Intuitive Human-Robot Interaction through Active 3D Gaze Tracking

Rowel Atienza and Alexander Zelinsky

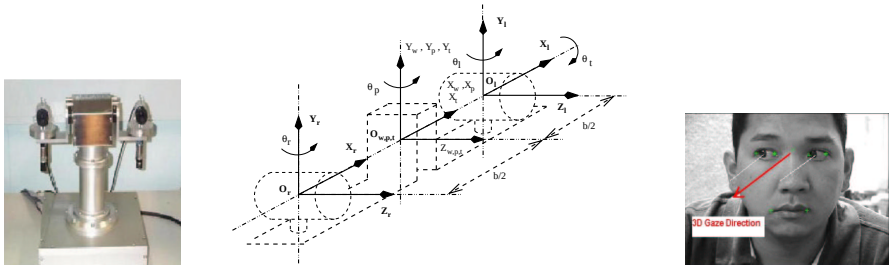
Research School of Information Sciences and Engineering
The Australian National University
Canberra, ACT 0200 Australia
{rowel,alex}@syseng.anu.edu.au

Abstract. One of the biggest obstacles facing humans and robots is the lack of means for natural and meaningful interaction. Robots find it difficult to understand human intentions since our way of communication is different from the way machines exchange their information. Our aim is to remove this barrier by creating systems that react and respond to natural human actions. In this research, we present a robotic system that identifies and picks up an arbitrary object in 3D space that a person is looking at. It is done through an active vision system that is able to understand the focus of attention of a user. Whenever the person is close enough, the gaze direction is determined and used to find the object of interest which is of unknown size, shape and color. A robot arm responds by picking up this object and handing it over to the person.

1 Introduction

Many present day robots are capable of performing ordinary and mundane human tasks. The problem in many situations is that robots do not understand human intentions. For example, a cleaning robot might struggle to comprehend its human master giving the command: “Clean that spilled milk on the floor.” while pointing and looking at the spilled milk. Although the robot is capable of accomplishing the cleaning task required, it is dumb from the point of view of understanding the context of the instruction. Speech understanding on its own is insufficient.

Our goal is to give robots the ability to see and understand humans by observing natural actions. We use an active vision equipped with zooming cameras (see Figure 1). Using zoom stereo cameras mounted on a movable head gives the user the freedom from wearing a tracking gadget and to move without restrictions. The active vision at first developed the ability to transfer and maintain its focus of attention on any object that is moving with color similar to human skin. This ability allows the tracking of the hand or the face creating the impression to the user that the robot is highly interactive and ready to receive commands. We then integrated a gaze tracking skill based on a detection algorithm our group previously developed [9]. This allows a robot to see where the user is looking at in 3D space (see Figure 1). By utilizing the gaze information provided, the active vision detects when a person is staring at an object and searches the gaze line to find this object of unknown shape, size and color. With these skills, we developed a robotic system where an active vision



(a) Active vision head

(b) Kinematic model

(c) 3D gaze vector

Fig. 1. Active vision, its kinematic model and gaze direction vector

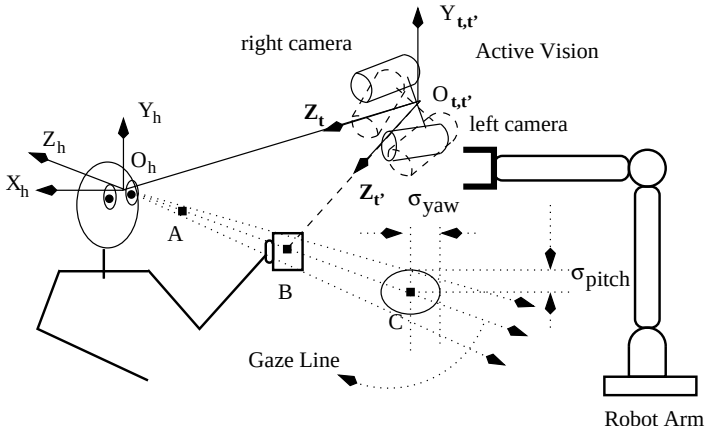


Fig. 2. The active vision traces the gaze line from gaze point A to C to find the object (the cup at gaze point B) using a zero-disparity filter for the robot arm to pick up. σ_{yaw} and σ_{pitch} are the standard deviations of gaze yaw and pitch respectively. They are used to detect steady gaze.

continuously tracks the face and gaze of a user. When the user focuses the gaze at an object, its 3D location is determined and a robot arm (a Whole Arm Manipulator or WAM) picks it up and hands it over. The scenario is illustrated in Figure 2. Such a capability allows tasks as the cleaning robot example posed earlier possible.

When robotic systems can already hear and see people [4,6,11], we will find a variety of robotic applications in our everyday tasks. For the remainder of this paper, we discuss the background material on how our system was developed. The topics include zoom camera calibration, active sensing, gaze tracking and interactive skills. Experimental results are also presented.

2 Zoom Camera Calibration

Before any meaningful Euclidean measurements on world objects can be made by the active vision, the zoom cameras must be calibrated first. A calibrated camera describes how a 3D world point is related to the camera coordinate system (extrinsic parameters) and how the camera coordinate system is related to image plane coordinate system (intrinsic parameters). Numerous calibration procedures have been proposed in the past two decades to measure the parameters of fixed-lens cameras (see [13] for a survey of camera calibration techniques). For zoom cameras the case is different. Due to the huge number of possible zoom-focus-aperture settings, zoom camera calibration is much more tedious and time consuming. To make zoom camera calibration more practical, we use Willson's method [12] and assign one focus setting for each zoom since for many practical purposes we only need one lens setting that gives a sharp image. Each intrinsic parameter can be modeled as:

$$p = K_0 + K_1 z + K_2 f + K_3 z^2 + K_4 f^2 + K_5 z f \tag{1}$$

where K_i = constant, f = focus and z = zoom using least squares techniques. The camera parameters f_x and f_y are the focal lengths in pixel dimensions along the x and y axes respectively, $o_x$and o_y are principal point coordinates also in pixel dimensions, and k_1 to k_3 are the radial distortion coefficients. To model the changes in the extrinsic parameters, the displacement of the camera center along the principal axis is represented by a second-degree polynomial of z. For a detailed discussion, please refer to our earlier work [1].

3 Active Sensing

In order to sense the human user, our active vision tracks any object that is moving and has a color resembling human skin color. The detection of skin color is done by comparing the chrominance of each representative pixel with a previously prepared chrominance chart of skin images that are contributed from different persons [5]. The result of skin detection is the probability that a pixel is representing skin, p_s. Motion detection uses optical flow techniques modified to include low-pass filtering [8] and a measure of confidence value [10]. The effect of camera ego-motion is also taken into consideration. The output is a flow vector $\mathbf{v} = \begin{bmatrix} v_x & v_y \end{bmatrix}^T$ on each image patch and its corresponding confidence value, p_m. These two visual cues are combined using a Particle Filter [7] to generate an estimate of the most probable location of the hand or the face in each image, $\mathbf{p}_{oc} = \begin{bmatrix} x_{oc} & y_{oc} \end{bmatrix}^T$. The scenario is illustrated on the left of Figure 3. A measure of confidence value that $\mathbf{p}_{oc}$ is from a hand or a face is proposed: $p_o = \frac{\sum_{n=1}^{N} \pi_t^{(n)}}{N}$, where π_t is the unnormalized weight of a pixel having both motion and skin color and N is the number of particles used. Using the centroid of the hand or face on each image, we can use triangulation in space to estimate the position of the tracking target. The estimate is further improved

with Kalman Filtering. The final output is a smoothed measure of the hand or face 3D location: ${}^{t}P'_{k} = \left[x_{ok}\ y_{ok}\ z_{ok} \right]^{T}$ which is used in tracking. Since the depth of the object being tracked is known, the zoom level can be adjusted to obtain high resolution images at all times.

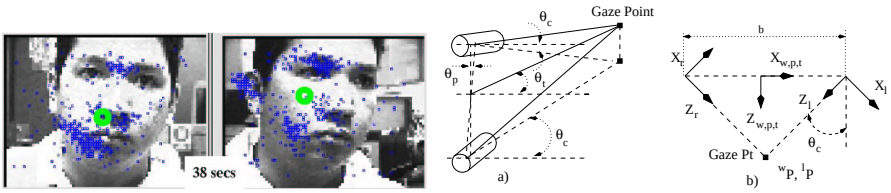


Fig. 3. *Left:* Particles during face tracking (circle = face centroid) from the experiment in Figure 6. *Right:* a) Stereo cameras directed at the gaze point, b) Coordinate systems assignment for a) when $\theta_p = \theta_t = 0$.

During tracking, we always maintain symmetrical configuration on our active stereo cameras (i.e. $\theta_r = \theta_c = -\theta_l$ where $\theta_r > 0$). This results to a simple expression for the inverse kinematics of the active vision as shown in Figure 3:

$$\begin{bmatrix} \theta_p \\ \theta_t \\ \theta_c \end{bmatrix} = \begin{bmatrix} \arctan\left(\frac{x}{z}\right) \\ -\arctan\left(\frac{y\cos\theta_p}{z}\right) \\ \arctan\left(\frac{b\cos\theta_p\cos\theta_t}{2z}\right) \end{bmatrix} \tag{2}$$

where θ_p , θ_t and θ_c are the pan, tilt and camera axis rotation angles and $\left[x\ y\ z \right]^{T}$ is the 3D gaze point. For a detailed discussion, please refer to our paper [3].

4 Active Gaze Tracking

We are using a gaze tracking algorithm called faceLab that our group has commercialized [9]. The system works by tracking stable facial features such as eye and lip corners. Based on the location of the iris, eye-gaze is determined. In the event that the iris can not be tracked reliably, the head pose is used to estimate gaze. Since the original algorithm was designed for a *fixed* stereo vision configuration, the camera parameters must be adjusted in real-time to track the gaze using active vision. This is possible since for certain zoom settings where the stereo pair is initially fully calibrated, moving to a new configuration requires changes in the extrinsic parameters only. The new set of extrinsic parameters can be derived using Figure 4. For camera A (camera B follows easily), the changes are made on the x and z components of ${}^{A}T_{W}$, and on ${}^{A}R_{W}$ only since the rotation from the default vergence angle is parallel to camera A Y axis. The new values are :

$$x = -\left\| {}^{A}T_{W} \right\| \sin\theta,\ z = \left\| {}^{A}T_{W} \right\| \cos\theta,\ and\ {}^{A}R_{W} = {}^{A}R'_{W} R_{Y}\left(\theta' - \theta\right) \tag{3}$$

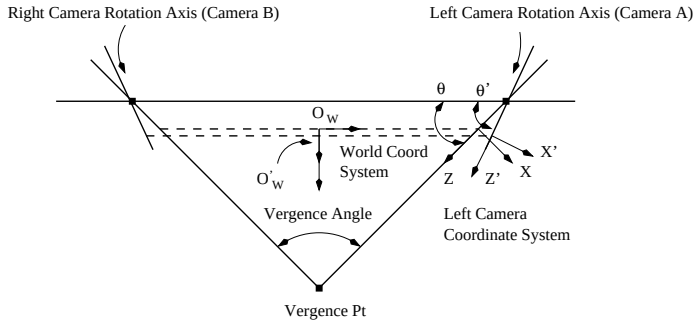


Fig. 4. The stereo cameras originally in default vergence position (symbols with prime) are moved to a new position (symbols without prime). Vergence angle and vergence/gaze point are also shown.

where $\left\| {}^{A}T_{W} \right\|$ is assumed to be approximately equal to the known $\left\| {}^{A}T'_{W} \right\|$ at the default vergence angle for practical purposes and ${}^{A}R'_{W}$ is the known rotation matrix at the default vergence angle position. When the extrinsic parameters are readjusted in real-time, the active gaze tracking system generates the following measurements: 3D eye gaze direction (${}^{w}G$), 3D head rotation (${}^{w}R_{h}$) and translation (${}^{w}H$) or 3D head pose only while the gaze is estimated from ${}^{h}G = \begin{bmatrix} 0 & 0 & -1 \end{bmatrix}^{T}$. Both measurements when available are sampled in real-time (>30Hz) with a corresponding confidence value, p_g, and are expressed with respect to the fixed world coordinate system, O_w in Figure 1b.

Given ${}^{w}H = \begin{bmatrix} x_0 & y_0 & z_0 \end{bmatrix}^{T}$, the head origin with respect to O_w, and ${}^{w}G + {}^{w}H = \begin{bmatrix} x_r & y_r & z_r \end{bmatrix}^{T}$, the gaze vector with respect to O_w translated at the head origin, the symmetric equations describing the 3D gaze line as shown in Figure 2 is:

$$\frac{x_1 - x_0}{x_r - x_0} = \frac{y_1 - y_0}{y_r - y_0} = \frac{z_1 - z_0}{z_r - z_0} \tag{4}$$

where we define ${}^{w}H$ as the origin , ${}^{w}G + {}^{w}H$ as the reference point and $\begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^{T}$ as any point on the line. In view of searching the object where a person is looking at, we can trace the gaze line by moving at an increasing distance, $d > 0$, away from the origin. The new gaze point in terms of world coordinates can be determined to be:

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} x_0 \pm \sqrt{x_k} \\ y_0 \pm \left(\frac{y_r - y_0}{x_r - x_0} \right) \sqrt{x_k} \\ z_0 \pm \left(\frac{z_r - z_0}{x_r - x_0} \right) \sqrt{x_k} \end{bmatrix} \tag{5}$$

where $x_k = \frac{d^2}{1+\left(\frac{y_r - y_0}{x_r - x_0}\right)^2 + \left(\frac{z_r - z_0}{x_r - x_0}\right)^2}$. The choice of sign ensures that the search moves toward the direction of ${}^{w}G + {}^{w}H$.

5 Interactive Skills

When the system detects steady gaze, it is concluded that the person is staring at a certain object. This object can be found by searching along the gaze line and triangulating in 3D space, thereby allowing the robot to pick it up. A steady gaze (SG) is assumed whenever the running standard deviations of gaze yaw and pitch are below a threshold:

$$SG = \begin{cases} 1 & \sigma_{yaw} \leq \sigma_{y0} \; and \; \sigma_{pitch} \leq \sigma_{p0} \\ 0 & otherwise \end{cases} . \tag{6}$$

Both σ_{yaw} and σ_{pitch} are computed from the gaze direction generated from $t - t_0$ to t with a normalized gaze confidence value $p_g(t)$ (i.e. $\sum_{t-t_0}^{t} p_g(t) = 1$). Using Equation 5 and a zero-disparity filter (ZDF), the object can be found along the gaze line. In our paper [3], we proposed that the 3D position is at the gaze point where the disparity between the left and right fovea is minimum (gaze point with maximum cross-correlation between left and right images). This technique fails in the scenario where the left and right cameras are both pointing at a blank wall. To avoid this problem, we instead use a fast edge detector (like a Laplacian) and shift both binarized edge images in the x-direction (horizontal) to find the maximum correlation: $\max \left\{ \mathbf{I}_{left}(x+n, y) \bullet \mathbf{I}_{right}(x+m, y) |_{n,m=-x_{shift}}^{n,m=x_{shift}} \right\}$, where $\mathbf{I}_{left}$ and $\mathbf{I}_{right}$ are the left and right binarized edge images of the fovea respectively. For speed of computation, the fovea is a rectangular sub-image centered at the image center when $n = m = 0$. The maximum amount of horizontal shift, $2x_{shift}$, approximates the threshold for the gaze yaw/pitch standard deviation. This is done on gaze points along the gaze line. Using this method, the object is located at the gaze point with maximum correlation. The centroids of the object on the left, $\mathbf{p}_l$, and right, $\mathbf{p}_r$, images are computed from the shifted edge pixels that coincide. Since we have fully calibrated stereo cameras, the object's 3D position with respect to the left camera can be estimated by triangulation: ${}^{l}\mathbf{X}_o = a\mathbf{P}_l + \frac{1}{2}c\left(\mathbf{P}_l \times^{l} R_r \mathbf{P}_r\right)$, where the constants a and c are computed from ${}^{l}\mathbf{T}_r = a\mathbf{P}_l + c\left(\mathbf{P}_l \times^{l} R_r \mathbf{P}_r\right) - b\,{}^{l}R_r \mathbf{P}_r$, ${}^{l}\mathbf{T}_r$ and ${}^{l}R_r$ are the translation vector and rotation matrix respectively of the right camera with respect to the left camera coordinate system, $\mathbf{P}_l = \begin{bmatrix} \mathbf{p}_l & f \end{bmatrix}^T$ ($\mathbf{P}_r = \begin{bmatrix} \mathbf{p}_r & f \end{bmatrix}^T$) and f is f_x or f_y. The object position, $\begin{bmatrix} {}^{WAM}\mathbf{X}_o \\ 1 \end{bmatrix} = {}^{WAM}T_w \, {}^{w}T_l \begin{bmatrix} {}^{l}\mathbf{X}_o \\ 1 \end{bmatrix}$, is then given to the WAM so it can pick up and hand over the object to the user. T is the transformation matrix.

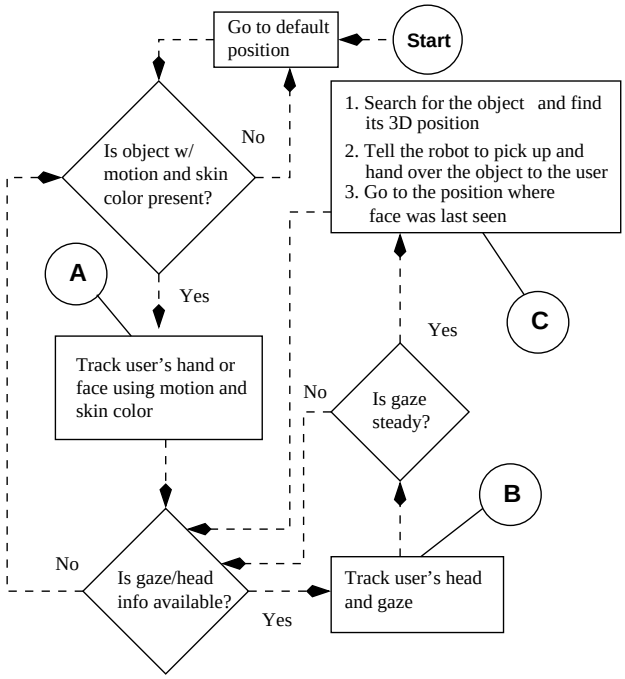


Fig. 5. General behavior of the active gaze interaction system

6 Results and Discussion

The summary of operations of the active vision system is shown in Figure 5. The summary describes how the active vision bootstraps itself when a user is located, finds the object a user is staring at, and tells the robot arm to pick it up and to hand it over. A demonstration on how the active vision gains attention through a waving hand and then continuously tracks the user's face is shown in Figure 6 (Processes A and B in Figure 5). The active vision decides whether a hand or a face is present or not is shown in Figure 7a (if $p_o > 0.1$, a user is present). Figure 7b shows the error during tracking. The mean error during tracking is -15.48 pixels for the left image (22.48 for the right image) while the standard deviation is 52.39 pixels (56.18 for the right image) in the x coordinate. An interesting demonstration on how we instructed a robot arm to pick up an object with a user steadily looking is shown in Figure 8 (Process C). At t = 2 secs, the user placed an object of unknown size, shape and color on the robot workplace. At t = 10 secs, the user was staring at the object. A steady gaze was detected and the object was searched along the gaze line as shown at t = 16 secs. Notice the upper left inset where the images from the cameras are shown. Sometime around t = 28 secs, the object was identified and its 3D location is found. The robot arm reacts, picks up the object and hands it over to the user (t > 28 secs).

Fig. 6. Active vision gaining attention and then tracking the user's face.

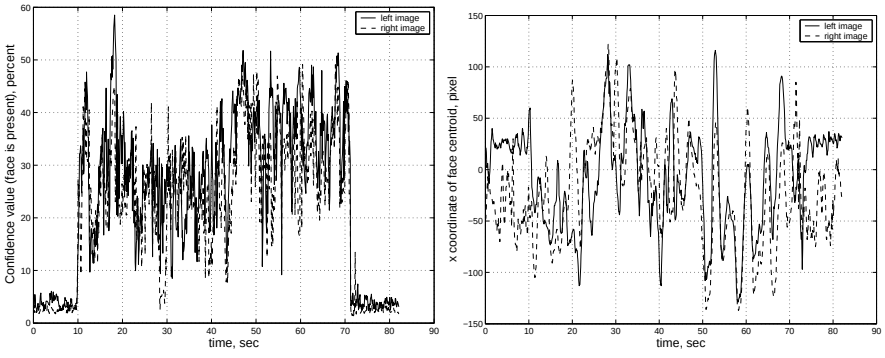


Fig. 7. *a)* Confidence value (face is present) vs time. *b)* Position controller error in x pixel coordinate vs time

The experiment videos for a different user and a different object are available in our website: http://www.syseng.anu.edu.au/rsl/rsl_demos.html .

7 Conclusion

We have created a robotic system that reacts and responds to our natural gaze. It is a step forward in our effort to make robots pervasive in dealing with our daily tasks. When combined with other human natural means of communications, we hope to see robots as an unrecognizable part of our daily lives.

Acknowledgment

The authors would like to thank Seeing Machines Inc., especially Dr. Sebastien Rougeaux, for the support on this research.

References

1. Atienza, R. and Zelinsky, A., A Practical Zoom Camera Calibration, Australian Conf on Robotics and Automation, Sydney, Nov. 2001.
2. Atienza, R. and Zelinsky, A., Active Gaze Tracking for Human-Robot Interaction, Intl Conf on Multimodal Interfaces (ICMI 2002), PA USA, 2002.
3. Atienza, R. and Zelinsky, A., Interactive Skills Using Active Gaze Tracking, ICMI 2003, Vancouver Canada, 2003.
4. Breazeal, C. and Aryananda, L., Recognition of Affective Communicative Intent in Robot-Directed Speech, Autonomous Robots, No. 12, 2002.
5. Cai, J. and Goshtasby, A., Detecting Human Faces in Color Images, Image and Vision Computing. 18: 63-75, 1999.
6. Hashimoto, S., et. al., Humanoid Robots in Waseda University - Hadaly-2 and WABIAN, Autonomous Robots, No. 12, 2002.
7. Isard, M. and Blake, A., Condensation-condensational density propagation for visual tracking, Intl Journal of Computer Vision, 29(1): 5-28 , 1998.
8. Fleet, D. and Langley, K., Recursive Filters for Optical Flow, IEEE Pattern Analysis and Machine Intelligence (PAMI), Vol. 17, No. 1, Jan 1995.
9. Seeing Machines Inc., www.seeingmachines.com.
10. Simoncelli, E. et. al., Probability Distributions of Optical Flow, IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR), 1991.
11. Waldherr, S., et. al., A Gesture Based Interface for Human-Robot Interaction, Autonomous Robots, No. 9, 2000.
12. Willson, R., Modeling and Calibration of Automated Zoom Lenses, Proc. of SPIE #2350: Videometrics III, Boston MA, October 1994.
13. Wong, K.Y., et. al., Camera Calibration from Surfaces of Revolution, IEEE PAMI, Vol. 25, No. 2, Feb 2003.

Fig. 8. Human-robot interaction experiment: A user instructs the WAM to pick up and hand over the object he is looking at. The object of arbitrary size, shape and color is placed on an unknown position in robot workplace. Upper left inset shows images from the cameras.