에 수행할 수 있음을 실험적으로 검증하였으며, 이는 스마트 워

커 시스템의 경제성과 보급 가능성을 동시에 확보할 수 있다는 점에서 의미가 있다. 다만 본 연구는 정상인 대상의 소규모 단일 환경 데이터셋을 기반으로 진행된 탐색적 연구로, 보행 장애가 있는 사용자나 다양한 보행 패턴을 가진 대상자에게 적용할 경우 성능 저하가 발생할 가능성이 있다. 이러한 한계를 해소하기 위해 향후에는 모터 궤적 데이터의 통합을 통해 자동 스마트 워커 제어의 기반을 마련하고, 다양한 사용자군 및 실제 사용 환경을 포함한 확장 데이터 수집을 통해 모델의 일반화 성능을 검증할 예정이다. 궁극적으로 본 연구는 저비용 센싱 기반의 인간-로봇 상호작용 기술 발전에 기여할 뿐 아니라, 고령사회에서 보행 안전성과 삶의 질을 향상시키는 사회적 파급 효과를 지닌다.

REFERENCES

- [1] M. Martins, C. Santos, A. Frizzera, and R. Ceres, "A review of the functionalities of smart walkers," *Medical Engineering and Physics*, vol. 37, no. 10, pp. 917–928, Oct. 2015. doi: <http://dx.doi.org/10.1016/j.medengphy.2015.07.006>
- [2] M. M. Martins, C. P. Santos, A. Frizzera-Neto, and R. Ceres, "Assistive mobility devices focusing on Smart Walkers: Classification and review," *Robotics and Autonomous Systems*, vol. 60, no. 4, pp. 548–562, Apr. 2012. doi: <http://dx.doi.org/10.1016/j.robot.2011.11.015>
- [3] S. D. Sierra M., M. Garzón, M. Múnera, and C. A. Cifuentes, "Human-Robot-Environment interaction interface for smart walker assisted gait: AGoRA Walker," *Sensors*, vol. 19, no. 13, pp. 2897, Jun. 2019. doi: <http://dx.doi.org/10.3390/s19132897>
- [4] C. Gonçalves, J. M. Lopes, S. Moccia, D. Berardini, L. Migliorelli, and C. P. Santos, "Deep learning-based approaches for human motion decoding in smart walkers for rehabilitation," *Computer Methods and Programs in Biomedicine*, vol. 241, pp. 107753, Feb. 2023.
- [5] D.-M. Ding, Y.-G. Wang, W. Zhang, and Q. Chen, "Fall detection system on smart walker based on multisensor data fusion and SPRT method," *IEEE Access*, vol. 10, pp. 80932–80948, Aug. 2022. doi: <http://dx.doi.org/10.1109/ACCESS.2022.3195674>
- [6] J. Kim, S. Lee, and H. Park, "Gait intention detection drive system for exoskeleton robots," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 29, no. 4, pp. 345–352, 2023.
- [7] G. Abou Haidar, R. Achkar, R. Maalouf, B. Mchick, H. Allam, and A. R. Moussa, "Smart walker," *Proc. of the International Conference on Future Internet of Things and Cloud (FiCloud)*, pp. 415–419, Aug. 2014. doi: <http://dx.doi.org/10.1109/FiCloud.2014.74>
- [8] X. Zhao, Z. Zhu, M. Liu, C. Zhao, Y. Zhao, J. Pan, Z. Wang, and C. Wu, "A smart robotic walker with intelligent close-proximity interaction capabilities for elderly mobility safety," *Frontiers in Neurobotics*, vol. 14, pp. 75, Oct. 2020. doi: <http://dx.doi.org/10.3389/fnbot.2020.575889>
- [9] C. Zhao, W. Guo, R. Wen, Z. Wang, and C. Wu, "Deep learning-driven front-following within close proximity: A hands-free control model on a smart walker," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 6, pp. 1040–1051, Dec. 2022. doi: <http://dx.doi.org/10.1109/ICRA46639.2022.9811910>
- [10] S. Page, M. M. Martins, L. Saint-Bauzel, C. P. Santos, and V.

- Pasqui, "Fast embedded feet pose estimation based on a depth camera for smart walker," *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4224–4229, May 2015.
- [11] N. Ejaz, S. J. Khan, F. Azim, M. Asif, E. Teuþan, A. Pleša, R. Păcurar, and S.-D. Stan, "Automatic gait classification model empowered by machine learning for people with and without osteoporosis using smart walker," *Applied Sciences*, vol. 14, no. 9, pp. 3874, Apr. 2024.
doi: <http://dx.doi.org/10.3390/app14093874>
- [12] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Automatic learning of gait signatures for people identification," *Pattern Recognition Letters*, vol. 98, pp. 219–226, Apr. 2017.
- [13] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Multimodal feature fusion for CNN-based gait recognition: an empirical comparison," *Neural Computing and Applications*, vol. 32, no. 18, pp. 14173–14193, Mar. 2020. doi: <http://dx.doi.org/10.1007/S00521-020-04811-Z>
- [14] D. H. Kim and J. M. Jeong, "AI CNN based defect inspection system of a governor using a rotating platform," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 30, no. 10, pp. 1082–1089, Oct. 2024.
- [15] M. S. Chae, M. H. Ha, and T. H. Park, "A hybrid method of ARIMA and transformer for forecasting the time series data," *Journal of Institute of Control, Robotics and Systems (in Korean)*, vol. 31, no. 7, pp. 722–727, Jul. 2025.
- [16] S. Thundiyil, S. S. Shalamzari, J. Picone, and S. McKenzie, "Transformers for modeling long-term dependencies in time series data: a review," in *Proc. of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4, Dec. 2023.
- [17] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, R. P. Ramachandran, and G. Rasool, "Transformers in time-series analysis: a tutorial," *Circuits, Systems, and Signal Processing*, vol. 42, no. 7, pp. 7433–7466, Jul. 2023.
doi: <http://dx.doi.org/10.1109/9.210138>
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6000–6010, Dec. 2017.
- [19] N. Mostofa, C. Feltner, K. Fullin, J. Guilbe, S. Zehabian, S. S. Bacanlı, L. Bölöni, and D. Turgut, "A Smart Walker for People with Both Visual and Mobility Impairment," *Sensors*, vol. 21, no. 10, p. 3488, May 2021, doi: <http://dx.doi.org/10.3390/s21103488>
- [20] C. Zhao, L. Guo, R. Wen, Y. Wang, and C. Wu, "Depth-temporal attention with dual modality data for walking intention prediction in close-proximity front-following," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 2, pp. 215–229, Feb. 2024.
- [21] T. N. Nguyen, H. H. Huynh, and J. Meunier, "Assessment of gait normality using a depth camera and mirrors," *Proc. of the IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 37–41, Mar. 2018.
doi: <http://dx.doi.org/10.1109/BHI.2018.8333364>
- [22] P. Limcharoen, N. Khamsemanan, and C. Nattee, "Gait recognition for 2-second walks using viewpoint normalization and sliding-window process," *IEEE Access*, vol. 11, pp. 37082–37095, Apr. 2023.
- [23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011.
- [24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, pp. 2623–2631, Aug. 2019.
- [25] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 2, pp. 316–322, Feb. 2006.
doi: <http://dx.doi.org/10.1109/TPAMI.2006.38>
- [26] H. Zhang, Z. Li, J. Yang, X. Wang, C. Guo, and C. Feng, "Revisiting hard negative mining in contrastive learning for visual understanding," *Electronics*, Vol. 12, No. 23, pp. 1–16, Dec. 2023.
- [27] Y. Li, J. Yang, X. Wang, and C. Guo, "DepthGait: Multi-scale cross-level feature fusion of depth gait sequences for human identification," *Sensors*, Vol. 23, No. 7, pp. 3241–3254, 2023.



최 원 석

2024년 신라대학교 융합기계공학부(학사). 2024년~현재 성균관대학교 지능형 로봇학과 석사과정. 관심분야는 인공지능, 로봇 제어, HRI.



최 준 혁

2024년 한국외국어대학교 정보통신공학과(학사). 2024년~현재 성균관대학교 전자전기컴퓨터공학과 석사과정. 관심분야는 강화학습, 딥러닝, 컴퓨터비전.



최 문 택

2000년 U. of Southern California 항공우주 공학과(공학박사). 2013년~현재 성균관대학교 지능형로봇학과 교수. 관심분야는 로봇 지능 학습, 헬스케어 빅데이터 분석, 근골격 재활분석, 외골격 로봇 제어.