Prototypical Network를 이용한 IMU 센서 데이터 기반의 손 제스처 인식

홍동진[1], 박형욱[2], 최지웅[2], 장원두[3*]
[1]국립부경대학교 인공지능연구소 박사후 연구원, [2]국립부경대학교 컴퓨터·인공지능공학부 학생,
[3]국립부경대학교 컴퓨터·인공지능공학부 교수

IMU-Based Hand Gesture Recognition Using Prototypical Networks

Dong-Jin Hong[1], Hyeong-Uk Park[2], Ji-Woong Choi[2], Won-Du Chang[3*]
[1]Researcher, Artificial Intelligence Institute, Pukyong National University
[2]Student, Division of Computer Engineering and Artificial Intelligence, Pukyong National University
[3]Professor, Division of Computer Engineering and Artificial Intelligence, Pukyong National University

요 약 최근 메타버스, 가상현실(VR), 스마트 헬스케어 등 비대면 인터페이스의 중요성이 증대됨에 따라, 사용자와 컴퓨터 간의 직관적인 상호작용을 위한 손 제스처 인식 기술이 주목받고 있다. 본 연구에서는 스마트 워치로 수집된 IMU 센서 데이터를 이용하는 Prototypical Network 기반의 손 제스처 인식 모델을 제안한다. 제안된 모델은 15명의 피험자로부터 수집된 10가지의 숫자 제스처 데이터를 분류하는 실험에서 평균 정확도 86.20%를 달성하였다. 본 연구는 스마트 헬스케어 및 메타버스 등 비대면 인터페이스가 필수적인 분야에서, 사용자가 별도의 복잡한 데이터 수집 과정 없이 개인화된 제스처를 사용할 수 있게 함으로써 사용자 편의성과 시스템의 실용성을 높이는 데 활용될 것으로 기대된다.

주제어 : 손 제스처 인식, IMU 센서, Prototypical Network, 시계열 합성곱 신경망, 퓨샷 러닝

Abstract With the increasing importance of contactless interfaces in fields such as the metaverse, virtual reality (VR), and smart healthcare, hand gesture recognition technology has gained significant attention for enabling intuitive human-computer interaction. This study proposes a hand gesture recognition model based on Prototypical Networks that utilizes IMU sensor data collected from smartwatches. In experiments classifying ten types of numeric gestures collected from 15 subjects, the proposed model achieved an average accuracy of 86.20%. This study is expected to contribute to enhancing user convenience and system practicality in fields requiring contactless interfaces, such as smart healthcare and the metaverse, by enabling users to utilize personalized gestures without complex data collection processes.

Key Words : Hand Gesture Recognition; IMU Sensor; Prototypical Networks; TCN; Few-shot Learning

본 논문은 국립부경대학교 자율창의학술연구비(2024년)에 의하여 연구되었음.
*교신저자 : 장원두(chang@pknu.ac.kr)
접수일 2025년 11월 11일 수정일 2025년 12월 01일 심사완료일 2025년 12월 18일

1. 서론

최근 가상현실(VR)과 증강현실(AR), 메타버스와 같은 실감형 콘텐츠 시장이 급성장함에 따라, HCI (Human-Computer Interaction) 기술은 기존의 마우스나 키보드와 같은 접촉식 입력 장치에서 벗어나 직관적이고 자연스러운 방식의 제어 인터페이스를 사용하는 방향으로 발전하고 있다[1, 2]. 손 제스처 인식(Hand Gesture Recognition)은 인간의 의사소통에 있어 비언어적 정보를 전달하기 위한 유용한 수단이다. 손 제스처 인식 기술은 스마트 헬스케어, 웨어러블 디바이스 제어, 인간-로봇 상호작용 등 다양한 분야에서 유용하게 활용되고 있다[3, 4].

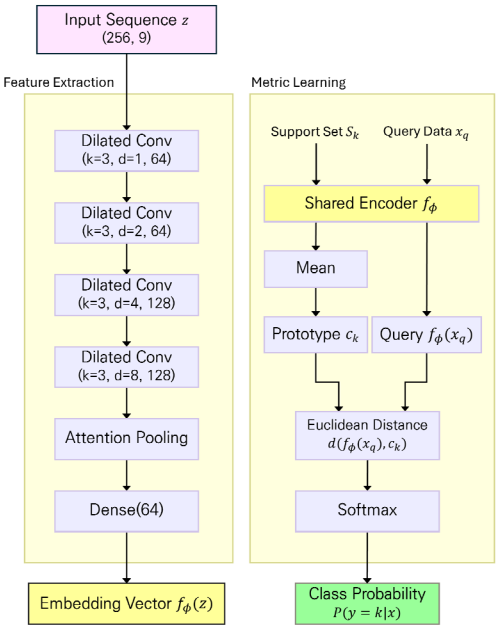
초기의 제스처 인식 연구는 주로 카메라를 활용한 비전 기반 방식으로 수행되었다[5]. 비전 기반 방식은 별도의 장비가 필요하지 않고, 사람이 보는 것과 동일한 직관적인 해석이 가능하다는 장점이 있지만, 카메라가 설치된 제한된 공간 내에서만 인식이 가능하다는 공간적 제약과 사용자의 영상 데이터 수집에 따른 프라이버시 침해 우려와 같은 문제가 있다[6, 7].

이에 대한 대안으로 IMU 센서를 활용한 제스처 인식 연구가 주목받고 있다[8, 9, 10]. 이러한 연구는 스마트 워치나 웨어러블 밴드의 IMU 데이터를 1D-CNN에 입력하여 제스처의 공간적 패턴을 추출하거나, LSTM을 통해 연속적인 동작의 흐름을 분석하기 위해 수행되었다. 직접적으로 손 제스처 인식을 위해 수행된 연구로는 3차원 공간에서의 손 움직임 궤적을 복원 및 분류를 위해 Amma 등이 제안한 관성 센서를 이용한 공중 필기(Air-writing) 인식 시스템에 관한 연구가 있다[11]. 해당 연구들은 96-99%대의 우수한 정확도를 달성하였으나, 1-3명의 적은 실험 참여자로부터 데이터를 수집하였으므로 사용자 간 변동성 문제가 발생할 수 있고, 이에 대한 보완이 필요하다는 것을 직접적으로 언급하였다.

이러한 문제를 극복하고 일반화된 분류 성능을 확보하기 위해 본 연구에서는 Prototypical Network를 이용한 손 제스처 인식 모델을 제안한다. Prototypical Network[12]는 임베딩 공간에서 각 클래스를 대표하는 프로토타입(Prototype)을 생성하고, 입력 데이터와의 유클리드 거리를 기반으로 분류를 수행하는 모델이다. 기존 퓨샷 러닝에서 주로 사용되는 샴 네트워크[13]가 입력 쌍 간의 이진 유사도 판별에 집중하는 것과 달리, Prototypical Network는 다중 클래스 분류 문제를 거리 기반으로 직관적이고 효율적으로 처리할 수 있다는 장점이 있다. 이에 더해, 노이즈가 많은 센서 데이터의 특징을 추출하는 데 효과적인 것으로 알려져 있다[14, 15]. 본 연구는 TCN 기반의 인코더와 Prototypical Network를 이용해 적은 수의 IMU 손 제스처 데이터만으로도 높은 정확도의 손 제스처 인식이 가능함을 보여주고자 한다.

2. 손 제스처 인식 모델

본 연구에서는 TCN(Temporal Convolutional Networks)[16] 인코더 구조로 구성된 특징 추출 모듈을 이용하는 Prototypical Network 기반의 손 제스처 인식 모델을 제안한다. 제안하는 모델의 전체적인 구조는 [Fig. 1]과 같다.



[Fig. 1] Overall architecture of the proposed model

제안하는 모델은 특징 추출(Feature Extraction) 모듈과 메트릭 러닝(Metric Learning) 모듈로 구성되어 있다. 특징 추출 모듈은 시계열 입력 시퀀스 z를 입력받아, 저차원의 특징 공간에 매핑하는 인코더(비선형 함수) f_ϕ에 해당한다. 메트릭 러닝 모듈은 서포트 집합과 쿼리 데이터에 대해 동일한 인코더 f_ϕ를 공유하며, 특징 벡터 간의 거리를 기반으로 클래스를 분류한다.

입력 시퀀스는 가속도, 선형 가속도, 자이로스코프의 3축 데이터를 포함하여 총 9개의 채널로 구성되며, 전처리를 통해 고정된 시간 길이로 정규화되어 모델에 주입된다. 본 연구에서는 256의 시간 길이를 가지도록 데이터를 전처리하였다.

특징 추출 모듈은 TCN 기반의 인코더로 설계되었으며, 4개의 팽창 합성곱 블록을 사용한다. 커널 크기 k는 3으로 고정하였으며, 필터 수는 초반 2개 블록은 64, 이후 2개 블록은 128로 설정하였다. 팽창 계수 d를 1, 2, 4, 8로 지수적으로 증가시켜 수용 영역을 점차 넓히는 구조로 설계되었다. 각 팽창 합성곱 블록 사이에는 배치 정규화와 ReLU 활성화 함수가 적용된다. 인코더를 통과한 특징 맵은 어텐션 풀링과 완전 연결 층(Dense)을 거쳐 64차원의 임베딩 벡터 $f_\phi(z)$가 된다. 특징 추출 모듈은 메트릭 러닝 모듈에서 서포트 데이터와 쿼리 데이터를 임베드하기 위한 공유된 인코더(Shared Encoder)로 사용된다.

메트릭 러닝 모듈은 서포트 집합을 활용하여 각 클래스의 기준점을 생성하고, 쿼리 데이터와의 거리를 측정하여 클래스를 예측하는 과정을 수행한다. 서포트 집합 S에 포함된 데이터들은 공유된 인코더를 통과한 후, 평균 연산을 통해 각 클래스 k를 대표하는 프로토타입 c_k로 변환된다. 쿼리 데이터 x 또한 동일한 인코더를 거쳐 임베딩 벡터 $f_\phi(x)$로 변환되며, 이후 각 클래스의 프로토타입 간의 유클리드 거리를 계산하여 유사도를 측정한다. 측정된 유사도에 대해 소프트맥스 연산을 적용하여 클래스별 확률 분포를 얻는다.

모델의 손실 함수는 클래스별 확률 분포를 바탕으로 정답 클래스에 대한 음의 로그 우도를 최소화하는 형태로 구성된다. 전체 쿼리 데이터 집합 Q에 대한 목적 함수 $J(\phi)$는 식 (1)과 같이 정의된다.

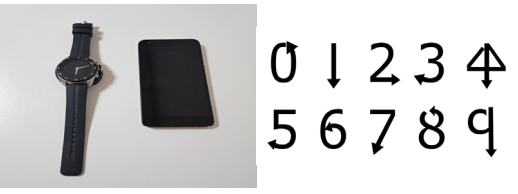
$$J(\phi) = \frac{1}{Q} \sum_{(x_q, y_q) \in Q} \log P(y = y_q | x_q) \qquad (1)$$

이를 통해 특징 추출 모듈(인코더) f_ϕ는 같은 제스처 클래스 내의 응집도는 높이고 서로 다른 클래스 간의 분리도는 최대화하는 방향으로 특징을 추출하도록 학습된다.

3. 모델 학습 및 성능 평가

3.1 데이터 구성

본 연구에서는 손 제스처 데이터 수집을 위해 가속도계와 자이로스코프가 내장된 스마트 워치(LG Watch W7)와 데이터 수신을 위한 스마트폰(Nexus 5X)을 활용하였다. 데이터셋 구축에는 총 15명의 피험자가 참여하였으며, 각 피험자는 스마트 워치를 착용한 상태에서 숫자 0부터 9까지의 10가지 클래스에 해당하는 손 제스처를 수행하였다. [Fig. 2]는 데이터 수집에 사용된 장비와 피험자가 시행한 각 숫자를 나타내는 손 제스처를 나타낸 그림이다.



[Fig. 2] Experimental setup for data collection: (a) Hardware devices; (b) Vocabulary of numeric hand gestures.

각 피험자는 양손으로 10가지 숫자 제스처를 2회씩 수행하였으며, 데이터셋은 총 600개(15명×2방향×숫자 10×2회)의 샘플로 구성되었다. 각 제스처 데이터는 60Hz의 샘플링 레이트로 약 3초간 수집되었다. 기록된 데이터는 각 센서 채널별로 평균이 0이고 분산이 1인 데이터 분포를 갖도록 조정되었다. 이후 원본 데이터를 중앙에 배치하고 시퀀스의 길이가 256이 되도록 양옆을 0으로 채우는 패딩을 수행하였다. 모든 데이터는 (256, 9)의 크기를 갖도록 전처리되어 모델에 입력되었다.

3.2 모델 학습 설정

제안된 모델의 학습 과정은 10개의 클래스를 분류하기 위해 각 클래스당 1개의 샘플만을 참조하는10-Way 1-Shot Learning 설정으로 진행되었다. 각 피험자가 수행한 2회의 제스처 중, 첫 번째 시퀀스를 서포트 집합으로 할당하고, 두 번째 시퀀스를 쿼리 집합으로 할당하였다. 따라서 서포트 집합과 쿼리 집합은 전체 600개의 샘플을 균등하게 나눈 300개의 샘플로 구성된다.

각 손의 움직임 특성을 효과적으로 학습하기 위해, 왼손 모델과 오른손 모델을 별도로 구축하여 독립적으로 학습을 진행하였다. 모델의 학습은 1,000 에피소드에 걸쳐 이루어졌다. 최적화 알고리즘으로는 Adam Optimizer를 사용하였으며, 학습률은 0.001로 설정하였다.

〈Table 1〉 Performance comparison of various backbones and learning methods

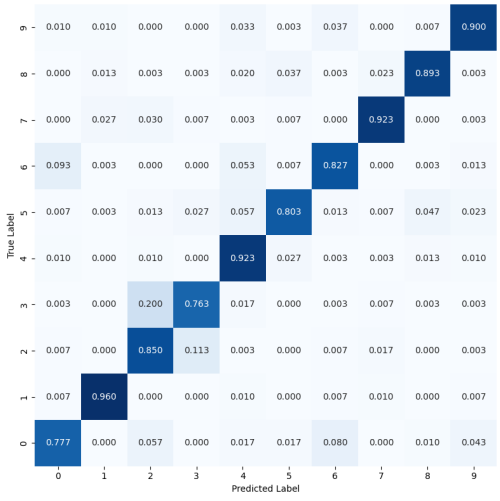
Method	Backbone	Left Hand Acc.(%)	Right Hand Acc,(%)	Avg. Acc.(%)
Traditional	DTW	56.00±0.00	72.67±0.00	64.33±0.00
Siamese Network	BiLSTM	35.40±2.12	44.80±2.29	40.10±1.74
	1D-CNN	36.43±2.96	48.27±3.19	42.20±2.36
	TCN	34.47±3.10	46.67±2.88	40.57±2.09
Prototypical Network	BiLSTM	43.33±9.68	52.60±9.44	47.97±7.07
	1D-CNN	**83.20±2.10**	87.67±1.96	85.43±1.12
	TCN(Ours)	82.00±1.26	**90.40±1.20**	**86.20±0.93**

3.3 모델 성능 평가

제안된 모델의 숫자 제스처 분류 성능을 객관적으로 검증하기 위해, 각 실험을 10회 진행한 후 왼손과 오른손 제스처에 대한 평균 정확도 및 표준편차를 측정하였다. 실험은 선행 연구들과의 실험 환경(센서 구성, 제스처 클래스 등) 차이에 따른 직접 비교의 한계를 극복하고 공정한 평가를 수행하기 위해 다양한 특징 추출 모듈(Backbone)과 메트릭 러닝 방법론(Method)을 조합하여 비교 수행되었으며, 결과는 〈Table 1〉과 같다.

DTW[17]와의 비교는 딥러닝을 통해 학습된 임베딩 공간의 거리 척도가, 단순히 신호를 시간 축으로 보정하여 측정한 물리적 거리 척도 대비 얼마나 유의미한 변별력을 갖는지를 검증하는 기준이 된다. 메트릭 러닝 방법에 대한 비교 모델로는 샴 네트워크를 사용하였다. 특징 추출 모듈로는 시계열 특징 추출에 주로 활용되는 BiLSTM[18], 1D-CNN, TCN을 각각 적용하였을 때의 성능을 비교하였다. 각 모델의 복잡도를 나타내는 파라미터 수는 BiLSTM 319,040개, TCN 114,561개, 1D-CNN 97,152개이며, 모두 동일한 전처리 및 하이퍼파라미터 조건에서 학습되었다.

실험 결과, 샴 네트워크를 메트릭 러닝 모듈로 사용한 경우, 특징 추출 모듈의 구조와 관계없이 40-42%의 낮은 평균 정확도를 보였다. 반면, Prototypical Network를 메트릭 러닝 모듈로 사용하는 경우에는 BiLSTM을 특징 추출 모듈로 사용하는 경우를 제외하면 모두 80% 이상의 평균 정확도를 달성하였다. 또한, 모든 실험에서 오른손 제스처 분류 모델이 왼손 제스처 분류 모델보다 높은 평균 정확도를 보이는 것을 확인할 수 있었다. 왼손 제스처 데이터에 대해서는 1D-CNN 기반 모델이 83.20%로 제안된 모델보다 1.20%p 더 높았으나, 오른손 제스처 데이터에 대해서는 제안된 모델이 90.40%로 가장 높은 정확도를 보였으며, 전체 평균 정확도 또한 제안된 모델이 가장 높았다.



[Fig. 3] Normalized confusion matrix of the proposed model

제안된 모델의 숫자 제스처 분류 결과에 대한 정규화된 혼동 행렬을 [Fig. 3]에 나타내었다. 대각 성분이 뚜렷하게 나타나 전반적인 분류 성능이 우수한 것으로 확인되었으나, 숫자 '0'과 '6'에 대하여 서로 잘못 인식하는 경우가 가장 많았으며, '2'와 '3'에 대해서도 비교적 빈번한 오분류가 발생하는 것을 확인할 수 있었다.

이러한 결과는 제안된 모델이 IMU 센서 신호로부터 제스처의 동적 특징을 충분히 추출할 수 있지만, 동작의 궤적이 기하학적으로 유사한 클래스들을 완벽하게 구분하는 데에는 한계가 있음을 보여준다. 실제로 허공에 쓰는 숫자 '0'과 '6'은 둥근 원을 그리는 회전 동작이 주를 이루며, '2'와 '3'은 가로 방향의 지그재그 혹은 곡선 움직임이 유사하여 일반적인 비전 기반 숫자 분류 모델에서도 혼동이 잦은 클래스에 해당한다. 센서 데이터상에

서도 두 제스처의 가속도 및 자이로 패턴이 매우 흡사하게 나타나기 때문으로 판단된다.

4. 결론

본 연구에서는 Prototypical Network 기반의 IMU 센서 데이터를 이용한 손 제스처 인식 모델을 제안하였다. 제안된 모델은 센서 데이터의 동적 특징을 효과적으로 포착하여, 적은 훈련 샘플만으로도 평균 정확도 86.20%를 달성하였다. 본 연구는 선행 연구들보다 많은 15명의 피험자로부터 데이터를 수집하여 다양성을 높였다.

모델의 성능 평가 과정에서 비교 모델과 제안된 모델 모두에서 오른손 제스처에 대한 인식 성능이 왼손보다 더 높은 정확도를 달성하였다. 이는 피험자가 주로 사용하는 손일수록 더 세밀하고 일관된 동작 제어가 가능하여 센서 신호의 패턴이 명확하게 형성되었기 때문으로 판단된다.

본 연구에서 데이터의 수집은 실제 사용성을 고려해 손가락 각도 등의 제약 없이 손목 궤적에 집중하도록 하여 다양성을 확보하였으나, 피험자 간 동작 방식에 편차를 발생시킬 수 있다.

향후 연구에서는 기하학적 궤적이 유사한 제스처 간의 오분류를 개선하고, 사용자별 숙련도와 생체역학적 특성을 반영한 개인화 적응(Personal Adaptation) 및 손의 방향이나 착용 위치 변화에 강건한 도메인 적응(Domain Adaptation) 기술을 도입하여 제스처를 수행하는 손의 방향에 구애받지 않는 통합 분류 모델로 확장하고자 한다.

REFERENCES

[1] P.Singhal, S.Verma, R.Gupta, R.Kumar and R.K.Arya, "Vision-based hand gesture recognition system for assistive communication using neural networks and GSM integration," *CICTN*, pp.891-895, 2025.

[2] M.Billinghurst, A.Clark and G.Lee, "A survey of augmented reality," *Foundations and Trends in Human-Computer Interaction*, Vol.8, No.2, pp.73-272, 2015.

[3] P.Cipresso, I.A.C.Giglioli, M.A.Raya and G.Riva, "The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature," *Frontiers in Psychology*, Vol.9, pp.2086, 2018.

[4] S.Mitra and T.Acharya, "Gesture recognition: A survey," TSMCC, Vol.37, No.3, pp.311-324, 2007.

[5] M.Oudah, A.Al-Naji and J.Chahl, "Hand gesture recognition based on computer vision: A review of techniques," *Journal of Imaging*, Vol.6, No.8, pp.73, 2020.

[6] J.Suarez and R.R.Murphy, "Hand gesture recognition with depth images: A review," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012.

[7] M.Ryoo, B.Rothrock, C.Fleming and H.J.Yang, "Privacy-preserving human activity recognition from extreme low resolution," in Proceedings of the *AAAI Conference on Artificial Intelligence*, Vol.31, No.1, 2017.

[8] Y.Alipouri, A.Bokharaeian and E.Monfared, "Construction labor activity recognition via an IMU-implemented wristband," *Journal of construction engineering and management*, Vol.151, No.11, 2025.

[9] S.Mekruksavanich, W.Phaphan and A.Jitpattanakul, "Sensor-based hand gesture recognition using one-dimensional deep convolutional and residual bidirectional gated recurrent unit neural network," *Lobachevskii J Math*, Vol.46, No.1, pp.464-480, 2025.

[10] DA.Dahiya, D.Wadhwa, R.Katti and L.G.Occhipinti, "Efficient hand gesture recognition using artificial intelligence and IMU-based wearable device," *LSENS*, Vol.8, No.12, pp.1-4, 2024.

[11] C.Amma, M.Georgi and T.Schultz, "Airwriting: Hands-free mobile text input by spotting and continuous recognition of 3d-space handwriting with inertial sensors," in *International Symposium on Wearable Computers (ISWC)*, pp.52-59, 2012.

[12] J.Snell, K.Swersky and R.Zemel, "Prototypical Network for few-shot learning," *Advances in neural information processing systems*, Vol.30, 2017.

[13] G.Koch, R.Zemel and R.Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, Vol.2, No.1, 2015.

[14] M.Mizuno and T.Hasegawa, "Deep metric learning for sensor-based human activity recognition," in Proceedings of the *2019 7th International Conference on Information Technology: IoT and Smart City*, pp.18-23, 2019.

[15] C.Finn, P.Abbeel and S.Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in Proceedings of the *34th International Conference on Machine Learning*, pp.1126-1135, 2017.

[16] S.Bai, J.Z.Kolter and V.Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint, 2018.

[17] L.Liu, W.Li and H.Jia, "Method of Time Series Similarity Measurement Based on Dynamic Time Warping," *Computers, Materials & Continua*, Vol.57, No.1, pp.97-106, 2018.

[18] S.Siami-Namini, N.Tavakoli and A.S.Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," in *2019 IEEE International Conference on Big Data (Big Data)*, pp.3285-3292, 2019.

홍 동 진(Dong-Jin Hong) [정회원]

- 2014년 2월 : 신라대학교 컴퓨터공학부 (공학사)
- 2016년 2월 : 부산대학교 컴퓨터공학과 (공학석사)
- 2025년 8월 : 부산대학교 정보융합공학과 (공학박사)
- 2025년 7월 ~ 현재 : 국립부경대학교 인공지능연구소 연구원

〈관심분야〉
신경망, 신호처리, 생성 모델, 영상처리, 딥러닝

박 형 욱(Hyeong-Uk Park) [준회원]

- 2025년 2월 : 부경대학교 컴퓨터·인공지능공학부 컴퓨터공학전공 (공학사)

〈관심분야〉
사물인터넷, 정보통신

최 지 웅(Ji-Woong Choi) [준회원]

- 2025년 2월 : 부경대학교 컴퓨터·인공지능공학부 컴퓨터공학전공 (공학사)

〈관심분야〉
사물인터넷, 정보통신

장 원 두(Won-Du Chang) [정회원]

- 2003년 2월 : 부산대학교 정보컴퓨터공학부 (공학사)
- 2005년 2월 : 부산대학교 컴퓨터공학과 (공학석사)
- 2011년 9월 : Aizu Univ.(일) Info. Sys. Dept. (컴퓨터이공학 박사)
- 2011년 8월 ~ 2013년 6월 : Mongolia Int. Univ. IT Dept. 조교수
- 2017년 3월 ~ 2020년 2월 : 동명대학교 전자및의용공학부 조교수
- 2023년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부 조교수/부교수

〈관심분야〉
패턴인식, 신호처리, 딥러닝, 영상처리, 헬스케어