The IMHG dataset: A Multi-View Hand Gesture RGB-D Dataset for Human-Robot Interaction

Dadhichi Shukla[1], Özgür Erkent[1] and Justus Piater[1]

***Abstract*— Hand gestures are one of the natural forms of communication in human-robot interaction scenarios. They can be used to delegate tasks from a human to a robot. To facilitate human-like interaction with robots, a major requirement for advancing in this direction is the availability of a hand gesture dataset for judging the performance of the proposed algorithms. We present details of the Innsbruck Multi-View Hand Gesture (IMHG) dataset recorded with two RGB-D cameras (Kinect). The dataset includes two types of referencing (pointing) gestures with the ground truth location of the target pointed at, two symbolic gestures, two manipulative gestures, and two interactional gestures. The dataset was recorded with 22 participants performing all eight hand gestures.**

I. Introduction

As robots are moving closer to popular deployment in our daily lives, there is increasing activity in Human-Robot Interaction (HRI) research. Amongst various forms of communications, hand gestures are a highly effective, general-purpose tool for interaction, thanks to the flexibility of the hands. Despite the advancements in HRI methodologies, lack of standardized evaluation hinders their application in sectors such as manufacturing, healthcare, and domestic helper.

Gesture recognition has recently received much attention in the HRI community [1]. We provide a multi-view hand gesture RGB-D dataset for quantitative evaluation of gesture recognition systems in a HRI context. A detailed description of the different types of recorded gestures is given in section II. Some example gestures collected in this dataset are shown in Fig. 1.

A. Motivation

In the early years of human-robot interaction, Dautenhahn [2] indicated that people prefer to interact with robots in a "natural" way. In human-robot interaction, camera-based vision serves as a natural, unencumbered, non-contact, and prop-free mode of interaction. Adopting hand gestures as an interface in HRI opens up new frontiers of research in a wide range of applications (e.g., surgical robotic nurse [3], swarm robots [4], assistive robotics [5], HRI in industrial scenarios [6], etc.). Some insights from human-computer interaction (HCI) may also prove to be valuable to HRI. These notwithstanding, the field of HRI needs to develop its own methods for interaction between a human and a robot. For example, in pointing gestures, a robot has to recognize the gesture as well as to estimate the pointing direction i.e. the pose of the hand. Moreover, it is difficult to perform quantitative evaluation for such *referencing gestures* because of a lack of ground truth.

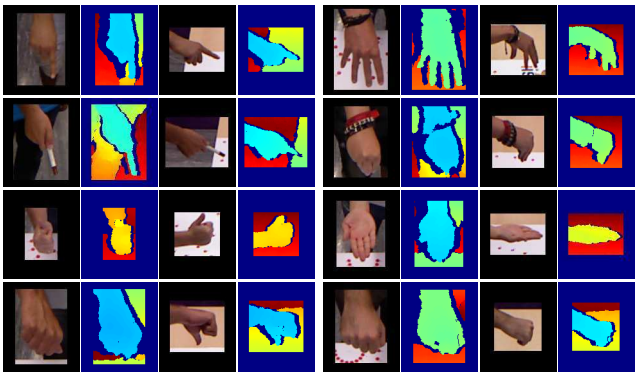


Fig. 1: Multi-view sample images from the IMHG dataset (*images cropped for visualization*). *Top-Bottom*: Finger pointing, Tool pointing, Thumb up (approve), Thumb down (disapprove), Grasp open, Grasp close, Receive, Fist (stop).

Many prominent methodologies [7], [8] use gesture recognition systems as an interface to interact with the robot. However, these gesture recognition systems mainly target robot guidance. The gestures are not conceived as commands for the robot to manipulate objects in the environment. This requires capabilities of recognizing human gestures and inferring intent (defined as the manipulative action expected from the robot). To address such scenarios of collaborative manipulation, we propose the novel, publicly-available IMHG dataset.

B. Related work

Human gesture recognition has been studied extensively to interact with robots. Several studies have addressed the topic to create a hand gesture dataset, although most of them are either American Sign Language (ASL) [9], [10] or human-computer interaction [11], [12], [13]. Other datasets [14], [15] capture full-body or upper-body of the participant. The intended domain of applicability of the dataset varies depending on the type of ground truth and chosen set of gestures.

A recent survey by Ruffieux et al. [16] provides a detailed overview of the most recent and/or popular publicly available vision based hand, upper-body, and full-body gesture datasets. Here, we briefly describe previously proposed hand gesture datasets. Kim et al. [11] released an RGB dataset with 9 classes of hand gesture with 900 image sequences. The target task of this dataset was to classify *shapes* as well

[1]The authors are with the Intelligent and Interactive Systems group, Institute of Computer Science, University of Innsbruck, Austria. Corresponding author: Dadhichi.Shukla@uibk.ac.at

as different *motions* at a time. The RGB-D dataset proposed by Liu et al. [13] consists of 10 categories of hand gestures representing shapes like circle, triangle, 'Z', etc. The gestures are performed with three hand postures with different backgrounds and varying illumination conditions. Kurakin et al. [9] proposed a RGB-D dataset of 12 dynamic American sign language (ASL) performed by 10 participants. More recently, Molina et al. [17] released a hand gesture dataset composed of English alphabets, Spanish sign language, and several miscellaneous annotated gestures captured by 11 participants and also, generated synthetically.

The RGB-D dataset collected by Ren et al. [12] can be related to our work. It consists of 10 types of gestures captured in cluttered background. Like the previous works it also addresses hand shape detection. Though the dataset is targeted for HCI applications, it can be applicable to HRI scenarios. One important difference with the former is that the hand gestures in the IMHG dataset are closely related to the semantic content of verbal language. A robot interprets these gestures as the command to be executed to interact with the environment.

We briefly summarize the previous datasets and the proposed IMHG dataset in Table I reviewing various characteristics.

C. Contribution

The main contributions of this dataset are:

- A multi-view RGB-D dataset with 22 participants performing 8 hand gestures.
- Two types of referencing gestures: (1) finger pointing, and (2) tool pointing with an elongated object in hand, are recorded.
- A corpus of 836 test images (704 referencing gestures with ground truth, and 132 other gestures).
- The data acquisition setup can be easily recreated to add new hand gestures in the future.

The IMHG dataset is dedicated to measuring the performance of recognition systems to do gesture understanding – that is, the interpretation of an indicative *robot manipulation* a user wishes to take place.

II. IMHG Dataset Description

There exist many semantic gestures in our daily lives. However, many of them are unsuitable for direct use in human-robot interaction. Nehaniv et al. [18] asserted five major categories of hand gestures in the context of human-robot interaction. One of the five categories, *expressive behaviour* that includes motions of hands, arms, and face, is excluded from the dataset. These types of gestures occur as part of the overall communicative behaviour, but without any specific interactive role of a robot.

According to the study conducted by Nehaniv et al. the eight hand gestures in the IMHG dataset can be grouped into the following four categories:

1. *Referencing hand gestures:* These gestures are used to refer to or to indicate objects (or loci) of interest.We record two types of pointing: (i) finger pointing, and (ii) tool pointing.
2. *Symbolic hand gestures:* The gestures in this category are defined by a prescribed set of interpretations. The static symbolic gestures are analogous to discrete actions on a user interface like Yes/No, Agree/Disagree, etc. In this dataset we capture two types of gestures: (i) Thumb up (approve), and (ii) Thumb down (disapprove), to indicate whether the task was understood/ performed correctly by a robot.
3. *Manipulative hand gestures:* These gestures involve displacement of or interaction with objects (e.g., pushing a box). We record two types of manipulative gestures: (i) *Grasp open* – The robot is to open the hand to grasp an object, (ii) *Grasp close* – The robot is to grasp the object of interest.
4. *Interactional hand gestures:* The gestures in this category are used to regulate interaction with a partner. They can be used to initiate, synchronize, or terminate an interaction. The emphasis on this category of gestures is neither reference nor communication, but for a cooperative action. The dataset includes two types of interactional gestures: (i) *Receive* – The robot is to hand the grasped object to the human, and (ii) *Fist (stop)* – The robot is to stop interacting with the environment. We include the *Fist* gesture because it is easily performed by a human within the camera view of our setup.[1]

Gesture	# Instances	Ground truth
Finger pointing	352	✓
Tool pointing	352	✓
Thumb up	22	×
Thumb down	22	×
Grasp open	22	×
Grasp close	22	×
Receive	22	✓
Fist (stop)	22	×

TABLE II: Summary of IMHG dataset. The ground truth of 'pointing' gesture and 'receive' gesture is the location of the pointed target and the location of the hand, respectively.

Some example images from the dataset are shown in Fig. 2. Each image in the dataset is labelled with its corresponding gesture. To evaluate *referencing* gestures we provide the location of the target pointed at as the ground truth. In the case of *receive* gesture the centroid of the palm is given as the ground truth. We summarize the IMHG dataset in Table II.

III. Dataset Acquisition Scenario

Camera setup

The IMHG dataset was captured using two RGB-D cameras (Kinect) placed orthogonally to record maximum information of the hand gesture. Figure 3a illustrates the multi-

[1] A raised, flat palm vertically facing the camera would certainly constitute a more intuitive *stop* gesture. However, since all other gestures in our scenario are performed low above the workspace, this would require a dedicated camera and eliminate the need for its explicit recognition.

Methods	#Classes	Views	RGB	Depth	Resolution	Pose of finger joints	Available	Application to HRI
Kim et al. [11]	9	T	✓	×	320 × 240	×	✓	×
Ren et al. [12]	10	F	✓	✓	640 × 480	×	✓	✓
Kurakin et al. [9]*	12	F	✓	✓	130 × 130	×	✓	×
Liu et al. [13]	10	T	✓	✓	320 × 240	×	✓	×
Molina et al. [17]*	55	F	×	✓	176 × 144	✓	✓	×
IMHG dataset	8	F, S	✓	✓	640 × 480	×	✓	✓

TABLE I: Summary of hand gesture datasets. Previous work is summarized based on the following characteristics: number of hand gesture classes; number of views (T - Top view, F - Front view, S - Side view); RGB data; depth data; resolution of images; pose of finger joints; availability of the dataset; application to HRI. *Sign language gestures.

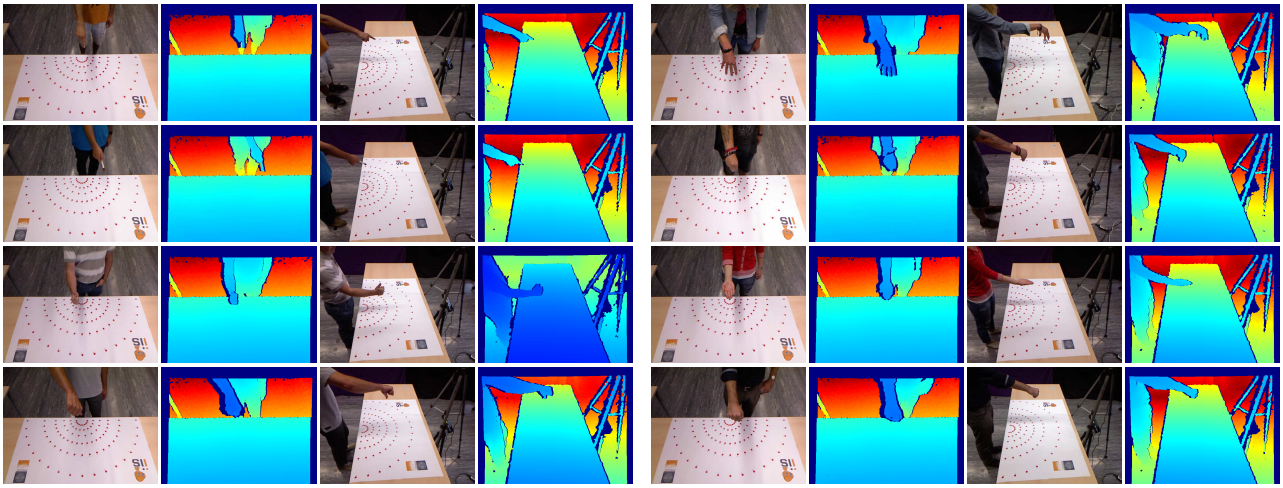


Fig. 2: IMHG dataset sample images.

view RGB-D camera setup. We captured 640×480 RGB images and `uint16` depth images. The depth sensing within Kinect is based on a structured infrared (IR) pattern. The simultaneous use of multiple depth cameras can extend the coverage of the vision system to a great extent. However, when multiple infrared patterns are projected at the same scene, the received depth signal degrades severely. To overcome this challenge Butler et al. [19] proposed the *Shake 'n' Sense* technique. The Kinect is minimally vibrated using an offset-weight vibration motion and thereby artificially introduces motion blur.

We address the depth interference problem in a different way. Instead of modifying the Kinect sensor we use the open-source *freenect* library to control the depth streaming of the Kinects. Using the freenect driver library it is possible to toggle the reading of the infrared pattern by controlling the flags of Kinects, thereby allowing multi-view RGB-D data to be captured. The extrinsic camera matrix between Kinects is estimated using ROS multiple camera calibration package. The calibration error can be up to 2 cm in 3D space.

Participants and Workspace

Participants from both genders were involved in the data acquisition process. They were asked to stand at a distance of approximately 1.3 m away from both the cameras, i.e. front view and side view, to perform eight classes of hand gestures. To avoid confusion, participants were shown different types of gestures prior to the recording, but no specific instructions were given to the participants on how to recreate the gesture, allowing their gestures to be recorded in a natural fashion.

We designed a polar coordinate system as shown in Fig. 3b with numbers marked at each *red* dot to capture the ground truth of the referencing gestures. The participants were asked to point at 16 randomly selected numbers. For the remaining six gestures we recorded only single instances, since they are not correlated with the location of an object. The workspace was configured such that hand gestures were visible from both cameras.

Dataset availability

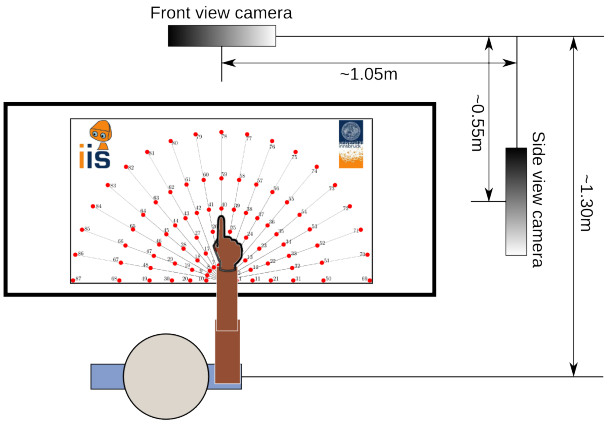
The IMHG dataset is available at this link[2]. New gestures can be added to the dataset by researchers, provided images are captured in a calibrated setup. Researchers can reproduce the data acquisition setup following the instructions given on the IMHG dataset page. It is to be noted that gestures should be recorded as static RGB-D images.

Researchers can contribute their work to the current dataset. The contributed set of images will be tested for calibration errors. Once accepted, they will be added on the page with an acknowledgement. It is encouraged to submit the hand gestures associated with a semantic content.

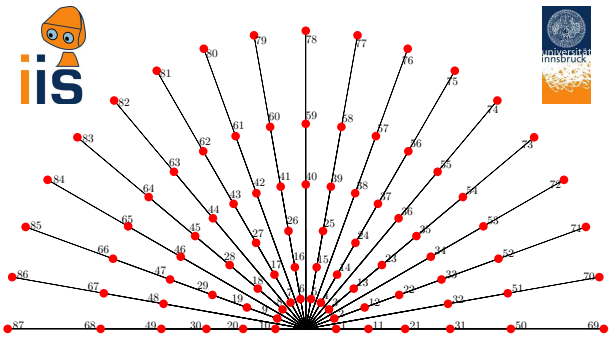
IV. Conclusions and Future Work

For a human-robot interaction to take place at close proximity it is necessary to measure the performance of

[2]`https://iis.uibk.ac.at/public/3rdHand/IMHG_dataset/`



(a) Illustration of the IMHG dataset acquisition setup.



(b) Polar coordinate system to record the ground truth location of the target pointed at.

Fig. 3: IMHG dataset acquisition setup.

gesture recognition systems independent of human body pose. We described a novel IMHG dataset from two RGB-D cameras with ground truth. The dataset comprises 8 classes of hand gestures with semantic meaning. The dataset mainly focuses on HRI scenarios. The data acquisition setup is easily reproducible for extension of the dataset with additional hand gestures. We are currently working on a baseline evaluation to detect hand gestures using a probabilistic framework.

Acknowledgement

The research leading to this work has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 610878, 3rd HAND.

References

[1] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, "Vision-based hand-gesture applications," *Communications of the ACM*, vol. 54, no. 2, pp. 60–71, 2011.

[2] K. Dautenhahn, "The art of designing socially intelligent agents: Science, fiction, and the human in the loop," *Applied artificial intelligence*, vol. 12, no. 7-8, pp. 573–617, 1998.

[3] J. P. Wachs, M. Jacob, Y.-T. Li, and G. Akingba, "Does a robotic scrub nurse improve economy of movements?" in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2012, pp. 83 160E–83 160E.

[4] J. Alonso-Mora, S. Haegeli Lohaus, P. Leemann, R. Siegwart, and P. Beardsley, "Gesture based human-multi-robot swarm interaction and its application to an interactive display," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5948–5953.

[5] M. Lopes, J. Peters, J. Piater, M. Toussaint, A. Baisero, B. Busch, O. Erkent, O. Kroemer, R. Lioutikov, G. Maeda, Y. Mollard, T. Munzer, and D. Shukla, "Semi-Autonomous 3rd-Hand Robot," in *Robotics in future manufacturing scenarios*, 3 2015, workshop at the European Robotics Forum, Vienna, Austria. [Online]. Available: https://iis.uibk.ac.at/public/papers/Lopes-2015-CogRobFoF.pdf

[6] P. Barattini, C. Morand, and N. M. Robertson, "A proposed gesture set for the control of industrial collaborative robots," in *RO-MAN, 2012 IEEE*. IEEE, 2012, pp. 132–137.

[7] C.-B. Park and S.-W. Lee, "Real-time 3d pointing gesture recognition for mobile robots with cascade hmm and particle filter," *Image and Vision Computing*, vol. 29, no. 1, pp. 51–63, 2011.

[8] M. Pateraki, H. Baltzakis, and P. Trahanias, "Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation," *Computer Vision and Image Understanding*, vol. 120, pp. 1–13, 2014.

[9] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 1975–1979.

[10] A. Barczak, N. Reyes, M. Abastillas, A. Piccio, and T. Susnjak, "A new 2D static hand gesture colour image dataset for ASL gestures," *Letters in the Information and Mathematical Sciences*, vol. 15, pp. 12–20, 2011.

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

[12] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 1093–1096.

[13] L. Liu and L. Shao, "Learning discriminative representations from rgb-d video data," in *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. AAAI Press, 2013, pp. 1493–1500.

[14] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante, "The chalearn gesture dataset (cgd 2011)," *Machine Vision and Applications*, vol. 25, no. 8, pp. 1929–1951, 2014.

[15] A. Sadeghipour, L. philippe Morency, and S. Kopp, "Gesture-based object recognition using histograms of guiding strokes," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2012, pp. 44.1–44.11.

[16] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled, "Gesture recognition corpora and tools: A scripted ground truthing method," *Computer Vision and Image Understanding*, vol. 131, pp. 72–87, 2015.

[17] J. Molina, J. A. Pajuelo, M. Escudero-Viñolo, J. Bescós, and J. M. Martínez, "A natural and synthetic corpus for benchmarking of hand gesture recognition systems," *Machine Vision and Applications*, vol. 25, no. 4, pp. 943–954, 2014.

[18] C. L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. Alami, "A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. IEEE, 2005, pp. 371–377.

[19] D. A. Butler, S. Izadi, O. Hilliges, D. Molyneaux, S. Hodges, and D. Kim, "Shake'n'sense: reducing interference for overlapping structured light depth cameras," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1933–1936.