Recognition of Arm Gestures Using Multiple Orientation Sensors: Repeatability Assessment

Martin Urban, Peter Bajcsy, Rob Kooper and Jean-Christophe Lementec

Abstract—**We present a solution to repeatability assessment of an arm gesture recognition system using multiple orientation sensors. We focus specifically on the problem of controlling Unmanned Aerial Vehicles (UAVs) in the presence of manned aircrafts on an aircraft deck. Our goal is to design a robust UAV control with the same gesture signals as used by current flight directors for controlling manned vehicles. Given the fact that such a system has to operate 24 hours a day in a noisy and harsh environment, for example, on a Navy carrier deck, our approach to this problem is based on arm gesture recognition rather than on speech recognition. We have investigated real-time and system design issues for a particular choice of active sensors, such as, the orientation sensors of the IS-300 Pro Precision Motion Tracker manufactured by InterSense. Our work consists of (1) scrutinizing sensor data acquisition parameters and reported arm orientation measurements, (2) choosing the optimal attachment and placement of sensors, (3) measuring repeatability of movements using Dynamic Time Warping (DTW) metric, and (4) testing performance of a template-based gesture classification algorithm and robot control mechanisms, where the robot represents an UAV surrogate in a laboratory environment.**

I. Introduction

With the current advancements of autonomous unmanned vehicles, there is a need to support a control of unmanned and manned vehicles without interfering with the current control mechanisms of manned vehicles. For instance, the current control mechanism for manned aircrafts is based on people, called flight directors or yellow shirts, performing gestures according to a pre-defined lexicon of gestures and pilots following the gesture corresponding commands. In order to avoid changes of standard control practices and accommodate newly developed unmanned aircrafts, a problem of unmanned vehicle control using standard control procedures arises. This problem motivated our work and development.

Manuscript received on March 31, 2004. This work was supported in part by the U.S. Navy under Grant N00014-03-M0321.

M. Urban, P. Bajcsy and R. Kooper are with the National Center for Supercomputing Applications (NCSA) at University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (corresponding author P. Bajcsy, e-mail: pbajcsy@ncsa.uiuc.edu, phone: 217-265-5387, fax: 217-244-7396).

J.C. Lementec is with CHI Systems, Inc., Fort Washington, PA, USA.

In this paper, we focus specifically on the problem of controlling Unmanned Aerial Vehicles (UAV's) in the presence of manned aircrafts on an aircraft carrier deck. Our goal is to design a UAV control with the same gesture signals as used by current flight directors for controlling manned vehicles. Given the fact that such a system has to operate 24 hours a day in a noisy and harsh environment, for example, on a Navy carrier deck, our approach to this problem is based on arm gesture recognition. Speech recognition systems were not recommended due to a very noisy background environment. Thus, our objective is to investigate real-time and system design issues for a particular choice of active sensors.

Our proposed gesture recognition system is based on IS-300Pro Precision Motion Tracker by InterSense [1], and an overview diagram in Fig. 1 describes the entire system. An operator (a yellow shirt) performs a gesture, during which the tracker sensors transmit acquired data to the IS-300Pro base unit and then to a PC. Sensor outputs are analyzed and classified into corresponding commands (gesture name). Gesture commands are converted into a set of robot instructions and sent to a robot. The robot surrogate, which in our case is Pioneer II from ActivMedia [4], executes robot instructions in our laboratory environment instead of a real UAV.

In the past, researchers have approached the problem of manual robot navigation via wireless control using handheld devices [8]. If the robot (or UAV) control should be driven by flight director's gestures then computer-vision based solutions might be considered. For example, a system with single or multiple cameras acquiring a video stream could analyze images and compare temporal signatures of gestures in video frames with a set of temporal templates [7]. Nevertheless, the computer-vision based approaches for gesture recognition are not robust enough for real-time UAV navigation on a cluttered Navy deck.

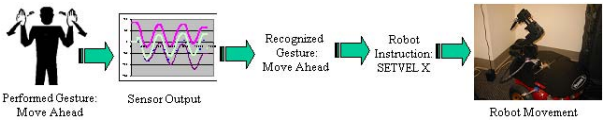


Fig. 1. Flow diagram of a developed system for robot control using hand gestures.

In this paper, we present several theoretical and experimental issues related to a design of a gesture

recognition system using the IS-300 Pro Precision Motion Tracker by InterSense. Our work consists of (1) scrutinizing sensor data acquisition parameters and reported arm orientation measurements in Section II, (2) choosing the optimal attachment and placement of sensors in Section III, (3) measuring repeatability of our experiments using Dynamic Time Warping (DTW) metric in Section IV, and (4) designing template-based gesture classification algorithm and robot control mechanism in Section V. Our work is summarized in Section VI together with an outline of challenges and future directions.

Fig. 2. a) IS300 Pro base device, b) Inertia Cube, and c) base device with 4 cubes on arm bands.

II. Sensor Data Acquisition Parameters

A. IS-300 Pro Parameters

The IS300 Pro Precision Motion Tracker is shown in Fig. 2 and it was developed for head movement tracking in virtual reality systems. The base unit of IS300 can track up to four sensor inertia cubes. We have scrutinized (1) acquisition rate (maximum tracking rate is 1200° per second, update rate is up to 500Hz), (2) measurement accuracy (RMS angular resolution is 0.02°, RMS angular accuracy is 1.0°, and RMS dynamic accuracy is 3.0°), (3) temperature range (0°C to 50°C), and (4) ruggedness (shock sensitive), in addition to size, weight and cost evaluation criteria mentioned before. All parameters were adequate for our application except from the sensor ruggedness. However, the ruggedness was not our major concern at this time. By using the IS300 Pro Precision Motion Tracker, we have also avoided the issues related to multiple sensor synchronization because the IS300-Pro base unit handles four sensors simultaneously.

B. Selection of Reported Orientation Measurements

One of the parameters scrutinized before running any experiments was the choice of reported orientation measurements, such as, (1) a 3x3 rotation matrix, (2) three Euler angles (yaw, pitch and roll), and (3) four-element quaternion. First, we decided not to use the rotation matrix because (a) it can be constructed from the Euler angles or quaternions, and (b) it requires transmitting larger number of bytes (matrix entries) than the other two representations and hence adds unnecessary communication and computational cost. Second, we evaluated the pros and cons of Euler angles and quaternions. The major advantage of Euler angle representation over quaternion representation is its easy comprehension for humans. The disadvantage of Euler angle representation is its singularity point when yaw is near 90 degrees. Third, we have investigated the transformation uniqueness between rotation matrices derived from Euler angles or quaternions, Euler angles, and four-element quaternions. This seemingly trivial issue is complicated by the fact that the IS300 Pro reports angular values in left-hand coordinate system while all computer graphics and java3D libraries use right-hand coordinate system. We have investigated conversions of (a) Euler angles to rotation matrix to Euler angles, (b) quaternions to rotation matrix to Euler angles, and (c) rotation matrix to Euler angles and quaternions. We have implemented two different methods, GEMS [6] and TRTA [5], for this purpose. To the best of our obtained knowledge, we could not find a method that would recreate identical angles in the above (a), (b) and (c) transformations.

Based on our scrutiny, we decided to (a) directly acquire Euler angles, (b) avoid any angular and coordinate system transformations by modeling gestures directly with a combination of absolute angles (roll and pitch) and relative angles (yaw), and (c) work in the left-hand coordinate system. The singularity point in Euler angle representation was compensated by an appropriate design of our classification algorithm.

III. Attachment And Placement of Sensors

Sensor placement is crucial for obtaining repeatable and gesture unique measurements. Repeatability of measurements was improved by a thorough design of tight attachment mechanisms between sensors and a sensor base, and a sensor base and human arm. Sensors were attached firmly to a plastic flat board by two screws and the board edges had openings for Velcro armband strips. To assure minimum movement of the sensors it is recommended to place armbands tightly around skin rather than around any sleeves or other clothing.

The issue of acquiring unique measurements for each gesture was addressed by investigating (1) different number of sensors per arm (two or three sensors per arm), (2) variable sensor locations and (3) several sensor orientations on a forearm or upper arm. The possible sensor placements are illustrated in Fig. 3.

We model a human arm by three connected (almost) rigid segments corresponding to the upper arm, lower arm, and wrist. A sensor mounted on each of these segments captures the full range of motion of that segment and three sensors together capture the full range of motion of the arm. However, the fact that the IS300 Pro unit can handle at most four sensors limits us to only two sensors per arm. A closer analysis of the NAVY gesture lexicon [2] suggests that the wrist segment gives the least amount of information about the whole arm orientation. Furthermore,

through numerous gesture experiments with three sensors on one arm, as shown in Fig. 3c, we determined that a user usually moves the wrist joint very rapidly with involuntary movements that are not part of the target gesture. Thus, we concluded that two sensors per arm, mounted on the upper and lower arm are adequate for majority of the gestures in the NAVY lexicon [2]. The placement of sensors along each respective arm segment is also critical to the information content of acquired data and to the gesture recognition accuracy and robustness. Placing the lower arm sensor near the wrist end allows for greater range of angles to be captured than by placing the sensor near the elbow, and this placement also gives a perfect representation for the facing direction of the palm. A larger rotational range at the wrist can be explained by the lower arm anatomy composed of two almost parallel bones. To capture the largest rotational range of the upper arm, a sensor is placed near the elbow because the flesh and muscles near the shoulder do not move as much as those close to the shoulder.

We also considered multiple initial sensor orientations with the sensors pointing either sideways (away from the body as shown in Fig. 3a, or forward (see Fig. 3b). The sideway pointing location is preferred, because the upward pointing sensor location hindered the movement of an arm as the lower arm would hit the upper arm's sensor while bending arm at the elbow by more than 90 degrees. Another reason for choosing the sideways placement is that sensors are more stable while resting on a flat surface (broader side of an arm segment) than on a highly curved surface (narrower side of an arm segment).

To prevent the weight of the connecting wires from pulling on attached sensors and making them move, it is recommended to hold the wires in a hand with enough wire between the hand and a sensor for free movement in all directions.

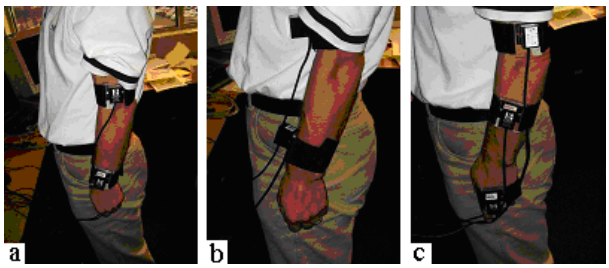


Fig. 3. Tested orientations were a) two sensors sideways, b) two sensors up, and c) three sensors sideways.

IV. Repeatability Assessment

In order to assess repeatability of gestures, we investigated characteristics of Euler angles under different sensor motions. We analyzed the outputs of individual sensors against each other, as well as, the sensor outputs of combined Euler angles from a gesture versus the sensor outputs from multiple repetitions of the same gesture, or multiple gestures. In addition to proposing a gesture repeatability metric to quantitatively measure system reliability, we also developed our own real time visualization software. The presented repeatability assessment is motivated by our effort to develop a very robust gesture recognition system that includes (a) training data selection, (b) gesture ranking based on similarity with other gestures for improving system robustness, and (c) detection of sensor failure, defect or performance deterioration.

A. Sensor Repeatability Analysis

First we assessed reliability of yaw, pitch and roll angles by comparing values from different sensors going through the same motion. For this purpose we attached two sensors to the cover of a hardback book, and positioned the book on a flat horizontal surface. We performed opening and closing of the book cover repeatedly under multiple book rotations. The average angular difference between two corresponding samples in the two sensors' data recordings was less than 2°; therefore we concluded that both sensors reported almost identical angular measurements.

To assess sensor repeatability of measurements obtained from arm motions, we experimented with mounting all sensors on the lower arm and moving the arm in order to compare multiple sensor outputs. Fig. 5 shows a couple of sensor placements, such as, one with the sensors on a side of the arm and the other with the sensors on a top of the arm. All performed arm motions consisted of raising and lowering an arm forward or sideway. Fig. 4 demonstrates the visual similarity of angular measurements obtained from repetitive motions of raising the arm sideways from vertical to horizontal position. We concluded that all sensors indeed report almost 100% identical readings when going through "theoretically" identical motion.

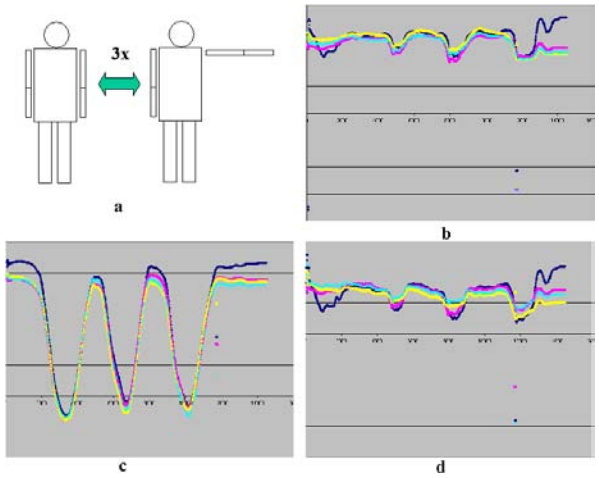


Fig. 4. Comparison of Euler angles when a) moving arm sideways up three times. Sensors were attached 4 in a row pointing up as in Fig. 5b. Angle graphs for all sensors show b) yaw, c) pitch, and d) roll.

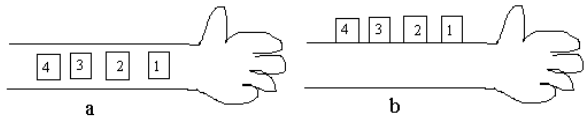


Fig. 5. Two placements of 4 sensors in a row on the lower right

B. Gesture Repeatability Metric

In order to quantitatively evaluate repeatability, we propose to use a Dynamic Time Warping (DTW) based metric that has been used in speech recognition domain [3] for matching words (sounds). The DTW algorithm accounts for different rates of the said words, which correspond to individual gestures in our case. First, the DTW algorithm finds the difference between two recordings of gestures one angle at a time, resulting in a numerical error for each angle. Second, the algorithm compares gesture-A (x-axis) with gesture-B (y-axis) by going only forward in time and making the best match at each sample pair (i, j), where i is a time sample from gesture-A and j is a time sample from gesture-B. To find the smallest value at each (i, j), the local distance is calculated first between the samples of gesture-A(i) and gesture-B(j), and then added to the lowest cumulative global distance from one of the three possible previous coordinates according to Eq. (1).

$$D_{i,j} = d_{i,j} + \min\left(D_{i-1,j}, D_{i-1,j-1}, D_{i,j-1}\right) \quad (1)$$

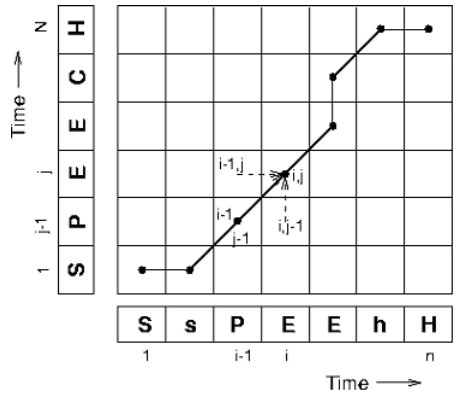


Fig. 6. An illustration of DTW algorithm used to compare two instances of the word "speech". In our case the letters are replaced by angle measures. The picture shows the shortest global path from beginning to end, as well as the calculation of error at coordinates (i, j).

Fig. 6 shows an illustration of DTW error computation for the in word "speech". In our case, the local distance $d_{i,j}$ is calculated according to (1) between one Euler angle from gesture-A and the corresponding Euler angle from gesture-B. $D_{i,j}$ is the overall error at times i and j for the chosen Euler angle in the two gestures. The final error E_{DTW} between two gestures for one Euler angle is the last computed $D_{i,j}$, where i and j are the final samples in their respective gestures. E_{DTW} corresponds to the upper right corner of the illustration in Fig. 6. By considering one Euler angle at a time, we obtain 12 different global errors E_{DTW} (3 angles for each of the 4 sensors). The total DTW based error E_T for a pair of gestures is then computed by summing all global errors according to (2).

$$E_T = \sum_{i=1}^{12} E_{DTW}(i) \quad (2)$$

TABLE I
RANKINGS OF 20 GESTURES FROM LEXICON
(1) MOST SIMILAR TO OTHERS, (20) MOST DIFFERENT FROM OTHERS

Rank	Gesture name
1	Turn To Right
2	Turn To Left
3	Launch Bar Up
4	Up Hook
5	Down Hook
6	Move Ahead
7	Disengage Nose-gear Steering Left
8	Fold Wings
9	Launch Bar Down
10	Move Back
11	Spread Wings
12	Engage Nose-gear Steering Left
13	Pivot To Left
14	Pivot To Right
15	I Have Command (Yellow shirt's left arm is up)
16	Brakes
17	Slow Down
18	Pass Control (To yellow shirt's left)
19	Stop
20	Slow Down Engines on Right

All directions refer to the orientation of the pilot in the aircraft, unless otherwise noted.

C. Experimental Results

While conducting gesture comparisons, an arbitrary percentile of the same gesture's recordings can be used as training data or as templates for classification. To select the best templates, six runs of each gesture were recorded first. Then, all gesture recordings are aligned by manually identifying their beginning and their end. Afterwards, DTW errors are computed for all pairs of gesture recordings. In this experiment, we set our objective to find three most similar gestures. All-possible combinations of triplets (6 choose 3) were evaluated by summing up the three pair-wise total errors for each triplet. For example, given the triplet of gesture sets 1, 2, and 3, the sum of total

TABLE II
CHART WITH TOTAL DTW ERRORS OF 6 "MOVE AHEAD" TRIALS

	2	3	4	5	6
1	21847	22242	26555	32440	28448
2		**10783**	18124	19220	**16395**
3			19880	18179	**16393**
4				22184	16918
5					18786

Bold number show that trials 2, 3, and 6 are most similar

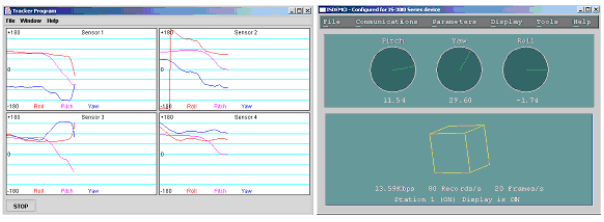
errors is equal to $E_T(1,2) + E_T(1,3) + E_T(2,3)$ (see Table II). Minimization of the sum of total errors leads to the optimal selection of training data.

We also compared all 20 gestures from the NAVY lexicon [2] to each using the DTW based metric. The results showed that gestures "turn to left" and "turn to right" appear to have the highest similarity to the rest of the gestures and hence might be misclassified more likely than other gestures. However, the E_T values for these two gestures are larger than the E_T values for gesture repetitions by a factor of at least three. The E_T value for a pair of the gestures "slow down" and "slow down engine on indicated side" leads to the largest value among E_T values for all pairs of gestures and therefore these two gestures should be classified with the highest confidence. We also ranked all 20 NAVY gestures based on the global error E_T. The ranking of gestures from least repeatable to most repeatable is shown in Table I and the error values are shown in Fig. 8. A lower value of E_T means that a gesture is more similar to the rest and thus less repeatable.

D. Data Visualization

All data are captured in sets, with each set containing a timestamp (milliseconds from power on or last reset), and 3 (Euler angles) or 4 (quaternion) values for each of the four sensors. The baud rate of the connection between the PC and the base unit determines the length of the time intervals, which are about 5-10 milliseconds at the highest setting of 115,600 baud. The lowest available transmission rate is 9,600 baud. It is reasonable to down-sample the collected data to about 10 samples per second since a human arm cannot make significantly large movement changes in such small time intervals. Down-sampling also reduces computational requirements for real-time gesture classification.

To help us assess the measurement repeatability, we created our own data visualization. Our visualization runs in a real-time capture mode and allows a visual inspection of any irregularities. The developed visualization software enables a user to choose and display any or all of the angles and sensors in multiple windows. This type of visual inspection is not possible with the IS300 demo software that comes on a CD with the hardware since it is only showing the orientation of one inertia cube at a time. An example of the visualizations is in Fig. 7.



a b

Fig. 7. Developed visualization (a) vs. Intersense demo visualization (b).

V. Classification And Robot Control

A. Classification

We tested a template-based gesture classification using the DTW similarity metric. Six repeated recordings were taken for each of 11 different gestures (turn to right, turn to left, launch bar up, move ahead, pivot to right, pivot to left, brakes, slow down, pass control, stop, slow down engines). All the recordings used were manually parsed to only contain the time period from the exact beginning to the exact ending of the performed gesture. In each set of six gesture repetitions, the DTW error was calculated for each pair of gestures and for each Euler angle separately, and then the sum of all DTW errors was compared against the other five gesture replicas. The gesture recording with the lowest global DTW values was chosen as the template for that gesture. Four more recordings of each of the eleven different gestures were acquired to form 44 angular streams of test data. These 44 recordings were compared to the 11 templates and classified using the DTW similarity metric with 91% accuracy. By inspecting the DTW values for the best matching templates, we selected a threshold of 65,000 for separating known and unknown gestures. By using this threshold and selecting only five of the eleven templates (16-20 in Table I or the best five according to gesture rankings: brakes, slow down, pass control, stop, slow down engines), the classification accuracy has improved to 95%. The results are shown in Table III.

TABLE III
CLASSIFICATION WITH 5 TEMPLATES, AND THRESHOLD OF 65,000

Rank	1	2	3	4	5	Unknown
1	4					
2		4				
3			4			
4				4		
5					4	
6			**1**			3
7						4
8			**1**			3
9						4
10						4
11						4

Bold values show incorrect classification.

The template based gesture classification method is very time consuming, and grows linearly with the number of templates used. This makes it practically unusable for real time gesture classification and it was used only for testing. We developed a real-time robust gesture classification technique that is not described here.

B. Robot Control

The last component of a gesture recognition system for controlling UAVs is the mapping between the NAVY gesture lexicon [2] and robot movement instructions. Given a set of robot instructions, this part was straightforward and a few examples of gesture mappings are shown in Table IV.

		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Brakes	1	80	90	128	94	87	87	85	81	108	108	95	99	120	168	100	110	80	67	70
Disengage Nosegear	2		65	71	81	78	85	61	67	78	104	67	96	96	143	92	142	54	63	67
Down Hook	3			72	73	97	69	60	85	69	102	99	82	104	102	76	124	72	66	48
Engage Nosegear	4				94	85	105	67	75	91	106	97	107	103	129	100	139	47	66	85
Fold Wings	5					115	65	46	78	92	94	75	101	93	157	67	92	77	73	72
I Have Command	6						106	85	97	88	101	91	110	112	77	109	165	63	81	74
Launch Bar Down	7							52	71	94	108	110	71	123	122	80	97	95	81	44
Launch Bar Up	8								62	79	104	93	93	94	113	72	101	69	60	58
Move Ahead	9									70	91	82	82	90	145	74	106	63	62	81
Move Back	10										110	84	77	78	103	80	142	66	60	93
Pass Control	11											110	116	107	152	105	105	97	112	95
Pivot to Left	12												108	87	127	96	137	40	85	104
Pivot to Right	13													109	126	86	140	73	44	71
Slow Down	14														150	68	106	86	97	126
Slow Down Engines on side	15															152	220	123	96	88
Spread Wings	16																88	76	70	91
Stop	17																	126	126	127
Turn to Left	18																		41	81
Turn to Right	19																			69
Up Hook	20																			

Fig. 8. A gesture dissimilarity matrix formed by comparing all pairs of gestures from the NAVY lexicon using the proposed DTW metric. Small values indicate high similarity. Color-coded entries show values below 70,000 (green); or below 60,000 (yellow); or below 50,000 (orange).

TABLE IV
EXAMPLES OF MAPPINGS FROM GESTURES TO ROBOT INSTRUCTIONS.

Gesture	Robot Instruction
Move Ahead	SETVEL 40
Turn To Left	SETVEL2 30 40
Turn To Right	SETVEL2 40 30
Brakes	STOP
Pivot To Left	SETVEL2 –30 30
Pivot To Right	SETVEL2 30 –30
Slow Down	MULTVEL 0.8
Move Back	SETVEL –40
Slow Down Engines on Left	MULTVELL 0.8

VI. CONCLUSION

From the analysis of different sensor placements we optimized the sensor orientation. We also showed that the gestures are very repeatable by comparing them using a DTW-based metric. Finally, we concluded that the gestures could be recognized with a template based classification method.

During the experiments, we have resolved several challenges related to (1) yaw variation due to flight director's orientation, (2) angular offset due to sensor attachment, and (3) singularity points of Euler angles. First, processing only relative values compensated the yaw variation. Second, the angular offset was detected by periodically running repeatability experiments and mending any loose sensor attachment. Third, an appropriate gesture modeling compensated the occurrence of singularity points.

In future, we will search for sensors that are wireless and smaller. These new sensor features could solve some problems with (1) tangled wires, (2) the need to be in the vicinity of the base device, and (3) the fatigue of a flight director caused by bulky sensors. We might conduct additional experiments to prove that the active sensing approach based on orientation sensors is height, size, shape, and gender of flight directors independent. The hand gesture tracking can be used anywhere where communication by sound is impossible, either due to requirements of silence such as in covert commando operations, or in loud places like construction sites. Another possible application of this technology in conjunction with human voice synthesis could help mute persons to better communicate with people that do not understand sign language.

REFERENCES

[1] Intersense web site: http://www.intersense.com/

[2] US Navy, "Field Manual FM1-564 Appendix A", web site: http://www.adtdl.army.mil/cgi-bin/atdl.dll/fm/1-564/AA.HTM

[3] Wrigley Stuart N. "Speech Recognition by Dynamic Time Warping", web site: http://www.dcs.shef.ac.uk/~stu/com326/

[4] ActivMedia support web site: http://robots.activmedia.com

[5] Gregory G. Slabaugh, "Computing Euler angles from a rotation matrix" (denoted as TRTA implementation from: http://www.starfireresearch.com/services/java3d/samplecode/FlorinEulers.html)

[6] Paul Heckbert (editor), Graphics Gems IV, Academic Press, 1994, ISBN: 0123361559 (Mac: 0123361567), (the GEMS method refers to a particular implementation of "Euler Angle Conversion" by Ken Shoemake, shoemake@graphics.cis.upenn.edu)

[7] Bobick A.F. and J. W. Davis, " The Recognition of Human Movement Using Temporal Templates," IEEE Trans. On PAMI, VOL. 23. NO. 3, March 2001, pp. 257-267

[8] Fong T., "Collaborative Control: A Robot-Centric Model for Vehicle Teleoperation," Ph.D. Dissertation, CMU-RI-TR-01-34, November 2001 (156 p).