HOLOGESTURE: A MULTIMODAL DATASET FOR HAND GESTURE RECOGNITION ROBUST TO HAND TEXTURES ON HEAD-MOUNTED MIXED-REALITY DEVICES

Jeongwoo Park *Je Hyeong Hong*⋆

Department of Electronic Engineering, Hanyang University, Seoul, South Korea

ABSTRACT

While the recent development of high performance mixed-reality (MR) devices is enabling its use in medical and industrial domains, this requires hand gesture recognition to be robust to different textures inflicted by gloves often worn for hygiene and safety purposes. Unfortunately, most existing hand gesture datasets are not captured using recent commercial MR devices, and none addresses the issue of wearing gloves in gesture recognition. We aim to fill these gaps by introducing a new dataset called HoloGesture, which comprises gesture clips acquired with and without latex gloves using Microsoft HoloLens 2. To leverage the multimodal nature of the latest MR device, we go beyond simply stacking RGB and depth frames and provide spatially aligned depth and RGB images. Experimental results show that i) incorporating gloves for training enhances robustness of gesture recognition to different hand textures and ii) spatial alignment of RGB and depth images enhances the recognition accuracy. Our code and dataset can be found at https://github.com/hellojpark/hologesture.

***Index Terms*—** hand gesture recognition, mixed-reality, multimodal dataset, gloves, depth

1. INTRODUCTION

Mixed Reality (MR) has been gaining popularity across various industries recently. In the medical field, MR's virtual holograms executed in the real world are of particular interest due to their potential to solve infection issues associated with displays in operating room. Moreover, the widened availability of high performance MR devices is also increasing their popularity in the field of industry and manufacturing.

In line with this growing interest, research on hand gesture recognition based on MR devices is being vigorously pursued. Particularly, with the recent demonstration of high accuracy by deep learning-based hand gesture recognition, the demand for datasets necessary for training has escalated. Most existing hand gesture datasets are third-person perspective data [1, 2, 3, 4, 5], primarily acquired for purposes such as webcam-based gesture recognition. The third-person perspective means the camera is positioned in front of the person performing the gestures. However, this perspective is not well-suited for hand gesture recognition with MR devices. Suitable datasets for MR devices should be ego-centric, where the camera's viewpoint aligns with the person's perspective. Zhang *et al.* [6] proposed a ego-centric dataset under various conditions. However, existing ego-centric datasets rarely involve data captured using commercially available MR devices [6, 7, 8], and none, to the best of our knowledge, considers gloves which are commonly used in a range of public sectors such as medicare and manufacturing industry. For industries where gloves are worn, such as in medical settings, datasets featuring gestures performed with gloves, in addition to bare hands, are necessary to maintain recognition performance currently based on bare hand data.

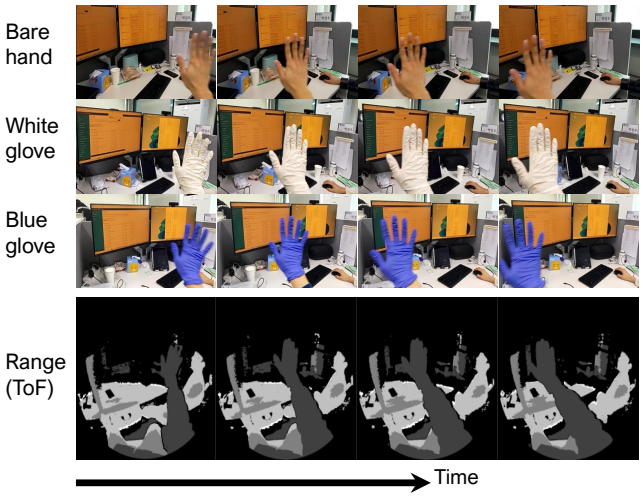


Fig. 1. Some of the gesture data performed with bare hands and while wearing white and blue medical gloves. Images are obtained using the RGB and ToF sensors of Hololens 2.

Additionally, current hand gesture datasets tend to offer multimodal data like RGB and depth. These data are acquired from different sensors, leading to disparities in frame size for each modality or varied hand positions within the images. Otherwise, some sensors may automatically adjust camera pose. Based on these multimodal datasets, prior research has improved hand gesture recognition accuracy by applying various fusion approaches [9, 10, 11]. However, prior research focuses on fusion methods, with a lack of analysis on the impact of spatial alignment data on accuracy. The fact

⋆Corresponding author

Table 1. Comparison of the existing datasets for hand gesture recognition. "CHMD" stands for "Commercially available Head-Mounted Device," representing to denote the use of commercially available MR devices. "View" means perspective of camera when capturing data. HoloGesture is an ego-centric hand gesture dataset acquired using commercially available MR devices and includes three modalities. Moreover, it encompasses gestures performed with bare hands and two latex gloves.

Dataset	RGB	Depth	IMU	View	Gloves	CHMD	# videos	# subjects	# gestures
Cambridge Hand Gesture [1]	O	O	X	3rd	X	X	900	2	9
ChAirGest 2013[2]	O	O	O	3rd	X	X	1,200	10	10
SKIG 2013 [3]	O	O	X	3rd	X	X	1,080	6	10
Jester [4]	O	X	X	3rd	X	X	148,092	1,376	27
nvGesture [5]	O	O	X	3rd	X	X	1,532	20	25
EgoGesture [6]	O	O	X	ego	X	X	24,161	50	83
GGesture [7]	O	X	O	ego	X	O	700	22	10
Interactive Museum[8]	O	X	X	ego	X	X	700	5	7
HoloGesture (ours)	**O**	**O**	**O**	**ego**	**O**	**O**	**4,050**	**10**	**27**

that existing datasets provide only spatially aligned or non-aligned data pairs makes it challenging to analyze the impact of spatial alignment on hand gesture recognition performance. In addition to fusion techniques, demonstrating that the spatial alignment of multimodal data affects recognition accuracy could aid future research in enhancing performance.

To address above limitations of existing datasets simultaneously, we propose a new hand gesture dataset called *HoloGesture*, which comprises 27 different gestures captured from each of various individuals using Microsoft HoloLens 2 [12]. The main contributions of our dataset is summarized below:

- the captured gestures are performed with bare hands and two types of medical latex gloves to better address medical and industrial environments with more strict regulations such as wearing hand gloves, and
- the dataset is captured using one of the latest commercial mixed-reality devices, allowing fusion of multimodal (RGB and depth) information and narrowing the gap between research and practical applications. The dataset additionally includes spatially aligned RGB and depth image-pairs as well their original forms to allow different means of leveraging multimodal data.

Each of above contributions is evaluated in Sec. 4 in which we demonstrate that incorporating gloves for training improves robustness of the hand gesture recogition models to different hand textures while the second contribution of utilizing multimodal information as well as spatial alignment both enhances the gesture recognition accuracy.

2. RELATED WORK

2.1. Datasets

Hand gesture datasets are collections of video that capture hand movements. These datasets can broadly be categorized into third-person perspective and ego-centric datasets.

Third-person perspective datasets [1, 3, 4, 5] are captured with a camera placed in front of the person. [1] comprises 9 classes, with each class consisting of 100 frames captured under 5 illumination conditions. [3] contains 1080 RGB and depth sequences from 6 people. It captured under 3 different backgrounds and 2 illumination conditions with 10 types of hand gestures. [4] is the largest hand gesture dataset acquired through crowd-sourcing via a webcam. It consists of 148,092 videos, capturing 27 classes from 1,376 people. [5] is a multimodal dataset, acquired considering hand gesture recognition in the interior environment of cars. The dataset consists of 1,532 videos, capturing 25 classes from 20 individuals.

Table 2. Descriptions of the 27 classes in HoloGesture

1 Move hand right	10 Call someone	19 Pull hand in
2 Move hand left	11 Open hand	20 Rotate fingers c.c.w.
3 Move hand up	12 Shaking hand	21 Rotate fingers c.w.
4 Move hand down	13 Show index finger	22 Push two fingers away
5 Move two fingers right	14 Show two fingers	23 Close hand two times
6 Move two fingers left	15 Show three fingers	24 Thumbs up
7 Move tow fingers up	16 Push hand up	25 Okay gesture
8 Move two fingers down	17 Push hand down	26 Move the edge of the hand left
9 Click with the index finger	18 Push hand out	27 Close hand one time

Ego-centric datasets are captured from the viewpoint of the person performing the gestures. [6] is captured using a Intel Realsense mounted on the head. The dataset is designed considering various conditions, including indoors, outdoors, stationary, and walking scenarios, resulting in 24,161 videos from 50 individuals across 83 classes. [7] is acquired with the MR device HoloLens 1. To offer diverse backgrounds, the data was captured in front of a green screen, allowing for the synthesis of various backgrounds. It provides RGB data, with 700 videos from 22 individuals across 10 classes.

We focus on hand gesture recognition in medical settings using the MR device which is related to ego-centric. Therefore, third-perspective datasets do not align with the objectives of our research. While existing ego-centric datasets captures considering various conditions, they do not account for the gloves employed in medical settings. Therefore, we provide a dataset that includes gestures performed with bare hands and two medical latex gloves, using HoloLens 2.

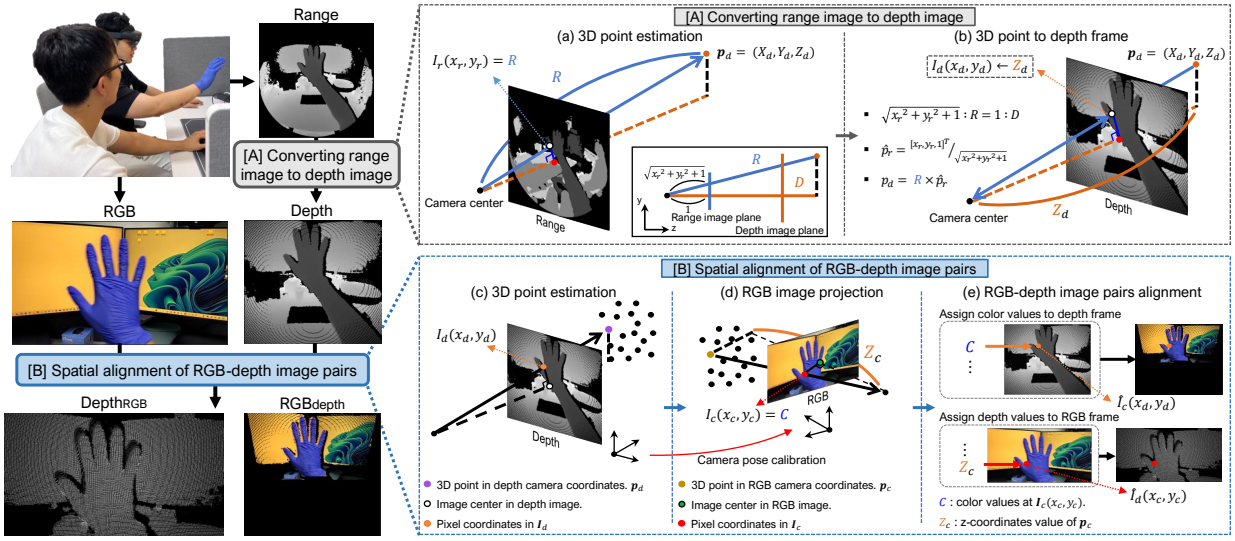


Fig. 2. The process from data acquisition to obtaining multi-domain data through spatially aligning preprocessing. Hololens 2 is used to capture RGB and range images. Then, the procedure labeled as 'Depth frame projection' in the figure is carried out to obtain depth images. Finally, the 'RGB-depth image pairs alignment' process is executed to acquire $\text{Depth}_{\text{RGB}}$ that is depth domain data aligned with the RGB frames and $\text{RGB}_{\text{depth}}$ that is RGB domain data aligned with depth frames. The methods for [A] 'Depth frame projection' and [B] 'RGB-depth image pairs alignment' are detailed in Sec. 3.3.

2.2. Algorithms

With the promising results in deep learning-based hand gesture recognition, there has been an increase in related studies. The tasks can primarily be categorized into two approaches. The first is unimodal training, which utilizes only one type of modality data. The second is multimodal training, which involves the combined use of different types of modality data.

The unimodal training involves extracting temporal and spatial features from the video and can be divided into two approaches: 1) a method processing spatio-temporal information simultaneously. [13, 14], the 3D ConvNets-based model, extract spatio-temporal features for adjacent frames composing a videos. This method operates similarly to standard ConvNets and appears to be a natural approach for video modeling. However, the extra dimension in 3D ConvNets, compared to 2D ConvNets, leads to more parameters and extended training time. 2) a method that first captures spatial features and then extract temporal information from the sequence of spatial features. [6, 15, 10] use 2D ConvNets to first capture the spatial information of frames composing a video, and then extracts temporal feature using models such as LSTM on the sequence of spatial features. This approach benefits from the usage of models pretrained on ImageNet such as VGG16, commonly used in image modeling. Additionally, it allows the use of models like Transformer, which have shown high accuracy in handling sequence data.

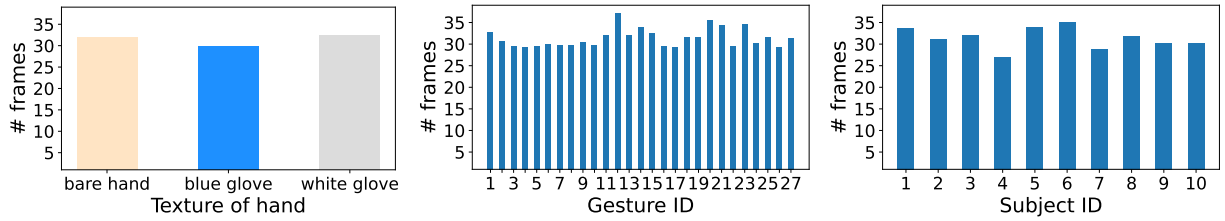
Multimodal training employs each type of modality data as input for separate networks. It aims to enhance accuracy by fusing or sharing the information from the networks corresponding to each modality. [9] used data level fusion to integrate RGB and optical flow data. Optical flow was integrated as an extra channel in the RGB image, serving as the network's input. [10] improved accuracy through late fusion, outperforming unimodal methods. It merges the probability distributions of each modality data obtained from their respective networks. [11] proposed a framework where individual networks learning each modality exchange information during the training process. Instead of relying solely on a single modality, it learns from multiple modalities.

For performance validation of glove and the spatially aligned data of HoloGesture, we have chosen baseline models [10, 13, 16] from the deep learning-based hand gesture recognition models. A detailed discussion about these baseline models will be provided in Sec. 4.1.

3. HOLOGESTURE DATASET

3.1. Dataset collection

For acquiring MR-based hand gesture dataset, we utilized Hololens 2. We access to the RGB and the phase-based Time-of-Flight (ToF) sensor of Hololens 2. These sensors provide us with range and RGB data. The range data, captured from 25 to 30 frame rate(fps), is a circular image in 512×512 size frame, as shown in Fig. 1. However, the circular image format is potentially not optimal for convolutional networks. Consequently, it is necessary to convert the range image to a depth image. Converting a range image to depth image is illustrated in Sec. 3.3. RGB data was acquired at a frame rate ranging from 25 to 30 fps, with a resolution of 760×428 pixels using RGB sensor (see Fig. 1 for a visual illustration). Due to variable frame-rate, we utilized the time information recorded during data acquisition to achieve temporal synchronization between RGB and depth video clips.



(a) Texture wise avg. frame length (b) Gesture wise avg. frame length (c) Subject wise avg. frame length

Fig. 3. Average number of frames comprising videos categorized by hand texture (a), gesture (b), and subject (c).

Data collection involved one person wearing Hololens 2 and performing gestures, while another individual provided guidance on the gestures. This data acquisition process is illustrated in the top left image in Fig. 2. We provide 4050 videos for each modality, with each video representing a single hand gesture. Additionally, the gesture performers executed all gestures with bare hands and while wearing white and blue medical latex gloves. We provide class labels for videos representing a single gesture, along with the frame numbers where the gestures start and end.

3.2. Dataset statistics

A total of 10 individuals contributed data, each performing 27 distinct hand gestures. Every individual repeated each hand gesture five times. Additionally, each individual performed all the aforementioned processes while either barehanded or wearing white and blue medical latex gloves.

To generate the train, validation, and test datasets, we employed a 3:1:1 ratio. Instead of randomly splitting the entire dataset, we ensure that the individuals contributing to each dataset are distinct. This decision was made to prevent potential bias, as gestures from the same individual may exhibit similarities. The age group of the contributors ranges from 20 to 30 years, and we captured gestures from the right hand.

The maximum number of frames in videos representing gestures is 47, while the minimum is 1. The majority of average frame count approach 30, indicating relatively low deviations. Specific numerical details are presented in Fig. 3.

3.3. Data preprocessing

We define $I \in \mathbb{R}^{H\times W\times C}$ as an image frame with H, W and C defined as image height, image width, and channel size respectively. Note $I_r \in \mathbb{R}^{512\times512\times1}$, $I_d \in \mathbb{R}^{320\times288\times1}$ and $I_c \in \mathbb{R}^{760\times428\times3}$ are the images from range, (virtual) depth and RGB frames respectively. Also, (x, y) is the pixel coordinates on the image I, while (X, Y, Z) represent the Cartesian 3D coordinates of the corresponding pixel.

Converting range image to depth image. For generating depth images, we adopt the framework proposed in [12]. Each pixel in the range image I_r comprises the distance (R) between the camera center and the corresponding 2D pixel's 3D location. (see (a) of [A] in Fig. 2). Then, the normalized 3D homogeneous coordinates ($\hat{\mathbf{p}}_r \in \mathbb{R}^3$) is computed for each pixel in the range image I_r. This is multiplied by the range depth R to retrieve 3D Cartesian coordinates of each pixel ($\mathbf{p}_d \in \mathbb{R}^3$). In terms of equations, $\hat{\mathbf{p}}_r := [x_r, y_r, 1]^\top / \sqrt{x_r^2 + y_r^2 + 1}$ and $\mathbf{p}_d := R\ \hat{\mathbf{p}}_r$ where (x_r, y_r) is the pixel coordinates in the range image I_r. Each 3D cartesian point $\mathbf{p}_d := (X_d, Y_d, Z_d)$ is then projected to the virtual depth image plane by using the camera intrinsics matrix for the virtual depth camera ($\mathtt{K}_d \in \mathbb{R}^{3\times3}$) whose default values are provided in [12]. In terms of equation, $[x_d, y_d]^\top := \pi(\mathtt{K}_d\mathbf{p}_d)$ where $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the projection operator such that $\pi([X, Y, Z]^\top) := [X/Z, Y/Z]$. Finally, the depth image I_d is rendered by assigning Z_d at the 2D pixel location (x_d, y_d): $I_d(x_d, y_d) \longleftarrow Z_d$. Above details are also depicted in[A] of Fig. 2.

Spatial alignment of RGB-depth image pairs. For aligning RGB and depth images, we utilize the camera intrinsics and extrinsic information consistently stored in HoloLens 2. Some of the stored data includes the transformation matrix from the RGB camera coordinates to world coordinates as $T_{c2w} \in \mathbb{R}^{4\times4}$ and the transformation matrix from the (virtual) depth camera coordinates to world coordinates as $T_{d2w} \in \mathbb{R}^{4\times4}$, where each transformation matrix $T \in \mathbb{R}^{4\times4}$ has form $\mathtt{T} = \begin{bmatrix} \mathtt{R} & 0 \\ \mathbf{t} & 1 \end{bmatrix}$, with $\mathtt{R} \in SO(3)$ defined as the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ defined as the translation vector. The device also stores the intrinsics matrix for the RGB camera defined as $\mathtt{K}_c \in \mathbb{R}^{3\times3}$.

For constructing a new depth image $\hat{I}_d \in \mathbb{R}^{H\times W}$ aligned to the RGB camera viewpoint, We use $\mathtt{T}_{d2w}$, $\mathtt{T}_{c2w}$ to transform each 3D point in the depth camera coordinates ($\mathbf{p}_d := [X_d, Y_d, Z_d]^\top$) to those in the RGB camera coordinates. In terms of equation, $[X_c, Y_c, Z_c, 1]^\top = \mathtt{T}_{c2w}^{-1}\mathtt{T}_{d2w}[X_d, Y_d, Z_d, 1]^T$, where $[X_c, Y_c, Z_c]^\top =: \mathbf{p}_c$ is the corresponding 3D point in the RGB camera coordinates. Hence, we can yield the 2D pixel coordinates $\mathbf{p}_c$ of the 3D point in the RGB image I_c as $[x_c, y_c]^\top := \pi(\mathtt{K}_c\mathbf{p}_c)$. We then assign Z_c (depth in the RGB camera pose) in each 2D pixel location (x_c, y_c). i.e., $\hat{I}_d(x_c, y_c) \longleftarrow Z_c$. For pixels without corresponding Z_c values, $I_d(x_c, y_c)$ is set to 0. Above creates depth images aligned to the RGB camera pose (denoted as $\text{Depth}_{\text{RGB}}$).

Alternatively, we can also align RGB images to the depth camera's viewpoint, creating $\hat{I}_c$ (denoted as $\text{RGB}_{\text{Depth}}$). This involves assigning the RGB values at the pixel location (x_c, y_c) in I_c to the pixel (x_d, y_d) in the aligned RGB image $\hat{I}_c$. i.e., $\hat{I}_c(x_d, y_d) \longleftarrow I_c(x_c, y_c)$. Above process is also illustrated in [B] of Fig. 2.

Table 3. Classification accuracy of single modality data. Subscript indicates the frame type, i.e. $\text{Depth}_{\text{RGB}}$ signifies depth data in RGB frame.

Model	Modality	Avg. accuracy
ReT [10]	Depth	93.46%
	$\text{Depth}_{\text{RGB}}$	89.38%
	RGB	90.86%
	$\text{RGB}_{\text{depth}}$	91.98%
C3D [13]	Depth	72.22%
	$\text{Depth}_{\text{RGB}}$	65.06%
	RGB	64.20%
	$\text{RGB}_{\text{depth}}$	70.49%
Zhou *et al.* [16]	Depth	94.46%
	$\text{Depth}_{\text{RGB}}$	86.71%
	RGB	**97.47%**
	$\text{RGB}_{\text{depth}}$	94.18%

4. EXPERIMENTAL RESULTS

4.1. Experimental settings

Dataset. To analyze the impact of glove type on each modality, we divided the HoloGesture into two datasets. The first dataset, henceforth referred to as the 'color-mixed,' includes data for bare hands, and white and blue latex gloves, with an equal number of videos. The second dataset, referred to as the 'bare-only,' is composed solely of bare hand gestures, similar to existing hand gesture datasets. For both datasets, we used 810 videos for training and 270 videos for validation. We also divided the test datasets into three different datasets consisting only of gestures with bare hands, white, and blue latex gloves. Each test dataset contains 270 videos.

When training on a single modality or using two modalities of data together, we utilized 2430 videos for training, and 810 videos each for validation and testing. Each dataset was ensured to include for bare hands and all types of glove data.

Baseline models. Among models that process spatio-temporal feature, we have selected C3D [13] because it is widely used in dataset evaluations [**?**]. In structures where spatial and temporal features are acquired separately, we select [10] which employs a common multimodal training approach. Additionally, to evaluate HoloGesture with a state-of-the-art model, we adopt [16], which demonstrates high accuracy in [5].

Implementation details. For the training of [10], hereafter referred to as ReT, an AdamW optimizer and categorical cross entropy loss were used. Learning rate and weight decay were both set to $1e^{-4}$. Augmentation was not utilized in this process. During the training of C3D [13], an SGD optimizer was employed to minimize categorical cross-entropy loss. The optimizer's learning rate was set to $3e^{-3}$, with a momentum of 0.9 and a weight decay of $1e^{-3}$. We utilized a C3D pretrained on Kinetics data. To use a pretrained model, the input data were resized to 112×112. In training Zhou *et al.* [16], we utilize label smoothing cross entropy loss and SGD optimizer. We set the learning rate to $1e^{-2}$, weight decay to $3e^{-4}$, and momentum to 0.9.

All models used frame-wise min-max normalized data as input and video samples in mini-batches of 4 during training. Additionally, the model with the highest validation accuracy was saved during the 50 epochs of training.

Table 4. Effect of gloves in hand gesture recognition accuracy when using RGB frames only. "Color-mixed" denotes all data and "bare-only" refers to data without wearing gloves.

Model	Train	Test		
		bare	blue	white
ReT [10]	color-mixed	88.67%	**89.41%**	**84.00%**
	bare-only	**89.78%**	28.96%	72.29%
C3D [13]	color-mixed	**65.70%**	**63.56%**	**56.22%**
	bare-only	61.70%	47.19%	47.85%
Zhou *et al.* [16]	color-mixed	**95.86%**	**96.20%**	**94.48%**
	bare-only	94.51%	60.08%	91.04%

Table 5. Effect of gloves in hand gesture recognition accuracy when using depth frames only. "Color-mixed" denotes all data and "bare-only" refers to data without wearing gloves.

Model	Train	Test		
		bare	blue	white
ReT [10]	color-mixed	**93.08%**	**92.15%**	**92.15%**
	bare-only	89.18%	90.67%	83.56%
C3D [13]	color-mixed	**71.48%**	**61.04%**	**65.85%**
	bare-only	61.26%	52.59%	54.52%
Zhou *et al.* [16]	color-mixed	**92.41%**	**91.70%**	**94.70%**
	bare-only	84.96%	78.06%	84.48%

4.2. Base experimental results of HoloGesture

To assess the capabilities of HoloGesture, we conduct evaluations using four modalities of data with baseline models. Among these, Zhou *et al.* scores highly with RGB and depth data. In processed data, ReT achieves the highest score as shown in Table 3.

4.3. Effect of gloves in hand gesture recognition accuracy

To compare the outcomes from datasets comprised solely fo bare hands with those considering gloves, we use the results from training on the 'bare-only' and testing on a dataset composed of bare hands as a baseline results. All metrics in Table 4 and 5 represent the average of five identical trials.

In the RGB domain, as shown in Table 4, when trained on the 'color-mixed' dataset, baseline models showed results within a 6% range of the baseline experimental results of 89.78%, 61.70%, and 94.36% for each type of glove in the test datasets. However, when trained on the 'bare-only' dataset, ReT showed up to a 60% difference, C3D showed about a 14% difference, and Zhou *et al.* presented about a 34% difference. The results from the experiments on 'color-mixed' and 'bare-only' demonstrate that training on the 'color-mixed' results in relatively smaller differences in test outcomes for bare hands and each glove data. This indicates that to achieve accuracy comparable to previous research based on bare hand

Table 6. Classification accuracy of intermediate feature fusion approach. Subscript indicates the frame type, i.e. RGB_{depth} signifies RGB data in depth frame.

Model	Modality	Avg. accuracy
ReT [10]	Depth + RGB	94.42%
	$Depth_{RGB}$ + RGB	92.44%
	Depth + RGB_{depth}	**94.69%**
C3D [13]	Depth + RGB	69.53%
	$Depth_{RGB}$ + RGB	69.11%
	Depth + RGB_{depth}	**72.20%**

Table 7. Classification accuracy of late fusion approach. Subscript indicates the frame type, i.e. RGB_{depth} signifies RGB data in depth frame.

Model	Modality	Avg. accuracy
ReT [10]	Depth + RGB	93.01%
	Depth + RGB_{depth}	**94.15%**
C3D [13]	Depth + RGB	70.91%
	Depth + RGB_{depth}	**71.41%**

data when recognizing hand gestures with gloves, datasets that consider gloves are necessary.

In the depth domain, as shown in Table 5, training on the 'color-mixed' either maintained similar accuracy or even showed an increase compared to training on the 'bare-only'. Conversely, training on the 'bare-only' generally resulted in a decline from the benchmark results of 89.18%, 61.26%, and 87.22%. This indicates that a dataset including glove data not only is essential for achieving accuracy comparable to prior research in hand gesture recognition while wearing gloves but also can be beneficial in enhancing performance.

4.4. Effect of spatial alignments in RGB-depth frames

For multimodal hand gesture recognition, we combine features prior to the linear classifier and used them as input to compute the probability distribution, a process we call intermediate feature fusion. Additionally, we have also tested the late fusion approach adopted in [10], which simply normalizes the sum of probability outputs from the RGB and depth branches to produce the final prediction.

We apply these two fusion approaches to two models for our experiments. We choose these methods because they are widely used multimodal training methods [17, 18, 10] and can be uniformly applied to all baseline models. The Zhou *et al.* already employs a specific fusion method, making it unsuitable for our fusion application. Consequently, we exclude Zhou *et al.* from the multimodal training baseline models to ensure fair comparison conditions.

We explore two main ways to demonstrate that spatial alignment can enhance hand gesture recognition accuracy. The first experiment involves utilizing data without performing spatial alignment while the second involves utilizing spatially aligned data. To identify the optimal spatial alignment between RGB and depth images, we applied spatial alignments in both RGB-to-depth and depth-to-RGB directions and conducted experiments for each scenario. All metrics in Table 6 and 7 represent the average of five identical trials.

As observed in Table 6, fusing data with spatial alignment performed on the depth data frames resulted in accuracy improvements in both baseline models compared to using unprocessed data. Moreover, to verify whether the same spatial alignment could enhance accuracy with different fusion techniques, we also conducted experiments on late fusion. The results, as indicated in Table 7, show that using data with spatial alignment on depth frames for fusion led to performance improvements compared to when it was not used.

Previous datasets have typically provided data that is spatially aligned or not, and prior research have applied different fusion methods to this existing data. However, our approach demonstrates that data spatial alignment can influence fusion accuracy. We conducted experiments with both preprocessed and unprocessed data. Furthermore, by performing experiments in two scenarios where spatial alignment was applied to depth and RGB frames, we show that accuracy improved when spatial alignment was performed on depth frames. This highlights that, in addition to traditional fusion methods, spatial alignment can also be a means to enhance accuracy.

5. CONCLUSION

We introduced HoloGesture, a new dataset capturing various gestures from multiple subjects using Microsoft HoloLens 2. The two main advantages of our dataset are i) each gesture is performed with and without medical latex gloves to account for different environmental settings and ii) the gestures are acquired using a recent MR device to benefit from multimodal (depth and RGB) data, which are optionally preprocessed for spatial alignment. Through our experiments using baseline models, we showed that the inclusion of glove data improves robustness of hand gesture recognition to varying hand textures, providing motivation for MR applications in medical and industrial domains. We also showed that the accuracy can be enhanced by using multimodal data and further via spatial alignment of RGB and depth data. This improvement was consistently observed in both baseline models across two different fusion methods. Through the release of our dataset and relevant code, we hope to accelerate future research in multimodal hand gesture recognition robust to hand textures.

6. ACKNOWLEDGEMENT

This work was in part supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2022R1A5A1022977), and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)).

7. REFERENCES

[1] Tae-Kyun Kim and Roberto Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009. 1, 2

[2] Simon Ruffieux, Denis Lalanne, and Elena Mugellini, "Chairgest: A challenge for multimodal mid-air gesture recognition for close hci," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, New York, NY, USA, 2013, ICMI '13, p. 483–488, Association for Computing Machinery. 1, 2

[3] Li Liu and Ling Shao, "Learning discriminative representations from rgb-d video data," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 08 2013, pp. 1493–1500. 1, 2

[4] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 2874–2882. 1, 2

[5] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4207–4215. 1, 2, 5

[6] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu, "Egogesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1038–1050, 2018. 1, 2, 3

[7] Tejo Chalasani, Jan Ondrej, and Aljosa Smolic, "Egocentric gesture recognition for head-mounted ar devices," in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2018, pp. 109–114. 1, 2

[8] Lorenzo Baraldi, Francesco Paci, Giuseppe Serra, Luca Benini, and Rita Cucchiara, "Gesture recognition in ego-centric videos using dense trajectories and hand segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 702–707. 1, 2

[9] Okan Köpüklü, Neslihan Köse, and Gerhard Rigoll, "Motion fused frames: Data level fusion strategy for hand gesture recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2184–21848. 1, 3

[10] Andrea D'Eusanio, Alessandro Simoni, Stefano Pini, Guido Borghi, Roberto Vezzani, and Rita Cucchiara, "A transformer-based network for dynamic hand gesture recognition," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 623–632. 1, 3, 5, 6

[11] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1165–1174. 1, 3

[12] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Xin Duan, Casey Meekhof, Jan Stuhmer, Thomas J. Cashman, Bugra Tekin, Johannes L. Schonberger, Bugra Tekin, Pawel Olszta, and Marc Pollefeys, "HoloLens 2 Research Mode as a Tool for Computer Vision Research," *arXiv:2008.11239*, 2020. 2, 4

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497. 3, 5, 6

[14] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733. 3

[15] Kenneth Lai and Svetlana N. Yanushkevich, "Cnn+rnn depth and skeleton based dynamic hand gesture recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3451–3456. 3

[16] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, Fan Wang, Du Zhang, Zhen Lei, Hao Li, and Rong Jin, "Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20154–20163. 3, 5

[17] Hilal Ergun, Yusuf Akyuz, Mustafa Sert, and Jianquan Liu, "Early and late level fusion of deep convolutional neural networks for visual concept recognition," *International Journal of Semantic Computing*, vol. 10, pp. 379–397, 09 2016. 6

[18] Yi Yang, Jingkuan Song, Zi Huang, Zhigang Ma, Nicu Sebe, and Alexander G. Hauptmann, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 572–581, 2013. 6