Real-time Translation of Upper-body Gestures to Virtual Avatars in Dissimilar Telepresence Environments

Supplementary Material

TABLE I
TARGET CONFIGURATIONS FOR EACH ACTION CATEGORY

Action Category (Number of Targets)	Grid	Azimuth (°)	Height (m)	Distance (m)
Pointing and gazing at a single target (1)	$9 \times 3 \times 1$	-70 to 70 (17.5)	-0.5, 0, 0.5	1.5
Pointing a single target with gaze shift (2)	$5 \times 3 \times 1$	-90 to 90 (45)	-0.5, 0, 0.5	1.5
Explaining with pointing (2)	$7 \times 3 \times 1$	-90 to 90 (30)	-0.5, 0, 0.5	1.5
Pointing and gazing at two targets in sequence (2)	$5 \times 3 \times 1$	-70 to 70 (35)	-0.5, 0, 0.5	1.5
Transitioning between a pointing gesture and a free-form gesture (1)	$9 \times 3 \times 1$	-90 to 90 (22.5)	-0.5, 0, 0.5	1.5
Touching and gazing at a single target (1)	$5 \times 3 \times 2$	-70 to 70 (35)	0, -0.3, -0.6	0.4, 0.5
Touching a single target with gaze shift (2)	$5 \times 3 \times 1$	-90 to 90 (45)	-0.4, -0.1, 0.2	0.45
Explaining with touching (2)	$7 \times 3 \times 1$	-90 to 90 (30)	-0.6, -0.3, 0	0.45
Touching and gazing at two targets in sequence (2)	$5 \times 3 \times 1$	-70 to 70 (35)	-0.6, -0.3, 0	0.45
Transitioning between a touching gesture and a free-form gesture (1)	$9 \times 3 \times 1$	-90 to 90 (22.5)	0, -0.3, -0.6	0.45

TABLE II
LENGTH OF EACH ACTION CATEGORY FOR DIFFERENT SUBJECTS.

Action Category	Subject					
	170cm	161cm	172cm	173cm	174cm	180cm
Pointing and gazing at a single target	201s	124s	110s	146s	127s	136s
Pointing at a single target with gaze shift	2169s	55s	59s	51s	56s	67s
Explaining with pointing	1530s	95s	98s	69s	71s	117s
Pointing and gazing at two targets in sequence	1772s	56s	47s	42s	51s	53s
Transitioning between a pointing gesture and a free-form gesture	1998s	222s	280s	202s	202s	240s
Touching and gazing at a single target	254s	162s	151s	142s	150s	149s
Touching a single target with gaze shift	2541s	77s	75s	60s	57s	76s
Explaining with touching	1529s	94s	83s	78s	71s	120s
Touching and gazing at two targets in sequence	2101s	53s	72s	49s	47s	55s
Transitioning between a touching gesture and a free-form gesture	1478s	226s	249s	218s	201s	257s

A. Motion Capture

We processed calibration to track the subjects' eyes, fingers, and joints. Once finished, the subject maintained a fixed position. They then performed movements toward targets that looked like simple spheres, displayed in a virtual environment created using the Unity3D game engine. The system tracked and captured joint transformations of the head, right hand, right index fingertip, left hand, left index fingertip, and gaze direction using an HTC Vive Pro Eye headset and Noitom Hi5 VR gloves. Additionally, the positions of the head and hand targets were recorded.

The training data was collected from only one single individual with a height of 170cm and an arm length (from shoulder to fingertip) of 74cm, and the test data was collected from five individuals with heights (and arm lengths) of 161cm (63cm), 172cm (70cm), 173cm (68cm), 174cm (70cm), and 180cm (75cm). Four individuals (161cm, 170cm, 172cm, and 180cm) used their right hand for the deictic gestures, while the remaining used their left hand. In terms of eye dominance, three subjects (170cm, 172cm, and 173cm) were right-eye dominant, and the others were left-eye dominant. Table II presents a summary of the length of captured data for all subjects.

B. Target Configuration

We arranged the targets for head and hand in a grid layout (azimuth $\times$ height $\times$ distance) to ensure that upper-body gestures in the dataset were uniformly distributed over the targets. Table I summarizes the target configurations for each action category. For instance, when pointing and gazing at a single target, the targets were arranged in a $9 \times 3 \times 1$ grid. The azimuth angle spanned 140°, ranging from $-70°$ to 70° with intervals of 17.5°. The height variations were set at -0.5m, 0m, and 0.5m from the eye level. The distance was set at 1.5m.

For the actions involving explaining with a deictic gesture (either pointing or touching), the single target for the deictic gesture was selected from the grid layout as detailed in Table I. The target for the explanatory gesture was chosen from a $7 \times 1 \times 1$ grid. The azimuth angle for this spanned 180°, ranging from $-90°$ to 90° with intervals of 30°. The height was fixed at 0m from the eye level, and to accommodate explanatory gestures made while looking at another person of various heights, random values between -0.25m and 0.25m were added. The distance was set at 1.5m. In actions with two targets, two different targets from the grid layout were selected in each trial. All combination cases from the grid were included in the training data, while the test data comprised 8 randomly selected cases.