Bayesian Model for Multimodal Sensory Information Fusion in Humanoid Robot

Received 23rd January 2010; accepted 15th September 2010.

Wei Kin Wong, Chu Kiong Loo, Tze Ming Neoh, Ying Wei Liew, Eng Kean Lee

Abstract:

In this paper, the Bayesian model for bimodal sensory information fusion is presented. It is a simple and biological plausible model used to model the sensory fusion in human's brain. It is adopted into humanoid robot to fuse the spatial information gained from analyzing auditory and visual input, aiming to increase the accuracy of object localization. Bayesian fusion model requires prior knowledge on weights for sensory systems. These weights can be determined based on standard deviation (SD) of unimodal localization error obtained in experiments. The performance of auditory and visual localization was tested under two conditions: fixation and saccade. The experiment result shows that Bayesian model did improve the accuracy of object localization. However, the fused position of the object is not accurate when both of the sensory systems were bias towards the same direction.

Keywords: *multimodal, Bayesian fusion, fixation, saccade, humanoid robot.*

1. Introduction

Realizing audiovisual object localization in humanoid robot is a challenging work. As the possible locations of the object are computed based on image and sound independently, the results need to be fused to form the ultimate perceptions. This Sensory information fusion has the advantages of reducing uncertainties of the information, providing information that is unavailable from single type of sensor and causing the system to be fault tolerant [8]. In years, researchers have developed sophisticated methods for multisensory information fusion for robot. For example, Yong-Ge Wu *et al.* [10] proposed an information fusion algorithm based on generalized Dempster-Shafer's theory of evidence (DSTE). Their result shows that higher accuracy is achieved when the vision evidence is dependent which violate the basic assumption of DSTE. J.A. Lopez-Orozco's *et al.* [11] proposed an enhanced Kalman filter multi-sensor fusion system that is used to calculate the position and orientation of an autonomous mobile robot. They focused on simplifying the Kalman filter multi-sensor fusion to lower its computational cost and solving the dependency on the selection of different sampling period and assimilation waiting interval. Kentaro Toyama's *et al.* [12] developed a Markov dynamic network model that integrates the analyses of multiple visual tracking algorithm to enable better head tracking. Their analysis shows that Bayesian system fusion usually outperforms any of its constituent systems, often making estimates close to the system estimate with the least error. Futoshi Kobayashi's *et al.* [13] proposed a Recurrent Fuzzy Inference (RFI) with recurrent inputs and applied it to a multi-sensor fusion system in order to estimate the state of systems. The membership functions of RFI are expressed by Radial Basis Function (RBF) with insensitive ranges and the shape of the membership functions can be adjusted by a learning algorithm. Lucy Y. Pao *et al.* [14] developed a multi-target tracking algorithm based on Joint Probabilistic Data Association (JPDA) for use in multi-sensor tracking situation. They found that the multisensory data association probability is actually the product of the single-sensor data association probability.

Instead of fusing spatial information of the object gained, several developed humanoid robot uses auditory information only as a guide for visual system. For example, Carlos Beltrán-González *et al.* [15] proposed a cross-modal perceptual architecture that can segment object objects based on visual-auditory sensorial cues, by constructing an associative sound-object memory and create visual expectation using a sound recognition algorithm. In Hans-H. Bothe's *et al.* paper [16], they described a hierarchically organized technical system performing auditory-visual sound source localization and camera control. In their fusion system, auditory maps that belongs to one video interval is fused. Then, the resultant map is filtered by three types of filters, and the results are fused again. Finally, the fovea position of the object is calculated using quadratic prediction. Hiromichi Nakashima *et al.* [17] proposed a learning model for sound source localization through interactions between motion and audio-visual sensing. The model consists of two modules, which are a visual estimation module consisting of a three-layer perceptron and an auditory estimation module consisting of a neural network with a Look-Up-Table algorithm.

Different from fusion models discussed above, in the field of neuroscience, recent researches [1-5] strongly suggest that multisensory cues, especially spatial cues may be combined in a much simpler statistically optimal manner in the brain. For example, Paola Binda *et al.* [2] conducted test on human's ability of localization of visual, auditory and audiovisual stimulus during the time of fixation and saccade. They assumed that saccade has little effect on auditory space perception and both auditory and visual spatial cues are conditionally independent. They showed that the result of fusion of auditory and visual spatial cue can be well modeled by Bayes's theorem and the precision of localization is better in multimodal presentation compare to unimodal presentation. At the same time, their result also showed that auditory signal becomes more important for localization at the time of saccades, suggesting that the visual signal has become transiently

noisy, and therefore receives less weight. In order to adapt this Bayesian model of multisensory data fusion into robot, the weights of all sensory systems under different conditions such as fixation or saccade must be determined. This can be done by conducting experiment to measure the level of erroneous of them.

In this paper, we focus on adopting the Bayesian fusion model into humanoid robot to fused spatial properties of an object detected by visual and auditory sensors. We also proposed a way to calculate weights of visual and auditory system based on the error of localization obtained through experiment, which is the main contribution of this paper. In order to reduce complexity of calculation, only azimuth position of the target object is considered.

The visual system of the robot locate the target object based on its dominant color, determined using color segmentation process that involve log-polar transformation, temporal difference and hue-saturation (H-S) histogram. The auditory system locate the target object using interaural time difference (ITD), which is the difference in time for the sound to reach the left and right microphone. ITD is determined using generalized cross-correlation (GCC) method performed in frequency domain. Note that unimodal object localization refers to object localization using only aidutory or visual system and bimodal object localization utilize information from both sensory system.

In section 2 of this paper, the Bayesian fusion model is described. The description of humanoid robot, the experiment and conclusion are available in section 3, 4 and 5, respectively.

2. Bayesian Fusion Model

In short, Bayes' theorem can be described as a way of converting one condition probability to other, by reweighting it with the relative probability of the two variables [1]. In this paper, the weights of sensory system during fixation and saccade were computed base on SD of localization error.

In order to derive the relationship between signal inputs, weights and the resultant output, let's assume that the independent sensor output is denoted by vector $X = (x_1, x_2, \ldots, x_m)$ and the object property (e.g. three-dimensional position) is denoted by $Y = (y_1, y_2, \ldots, y_m)$, so that $p(X|Y)$ is the probability of sensor output being X given that the object property is Y, and $p(Y|X)$ is called the posterior probability of object property being Y given that the sensor output is X. In this particular application, $p(X|Y)$ can be computed from data collected during tests, while $p(Y|X)$ is the desired outcome. These two probabilities are related by Bayes' theorem as follow [8]:

$$p(Y|X) = \frac{p(Y|X)p(Y)}{p(X)} \tag{1}$$

where the marginal probability $p(X)$ and the prior probability $p(Y)$ are the unconditional probabilities of the sensor output and object property being X and Y respectively. Then, assume that in the system there are k sensors, which give the following readings: $\chi = (X_1, X_2, \ldots, X_k)$ and the best estimate of the object property Y can be developed using these k sensors reading. This can be achieved by using the likelihood estimate. In the likelihood estimate Y is computed such that the following is maximized:

$$p(\chi|Y) = \Pi_{i=1}^{k} p(X_i|Y) \tag{2}$$

It is usually easier to deal with the logarithm of the likelihood than the likelihood itself, because the product can be changed to the sum, and the term involving exponents can be simplified. Let $L(Y)$ be the log-likelihood function:

$$L(Y) = \log p(\chi|Y) = \Sigma_{i=1}^{k} \log p(X_i|Y) \tag{3}$$

Assume that the reading from the sensors follow Gaussian density function, so that $p(X_i|Y)$ is given by:

$$p(X_i|Y) = \frac{1}{(2\pi)^{n/2}|C_i|^{1/2}} e^{(-\frac{1}{2}(x_i-Y)^t C_i^{-1}(x_i-Y))} \tag{4}$$

where C_i is the variance-covariance matrix, t denotes the transpose, and $||$ denote the determinant. Now, the expression for likelihood (Eq. 3) becomes:

$$L(Y) = \Sigma_{i=1}^{k} - \tfrac{1}{2}\log[(2\pi)^n|C_i|] - \tfrac{1}{2}(X_i - Y)^t C_i^{-1}(X_i - Y) \tag{5}$$

The best estimate $\bar{Y}$ of Y can be found by differentiating L with respect to Y, equating the result to zero, and computing the value of Y, as follows:

$$\bar{Y} = \frac{\Sigma_{i=1}^{k} C_i^{-1} X_i}{\Sigma_{i=1}^{k} C_i^{-1}} \tag{6}$$

Consider a system that contains only two sensors and let both $\bar{Y}$ and X_i in Eq. 6 be scalar measurements and let C simply be the SD of localization error of the sensor, Eq. 6 is then becomes:

$$y = \frac{\frac{1}{\sigma_1^2}x_1 + \frac{1}{\sigma_2^2}x_2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \tag{7}$$

In our case, symbol S is used to denote the azimuth position, so that $y = S_{v,a}$ be the best estimation of object azimuth position, $x_1 = S_v$ and $x_2 = S_a$ be the visual and auditory spatial information in azimuth plane computed by visual and auditory system. Then, the Eq. 7 is rewritten as:

$$S_{v,a} = \frac{\frac{1}{\sigma_v^2}S_v + \frac{1}{\sigma_a^2}S_a}{\frac{1}{\sigma_v^2} + \frac{1}{\sigma_a^2}} \tag{8}$$

Clearly, Eq. 8 shows that result of bimodal localization is actually the weighted sum of the results of two unimodal localization, where the weight of visual and auditory system (w_v and w_a) [2] are defined as:

$$w_v = \frac{\frac{1}{\sigma_v^2}}{\frac{1}{\sigma_v^2} + \frac{1}{\sigma_a^2}} \tag{9}$$

and

$$w_a = \frac{\frac{1}{\sigma_a^2}}{\frac{1}{\sigma_v^2} + \frac{1}{\sigma_a^2}} \tag{10}$$

Thus, Eq. 8 can be further simplified into

$$S_{v,a} = w_v S_v + w_a S_a \quad (11)$$

The predicted bimodal threshold $\sigma_{v,a}$ is then given by

$$\sigma_{v,a} = \sqrt{\frac{\sigma_v^2 \sigma_a^2}{\sigma_v^2 + \sigma_a^2}} \quad (12)$$

Clearly, according to Eq. 11, SD denotes the reliability of the sensory system. A reliable sensory system with lower SD will be assigned a larger weight and will become more important in estimating the target's properties under certain condition (e.g. fixation and saccade).

3. E-Bee the humanoid robot

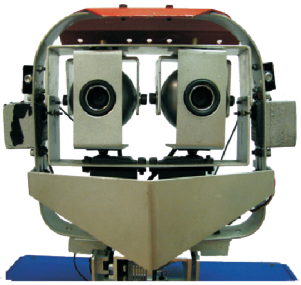


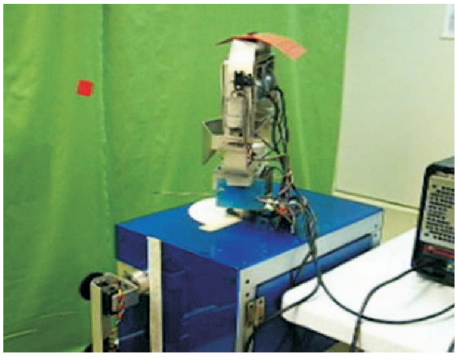
Fig . 1. E-Bee's head.

The humanoid robot used in the experiment is called E-Bee, shown in Fig. 1. Its head consists of 5 degree of freedom (D.O.F) and is driven by servomotors. These servomotors are powered by a unit of D.C. power supply and are control by a SSC-32 servomotor controller connected to the computer. The computer contains an Intel Xeon E5335 Quad-core processor, 2 GB RAM, 100 Mbps network controller, a standard 44.1kHz soundcard and a 128 MB graphic card. A pair of Logitech Quickcam Express and a pair of Sony ECM PC50 omni-directional microphones were used in the robot to grab image and sound. The Linux base asymmetric multiprocessing (AMP) system is equipped with Intel open source computer vision library (OpenCV) and Intel Integrated Performance Primitives (Intel IPP) which contains C/C++ Compiler, Math Kernel Library and Signal Processing Library.

4. Experiment

In order to study the Bayesian model for sensory information fusion and to compute the weight of visual and auditory system during the time of fixation and saccade, tests were conducted to obtain the SD of these two sensory systems. In these tests, only azimuth position of the stimulus was considered. We assumed that error and sampling period of auditory and visual system were independent and time lag between visual and auditory signal can be ignored. Background of image was a static color background in green color, located in front of the humanoid robot, with a distance of 60 cm from the eye-center, as shown in Fig. 2. The visual stimulus was presented as a 2 cm x 2 cm red square on the green background and the auditory signal was presented as continuous white noise, played through a normal speaker. Fluorescent light was used with no additional light source such as sun light. Due to hardware limitation, the neck was turned relatively to the stimulus to make the audiovisual stimulus move horizontally in the image captured, instead of driving the stimulus to move relative to the head. The range of head's motion was limited to 30° to the left and right. The sampling rate of the program was 6 Hz and the eyes and ears of the robot share the common center in horizontal plane.

(a)



(b)

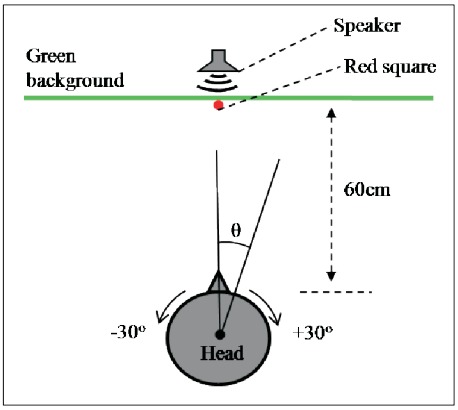


Fig. 2. (a) The robot facing forward which is the direction of the red square and white noise and (b) top view of the experiment setup, θ indicates the azimuth angle with respect to head center (modified from [9]).

Total of five tests were conducted in this experiment. They are fixation test, 20 deg/sec saccade test, 30 deg/sec saccade test, 60 deg/sec saccade test and condition specific calibration saccade test. During fixation test, the robot's head and eyes were initially fixated towards the background at position (0°, -20°) (elevation angle of 0o and azimuth angle of -20°; positive value indicates right and up of the humanoid robot while negative value indicates left and down). The head was then turned with 1o stepping and fixate for 1 sec after each step, until it reach (0°, +20°) and back to (0°, -20°). The test was repeated three times. Perceived positions of auditory and visual stimulus were recorded each time the robot head fixated. During saccade test, the neck was turned by three different angular speeds (20 deg/sec, 30 deg/sec and 60 deg/sec). The robot head was firstly fixate at (0°, -30°). Once the tests start, the head was turned between position (0°, +30°) and (0°, -30°) and only the data that fell within the range (20° to the left and right) were recorded. Each test was repeated 50 times in order to minimize the influence of random hardware error and the motion range was extended to 30° at both sides during sac-cade test to avoid obtaining data during saccade onset and changes of direction of motion because images was noisy during these conditions. After the results of these saccade test were obtained, the robot system was recalibrated based on the amount of error obtained during 60 deg/sec saccade test. Then, the condition specific cali-

bration test was car-ried out by using similar setting with 60 deg/sec saccade test.

The robot head was assumed to turn with a constant speed until the end of motion after a short time from saccade onset and changes of direction of motion. The actual azimuth position of robot head was estimated using a simple mathematic equation shown below:

$$\theta = n(30 - \omega t) \tag{13}$$

$$n = \begin{cases} -1 & \textit{for motion from left to right} \\ 1 & \textit{for motion from right to left} \end{cases}$$

where θ denote the azimuth position of the stimulus, ω denote the angular speed and t denote the time relative to saccade onset of each turn.

Fig. 3 demonstrates the result of all five tests. The result of fixation test indicates that during fixation, auditory and visual localization was accurate with little error of object localization. According to the results of 20 deg/sec, 30 deg/sec and 60 deg/sec saccade test, the error of unimodal object localization increased as the angular speed of robot head was increased. This is because the images captured were blur and not reliable during the time of saccade. This phenomenon affected the result more and more significantly as the angular speed of the robot head increased, especially during the time saccadic motion starts and during changes of direction of motion.

Table 1 and Fig. 4 summarize the calculated mean and SD of the error of unimodal and bimodal object localization. Table 1 also summarize the calculated weights of unimodal object localization under different conditions. The bimodal threshold, $\sigma_{v,a}$ was calculated by comparing the result of bimodal localization with the ideal position, rather than using Eq. 12. According to the results in Table 1, it is clear that Bayesian model of sensory information fusion did improve the accuracy of object localization as the mean error of bimodal object localization were smaller then the mean error of object localization by visual system at all time. Also, during all three saccade tests, auditory system has become more reliable as it carried lower error than visual system. However, the results of object localiza-tion by auditory system during all saccade tests were still less reliable since they deviated greatly (2° to 7°) from ideal position throughout the tests. As a result, the result of bimodal object localization was distorted greatly when both sensory systems were not accurate. This has clearly demonstrates the disadvantage of Bayesian model that, during the absence of reliable sensory system such as auditory system in human [2], the result of Bayesian model could become inaccurate as well.

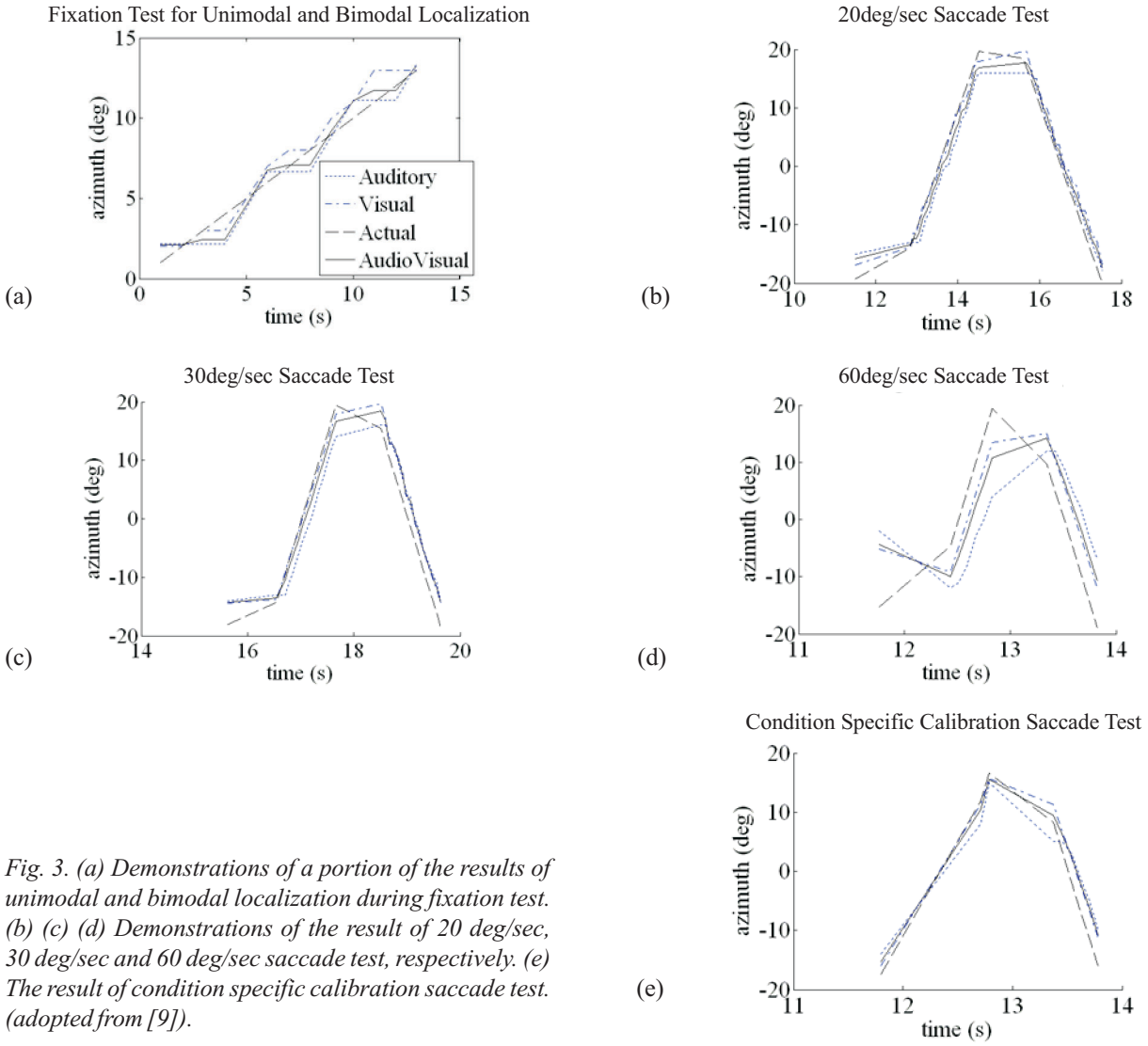
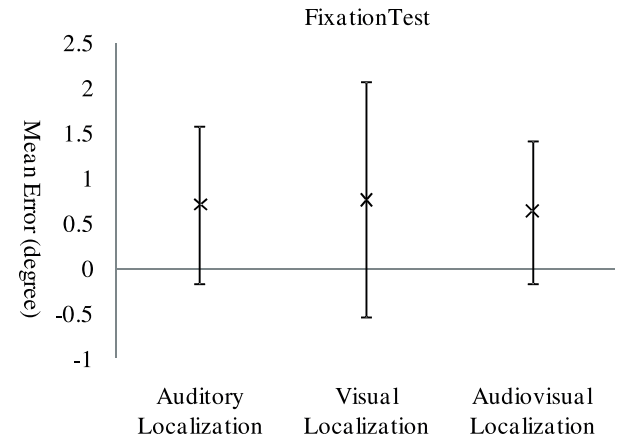


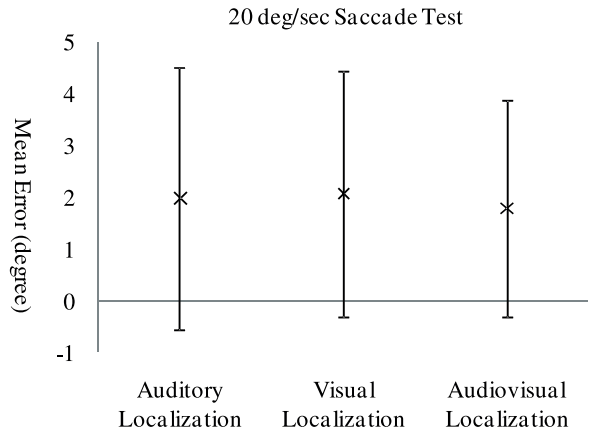
Fig. 3. (a) Demonstrations of a portion of the results of unimodal and bimodal localization during fixation test. (b) (c) (d) Demonstrations of the result of 20 deg/sec, 30 deg/sec and 60 deg/sec saccade test, respectively. (e) The result of condition specific calibration saccade test. (adopted from [9]).

Table 1. Comparison of mean and SD of error of localization of auditory, visual and Bayesian audiovisual localization, as well as the weight of unimodal localization [9].

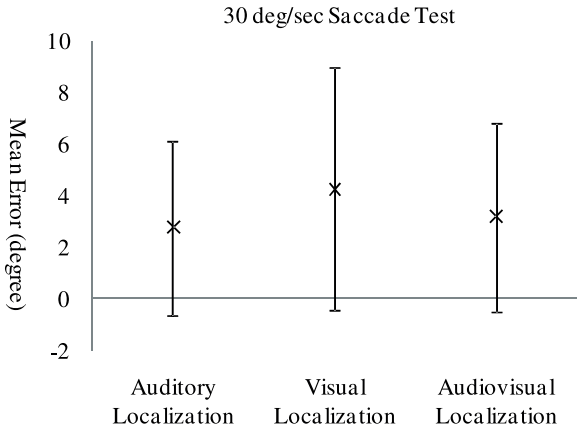
Test	Auditory localization			Visual localization			Bayesian audiovisual localization	
	Mean ($\bar{x}_a$)	SD (σ_a)	Weight (w_a)	Mean ($\bar{x}_v$)	SD (σ_v)	Weight (w_v)	Mean ($\bar{x}_{v,a}$)	SD ($\sigma_{v,a}$)
Fixation	0.7159	0.8732	0.6897	0.7683	1.3020	0.3103	0.6341	0.7905
20deg/sec saccade	1.9910	2.5389	0.4636	2.0713	2.3605	0.5364	1.7941	2.1015
30deg/sec saccade	2.7428	3.3786	0.6584	4.2705	4.6905	0.3416	3.1689	3.6338
60deg/sec saccade	6.6249	6.9261	0.7239	10.7557	11.2159	0.2761	7.7643	7.9769
Condition specific calibration saccade	1.9197	2.3324	0.7132	2.9261	3.6779	0.2868	1.9492	2.3861



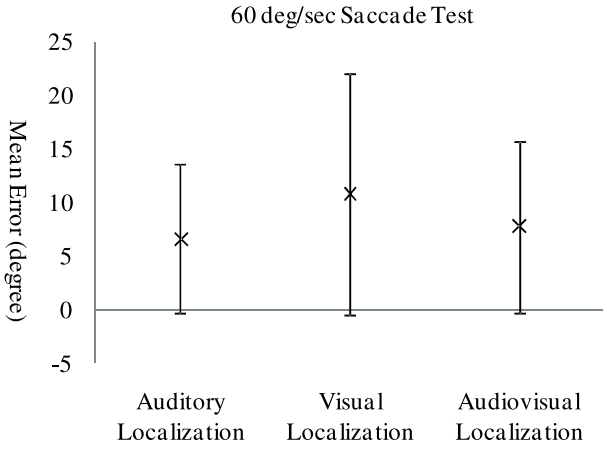
(a)



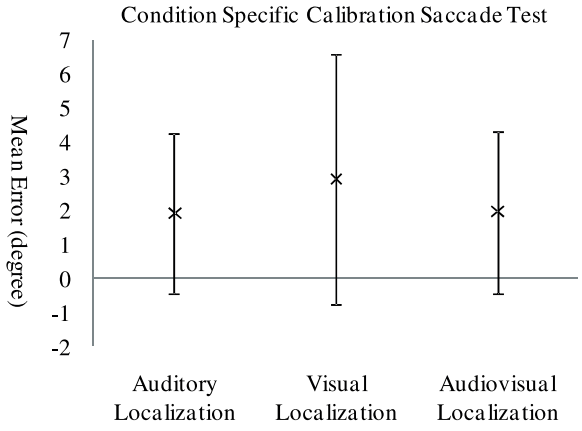
(b)



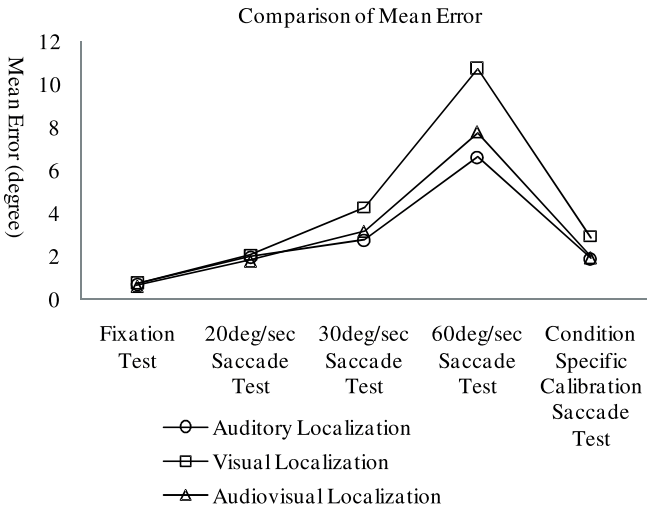
(c)



(d)



(e)



(f)

Fig. 4. (a)(b)(c)(d)(e) Demonstrations of mean and SD of auditory, visual and audiovisual localization under different conditions. (f) Comparison of mean error of localization. The mean error increased when saccade speed increased in all tests, except in the last test where condition specific calibration helped to lower the mean error.

Hence, in order to overcome this weakness of Bayesian model, condition specific calibration was proposed. After the condition specific calibration, the mean and SD of error of unimodal object localization became much smaller (see Fig. 4) compare to the 60 deg/sec saccade test. However, the disadvantage of this solution was that, the humanoid robot must always be aware of the angular speed of its head and condition specific calibration value must be known. Similar to biological system, altering robot design will eventually alter the weight of all sensory system.

5. Conclusion

In this paper, the Bayesian model for multimodal sensory information fusion is described. We show that Bayesian fusion of multisensory information is simple and can be applied in humanoid robot to increase the accuracy of object localization. However, current model requires prior knowledge on the reliability of sensory system obtained through a series of experiments. Besides, the Bayesian model failed to generate accurate spatial position of audiovisual stimulus when both of the sensory systems were not reliable, especially during saccade. Future work should focus on using neural network and reinforcement learning to provide the Bayesian fusion model the ability to learn the weights online.

ACKNOWLEDGMENTS

This research was funded by Intel Research Grant from year 2007 to 2009.

AUTHORS

Wei Kin Wong, Tze Ming Neoh, Chu Kiong Loo* - Centre for Robotics and Electrical Systems, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia. E-mails: kin2031@yahoo.com, neohtm@gmail.com, ckloo@mmu.edu.my.

Ying Wei Liew, Eng Kean Lee - Intel Malaysia Sdn. Bhd., Bayan Lepas Free Industrial Zone, 11900 Penang, Malaysia. E-mails: ying.wei.liew@intel.com, eng.kean.lee@intel.com.

* Corresponding author

References

[1] Knill D.C., "Bayesian models of sensory cue integration". In: Kenji Doya, Shin Ishii, Alexandre Pouget, Rajesh P. N. Rao, *Bayesian Brain: Probabilistic Approach to Neural Coding,* The MIT Press, Cambridge, 2007, pp.189-206.

[2] Binda P., Bruno A., Burr D.C., Morrone M.C., "Fusion of visual and auditory stimuli during saccades: a Bay-esian explanation for perisaccadic distortions". *The Journal of Neuroscience,* vol. 27, 2007, pp. 8525-8532.

[3] Sophie Deneve, Alexandre Pouget, "Bayesian multisensory integration and cross-modal spatial links". *Journal of Physiology-Paris,* vol. 98, 2004, pp. 249-258.

[4] Burr D.C., Alais D., "Combining visual and auditory information". *Progress in Brain Research,* vol.155, 2006, pp. 243-258.

[5] Battaglia P.W., Jacobs R.A., Aslin R.N., "Bayesian integration of visual and auditory signals for spatial localization". *Journal of the Optical Society of America,* vol. 20, 2003, pp. 1391-1397.

[6] Bolognini N., Rasi F., L'adavas E., "Visual localization of sounds". *Neuropsychologia,* vol. 43, 2005, pp. 1655-1661.

[7] Sommer K.-D., Kuhn O., Puente Leon F., Bernd R.L. Siebert, "A Bayesian approach to information fusion for evaluating the measurement uncertainty". *Robotics and Autonomous Systems,* vol. 57, 2009, pp. 339-344.

[8] Hackett J.K., Shah M., "Multisensor fusion: a perspective". In: *Proc. of IEEE International Conference on Robotics and Automation*, vol. 2, 1990, pp.1324-1330.

[9] Wei Kin Wong, Tze Ming Neoh, Chu Kiong Loo, Chuan Poh Ong, "Bayesian fusion of auditory and visual spatial cues during fixation and saccade in humanoid robot". *Lecture Notes in Com-puter Science*, vol. 5506, 2008, pp. 1103-1109.

[10] Yong-Ge Wu, Jing-Yu Yang, Ke Liu, "Obstacle detection and environment modeling based on multisensor fusion for robot navigation". *Artificial Intelligence in Engineering,* vol. 10, 1996, pp. 323-333.

[11] Lopez-Orozco J.A., de la Cruz J.M., Besada E., Ruiperez P., "An asynchronous, robust, and distributed multisensor fusion

system for mobile robots". *The International Journal of Robotics Research,* vol. 19, 2000, pp. 914-932.

[12] Toyama K., Horvitz E., "Bayesian modality fusion: probabilistic integration of multiple vision algorithms for head tracking". In: *Proc. of 4th Asian Conference on Computer Vision,* 2000.

[13] Kobayashi F., Arai F., "Sensor fusion system using recurrent fuzzy inference". *Journal of Intelligent and Robotic Systems,* vol. 23, 1998, pp. 201-216.

[14] Pao L.Y., O'Neil S.D., "Multisensor fusion algorithms for tracking". *American Control Conference,* 1993, pp. 859-863.

[15] Beltràn-Gonzàlez C., Sandini G., "Visual attention priming based on crossmodal expectations". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems,* 2005, pp. 1060-1065.

[16] Bothe H.-H., Persson M., Biel L., Rosenholm M., "Multivariate sensor fusion by a neural network model". In: *Proc. 2nd International Conference on Information Fusion,* 1999, pp. 1094-1101.

[17] Hiromichi Nakashima, Noboru Ohnishi, "Acquiring localization ability by integration between motion and sensing". In: *Proc. of IEEE International Conference on Systems, Man, and Cybernetics,* vol. 2, 1999, pp. 312-317.

[18] Kòrding K.P., Beierholm U., Ma W.J., *et al.*, "Causal Inference in Multisensory Perception". *PLoS ONE,* vol.2, 2007, e943.

[19] Mishra J., Martinez A., Sejnowski T.J., Hillyard S.A., "Early Cross-Modal Interactions in Auditory and Visual Cortex underlie a Sound-Induced Visual Illusion". *The Journal of Neuroscience,* vol. 27, 2007, pp. 4120-4131.

[20] Yoshiaki Sakagami, Ryujin Watanabe, Chiaki Aoyama, *et al.*, "The intelligent ASIMO: system overview and integration". In: *IEEE/RSJ International Conference on Intelligent Robots and System,* vol. 3, 2002, pp. 2478-2483.

[21] Metta G., Sandini G., Vernon D., *et al.*, "The iCub humanoid robot: an open platform for research in embodied cognition". *Performance Metrics for Intelligent Systems Workshop,* Gaithersburg, USA, 2008.

[22] Metta G., Gasteratos A., Sandini G., "Learning to track colored objects with Log-Polar vision". *Mechatronic,* vol. 14, 2004, pp. 989-1006.

[23] Berton F., *A brief introduction to log-polar mapping.* LIRA-Lab, University of Genova, 2006.

[24] Natale L., Metta G., Sandini G., "Development of auditory-evoked reflexes: visuoacoustic cues integration in a binocular head". *Robotics and Autonomous Systems,* vol. 39, 2002, pp. 87-106.

[25] Kee K.C., Loo C.K., Khor S.E., *Sound localization using generalized cross correlation: Performance Comparison of Pre-Filter.* Center of Robotics and Automation, Multimedia University, 2008.