

A Practical Zoom Camera Calibration Technique: An Application on Active Vision for Human-Robot Interaction

Rowel Atienza and Alex Zelinsky

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

Abstract

Knowing where a person is looking at and understanding the user's head gestures provide an effective and safe means to communicate with a robot. This type of human-robot interaction becomes even more effective when the vision system implemented simulates a human vision system enhanced by a pair of zoom cameras. But before zoom cameras can be effectively used, they are beset by a tedious and time-consuming calibration procedure. In this paper, we present a practical zoom camera calibration technique that is simple and can be effectively used not only for enhancing active vision systems but also for many other applications.

1 Introduction

The ability to detect a person, look at a face, know where the user is looking at, understand any head gesture and observe the task that is being executed provide a simple and effective way for a robot to interact with and learn from a person on its environment. We have an experimental platform used to accomplish this goal. The platform spans the integration of a robot arm (a whole arm manipulator or WAM) and an active vision system enhanced by a pair of zoom cameras. The WAM motion controller has been designed such that the amount of force on impact during collision with any object in its workplace is limited to a safe value [Heinzmann and Zelinsky, 1999]. Thus, a person interacting with the WAM is never harmed in case of accidental collisions. The active vision system uses an active head developed in our lab [Sutherland *et al.*, 2001] retrofitted with a pair of zoom cameras. We plan to implement an eye gaze detection algorithm developed by Seeing Machines [2001] and an algorithm for detecting simple head gestures such as nodding (agree) or a head-shake (disagree) [Heinzmann, 2001] to enable the WAM to interact with a person in front of it. The zoom cameras, though more difficult to use, are chosen to provide a higher performance over standard static cameras. For example, if we want to track the face of moving person while maintaining its image resolution

high for accurate eye gaze tracking, it is not enough to have a real-time controller dynamically positioning the active head. The zoom setting must also be adjusted at the same time. This is something static cameras cannot obviously do. Real time adjustment of zoom setting will require that the camera calibration can be determined in real-time. Otherwise, many important physical measurements like eye gaze tracking are not possible. Camera parameters over the entire zoom settings must be known. Calibrating zoom camera on all its possible setting is impractical because of the enormous amount of data required. In this paper, we present an easier zoom camera calibration technique. The main motivation in our method is not all operating ranges of a zoom camera are useful in common applications. Therefore, if the zoom camera is calibrated only on carefully selected zoom-focus-aperture points then a function can be fitted for each camera parameter. The function can be used to approximate intermediate values and neighboring zoom-focus-aperture points. Details of the proposed zoom camera calibration procedure and its experimental results are discussed after some additional information of our platform for human-robot interaction is described.

2 An Overview of the Experimental Platform for Human-Robot Interaction

Figure 1a explains our experimental platform used for human-robot interaction. It is made of three subsystems. The first is a 7-dof Barrett WAM with a three-finger hand used to execute simple pick and place of objects around its workplace. It is fitted with a safe impact controller [Heinzmann and Zelinsky 1999] to limit the amount of force during collision of any of its part with any object. The safe control architecture makes it suitable for a safe direct human interaction thus we call the robot human-friendly. The second subsystem is a 4-dof active head called HyDrA retrofitted with two zoom cameras. HyDrA can control the two cameras in vergence and version as well as tilt. The vision system will have the ability to detect eye gaze and simple head gestures in real-time. It will use the eye gaze detection software developed by Seeing Machines [2001]. An example of eye gaze

detection is illustrated in Figure 1b. It is crucial in determining which object a person is looking at without reconstructing the 3D view of the workspace. One major improvement of the active vision is its ability to precisely estimate eye gaze and detect head gestures because of its zoom cameras. A person's face can also be easily tracked since the active vision can fixate anywhere in a 3D volume that is almost equivalent to a 5m in diameter hemisphere. Therefore, from the user's point of view, the interface is more natural because the person is not restricted to a fixed location in space. The last subsystem is a person acting as a skill demonstrator who will interact with and teach the robot. The future objective is to let the robot learn from the skills the person is demonstrating. Since the robot knows which object a person is looking at and how the user manipulates this object with respect to other objects in the scene, it should be able to deduce some hypotheses on how a task is accomplished. Hypotheses are generated after several demonstration of the same skill. Afterwards, these hypotheses can be used to accomplish any future task that is similar to what the robot was trained for. Feedback from the human demonstrator while teaching the robot a certain skill may be in the form of simple head gestures such as nodding (agree) or a head-shake (disagree) [Heinzmann, 2001].

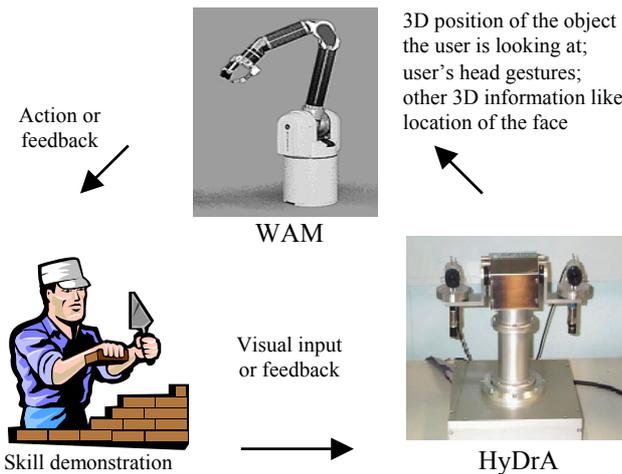


Figure 1a. The experimental platform for human-robot interaction using active vision



Figure 1b. Example of eye gaze detection [from Matsumoto and Zelinsky, 2000]

3 Camera Parameters – An Introduction

Using a pinhole model, camera parameters are grouped into two: 1) intrinsic parameters and 2) extrinsic parameters. Intrinsic parameters describe the physical properties of the image plane with respect to the camera coordinate system and can be written as:

$$\mathbf{A} = \begin{bmatrix} f_x & c & o_x \\ 0 & f_y & o_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (1)$$

where: f_x and f_y are the focal lengths along the x and y axes of the image plane in pixel dimensions, o_x and o_y are the x and y coordinates of the image plane origin (also known as the principal point) in pixel dimensions, and c is the skewness of the x and y axes of the image plane. For many practical purposes, it can be assumed to be equal to zero.

On the other hand, extrinsic parameters describe the transformation between the camera and world coordinate systems and can be represented by:

$$\mathbf{T} = [\mathbf{R} \ \mathbf{t}], \quad (2)$$

where: \mathbf{R} and \mathbf{t} are the rotation and translation matrices that relate the world coordinate system to the camera coordinate system.

Therefore, a point $\mathbf{M}=[X, Y, Z, 1]^T$ in 3D space is represented as $\mathbf{m}=[x, y, 1]^T$ in the image plane. The two points are related as:

$$\mathbf{s}\mathbf{m} = \mathbf{A}\mathbf{T}\mathbf{M}, \quad (3)$$

where s is an arbitrary scale factor.

The effects of lens distortion are also included as part of the intrinsic parameters. Radial distortion tends to have the biggest effect on the overall distortion and is normally measured by distortion coefficients k_1, k_2, k_3 and k_4 . Once the \mathbf{A} and \mathbf{T} matrices have been estimated, the radial distortion coefficients are measured by iteratively undistorting the images generated by the camera.

3.1 Complexity of Zoom Camera Calibration

The complexity of calibrating zoom cameras comes from the fact that both intrinsic and extrinsic parameters are dependent on zoom, focus and aperture settings. In contrast, static cameras have only one zoom, focus and aperture setting. To get an idea of the magnitude of the complexity, if a static camera requires 10 sets of data to calibrate, a zoom camera requires 1,250,000 ($10 \times 50 \times 50 \times 50$) assuming each zoom, focus and aperture settings has 50 points from its minimum value to its maximum value - an underestimation of the features of commercially available zoom cameras. Even if we assume that the effect of aperture is minimal, the number of points is still a staggering 25,000.

Various techniques have been proposed to reduce the number of data points and at the same time generate a useful zoom camera model. The most notable is the work of [Wilson 1994; Wilson and Shafer 1993] wherein the intrinsic and extrinsic parameters are estimated for a constant aperture setting by getting a sample at fixed interval from the number of data points required. Afterwards, up to a five degree polynomial is used to

approximate the camera parameters in continuous mode for other zoom-focus combinations. The work of [Li and Lavest 1996] on zoom camera calibration for an active head essentially needs the same amount of data. A similar effort to Wilson's work using an artificial neural network (ANN) to closely approximate the camera model is the work of [Ahmed and Farag 2000]. It uses more data since the ANN requires a significant amount of information to converge.

In this paper, we propose a procedure to make zoom calibration less tedious and time-consuming. The basic idea is that not all zoom-focus-aperture combinations are useful in common applications. In most cases such as object recognition or 3D reconstruction, the zoom camera operates at points where images of the desired object are sharply focused. At the same time, we use the results of Wilson, Li and Lavest to note that: 1) thin lens and pinhole approximation can be used for each camera setting, 2) variations of camera parameters with respect to aperture is minimal and can be neglected for practical purposes and, 3) the only parameters that are dependent on zoom and focus setting are the internal parameters and the external parameter T_z or the translation on the Z axis of the camera with respect to a world coordinate system. T_z can in fact be approximated as function of zoom only because the change in T_z is primarily due to the repositioning of lens components when the camera zoom is adjusted [Wilson and Shafer 1993]. Having these constraints, we further note that for each zoom setting, we only need one focus setting to obtain a clear image if we know the distance of the object we are looking at. For each zoom value, the focus setting is the one obtained when we manually set the camera to defocus mode or zero focus level and let its electronic auto-focus controller find the best focus. Given all these simplifying assumptions we will need approximately less than 10% of the original number of data required in Wilson's method.

In the next section, we provide details of the proposed method.

3.2 A Practical Zoom Camera Calibration Technique

Using thin lens and pinhole assumptions, we can use traditional calibration techniques for each zoom-focus setting. We use Zhang's [2000] algorithm to calibrate the zoom camera on each zoom-focus setting since it is easy to use and the source code is freely available from the internet [Intel OCVL, 2001]. A chessboard grid made of 10×7 squares of 1.0×1.0 inch in dimension is used to calibrate the camera that is a Sony CCB-EX37. The camera has been set to auto-exposure mode. It has 1440 zoom levels or steps and 1793 focus levels or steps. The camera is fully controllable via RS232. In the calibration we did, only the first 1000 zoom levels are used to see if the function fitted will extrapolate well. The electronic auto-focus can select any value from the 1793 levels.

Before the actual calibration is done, a lookup table for focus as a function of zoom and distance is created. The idea is to generate a function that will estimate the required amount of focus level given the zoom and the distance of the desired object. In other words, if we are looking at some object like the chessboard pattern, since

we normally know the zoom level we are using and the approximate distance of the object, all we have to do is to use the lookup table/function to set the correct focus to get a clear image. Thus, focusing can happen within a few tens of milliseconds compared to an average of five seconds if we use the auto-focus feature of the camera. We propose to use this scheme to focus on the face of a moving person or an object of interest in front of the robot. In this case, distance can be measured using the two views from the left and right cameras.

To prepare the lookup table, the chessboard is placed in front of the camera while zoom is adjusted from level 0 to level 1000 at interval of 25. This is done from 30 cm to 130cm at 10cm interval. At each zoom level interval, the focus is first set to level 0 (defocus) and the camera is then set to search for the best possible focus using its electronic auto-focus. Experience has shown that this technique is more effective rather starting at an arbitrary focus level. The value at the end of the auto-focus search is then saved in a lookup table. After the lookup table has been created, a second-degree polynomial:

$$f = K_0 + K_1z + K_2d + K_3z^2 + K_4d^2 + K_5zd, \quad (4)$$

where $K_i = \text{constant}$, $f = \text{focus}$, $z = \text{zoom}$ and $d = \text{distance}$, is fitted using least squares techniques for fast focusing. The curve for the focus vs. zoom and distance called focus function is shown in Figure 2.

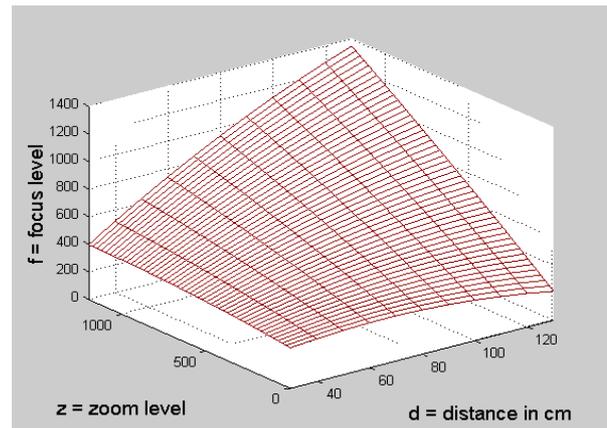


Figure 2. Focus as a function of zoom and distance

After the focus function has been determined, calibration is done from zoom level 0 to 1000 at step intervals of 25. The focus level used is derived from the focus function using the approximate distance of the chessboard pattern from the camera. The calibration data gathered is used to fit a second-degree polynomial using least squares techniques to get each intrinsic parameter as a function of zoom and focus. A second-degree polynomial was chosen since the improvement in sum of squared errors (SSE) or differences if a third-degree polynomial is used is not significant. Therefore, each intrinsic parameter is of the form:

$$p = K_0 + K_1z + K_2f + K_3z^2 + K_4f^2 + K_5zf, \quad (5)$$

where $p = \text{intrinsic parameter}$, $K_i = \text{constant}$, $z = \text{zoom}$ and $f = \text{focus}$. The graphs for the intrinsic parameters are shown in Figures 3a to 3f. Due to limitation of space, the graphs of k_3 and k_4 are not shown.

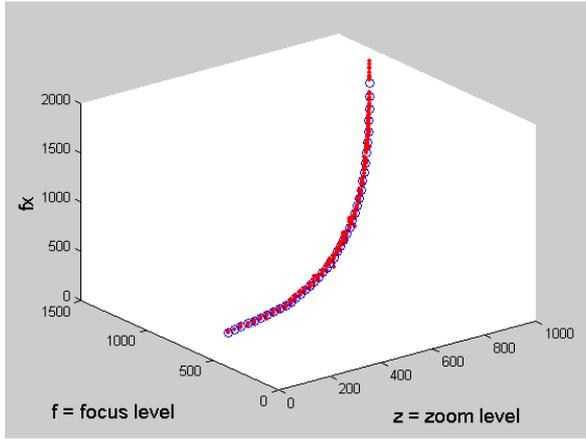


Figure 3a. f_x as a function of zoom and focus

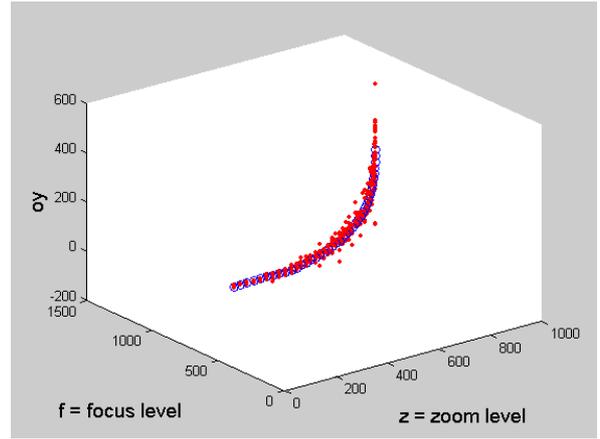


Figure 3d. o_y as a function of zoom and focus

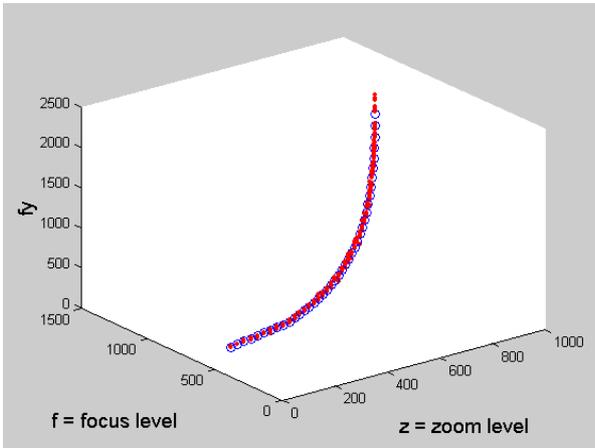


Figure 3b. f_y as a function of zoom and focus

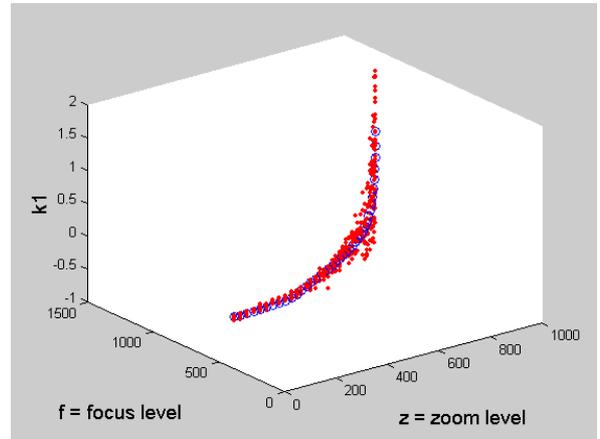


Figure 3e. k_1 as a function of zoom and focus

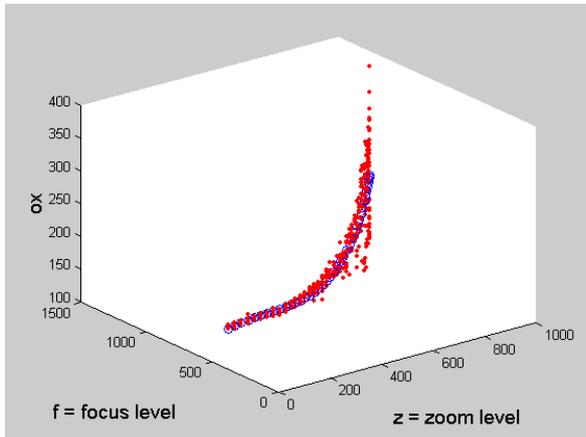


Figure 3c. o_x as a function of zoom and focus

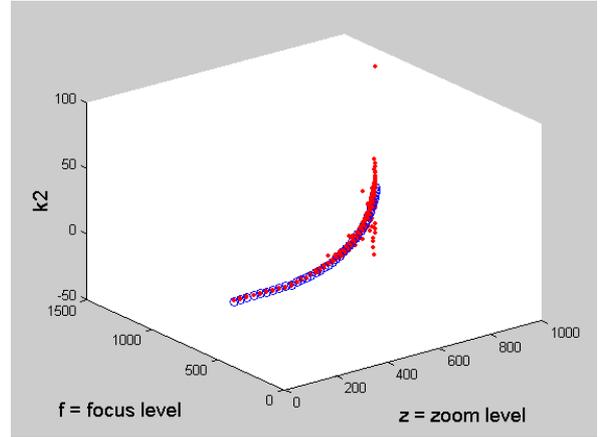


Figure 3f. k_2 as a function of zoom and focus

Since the chessboard pattern moves during calibration, it is not possible to estimate the variation in T_Z . To determine the variation in T_Z , we fixed a square plane in front of the camera and parallel to the image plane. The square plane was of known size and was large enough to see from zoom level 0 to 1000. Since the square plane dimensions are known, we can estimate T_Z and its variation when the zoom is adjusted. It should be noted

that the focus level is adjusted at the same time since the approximate distance of the plane can be measured manually. Furthermore, before any pixel level measurement is obtained, the image is undistorted using the intrinsic parameters obtained. T_Z can be approximated by using the average of four T_Z 's obtained for each side of the square. Each T_Z is estimated by:

$$T_Z = (\text{side} \bullet f) / \Delta \text{pixel}, \quad (6)$$

where $side$ = length of each side of the square, f_i = focal length f_x (f_y) by the x axis (y axis) and $\Delta pixel$ = distance between two corners along x axis (y axis) in sub-pixel accuracy. The plot of T_Z and its third-degree polynomial approximation:

$$T_Z = K_0 + K_1z + K_2z^2 + K_3z^3, \quad (7)$$

where K_i is a constant and z is zoom level, are shown in Figure 4. It should be noted that the most important parameter here is ΔT_Z or the change in T_Z as zoom is adjusted. It can be obtained by subtracting T_Z at zoom level 0.

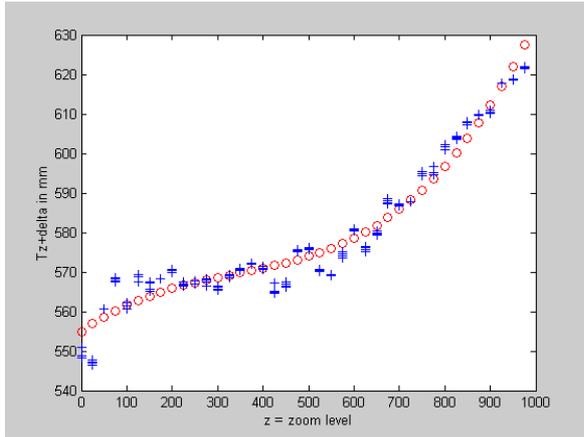


Figure 4. Plot of T_Z as a function of zoom

3.3 Experimental Verifications

To verify the accuracy of the proposed zoom calibration method, we used each intrinsic parameter function, the ΔT_Z offset function and equation 6 to: 1) estimate the dimension of a 25.22 x 25.46 mm square plane and 2) the distance between the inner corners of the left and right eyes of a head model. Both are placed in front of the camera of known distance (with respect to zoom level 0) and as much as possible parallel to the image plane as zoom level is adjusted.

Figures 5a and 5b show that the SSE improves and in many cases approaches a zero value as the resolution of the image is improved by increasing the zoom level. In both figures, before any pixel level measurement is done, the image is first undistorted using the intrinsic parameters. There are two cases for each graph. Case 1 is using the focus level that was generated in the calibration procedure (i.e. we know the zoom therefore we can determine the focus level used by fitting another polynomial on the calibration data). This focus level does not necessarily give a very clear image on all zoom levels because it does not take distance into consideration. Case 2 is using the focus function obtained previously which gives a clear image on each zoom level. However, when we tried using the camera parameters from the second case, we obtained large SSE on estimating the dimension. Case 2 corresponds to points outside the line in Figures 3a to 3f. That means the graphs of the intrinsic parameters are only useful if the zoom and focus levels stay along the line. Therefore, even though we are using the focus function to get a clear image of the plane, we plug in the

focus level from the calibration data to compute the camera parameters. Although we used the focus level from calibration data where it should have been from the focus function, the accuracy of measurement is not jeopardized as indicated by the increase in SSE of 2.4% in Figure 5a and 10.8% in Figure 5b. A further note on the results, it will be noticed that in Figure 5a, the error jumps as the zoom level goes beyond 1000 in Case 1. It is because the calibration was done from zoom levels 0 to 1000 only. However, when the focus function is used, it can be seen that the our estimation of dimension extrapolate well up to zoom level 1100 which is the maximum level in the focus lookup table created earlier. It can also be noted in Figure 5b that while case 1 can longer produce useful data beyond zoom 1025 because the image becomes blurred, case 2 is still generating measurements up to zoom level 1100. The two focus levels are compared in Figure 6.

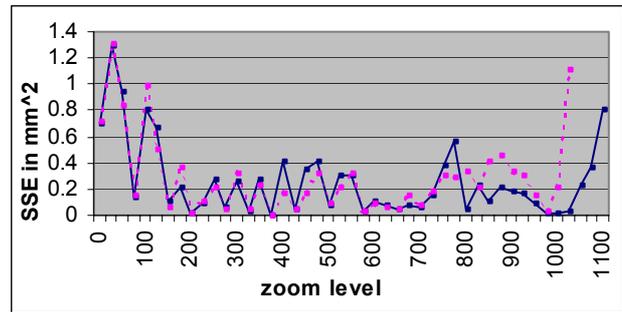


Figure 5a. SSE vs. zoom level. Case 1: broken line, Case 2: solid line. Sum of all SSE for Case 1 = 10.94 mm², sum of all SSE for Case 2 = 11.20 mm², distance of square plane = 50.0 cm

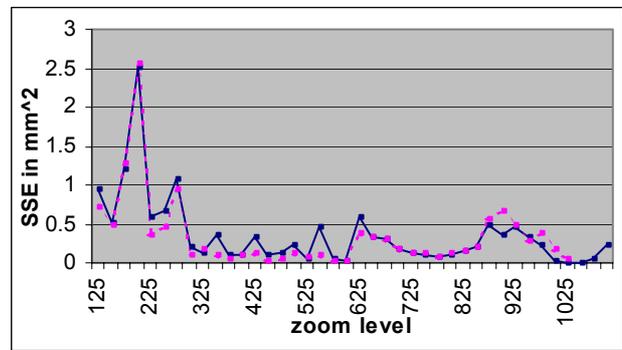


Figure 5b. SSE vs. zoom level. Case 1: broken line, Case 2: solid line. Sum of all SSE for Case 1 = 13.88 mm², sum of all SSE for Case 2 = 12.52 mm², distance of square plane = 70.0 cm.

The second part of experimental validation takes a more realistic object – a mannequin head model as shown in Figure 7. Since we know the distance d between the inner corners of the left and right eyes is 32.2 mm, we use equation 6 and each intrinsic parameter function again to verify if zooming in at the face increases the accuracy of measurements and consequently verifying our proposed calibration method. We placed the head model on two places: 1) 70 cm and 2) 120 cm and the distance between

the two eye corners is tabulated as each zoom level is increased by 50. All throughout the verification, we use Case 2 technique discussed earlier. The experimental data recorded in Figure 8 shows that the error is almost approaching zero as the zoom level goes beyond 450.

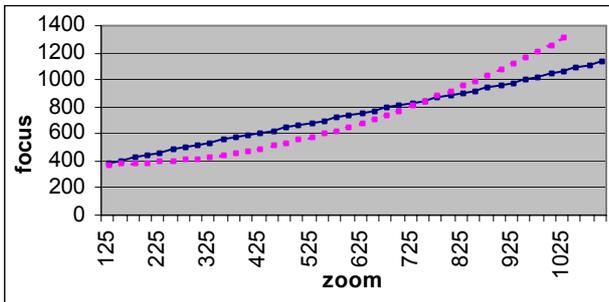


Figure 6. focus level vs zoom level
Case 1: broken line, Case 2: solid line

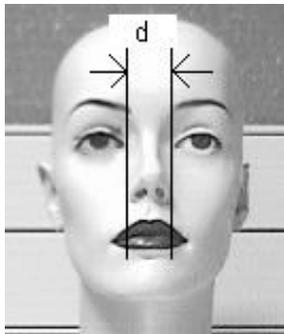


Figure 7. The head model showing the distance, d , between the inner corners of left and right eyes.

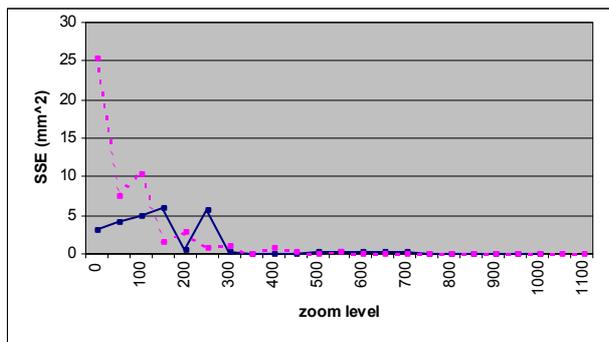


Figure 8. SSE vs. zoom level. Solid line – 70 cm. Broken line – 120 cm.

Conclusion

We presented here an experimental platform for human-robot interaction using active vision. Since we want to improve the performance of the vision system, we used zoom cameras instead of ordinary cameras to get more

meaningful information on a wider area around the robot workspace. Using zoom cameras comes with a price of more complex camera modeling and thus we present a more practical calibration technique. Experimental results show that at higher zoom levels, SSE approaches near zero in many cases. Thus, verifying the validity of our and develop a head gesture detection algorithm to aid the robot proposed calibration method. In the future, we plan to integrate the eye gaze tracking made by Seeing Machines in learning from demonstration.

Acknowledgements

Sebastien Rougeux, Richard Hartley, and David Liebowitz for their suggestions.

References

- [Ahmed and Farag, 2000] M. Ahmed and A. Farag. A Neural Optimization Framework for Zoom Lens Camera Calibration. *IEEE CVPR '00*, June, 2000.
- [Heinzmann and Zelinsky, 1999a] J. Heinzmann and A. Zelinsky. A Safe-Control Paradigm for Human-Robot Interaction. *Journal of Intelligent and Robotic Systems*, Kluwer Academic Publishers, The Netherlands, 1999.
- [Heinzmann and Zelinsky, 1999b] J. Heinzmann and A. Zelinsky. Building Human-Friendly Robot Systems. *Proc. Intl. Symp. of Rob. Res., USA*, 9-12 Oct. 1999.
- [Heinzmann, 2001] J. Heinzmann. Ph.D thesis. RSISE, The Australian National University, 2001.
- [Intel OCVL, 2001], Intel Open Computer Vision Library. <http://sourceforge.net/projects/opencvlibrary>, 2001.
- [Li and Lavest, 1996] M. Li and J. M. Lavest. Some Aspects of Zoom Lens Camera Calibration. *IEEE Trans. on PAMI*, vol. 18, no. 11, Nov. 1996.
- [Matsumoto and Zelinsky, 2000] Y. Matsumoto and A. Zelinsky. An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement, *IEEE 4th Intl. Conf. on Face and Gesture Recognition (FG'2000)*, France, March, 2000.
- [Seeing Machines, 2001] Seeing Machines. <http://www.seeingmachines.com>, 2001.
- [Sutherland et. al., 2000] O. Sutherland, S. Rougeux, S Abdallah and A. Zelinsky. "Tracking with Hybrid-Drive Active Vision". *ISER 2000*, Honolulu, December 2000.
- [Wilson and Shafer, 1993] R. Wilson and S. Shafer. A Perspective Projection Camera Model for Zoom Lenses. *Proc. Second Conf. on Optical 3-D Measurement Techniques*, Switzerland, October 1993.
- [Wilson, 1994] R. Wilson. Modeling and Calibration of Automated Zoom Lenses. *Proc. of SPIE #2350: Videometrics III*, Boston MA, October 1994.
- [Zhang, 2000] Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE Trans. on PAMI*, vol. 22, no. 11, 2000.