A Multi-View Hand Gesture RGB-D Dataset for Human-Robot Interaction Scenarios

Dadhichi Shukla*, Özgür Erkent* and Justus Piater*

***Abstract*— Understanding semantic meaning from hand gestures is a challenging but essential task in human-robot interaction scenarios. In this paper we present a baseline evaluation of the Innsbruck Multi-View Hand Gesture (IMHG) dataset [1] recorded with two RGB-D cameras (Kinect). As a baseline, we adopt a probabilistic appearance-based framework [2] to detect a hand gesture and estimate its pose using two cameras. The dataset consists of two types of deictic gestures with the ground truth location of the target, two symbolic gestures, two manipulative gestures, and two interactional gestures. We discuss the effect of parallax due to the offset between head and hand while performing deictic gestures. Furthermore, we evaluate the proposed framework to estimate the potential referents on the Innsbruck Pointing at Objects (IPO) dataset [2].**

I. Introduction

For human-robot interactions to take place naturally robots need to recognize intuitive hand gestures performed by the user. Burger et al. [3] suggested that robots in such scenarios need to be equipped with transactional intelligence which means being able to communicate meaningfully with a human user. Amongst different forms of communications, hand gestures are a highly effective and universal tool for interaction, thanks to the flexibility of the hands. They can be used to delegate tasks from a human to a robot.

In this work we focus on evaluating the IMHG dataset [1] to recognize hand gestures. We extend our probabilistic, appearance-based, deictic gesture detection framework [2], which uses only single camera, to multiple hand gesture detection and pose estimation using multiple cameras. Based on Quek's taxonomy [4], the different types of hand gestures in the IMHG dataset can be categorized as shown in Fig. 1. *Modalizing* gestures in the taxonomy tree are associated with speech; we deliberately exclude them from the IMHG dataset. Figure 2 illustrates the IMHG data acquisition setup. Quek suggested his taxonomy of hand gestures for human-computer interaction (HCI), but it is equally suitable for HRI scenarios.

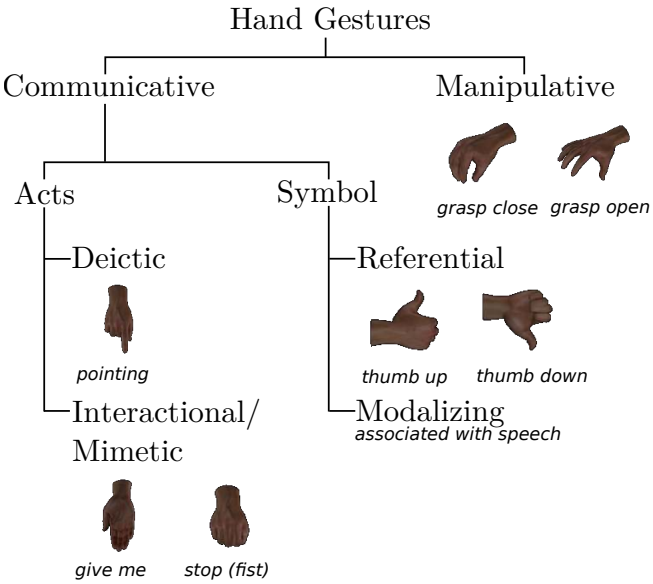


Fig. 1: Taxonomy of hand gestures.

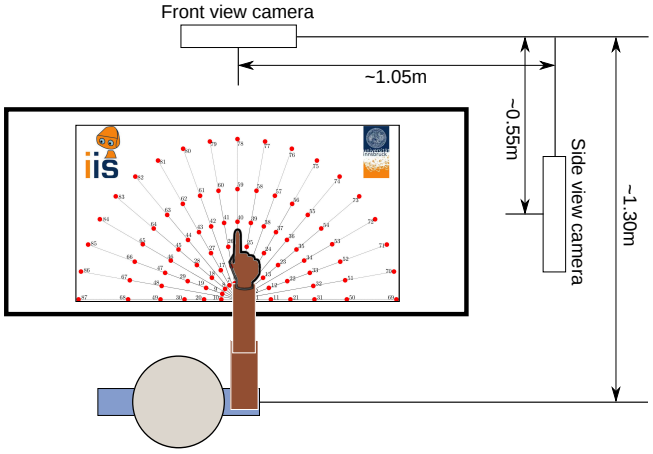


Fig. 2: Illustration of the IMHG dataset acquisition setup

A. Motivation and Contribution

Previous studies [5], [6] have demonstrated that gesture recognition and pose estimation can be improved by using multiple cameras. However, these gesture recognition systems require full-body or upper-body human pose. They are mainly targeted towards robot guidance. The gestures are not conceived as commands for the robot to manipulate objects in the environment.

If robots are supposed to collaborate with humans in close proximity, for example, furniture assembly [7], socially assistive robots [8], etc., it is likely that a robot (with a limited field of view) can only see user's hand and not the whole body. In such scenarios it is irrelevant to observe full-body human pose since people perform gestures with hand and fingers, not with body and arm [9]. To detect hand gesture some methods either define a bounding box [10] or use information from hand tracker [11]. The proposed hand gesture detection and pose estimation framework is

*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

independent of any prior information regarding human body pose.

Bangerter et al. [12] conducted human (pointer) – human (guesser) studies to examine the perceptual accuracy of deictic gestures. Their study demonstrates that people point by aligning the tip of their pointing finger with the eye-object line. In other words, rather than the finger axis intersecting with the target, people tend to raise the fingertip from the finger-object line towards the eye-object line. The authors call this *pointer bias*, i.e., the *intended target* is predictably different from the *estimated target*. It is plausible that guessers may also be subject to dominant eye bias which may be induced automatically [13].

Similarly, we examine the accuracy of pointing gestures in human-robot interactions. One important difference is that in our study the guesser is the robot vision system which is free from dominant eye bias. Our results were found to be in accordance with the findings of Bangerter et al. In section III-A.1 we discuss the effect of parallax and experimentally quantify pointer bias whilst pointing with the index finger and pointing with the tool in hand.

The main contributions of this work are:

- A probabilistic appearance-based framework to detect hand gestures and estimate their pose using the information from multiple RGB-D cameras.
- A baseline evaluation of the IMHG dataset consisting of 836 test sample pairs (one from each camera) captured with 22 participants (704 deictic gestures with the ground truth, and 132 other gestures).
- Insights into the effect of parallax in performing deictic gestures.
- Quantitative analysis of the publicly-available Innsbruck Pointing at Objects (IPO) dataset [2][1] in estimating potential referents from deictic gestures.

B. Related work

Several prominent studies [14], [15], [16], [17] in HCI have addressed the topic of creating a hand gesture dataset. Other datasets [18], [19] capture the full body or the upper body of the participant. 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 *deictic gestures* because of a lack of ground truth.

A recent and comprehensive survey by Ruffieux et al. [20] reviews publicly-available, vision-based hand, upper-body, and full-body gesture datasets. One of the initial efforts by Kim et al. [14], the Cambridge Hand Gesture Dataset (CHGD), is an RGB dataset with 9 classes of hand gestures. The dataset was conceived for classification of hand shapes and hand motions. With the availability of RGB-D sensors, Liu et al. [16] released the Sheffield Kinect Gesture (SKIG) dataset, which consists of 10 categories of hand gestures representing shapes like circle, triangle, 'Z', etc. Another notable RGB-D dataset, the Microsoft Research Gesture dataset (MSR) proposed by Kurakin et al. [21], consists of 12 dynamic American sign language (ASL) gestures performed by 10 participants. More recently, Molina et al. [17] released a hand gesture dataset composed of the Spanish sign language alphabet and several miscellaneous annotated gestures captured by 11 participants and also generated synthetically.

The hand gesture dataset by Ren et al. [15] is somewhat related to the IMHG dataset. Similar to previous work it also addresses hand shape detection. A pertinent difference with the former is that in the IMHG dataset the hand gestures 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. Additionally, the test scenes are captured using multiple Kinect cameras. We briefly summarize the aforementioned datasets in Table I reviewing various characteristics.

There have also been several studies to estimate the direction of pointing gestures. Here we briefly summarize some previous studies comparable to our method. Contrary to our method, the pose of the user's body is an essential requirement for all these studies. These state-of-the-art methods can be categorized into three main types of pointing strategies:

1) *Elbow-Hand line* – only the forearm is used to point. Großmann et al. [22] used the open-source, Kinect-based OpenNI NITE skeletal tracking library to estimate the pose of the elbow and the hand. They then trace a 3D line along the forearm to estimate the location of the target.
2) *Shoulder-Hand line* – user is supposed to stretch the whole arm to point. Huber et al. [23] tracked body features in proximity spaces and computed the *shoulder-hand* line. Droeschel et al. [24] proposed an approach based on Gaussian Process Regression (GPR) to estimate the target location. They segmented body parts from depth data to extract the pose of the whole arm. They also investigated the accuracy of *elbow-hand line* strategy.
3) *Head-Hand line* – the fingertip is aligned between the eye and the target. The approach proposed by Nicket et al. [25] combines skin color information with 3D depth skin color clusters. They train Hidden Markov Models (HMM) on different phases of pointing gestures and estimate the location of the target using the *head-hand* line.

A qualitative comparison of the proposed work with state-of-the-art methods is discussed in section III-B.

II. Hand Gesture Detection and Pose Estimation

Libhand [26], a synthetic hand model proposed by Romero et al. [27], was used to generate the training data for the proposed gesture detection framework. An illustration of a set of training viewpoints on the viewing sphere with some of the training images is shown in Fig. 3. The hand gesture detection and pose estimation framework used is based on

[1] https://iis.uibk.ac.at/datasets/ipo

Methods	#Classes	Views	RGB	Depth	Resolution	Pose of finger joints	Available	Application to HRI
Kim et al. [14]	9	TV	✓	×	320 × 240	×	✓	×
Ren et al. [15]	10	FV	✓	✓	640 × 480	×	✓	✓
Kurakin et al. [21]*	12	FV	✓	✓	130 × 130	×	✓	×
Liu et al. [16]	10	TV	✓	✓	320 × 240	×	✓	×
Molina et al. [17]*	55	FV	×	✓	176 × 144	✓	✓	×
IMHG dataset [1][2]	8	FV, SV	✓	✓	640 × 480	×	✓	✓

TABLE I: Summary of hand gesture datasets based on the following characteristics: number of hand gesture classes; number of views (TV - top view, FV - Front view, SV - Side view); RGB data; depth data; resolution of images; pose of finger joints; availability of the dataset; application to HRI. *Sign language gestures.

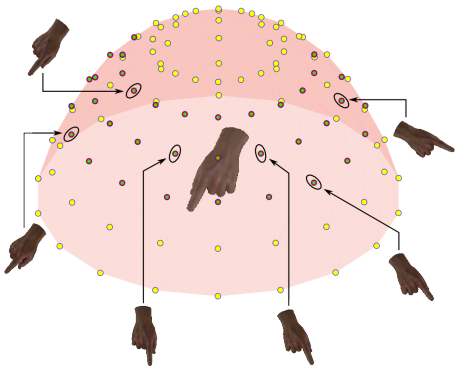


Fig. 3: Training images captured from different viewpoints for pointing gesture. The viewpoints marked in *red-green* are a subset of the training data used to learn the model of the pointing gesture. Similarly, training images are captured for non-deictic gestures.

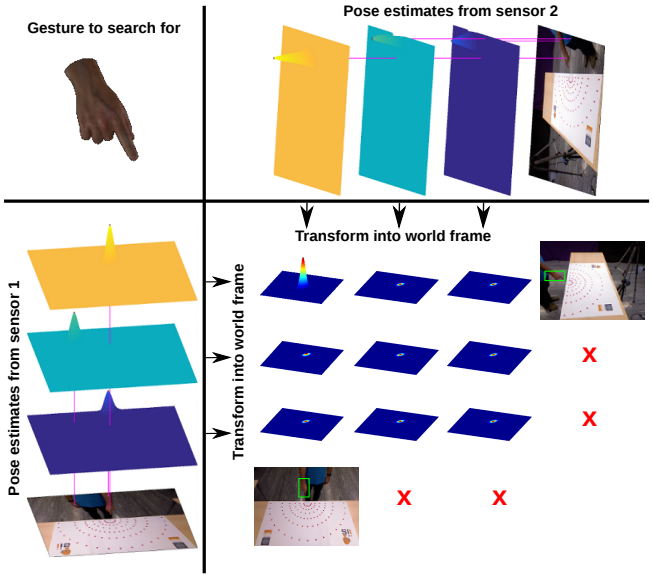


Fig. 4: Hand pose estimation using two cameras. Pose estimates are integrated from both views in 6D pose space. The pose with the highest score is retained as the best estimated pose.

probabilistic representations of both the training and the test data [28]. An advantage of this framework is that the method can be trained with a generic hand model to detect gestures performed by users with different shapes and sizes of the hand.

To detect hand gestures and their pose (location, azimuth angle and elevation angle) we adopt the same methodology as described in our previous work of detecting pointing gestures [2]. Here, we describe the method to combine information from multiple cameras to overcome the ambiguity in pose angles that may occur in individual cameras.

Pose Estimation with Multiple Cameras

The multi-view integration process is illustrated in Fig. 4. The process is carried out in two steps. First, we obtain the pose estimates from two cameras. Each pose estimate is transformed from camera frame to world frame. Second, the pose estimates from both cameras are integrated probabilistically in the world frame. The pose with the highest score is considered the correct estimate (in Fig. 4, the first element of the diagonal). The details of the process are as follows.

We have two sets of estimated poses $\mathcal{S}_1$ and $\mathcal{S}_2$ from camera 1 and camera 2, respectively. First, we consider all possible pose estimation combinations from both cameras.

[2]https://iis.uibk.ac.at/datasets/imhg

Let N be the set of all possible pose estimation combinations. Each element $n_j \in N$ contains a set of pose estimations consisting of one estimate from each camera. We obtain a 6D distribution for each pose estimation in the i^{th} camera as

$$\Phi(\mathcal{S}_i) = \mathcal{N}(\mathcal{S}_i, \Sigma) \tag{1}$$

which is a Gaussian approximation of the estimate in the i^{th} camera with a covariance of Σ. We use this distribution to define a function of the combined pose estimates n_j,

$$\varphi(n_j) = \prod_{\mathcal{S}_i \in n_j} \Phi(\mathcal{S}_i). \tag{2}$$

The final estimated pose of the hand can be obtained by finding the maximum of scores among $\varphi(n_j)$,

$$s^* = \underset{j}{\operatorname{argmax}}\, \varphi(n_j). \tag{3}$$

III. Results and Discussion

The IMHG dataset consists of 8 types of gestures; namely: finger pointing, tool pointing, receive (give me), thumb up (approve), thumb down (disapprove), grasp open, grasp

close, and fist (stop) [1]. To evaluate our method, we compare the estimated hand pose with the manually labelled ground truth computed using depth data. For deictic gestures the ground truth is the target location (red dot) on the table as shown in Fig. 2. We compute error as the distance between the estimated target location and the ground truth location in the world frame. The estimated target location is the intersection of the line of direction of pointing gesture (pose of the pointing gesture) and the planar surface (table). For the non-deictic gestures the estimated pose is compared only with the ground truth location of the hand.

A multiple-camera setup is prone to calibration errors. The average calibration error in the IMHG dataset acquisition system was estimated to be around 1.8 cm which is in accordance with other multiple-camera setups. For instance, Macknojia et al. [29] estimated the extrinsic camera calibration error to be around 2.5 cm in a network of 5 Kinect sensors.

To speed up the estimation process we mask out the edge points of the background using depth information. The gestures in the dataset are performed with either of the hands. We therefore search for two hands and accept the pose estimate closer to the camera as the *dominant hand*. Additionally, the method can be used to detect a novel gesture performed by both hands as long as they are spatially distinct in the image.

We performed two sets of experiments. The first is aimed at a quantitative analysis of the parallax effect in pointing at the *true* target location. We evaluated our framework on the IPO dataset [2] to estimate the potential target objects. Then, we perform baseline evaluations of the IMHG dataset. Here, we compute the accuracy and the standard deviation of the highest-scoring pose estimate. We also compare our method with state-of-the-art methods.

A. Experiment 1: Effect of parallax in deictic gestures

A deictic gesture can only be used in the cases of what Clark et al. [30] refer to as *physical copresence* – both participants are able to view the referent in the situation in which the gesture occurs. Furthermore, Clark et al. [31] asserted in their study that the precise target point indicated by the user is in most cases spatially distant from the object the user intends to indicate. This is due to two main factors. One is a simple geometric error on the part of the human due to parallax. Secondly, deictic gestures in 3D contain no inherent information regarding the distance. Instead, a deictic gesture is typically constrained to a set of spatially spaced potential referents. We quantitatively examine the above two factors as follows.

1) Parallax error: In the IMHG dataset each participant was asked to point at different target locations on a polar coordinate system as shown in Fig. 2. To study the effect of parallax we first calculate the line of the pointing direction with manually labelled points along the index finger or the pointing tool using depth data. Next, we find *the targeted location* as the intersection between the line along the pointing object (index finger or pointing tool) and the planar surface (table). Finally, we compute the distance between the *intended target*, i.e. the ground-truth target location, and the *targeted location*, i.e. the point of intersection.

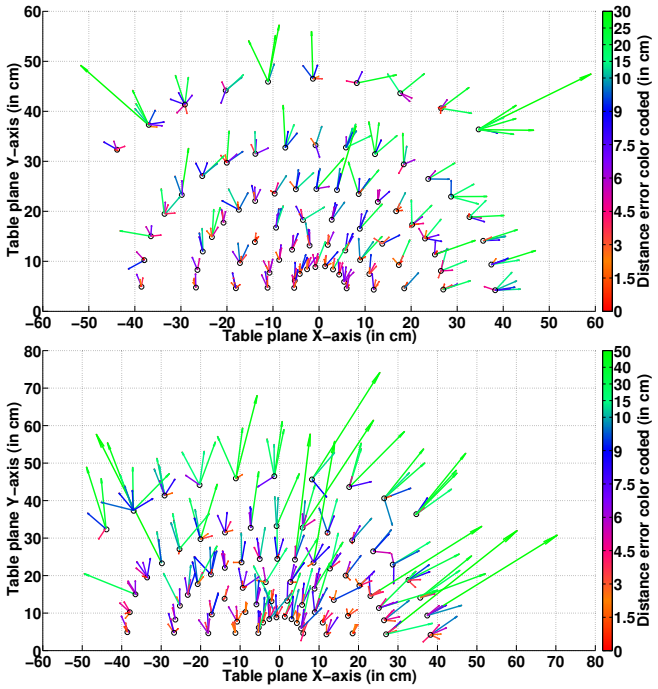


Fig. 5: Pointer bias observed (*top*) while pointing with the index finger, (*bottom*) while pointing with the tool in hand. The pointer bias is represented by vectors from the intended target (black circle) to the targeted location. The magnitude of the vector i.e. error in distance, is color coded.

Evidently, the intended target and the targeted location are typically found to be significantly apart. The *pointer bias* as described in section I-A was found to be systematically outward i.e., away from the participant. It can be seen from the polar chart as shown in Fig. 5. The vectors represent the error between the intended target and the targeted location.

Table II shows the mean errors in the horizontal dimension (X-axis) μ_{X}, in the vertical dimension (Y-axis) μ_{Y}, and the absolute error μ_{a} with corresponding standard deviations while pointing with the index finger and the tool in hand. The mean errors, μ_{X} and μ_{Y}, are computed as targeted location minus intended location. Their values indicate that the line of pointing direction overshoots the intended target. In the horizontal dimension the bias was found to be stronger for the ipsilateral (on the side of participant's dominant hand) target points as compared to the contralateral target points. For example, the bias for a right-handed participant pointing at the target on the right is stronger than for a target on the left. We speculate that this is because, while pointing at contralateral targets, the hand crosses below the eye, reducing the parallax effect. In the vertical dimension the effect of parallax was observed to be much larger. Overall it can be seen from the absolute error μ_{a} and standard deviation σ_{a} that pointing with the tool induces a greater parallax effect compared to pointing with the index finger.

Finger pointing	μ_X	σ_X	μ_Y	σ_Y	μ_a	σ_a
Ipsilateral targets	2.38	2.95	1.78	3.81	4.37	4.83
Contralateral targets	0.35	1.43	2.56	1.85	3.47	2.21

Tool pointing	μ_X	σ_X	μ_Y	σ_Y	μ_a	σ_a
Ipsilateral targets	3.79	5.29	5.36	6.13	8.59	8.74
Contralateral targets	−0.08	2.53	5.27	4.32	6.78	5.82

TABLE II: Pointer bias comparison (*top*) while pointing with the index finger, (*bottom*) while pointing with the tool in hand. μ_X, μ_Y, μ_a are mean errors along X-axis, Y-axis and absolute error and σ_X, σ_Y, σ_a, are corresponding standard deviations, respectively, in cm.

The estimated error due to parallax varies among different users. It depends on various factors, including the way the pointing tool is held, the shape of the index finger, or the pointing style itself. For example, a very tense or flimsy index finger leads to large errors.

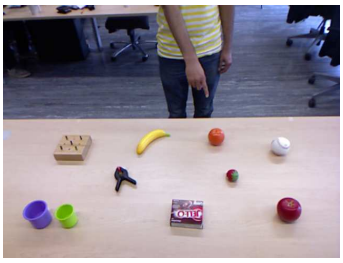
2) Potential referents: A deictic gesture essentially indicates a targeted object. The probability of the object to be selected as the estimated target depends on its distance to the estimated direction. The nearer the object is to the estimated direction more likely it is to be considered as the target. In human-robot interaction it is feasible to associate the direction of the pointing gesture with *potential referents*. Such a framework enables a robot to interact with the human in case of ambiguity between the *intended target* and the *estimated target*.

We test our method on the IPO Dataset. A sample test scene from the dataset is shown in Fig. 6a. The objects are placed 5° to 20° apart. We compute a confidence matrix which indicates the potential referents associated with the highest-scoring estimation of the pointing gesture. The confidence matrix shown in Fig. 6b illustrates the probability of an object to be selected as the estimated target while it is an intended target. The rows of the matrix represent the object intended by the user, and the columns represent the estimated object.

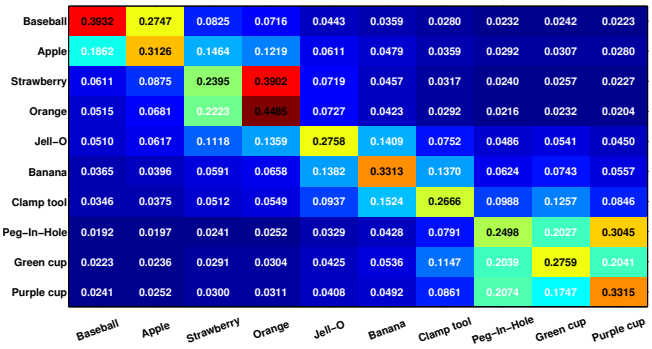
For example, when the user is pointing at the *orange*, the potential referents based on confidence are *orange* and *strawberry*. Since the confidence of *orange* is higher than that of *strawberry*, it is selected as the targeted object. But in the case where the user points at the *strawberry*, *orange* is selected over *strawberry* because of the ambiguity in one of the pose angles and *orange* being closer to the line of the direction of pointing. In such situations, the robot can interact with the user to verify which is the intended target among the estimated potential referents.

B. Experiment 2: Baseline evaluations of the IMHG dataset

The *pose/appearance* space to learn a hand gesture is created from a set of training images on the training sphere. Training images are spaced at 10° intervals in pose angles (azimuth and elevation). The kernel parameters described in [2] are maintained constant for each gesture across all participants.



(a)



(b)

Fig. 6: (a) Sample scene from the IPO dataset. (b) Confidence matrix showing the probability of an object to be selected as the target object. Rows: ground-truth objects pointed at by the user; columns: objects at estimated target locations.

Based on the variability in shape and size of the hand, a pose estimate is accepted only if the location of the estimated pose (i.e. centroid of the hand model) is within a radius of 7 cm of the ground location of the hand. There are 704 test sample pairs of deictic gestures and 22 test sample pairs of the 6 non-deictic gestures in the IMHG dataset. We compare pose estimation results using a single camera (i.e. frontal view), and multiple cameras (frontal and side views). Table III shows that our method exhibits a mean error of 8.33 cm and a standard deviation of around 5.48 cm in estimating the target location for deictic gestures with a detection rate of 76.42% when two cameras are used; this is substantially stronger than single-camera performance.

In the case of non-deictic gestures the standard deviation in estimating the location of the hand ranges from 1.45 cm to 2.21 cm, with an average detection rate of 86.36% using 2 cameras. When only a single camera is considered the method achieves an average gesture detection rate of 78.01%. It can be seen that using multiple cameras improves the performance of the system except in the case of the *Fist (stop)* gesture where the detection rate is comparable. The fist gesture shows high similarities with the forearm for some participants. The pose estimates from two views are therefore not in agreement, which results in a comparatively low detection rate in a multi-camera setup.

Figure 7 shows examples of true detections within the error range. The ground truth target locations (only for deictic gestures) are marked with a *green* circle, and the estimated locations are marked with a *blue* circle. The

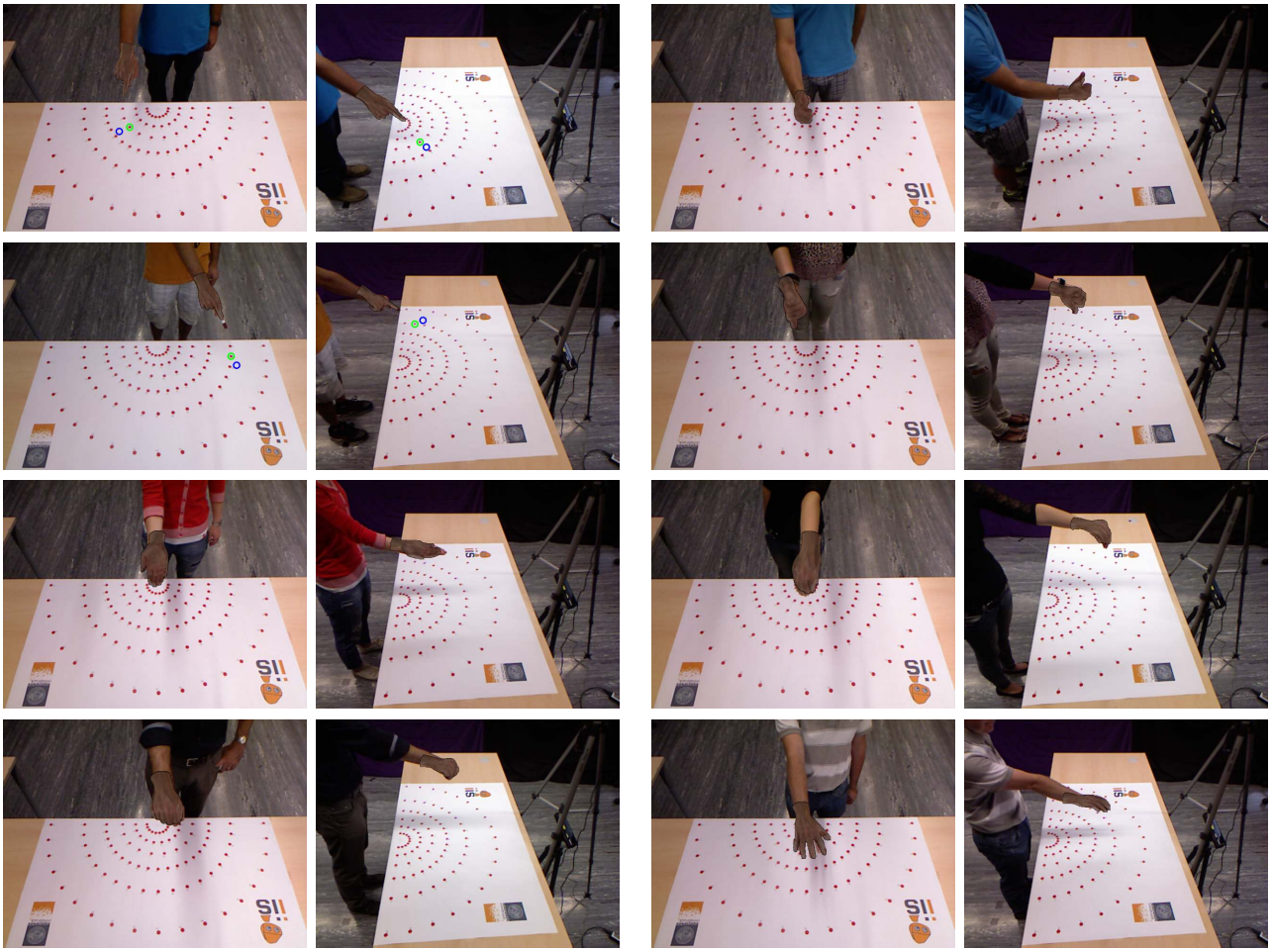


Fig. 7: Examples of true detections on the IMHG dataset. For *deictic gestures* the ground truth location is marked with a green circle, and the estimated location is marked with a blue circle. The learned gesture model is overlaid on the image at the detected hand location. *Top to bottom columnwise*: Finger pointing, Tool pointing, Give me, Fist (stop), Thumb up (approve), Thumb down (disapprove), Grasp close, Grasp open.

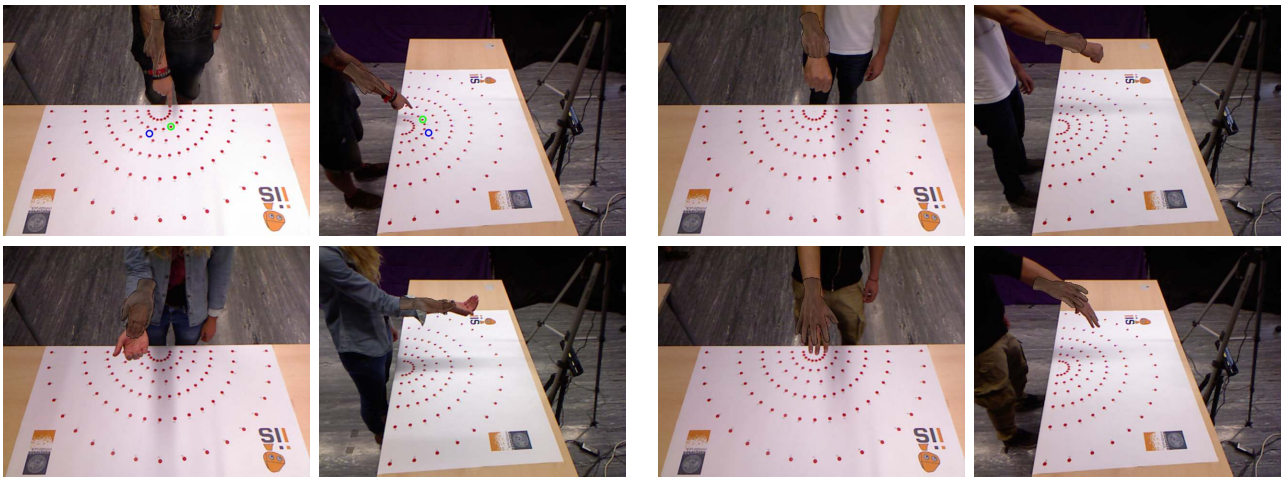


Fig. 8: Examples of false detectins on the IMHG dataset. For *deictic gestures* the ground truth location is marked with a green circle, and the estimated location is marked with a blue circle. The learned gesture model is overlaid on the image at the detected hand location. *Top to bottom columnwise*: Finger pointing, Give me, Fist (stop), Grasp open.

	Single camera			Two cameras		
	$d\,\%$	μ	σ	$d\,\%$	μ	σ
Deictic	64.20	8.87	7.52	76.42	8.33	5.48
Give me	86.36	3.13	1.89	95.45	3.65	1.49
Fist (stop)	90.91	2.47	1.14	86.36	3.34	1.82
Approve	77.17	4.23	2.04	86.36	3.57	2.21
Disapprove	68.18	4.73	1.60	81.82	3.34	1.45
Grasp close	81.82	3.28	1.12	86.36	3.57	1.45
Grasp open	63.64	3.82	1.95	81.82	3.92	1.96
Average	76.04	4.36	2.46	84.94	4.24	2.26

TABLE III: Mean error (μ) and standard deviation (σ) in cm at $d\,\%$ detection rate.

Methods	μ	σ
Elbow-hand [24]	49.0	28.0
Elbow-hand [22]	24.0	-
Shoulder-hand [23]	41.0	17.0
Shoulder-hand with GPR [24]	17.0	12.0
Head-hand [25]	39.0	17.0
Our method (hand only)	8.33	5.48

TABLE IV: Qualitative comparison with state-of-the-art methods. Mean error (μ) and standard deviation (σ) in cm.

learned gesture model is overlaid on the test scenes at the detected hand location. Results shown are 2D illustrations of the 3D estimations. Some false detection results can be seen in Fig. 8. Our method fails in the cases of high but spurious similarities between the appearance of a training model and other body parts such as the forearm.

An overall comparison of the proposed pointing gesture detection framework with the state of the art is summarized in Table IV. The test scenarios presented in the compared works are similar to that of the IMHG dataset. For example, participants are standing at a distance of approximately 1-2 m from the camera, test scenes captured under natural lighting conditions, presence of the clutter in the background, and users do not use any props like hand gloves. Qualitatively, we achieve a higher accuracy in estimating the location of the target as compared to other methods. Please note that no prior information of the object or the body pose is provided to our framework – an advantage in human-robot interactions taking place in close proximity. Moreover, the spatial area covered by target locations (*red* dots) in the IMHG dataset is comparatively smaller which further enhances the challenge of target estimation.

IV. Conclusions

We propose a probabilistic framework to detect static hand gestures and estimate their pose for human-robot interaction scenarios. The focus of this paper is to evaluate the proposed framework on the IMHG dataset as a baseline. We found that the effect of parallax can lead to false estimation of the target by the guesser. Furthermore, we evaluate the ability of our method to estimate potential referents to overcome the issue of pose ambiguities in deictic gestures. A robot can interact with the user in such situations and can react according to the next gesture like approval or disapproval. The robot can be stationary or mobile and equipped with one or multiple vision sensors.

Our framework can accommodate variability in size, shape and color of hand gestures. This enables a robot to detect gestures performed by various participants with a single hand model for each gesture. Therefore, we are independent of acquiring new training data for each participant.

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. We would like to thank Dr. Emre Ugur (University of Innsbruck) for discussions and suggestions.

REFERENCES

[1] D. Shukla, O. Erkent, and J. Piater, "The IMHG dataset: A Multi-View Hand Gesture RGB-D Dataset for Human-Robot Interaction," in *Towards Standardized Experiments in Human Robot Interactions*, 10 2015, workshop at IROS. [Online]. Available: https://iis.uibk.ac.at/public/papers/Shukla-2015-StandardHRI.pdf

[2] ——, "Probabilistic detection of pointing directions for human robot interaction," in *International Conference on Digital Image Computing: Techniques and Applications*, 11 2015. [Online]. Available: https://iis.uibk.ac.at/public/papers/Shukla-2015-DICTA.pdf

[3] B. Burger, I. Ferrané, F. Lerasle, and G. Infantes, "Two-handed gesture recognition and fusion with speech to command a robot," *Autonomous Robots*, vol. 32, no. 2, pp. 129–147, 2012.

[4] F. K. Quek, "Eyes in the interface," *Image and vision computing*, vol. 13, no. 6, pp. 511–525, 1995.

[5] E. Sato, T. Yamaguchi, and F. Harashima, "Natural interface using pointing behavior for human–robot gestural interaction," *Industrial Electronics, IEEE Transactions on*, vol. 54, no. 2, pp. 1105–1112, 2007.

[6] K. Hu, S. Canavan, and L. Yin, "Hand pointing estimation for human computer interaction based on two orthogonal-views," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 3760–3763.

[7] L. Rozo, S. Calinon, D. Caldwell, P. Jiménez Schlegl, C. Torras, *et al.*, "Learning collaborative impedance-based robot behaviors," 2013.

[8] D. Michel, K. Papoutsakis, and A. A. Argyros, "Gesture recognition supporting the interaction of humans with socially assistive robots," in *Advances in Visual Computing*. Springer, 2014, pp. 793–804.

[9] L. Jensen, K. Fischer, D. Shukla, and J. Piater, "Negotiating Instruction Strategies during Robot Action Demonstration," in *10th ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 3 2015, pp. 143–144. [Online]. Available: https://iis.uibk.ac.at/public/papers/Jensen-2015-HRI.pdf

[10] F. A. Kondori, S. Yousefit, A. Ostovar, L. Liu, and H. Li, "A direct method for 3d hand pose recovery," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 345–350.

[11] J. Lambrecht, H. Walzel, and J. Kruger, "Robust finger gesture recognition on handheld devices for spatial programming of industrial robots," in *RO-MAN, 2013 IEEE*. IEEE, 2013, pp. 99–106.

[12] A. Bangerter and D. M. Oppenheimer, "Accuracy in detecting referents of pointing gestures unaccompanied by language," *Gesture*, vol. 6, no. 1, pp. 85–102, 2006.

[13] S. R. Langton and V. Bruce, "You must see the point: Automatic processing of cues to the direction of social attention." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 2, p. 747, 2000.

[14] 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.

[15] 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.

[16] 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.

[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] 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.

[19] 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.

[20] 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.

[21] 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.

[22] B. Großmann, M. R. Pedersen, J. Klonovs, D. Herzog, L. Nalpantidis, and V. Krüger, "Communicating unknown objects to robots through pointing gestures," in *Advances in Autonomous Robotics Systems*. Springer, 2014, pp. 209–220.

[23] E. Huber and D. Kortenkamp, "A behavior-based approach to active stereo vision for mobile robots," *Engineering Applications of Artificial Intelligence*, vol. 11, no. 2, pp. 229–243, 1998.

[24] D. Droeschel, J. Stückler, and S. Behnke, "Learning to interpret pointing gestures with a time-of-flight camera," in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 481–488.

[25] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," in *Proceedings of the 5th international conference on Multimodal interfaces*. ACM, 2003, pp. 140–146.

[26] M. Šarić, "Libhand: A library for hand articulation," 2011, version 0.9. [Online]. Available: http://www.libhand.org/

[27] J. Romero, H. Kjellström, and D. Kragic, "Monocular real-time 3d articulated hand pose estimation," in *Humanoid Robots, 2009. Humanoids 2009. 9th IEEE-RAS International Conference on*. IEEE, 2009, pp. 87–92.

[28] D. Teney and J. Piater, "Multiview feature distributions for object detection and continuous pose estimation," *Computer Vision and Image Understanding*, vol. 125, pp. 265–282, 8 2014. [Online]. Available: https://iis.uibk.ac.at/public/papers/Teney-2014-CVIU.pdf

[29] R. Macknojia, A. Chávez-Aragón, P. Payeur, and R. Laganiere, "Calibration of a network of kinect sensors for robotic inspection over a large workspace," in *Robot Vision (WORV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 184–190.

[30] H. H. Clark and C. R. Marshall, "Definite reference and mutual knowledge," 1981.

[31] H. H. Clark, R. Schreuder, and S. Buttrick, "Common ground at the understanding of demonstrative reference," *Journal of verbal learning and verbal behavior*, vol. 22, no. 2, pp. 245–258, 1983.