Journal of KIIT. Vol. 23, No. 7, pp. 167-170, Jul. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.7.167

Watch My Wrist: 증강현실 환경에서 카메라 시야를 벗어난 손동작 인식을 위한 손목 근육 움직임 기반 예측

이민재*[1], 배지호*[2], 이수원*[3]

Watch My Wrist: Wrist Muscle Movement-based Hand Gesture Recognition for Out-of-Sight in Augmented Reality

Minjae Lee*[1], Jiho Bae*[2], and Suwon Lee*[3]

요 약

증강현실(AR, Augmented Reality) 기술의 발전으로 사용자들은 컨트롤러와 손을 이용하여 AR 장비와 상호작용하며 다양한 콘텐츠를 소비하고 있다. 그러나 손을 이용한 상호작용 과정에서 손동작이 AR 장비의 시야 안에 들어와야 한다는 한계가 존재한다. 본 논문에서는 이를 해결하기 위해 손목을 비롯한 팔의 일부 정보만을 활용하여 손동작을 예측할 방법과 이에 적합한 데이터셋의 구성을 제안하며, 실험을 통해 그 가능성을 검증한다. 이를 통해 AR 장비로 인한 동작 제약을 줄이고, 사용자가 더욱 자연스럽고 몰입감 있는 증강현실 경험을 할 수 있도록 하는 것을 목표로 한다.

Abstract

With the advancement of Augmented Reality (AR) technology, users interact with devices using controllers and hand gestures. However, these gestures must remain within the device's field of view. To overcome this, this paper proposes a method to predict hand gestures using partial arm information, including the wrist. A suitable dataset is introduced, and the method is validated through experiments. This paper aims to reduce movement constraints and enhance the AR experience.

Keywords

augmented reality, gesture recognition, muscle movement, contrastive learning, human-computer interaction

Ⅰ. 서 론

증강현실(AR, Augmented Reality) 기술은 발전을 거듭[1]하며 사용자와의 상호작용 방식도 다양해졌다. 대표적으로 컨트롤러 기반 조작과 손 인식 기반 인터페이스가 있다. 컨트롤러는 정밀한 조작이 가능하지만 물리적 장치가 필요하며, 손 인식 방식은 직관적이지만 손이 AR 기기의 카메라 시야(FoV, Field of View) 내에 있어야만 동작 인식이 가능하다는 한계가 있다. 이는 사용자의 자유로운 움직임을 제한하고 불편한 자세를 유발해 몰입도를 저하시킨다. 특히 AR 작업 환경에서는 자연스러운 동작 보장이

* 경상국립대학교 컴퓨터공학과(*[3] 교신저자)
- ORCID[1]: https://orcid.org/0009-0005-2796-3207
- ORCID[2]: https://orcid.org/0009-0004-5942-4972
- ORCID[3]: https://orcid.org/0000-0003-2603-1385

· Received: Apr. 04, 2025, Revised: Apr. 17, 2025, Accepted: Apr. 20, 2025
· Corresponding Author: Suwon Lee
Department of Computer Science and Engineering, Gyeongsang National University, Republic of Korea
Tel.: +82-55-772-1394, Email: leesuwon@gnu.ac.kr

중요하지만, 기존 손 인식 기술은 이를 충분히 고려하지 못하고 있다. 하지만 그림 1과 같이 손목이나 팔목 움직임은 손동작과 밀접하게 연관되어 있다는 것을 알 수 있는데, 이 사실을 활용한다면 손 전체가 보이지 않더라도 주변 신체 부위만을 활용해 손동작을 예측할 수 있을 것이다. 즉, 손이 적절한 활동 범위 안에 있다면 기존의 문제를 해결하여 자연스럽고 자유롭게 AR 상호작용을 실현할 수 있을 것이다.

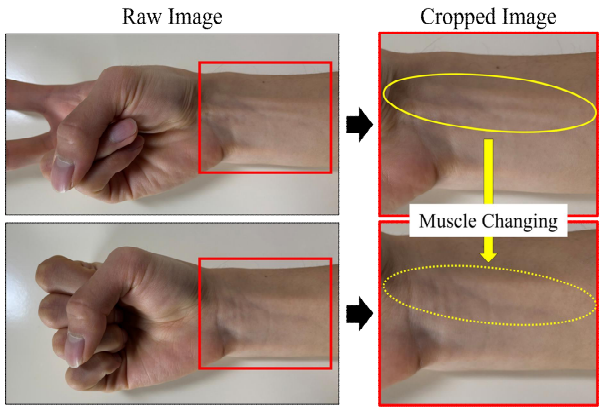

그림 1. 손가락의 변화에 따른 손목 근육의 변화 및 학습/추론을 위한 원본/크롭 이미지
Fig. 1. Changes in wrist muscles according to finger movements and raw/cropped images for training and inference

II. 관련 연구

2.1 카메라 기반의 손동작 인식

증강현실뿐만이 아닌 여러 컴퓨팅 디바이스에서 사용자의 손동작을 인식하기 위해선 손동작을 포착할 수 있는 장치가 필요하며 웹캠 등의 카메라 장비가 대표적이다. 최근에는 딥러닝 기반의 방법[2][3]이 주류를 이루고 있으며 우수한 성능을 보인다. 다만, 이 방법은 손가락을 포함하는 손 전체가 카메라 안에 정확히 들어와 있다는 것을 전제로 하는 한계에 부딪힌다.

2.2 비-카메라 기반의 손동작 인식

카메라를 사용하지 않고 손동작 인식하기 위한 연구로 적외선[4], 근적외선[5], WiFi 신호[6], 근전도 검사(EMG, Electromyography)[7] 등의 여러 방식이 사용되고 있다. 2.1절의 카메라 기반 방법과 유사한 성능이 보고되나 적외선, 근적외선, WiFi 등의 신호를 생성하고 인식하는 추가 장비가 필요하다. EMG 방식의 경우 피부 접촉형 센서 또는 피부에 꽂는 바늘 등 현실적으로 AR 시스템에서 사용자에게 손동작을 인식하기 위해 사용하기에는 불편하거나 위험하다. 따라서 AR 디바이스에서 기본적으로 제공하는 카메라 기반의 방식을 활용하는 것은 비용과 신체 보호를 위해 유리하다.

III. 데이터 구성

그림 2는 다양한 시간대와 조명 환경에서, 여러 장소에서 촬영된 전체 데이터 중 대표적인 예시 데이터 8개를 시각화한 것이다. 총 10명의 참여자가 왼손과 오른손을 사용하여 브이, 오케이, 손바닥, 주먹 동작을 반복적으로 바꿔가며 수행하게 하였으며, 그림 1과 같이 원본 이미지에서 손목 부분이 포함된 영역을 잘라내어 추후 딥러닝 모델의 추론 데이터로 구성하였다. 전체 데이터에 대한 자세한 구성은 표 1에서 확인할 수 있으며 성별, 손동작, 왼손/오른손에 따른 데이터를 균일하게 구성하였다. 학습과 추론을 위해 전체 데이터를 7:3으로 분리하여 준비하였다.

그림 2. 다양한 환경, 성별, 밝기, 장소에서 촬영한 손동작 이미지 집합: 왼손/오른손, 오케이/브이/주먹/손바닥으로 구분
Fig. 2. Collection of hand gesture images captured in various environments, genders, lighting conditions, and locations: Categorized by left/right hand and gestures, oK, vee, fist, and palm

표 1. 성별, 손동작, 왼손/오른손 구분에 따른 데이터 구성 개수 및 비율. 밑줄은 기준 비율을 의미함
Table 1. Data distribution by gender, hand gesture, and left/right hand. The underline indicates the reference ratio

Category		#Frame	#Valid frame	Ratio
Gender	M	16,228	10,744	1.00
	F	18,408	12,159	1.13
Hand gestures	Fist	-	6,103	1.00
	Palm	-	5,912	0.97
	Vee	-	5,537	0.91
	Okay	-	5,351	0.88
Handedness	L	17,335	11,464	1.00
	R	17,301	11,439	0.99

Ⅳ. 방 법

본 연구에서는 손목과 그 주변의 특징을 활용하여 손동작을 예측하는 태스크와 이를 해결할 수 있는 한 방식을 제안한다. 고성능 인식기(ViT, Vision Transformer)와 비교적 가벼운 실시간 예측기(MobileNet)로 구성된 학습 구조를 적용하였다. 특히, AR 시스템에서는 실시간 처리가 필수적이므로, 계산량을 최소화하기 위해 경량화된 모델을 채택하였다. 하지만 실시간 예측기는 근육의 섬세한 움직임을 포착하기 어려워 대조 학습을 통해 교사 모델로부터 지식을 전달받도록 설계하였다.

입력 데이터는 전체 손이 보이는 이미지와 손목 및 팔목만 보이는 크롭 이미지로 구성된다. 기존 카메라 기반의 손동작 인식 모델이 손 전체를 기반으로 특징을 추출하는 반면, 본 연구에서는 손목 주변만의 정보를 활용하여 손동작을 유추한다. 그림 3에서 고성능 인식기인 교사 모델이 전체 이미지를 입력으로 받아 강력한 특징 표현을 학습하며, 실시간 예측기인 학생 모델은 크롭 이미지를 입력으로 받아 경량화된 형태로 특징을 학습한다. 이때, 두 모델의 출력 벡터를 비교하기 위해 우선 투영(Π) 과정을 거쳐 같은 크기로 변환하고, 학생 모델이 교사 모델의 특징 공간을 닮도록 학습하게 된다.

손실 함수는 손동작 손실(L_G), 왼손/오른손 손실(L_H), 그리고 식 (1)에 따른 대조 손실(L_{CL})[8]의 선형 결합으로 정의되는데, 학생 모델이 손목만 보고 추출한 특징 벡터(z_S^i)를 교사 모델이 손 전체를 보고 추출한 특징 벡터(z_T^i)와 유사해지도록, 즉 교사의 판단을 모방하도록 학습시키게 유도한다.

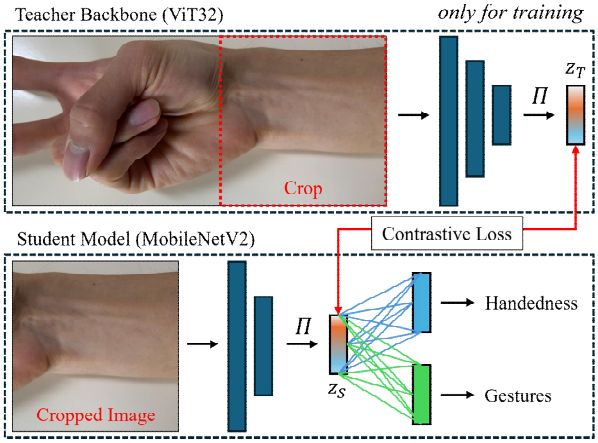

그림 3. 손동작 및 손 방향 예측을 위한 모델 구조
Fig. 3. Model architecture for hand gesture and handedness recognition

$$L_{CL} = -\sum_i \log \frac{\exp(z_T^i \cdot z_S^i/\tau)}{\sum_j \exp(z_T^i \cdot z_S^j/\tau)} \quad (1)$$

Ⅴ. 실험 및 논의

본 연구에서는 2.1의 기존 연구와 동일하게, 손 전체 이미지를 입력으로 사용하는 학생 모델과 손목 주변만을 크롭한 이미지를 사용하는 학생 모델 간의 성능 차이를 비교하였다. 손방향 분류 과제에서는 팔목의 각도나 자세 정보가 주요 결정 요인으로 작용하여, 입력 이미지 범위에 관계없이 F1-Score 및 Accuracy 모두 0.9 이상의 높은 성능을 나타내었다. 반면, 제스처 분류 과제에서는 손 모양 정보의 영향이 크기 때문에 손목 크롭 이미지를 사용하는 경우 클래스 평균 Accuracy는 36.5%, 클래스 평균 F1-Score는 36.7% 감소하는 것으로 나타났다. 이후 대조 학습을 통해 교사 모델로부터 지식을 전달받은 학생 모델은 각각 5.0%, 4.7%의 성능 개선을 보였으나, 손목 주변 정보만으로는 실용화 수준의 성능 확보에는 한계가 있었다. 추후, 다양한 환경에서의 데이터 수집을 진행하고, 손목 주변의 정보를 강조하는 등 기술적 보완이 필요할 것이다.

표 2. 사용 데이터 및 방법에 따른 모델 성능 비교. 원본, 크롭 이미지(교사 없음), 크롭 이미지(교사 있음) 간 성능 비교
Table 2. Model performance comparison by data and method. Comparison of original, cropped (without teacher), and cropped (with teacher) images

Method \ Metric	Gesture								Handedness			
	F1-Score				Accuracy				F1-Score		Accuracy	
	Fist	Okay	Palm	Vee	Fist	Okay	Palm	Vee	Left	Right	Left	Right
Student(Raw)	.9927	.9753	.9909	.9803	.9864	.9619	.9954	.9950	1.0	1.0	1.0	1.0
Student(Cropped) w/o teacher	.7853	.4341	.6492	.6012	.7074	.3636	.7844	.6209	.9761	.9861	.9650	.9927
Student(Cropped) w teacher	.8788	.5109	.6885	.5929	.8542	.5571	.7627	.4905	.9315	.9603	.9204	.9671

Ⅵ. 결 론

본 연구는 증강현실 환경에서 시야 밖 손동작을 인식하기 위해 손목 주변 정보만으로도 제스쳐와 손 방향 분류가 가능한 모델을 제안하였다. 이를 위해 직접 데이터셋을 구축하고 모델 구조를 제시하며 실험을 진행하였으나, 성능적 한계가 존재하였다. 실질적 활용을 위해 참여자 수를 확대하고 손목 주변 정보 강조 방법 등을 제시하였다. 해당 연구는 추후 연구를 통해 정확도와 실시간성을 모두 충족하는 손동작 인식 기법으로 발전될 것이다.

References

[1] M. Lee, J. Bae, U. Kim, S. M. Choi, and S. Lee, "Shrunken Reality: Augmenting Real-World Contexts in Real-Time on Realistic Miniature Dioramas", SIGGRAPH Asia 2024 Technical Communications, Tokyo, Japan, No. 26, pp. 1-4, Dec. 2024. http://doi.org/10.1145/3681758.3697983.

[2] A. Mujahid, et al., "Real-time hand gesture recognition based on deep learning YOLOv3 model", Applied Sciences, Vol. 11, No. 9, pp. 4164, May 2021. https://doi.org/10.3390/app11094164.

[3] O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition", Neural Computing and Applications, Vol. 28, No. 12, pp. 3941-3951, Apr. 2016. https://doi.org/10.1007/s00521-016-2294-8.

[4] L. Yu, H. Abuella, M. Z. Islam, J. F. O'Hara, C. Crick, and S. Ekin, "Gesture recognition using reflected visible and infrared lightwave signals", IEEE Transactions on Human-Machine Systems, Vol. 51, No. 1, pp. 44-55, Feb. 2021. https://doi.org/10.1109/THMS.2020.3043302.

[5] Q. Zhang, et al., "Airfinger: Micro finger gesture recognition via NIR light sensing for smart devices", 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS), Singapore, Singapore, pp. 552-562, Nov. 2020. https://doi.org/10.1109/ICDCS47774.2020.00073.

[6] S. Tan and J. Yang, "WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition", Proce. of the 17th ACM international symposium on mobile ad hoc networking and computing, Paderborn, Germany, pp. 201-210, Jul. 2016. https://doi.org/10.1145/2942358.2942393.

[7] K. H. Lee, J. Y. Min, and S. Byun, "Electromyogram-based classification of hand and finger gestures using artificial neural networks", Sensors, Vol. 22, No. 1, pp. 225, Dec. 2021. https://doi.org/10.3390/s22010225.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", International conference on machine learning, Vienna, Austria, pp. 1597-1607, Jul. 2020. https://doi.org/10.48550/arXiv.2002.05709.