Regular paper | Journal of the KIECS. pp. 397-402, vol. 19, no. 2, Apr. 30. 2024, t. 124, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.2.397

생성형 AI 기술을 적용한 음성 및 모션 인식 기반 양방향 대화형 알고리즘

장대성* · 김종찬**

Two-way Interactive Algorithms Based on Speech and Motion Recognition with Generative AI Technology

Dae-Sung Jang* · Jong-Chan Kim**

요 약

음성 인식과 모션 인식 기술은 다양한 스마트 디바이스에 적용되어 사용되고 있으나, 단순한 명령어 인식 형태로 구성되어 단순 기능으로 사용되고 있다. 인식 데이터에 대한 단순 기능에서 벗어나 다양한 분야에서 학습된 데이터를 기반으로 전문적인 명령어 수행 능력이 요구되고 있다. 현재 세계적으로 경쟁이 이루어지고 있는 생성형 AI를 활용하여 사용자에게 최적의 데이터를 제공하고, 음성 인식과 모션 인식을 통해 상호작용할 수 있는 시스템 플랫폼에 대한 연구가 진행되고 있다. 본 연구를 위해 설계한 주요 기술 프로세스는 음성 및 모션 인식 기능, AI 기술 적용, 양방향 커뮤니케이션 등 기술을 이용한 설계하였다. 본 논문에서는 AI 기술을 적용한 디바이스와 음성인식과 모션 인식 기술을 통해 디바이스와 사용자 간 양방향 커뮤니케이션을 다양한 입력방식에 의해 이루어질 수 있도록 하였다.

ABSTRACT

Speech recognition and motion recognition technologies are applied and used in various smart devices, but they are composed of simple command recognition forms and are used as simple functions. Apart from simple functions for recognition data, professional command execution capabilities are required based on data learned in various fields. Research is being conducted on a system platform that provides optimal data to users using Generative AI, which is currently competing around the world, and can interact through voice recognition and motion recognition. The main technical processes designed for this study were designed using technologies such as voice and motion recognition functions, application of AI technology, and two-way communication. In this paper, two-way communication between a device and a user can be achieved by various input methods through voice recognition and motion recognition technology applied with AI technology.

키워드

AI(Artificial Intelligence), Motion, Recognition, Voice, Interactive
인공지능, 모션, 인식, 음성, 대화형

* 순천대학교 컴퓨터공학과 연구원(Email)
** 교신저자 : 순천대학교 컴퓨터공학과
· 접 수 일 : 2024. 02. 29
· 수정완료일 : 2024. 03. 21
· 게재확정일 : 2024. 04. 12
· Received : Feb. 29, 2024, Revised : Mar. 21, 2024, Accepted : Feb. 12, 2024
· Corresponding Author : Jong-Chan Kim
Dept. Sunchon National University,
Email : seaghost@scnu.ac.kr

Ⅰ. 서 론

4차 산업혁명은 단순 디바이스와 시스템 간의 연결과 스마트화하는 목적을 기반으로 다양한 범주까지 영향을 끼치고 있다. 유전자 염기서열분석, 나노기술, 재생가능에너지, 양자 컴퓨팅까지 다양한 기술 분야에서 연구를 통한 유의미한 발전을 이루고 있으며, 이 모든 기술이 융합하여 물리적 영역 디지털 영역 생물 영역이 상호교류하는 것에 집중해야 한다.

2024년 IT 기술 트렌드에는 기존 데이터를 기반으로 유저의 특정 요구에 따라 새롭고 독창적인 콘텐츠를 생성하는 생성형 AI, 보안 위협에 대응하는 기술로, 데이터와 시스템을 보호하는 사이버 보안, 명 과학과 기술을 융합하여 의료, 환경, 식품 등 다양한 분야에서 활용되는 바이오 기술, 데이터 저장 및 처리를 클라우드 환경에서 수행하는 기술 클라우드, 인공지능과 로봇 기술을 결합하여 생산성을 향상시키는 기술의 로봇 자동화, 스마트 팩토리 등의 기술이 주목받고 있다[1].

본 논문에서는 음성 및 모션 인식 기술에 생성형 AI 기술을 적용하여 유저와 시스템 간의 음성 및 모션 인식 데이터 등을 통해 이루어지는 양방향 커뮤니케이션이 가능한 시스템 플랫폼에 대한 연구로 유저와 커뮤니케이션 중에 발생되는 다양한 이벤트 데이터를 가지고 최적화된 데이터를 생산과 학습을 위한 연구를 하는데 목적을 가지고 있다. 이러한 시스템을 구현하기 위해 Open API를 이용한 대화형 시스템 플랫폼 설계와 모션과 음성 인식 등 측정되는 데이터를 분석 및 가공하는 알고리즘을 제안했다.

Ⅱ. 선행 연구

인공지능(AI: Artificial Intelligence), 생성형 AI, 모션 및 음성 인식 기술에 대해 살펴보고, 본 논문을 진행하기 위해 구성한 H/W 시스템 구성과 Open API를 이용한 양방향 커뮤니케이션 알고리즘 등의 현재 기술 시장 상황 및 전망 등을 분석했다[2].

2.1 생성형 AI

인공지능은 컴퓨터 과학의 한 분야로, 인간의 행위 형태 중 학습, 추론, 지각 능력 등을 인공적으로 구현하려는 기술이다. 이는 정보공학 분야에서도 중요한 인프라 기술이며, 인간을 포함한 동물이 갖고 있는 지능(Natural Intelligence)와는 다른 개념이다. 인간의 지능을 모방한 기능을 갖춘 컴퓨터 시스템으로, 인간의 지능을 기계 등에 인공적으로 구현한 것이다[3].

생성형 AI(Generative AI)는 새로운 콘텐츠와 아이디어를 생산 및 창작할 수 있는 컴퓨터 모델을 중심으로 하는 인공지능의 한 분야이다. 대화, 이야기, 이미지, 동영상, 음악 등 프롬프트 입력을 통해 요구에 맞는 콘텐츠를 다양한 형태로 생산할 수 있다.

이 기술은 이미지 인식, NLP: Natural Language Processing(자연어 처리), 번역과 같이 새로운 컴퓨팅 작업에서 인간 지능을 모방하려고 시도한다. 생성형 AI는 인공지능의 다음 단계로, 인간 언어, 프로그래밍 언어, 예술, 화학, 생물학 또는 복잡한 주제를 학습하도록 AI를 훈련 시킬 수 있다. 이를 통해 챗봇, 미디어 생성, 제품 개발과 설계 등 다양한 목적을 위해 생성형 AI를 사용할 수 있다[4].

생성형 AI를 사용한 제품과 적용 사례를 살펴보면, OpenAI에서 개발한 챗GPT (ChatGPT) 챗봇 서비스로, 질문에 답변하거나 글 작성, 코드 생성 등 다양한 작업을 수행할 수 있습니다. 무료 버전도 제공되며, 유저 정의도 가능하며, 구글이 개발한 바드(Bard)는 대화형 인공지능으로, 창의성과 자연스러움에 강점을 두고 있습니다. 음유시인이라는 뜻으로, 예술과 문학 분야에서 활용될 수 있다. 그리고 메타에서 발표한 대화형 인공지능 라마(LLaMA)는 다양한 언어 데이터를 기반으로 하는데 성능이 뛰어나며, 넓은 범위의 전문적인 분야에서 활용될 것으로 기대된다[5].

우리나라에서도 고도화된 생성형 AI 개발하기 위해 많은 연구 진행되고 있는 네이버에서 개발한 하이퍼클로바X(HyperClovaX)는 최대규모 인공지능으로, 이미지 설명, 기사 작성, 데이터 분석 등 다양한 작업에 활용된다.

2.2 멀티 모달(Multi Modal)

멀티 모달(Multi Modal)은 다양한 형태와 의미로 컴퓨터와 대화하는 환경을 뜻한다[6].

'모달'은 모달리티(modality)를 의미하는데 '모달리티란 인터랙션 과정에서 사용되는 의사소통 채널'을 말한다[7].

멀티 모달 인터페이스는 전통적 텍스트 외에 음성, 제스처, 시선, 표정, 생체신호 등 여러 입력 방식을 융합하여 인간과 컴퓨터 사이에 자연스러운 의사소통이 가능한 유저 친화형(user-friendly) 기술로서 과거의 기계 중심 입력에서 현재는 휴먼 중심의 자연스러운 입력 방식으로 변화하고 있다. 멀티 모달 인터페이스 구성도는 그림 1과 같다[8].

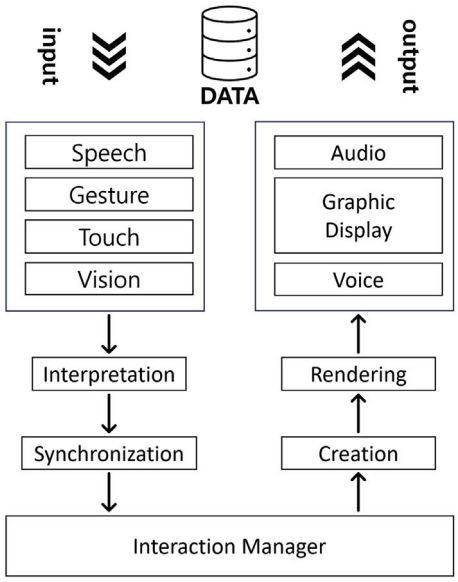

그림 1. 멀티 모달 인터페이스 구성도
Fig. 1 Multi-modal interface configuration diagram

2.3 음성 및 모션 인식

음성 인식 기술은 사람의 음성을 컴퓨터가 이해하고 처리할 수 있도록 하는 기술이다. 이 기술은 음성을 스캔하고 필요한 음성과 일치시켜 사람의 음성 생체 인식을 평가한다. 음성 명령을 이해하고 처리하는데 사용되며, 다양한 응용 분야에서 활용된다.

모션 인식 기술은 유저의 동작을 감지하고 분석하는 기술이다. 카메라나 센서를 통해 유저의 움직임을 인식하고, 이를 데이터로 변환하여 처리한다[9].

모션 제스처, 포즈, 움직임 등을 인식하여 다양한 응용 분야에서 활용된다.

이러한 기술들은 음성 및 동작을 통해 유저와 상호작용하고, 다양한 분야에서 편의성을 제공하며, 보안, 의료, 게임, 로봇, 가상 현실 등 많은 분야에서 기술 고도화와 융합 기술 개발을 위해 연구되고 적용되고 있다[10].

음성 및 모션 인식 기술은 다양한 분야에서 활용되고 있다. 몇 가지 주목할 만한 분야를 살펴보면, 자율주행 시스템, 번호판 인식, 내비게이션 등에서 음성 및 모션 인식 기술이 사용되는 교통 분야, 공장 자동화, 로봇 제어 기술의 제조 분야, 학습 분석, 학습 솔루션 제공 등의 학업 분야, 문학, 미술, 음악 등에서 활용되어 지는 예술 분야, 의료, 법, 언론 등 전문 지식을 필요로 하는 전문 분야에서도 음성 및 모션 인식 기술이 사용된다[11].

III. 생성형 AI 기술을 적용한 음성 및 모션 인식 기반 양방향 알고리즘

'생성형 AI 기술 적용 음성 및 모션 인식 기반 양방향 대화형 시스템 플랫폼 설계'는 멀티 모달 인터페이스 개념을 적용한 연구로 음성 및 모션 인식된 데이터를 해석하고 변환한 데이터를 동기화하여 인터랙션 매니저를 통해 동기화된 데이터 처리를 통해 유저에게 제공할 데이터를 생성한다. 생성된 데이터를 출력 형태에 따라 렌더링을 진행하여 가공된 데이터를 유저가 원하는 출력 형태로 제공하는 형태의 시스템 플랫폼을 제안하려고 한다. 본 논문에서 제안하는 양방향 대화형 시스템 플랫폼 구성도는 그림2와 같다.

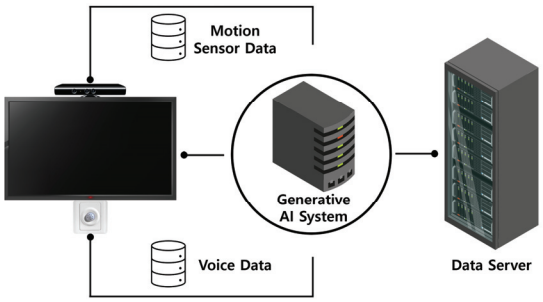

그림 2. 양방향 대화형 시스템 플랫폼
Fig. 2 Interactive interactive systems platform

시스템 설계에 대한 구성으로 생성형 AI 시스템의 운영과 음성 및 모션 센서에서 측정되는 이벤트 데이터를 해석 및 변환할 중앙 관리 하드웨어 디바이스,

AI 데이터 학습과 데이터 변환 생성 데이터 저장 및 관리를 위한 데이터 서버, 유저와의 원활한 커뮤니케이션을 위한 시각화 데이터를 출력할 디스플레이 등으로 구성된다.

유저가 요구하는 최적의 데이터를 추출하기 위해 입력되는 음성 및 모션 이벤트 데이터를 해석하여 추출된 데이터를 생성형 AI 언어 모델과 동기화 후 학습된 데이터에서 유저가 요구하는 데이터를 추출하고 최적화된 데이터를 제공하기 위한 검증 작업을 진행한다. 알고리즘을 통해 검증된 데이터를 최종적으로 유저에게 제공하기 위해 데이터를 생성하고 별도의 출력 형태에 맞는 가공 과정을 통해 유저에게 데이터를 제공한다.

3.1 양방향 대화형 알고리즘

'생성형 AI 기술 적용 음성 및 모션 인식 기반 양방향 대화형 알고리즘'의 설계 구성은 생성형 AI 시스템을 운영할 PC 하드웨어를 중심으로 이벤트 데이터를 입력받을 센서, 데이터의 저장과 AI 학습용 데이터를 저장 및 관리할 스토리지 서버, 유저에게 시각화된 출력 데이터를 제공할 디스플레이 등으로 구성된다.

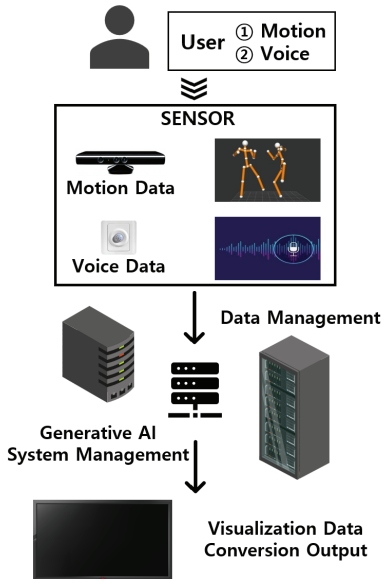

그림 3. 시스템 플랫폼 운영 프로세스
Fig. 3 System Platform Operations Process

그림3은 본 논문에서 제안하는 시스템 플랫폼 운영 프로세스이다. 양방향 대화형 시스템 플랫폼은 유저가 입력하는 음성 및 모션 데이터를 생성형 AI를 운영하는 관리 시스템에 전송하면, 관리 시스템은 해당 데이터를 데이터 관리 서버에 저장과 동시에 학습 알고리즘과 데이터 추출 알고리즘을 통하여 최적화된 데이터를 생성한다. 그리고 생성된 데이터를 텍스트, 이미지, 모션 등의 시각화 된 형태로 렌더링을 거친 후 유저에게 제공한다.

3.2 생성형 AI 챗봇 시스템 설계

본 논문에서 설계하는 시스템 플랫폼을 제작하기위해 Open AI-Chat-3, 메타-라마, 구글의 PaLM2 등의 API 중 메타-라마를 기반으로 설계를 진행하였다.

생성형 AI 시장에는 Open AI, 구글의 모델들이 있으나 이와 같은 모델은 유료 사용모델이며, 메타에서 오픈한 라마의 경우 무료로 배포하여 유료 서비스 모델보다 성능은 조금 떨어지지만 학술 및 상업용 서비스를 개발하는 부분에서 매력적인 선택이라고 생각한다.

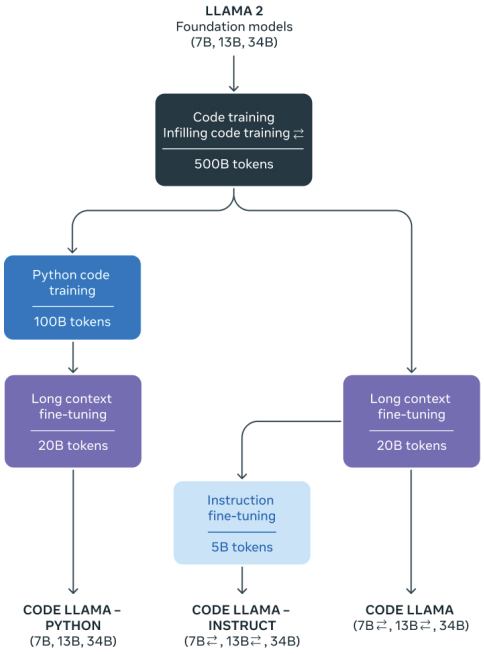

그림 4. 메타 AI - 라마 모델
Fig. 4 Meta AI - Lamar Model

그림4는 메타AI LLAMA 2 파운데이션 모델 구성도이다. 메타에서 배포한 라마는 1조 600억개의 파라미터를 가진 LLM이며, API를 사용할 경우 무료로 제공되지만, 하루 1000번 요청 제한과 텍스트 이미지 크기도 제한된다. 그러나 라마 모델 자체를 다운로드하면 이러한 제한이 사라지나, 라마 모델의 파라미터를 사용하기 위해서는 최소 16GB GPU 메모리가 필요하다. 라마의 모델을 구현하기 위해서는 파이썬과 텐서플로우 등의 라이브러리 설치가 요구된다.

표1은 알고리즘 설계를 진행하기 위해 CPU 14th - i7, RAM 64G, GPU RTX 4070 Ti 사양의 환경을 구성하였다.

표 1. 연구 시스템 환경
Table 1. Research System Environment

H/W	CPU	INTEL 14th-i7
	RAM	64G
	GPU	RTX 4070 Ti
S/W	Programming Language	Python
	Generative AI	LLAMA

3.3 음성 및 모션 인식 데이터 처리 설계

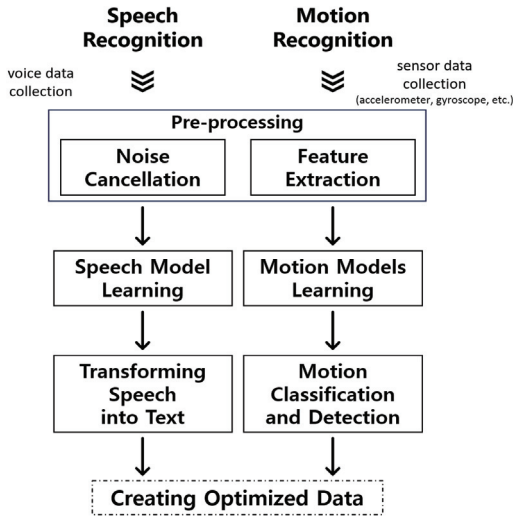

그림 5. 음성 및 모션 인식 기술 알고리즘
Fig. 5 Speech and motion recognition technology algorithms

본 논문에서 설계하는 '생성형 AI 기술 적용 음성 및 모션 인식 기반 양방향 대화형 알고리즘'은 플랫폼에 장착된 각각의 센서에서 수집되는 입력 데이터를 생성형 AI 시스템에서 수행할 수 있는 언어 또는 데이터로 변환하는 프로세스이다. 그림5는 음성 및 모션 인식 데이터 처리 설계에 대한 알고리즘 구성도이다.

음성 및 모션 센서에서 수집된 데이터는 시스템과 연결된 포트를 통해 데이터가 전송이 되어 노이즈 제거, 특징 추출, 음성 및 모션 모델 학습, 음성 데이터 텍스트 변환, 모션 분류 및 데이터 검출 등의 알고리즘을 통해 언어를 생성하여 생성형 AI 시스템에 해당 언어를 전송한다. 수집된 데이터와 변환 및 검출된 데이터를 저장 및 관리를 위한 처리 작업을 수행한다.

Ⅳ. 결 론

본 논문에서는 생성형 AI 기술에 대한 발전과 기술 확보에 대한 중요성이 증가하는 환경에 맞춰 연구를 기획 및 설계를 진행하였다. 아이디어 스케치로 멀티 모달의 개념을 시작으로 다양한 기술의 융합을 위해 노력하였으며, 음성 및 모션 인식 기술에 생성형 AI 기술을 융합하여 유저가 요구하는 데이터를 제공하는 시스템 프로세스 형태로 본 논문 '생성형 AI 기술 적용 음성 및 모션 인식 기반 양방향 대화형 시스템 플랫폼' 설계를 제안하였다.

이를 통해 기존 음성 및 모션 인식 기술에 생성형 AI 기술을 적용하여 융합된 기술을 통해 다양한 콘텐츠를 생산할 수 있을 것으로 생각되며, 개발이 완료될 경우 각각의 분야에 커스터마이징된 형태로 세분화하여 적용할 수 있을 것이다. 본 연구에서는 전통적 텍스트 외에 음성, 제스처, 시선, 표정, 생체신호 등 여러 입력 방식을 통한 유저와 컴퓨터의 자연스러운 의사소통을 통해 정확하고 최적화된 데이터를 제공하는 것이 목표이다.

향후 메타의 LLAMA 오픈 생성형 AI 모델을 적용한 시스템을 구성하고 음성 및 모션 센서를 활용한 실제 구현으로 진행할 예정이다.

References

[1] Editor-in-chief of Convergence Management Review, "The Fourth Industrial Revolution to be opened by Artificial Intelligence," *Convergence Management Review*, vol. 21 No.-2021, pp. 1-1.

[2] Seo Ji-hoon, and Ju Ji-hong, "Korean AI learning model based on e-learning for artificial intelligence education," *Artificial intelligence research paper*. vol. 3, No. 2-2022, pp. 14-22.

[3] Hyun Jung-woo, Kim Chul-hoo, and Gil Hyung-bae. "An analysis of trends and implications of intelligent robots and Generative AI," *mechanical technology policy magazine*, vol.3m no. 114 2024. pp. 1-35.

[4] Lee Hyo-seop, Shim Ho-seok. "A Study on the Design of a Metaverse Based Knowledge Management Model Using Generative AI," *Journal of the Korean Society for Industrial Technology Convergence*. vol. 2023. no. 2, pp. 21-32.

[5] Park Young-hyun. "A study on the use and cases of the Generative AI model (GAN) in international management following the development of AI technology," *a study on trade management in Korea*. vol. 2. no. 33 2024. pp. 39-54.

[6] Yoon Yeo-chan. " trends in multimodal Generative AI technology." *Journal of Information Science*, vol. 42. no. 1 2024. pp. 42-47.

[7] Moon Yi-sun, and Kim Hyung-seok. " biological signal and voice data - based multimodal emotion recognition," *a collection of papers at a national academic conference of the society of control robot systems*. vol. 2023. no. 6 2023, pp. 635-636.

[8] Lee Soo-min, Lee Mi-ran, Wei Qun, and Park Hee-jun. " a research on the development of a heart sound analysis wearable device capable of measuring multi-modal bio-signals," *Journal of the Multimedia Society*. vol. 25, no. 9 2022, pp. 1251-1256.

[9] Park Yang-woo, Lee Woo-jae, Kim Min-seop, Jung Myung-jin, Kang Min-jae, and Yeom Sang-ho. "Implementation of Input Devices Using Motion Recognition and Voice Recognition," *Korean Society for Computer and Information*, vol. 31. No. 1 2023. pp. 287-288.

[10] Park Ki-chang, Seo Sung-chae, Jung Seung-moon, Kang Im-cheol, and Kim Byung-ki. "Designing a Gesture Interface Model for GUI Application Control," *Korean Content Society Journal,* vol. 13. No. 1 2013. pp. 55-63.

[11] Park Yang-woo, Lee Woo-jae, Kim Min-seop, Jung Myung-jin, Kang Min-jae, and Yeom Sang-ho, "Implementation of Input Devices Using Motion Recognition and Voice Recognition," *academic presentation collection of the Korean Computer Information Society*. vol. 31 Nno. 1 2023, pp. 287-288.

저자 소개

장대성(Dae-Sung Jang)

2018년 우석대학교 정보보안학과 졸업(공학사)
2020년 순천대학교 대학원 컴퓨터공학과 (공학석사)
2023년 순천대학교 대학원 컴퓨터공학과 (공학박사 수료)
※ 관심분야 : 콘텐츠, 컴퓨터그래픽스, 임베디스시스템, 3D Printer, IoT

김종찬(Jong-Chan Kim)

2000년 순천대학교 컴퓨터과학과 졸업(이학사)
2002년 순천대학교 대학원 컴퓨터과학과 졸업(이학석사)
2007년 순천대학교 대학원 컴퓨터과학과 졸업(이학박사)
2013년 서울대학교 자동화 시스템 연구소 선임연구원
2021년 ~현재 순천대학교 컴퓨터공학과 조교수
※ 관심분야 : 영상처리, HCI, 콘텐츠, 컴퓨터그래픽스, 기계학습, 객체추적, 컴퓨터비전, 데이터 분석 및 예측