On the Performance of State Estimation for Visual Servo Systems

Brad Bishop
Coordinated Science Lab
University of Illinois
Urbana, IL 61801

Seth Hutchinson
Elect. and Comp. Eng.
University of Illinois
Urbana, IL 61801

Mark Spong
Coordinated Science Lab
University of Illinois
Urbana, IL 61801

Abstract

In this paper we discuss the use of computer vision for real-time state estimation in feedback control systems. To this end, we construct a system for visual state estimation of simple state vectors and study the effects of various real-world disturbances on the state estimates. Simulations are performed using a detailed camera model to study the performance of an image plane position estimation algorithm for a single circular feature. Various disturbances, such as lens distortion, noise, defocus, and blurring are simulated and analyzed with respect to this estimation routine and visual state estimation in general.

1 Introduction

A common characteristic of the visual servo control schemes reported to date is that the vision system is used in an outer "command loop", which generates reference inputs to an inner "robot control loop" (e.g. [1], [3], [4], [5], [8], [9], [13]). This arrangement, which we shall refer to as a *dual-loop* visual servo controller, is illustrated in Figure 1. In dual-loop controllers, the vision loop typically runs at a frequency much lower than that of the robot controller. This difference in sampling rates is typically due to limitations of the vision system, which include limits on the sampling time for vision hardware, and the computing time required by various vision algorithms.

By keeping the vision sensing outside the servo level control loop, these hierarchical control schemes possess certain inherent robustness properties. However, such robustness may be achieved at the expense of performance. An alternative controller architecture is illustrated in Figure 2, which we shall refer to as a *direct* visual servo system.

In the direct visual servo system, the vision system is directly feeding back state information (instead of reference inputs) to the robot controller. Such an architecture, where the vision system senses part or all of the system state, may be useful in several situations. For example, in a *task space feedback linearization scheme*[7], [11], one can define an output error as the difference between the position of the object to be tracked (environment) and the robot end-effector, and use this error directly in the output feedback control scheme. In this case, the only means of sensing the object may be with the vision system.

More generally, we may wish to consider the robot/environment as the complete system to be controlled, rather than just the robot itself, as is typical in most of robot control theory. In this case, the system state consists of the state of the robot and the state of the environment to which it is coupled. Hence, the vision system is really measuring part of the system

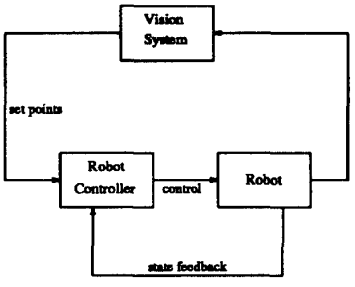

Figure 1: Dual-loop visual servo

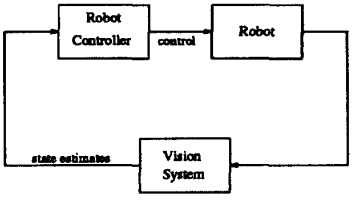

Figure 2: Direct visual servo

state when it views the environment of the robot.

In order to design robust direct visual servo control systems, it is important to fully understand the interactions between the image formation process and state estimation. To this end, in this paper we present a detailed model of the image formation process, including a set of parameters that characterize the departure of the system from the ideal (which in the computer vision literature is typically considered to be an ideal pinhole lens system). To date there has been no analysis of the relationship between distortions in the imaging process and the performance of visual servo systems. Therefore, although our work is done specifically in the context of direct visual servo systems, we expect the results to also be of benefit to designers of dual-loop visual servo systems.

The remainder of the paper is organized as follows. In Section 2 we present a detailed model of the image formation process, and a set of parameters that characterize various distortions that can occur. In Section 3 we present a qualitative analysis of the effects of the distortions on state estimation and discuss a number of simulation studies that quantitatively illustrate these effects.

2 A model of the image formation process

Once an ideal camera model is determined, we will be primarily concerned with two classes of disturbances: those which affect imaging geometry, and those which affect image intensity. Geometric disturbances affect any state estimator based on computer vision. Disturbances that affect image intensity will have varying effects on state estimation, depending on the particular vision algorithms utilized.

2.1 Imaging geometry

As is standard in the computer vision literature, we assume that the underlying model of the camera is that of an ideal pinhole. With this model, a scene point whose coordinates in the camera frame are (x_c, y_c, z_c) projects onto the image plane as follows

$$u_i = f x_c / z_c \tag{1}$$

$$v_i = f y_c / z_c \tag{2}$$

where f is the focal length of the camera and the subscript i indicates the ideal image coordinates, with no distortion. In a real imaging system, there is a class of distortions that can affect the geometric correspondence between points in the camera's field of view and points in the image. The primary members of this class are typically known as lens distortions or aberrations. There exist camera calibration schemes which seek to identify a set of lens distortions for off-the-shelf cameras such as those to be used with this system (e.g., [12], [14]). We have chosen the work of [14] as the basis of our model. The three types of distortion that are documented there are: *radial*, *decentering*, and *thin prism* distortions. We now give a brief overview of each of these distortions and discuss their physical causes.

Radial distortion causes a displacement of image points toward (pincushion distortions) or away from (barrel distortions) the optical axis, and arises from flaws in the lens construction, typically in the curvature of one or more of the lenses in the system. While it cannot be corrected, it does not change over time, and once isolated can be considered a known parameter. The result of radial distortion is a displacement of $(\delta_{ur}, \delta_{vr})$ to the ideal image coordinates, given by

$$\delta_{ur} = k_1 u(u^2 + v^2) + O[(u,v)^5] \tag{3}$$

$$\delta_{vr} = k_1 v(u^2 + v^2) + O[(u,v)^5], \tag{4}$$

where we use the notation $O[(u,v)^n]$ to denote an $n-th$ order term in the image coordinates u, v. The parameter k_1 is positive for barrel and negative for pincushion distortion.

The second type of lens distortion is known as decentering and occurs when the elements in a lens system are not aligned properly, so that the optical axes of the lenses may differ slightly. This type of distortion can easily appear after a camera system has been moved or disassembled, even though it was not present earlier. Thus, this distortion is one of the hardest to isolate in a real system, unless calibration is performed before every use of the camera. The effect of decentering distortion, parameterized by (p_1, p_2), is a displacement of $(\delta_{ud}, \delta_{vd})$ to the ideal image coordinates, given by

$$\delta_{ud} = p_1(3u^2 + v^2) + 2p_2 uv + O[(u,v)^4] \tag{5}$$

$$\delta_{vd} = 2p_1 uv + p_2(u^2 + 3v^2) + O[(u,v)^4]. \tag{6}$$

The final lens distortion that we consider arises from improper lens design or construction of the lens array. Typical examples include variations in the radius of curvature over a single side of a lens, slight tilt of a lens in the array, or a fundamental miscalculation of the characteristics of the designed lens system. This type of distortion can be modeled by adding a parameterized thin prism to the system. The resulting

distortion is given by

$$\delta_{up} = s_1(u^2 + v^2) + O[(u,v)^4] \quad (7)$$
$$\delta_{vp} = s_2(u^2 + v^2) + O[(u,v)^4], \quad (8)$$

where s_1 and s_2 are the parameters of the thin prism.

Combining the effects of these three types of distortion gives the following

$$u = u_i + \delta_{u,r} + \delta_{u,d} + \delta_{u,p} \quad (9)$$
$$v = v_i + \delta_{v,r} + \delta_{v,d} + \delta_{v,p}, \quad (10)$$

where (u_i, v_i) are the ideal image coordinates of the point of interest. For modeling of overall distortion, we ignore all but the lowest order terms in each of the distortion equations. This is justified by a restriction of the image plane to $[-1, 1]$ in both u and v, together with simulation results which indicate these lowest order terms offer a good representation of the distortions with only five parameters.

2.2 Image intensity

The intensity of an image pixel is a function of a number of parameters, including the illumination of the scene, and the reflectivity and orientation of surfaces in the scene. Here, we are not concerned with the exact relationships between the scene and the light pattern incident on the lens system, but rather with how the camera parameters affect the image intensity based on this incident light pattern, and thereby affect the performance of state estimation algorithms. In particular, in this section we will consider the effects of blur, camera sensitivity, defocusing and noise.

We assume that the system uses a standard CCD camera. Each pixel's gray level is generated by a photo-transistor and a single CCD element. The level of charge generated in each CCD is determined by integrating the intensity of the incoming light on that pixel over the sample period. Therefore, if the features of interest are moving during the sample period, the resulting image will be blurred.

Camera sensitivity is defined as the responsiveness of each camera sensing element to the light incident on the lens, which is focused onto the elements of the image array. An imaging system may show excellent sensitivity near its optical axis, but may suffer from a decrease of transmitted light from the lens system near the edges of the CCD array, causing what should be identical intensities at the optical axis and the edge of the image to differ by several gray scale levels. Thus, camera sensitivity is often a result of the lens system and not the CCD array. Under the assumption that this is the case, we model this distortion as

$$I = I_0 \cos^{\beta}(\tan^{-1}(r/f)) \quad (11)$$

where I is the distorted image intensity, I_0 is the original, undistorted image intensity, $r = \sqrt{u^2 + v^2}$, f is the camera focal length and β is a non-negative integer.

Another problem that can arise in most visual systems is that of defocus. If the system is not in proper focus, the image, even with no blurring, will be smeared out. Defocus is represented by a two-dimensional convolution of the image intensity profile with a radially symmetric function (known as the *point spread function*). We choose to represent this effect as a discrete convolution applied multiple times to the image. For this model, defocused images are generated by calculating a number of intermediate images from a set of ideally focused pixel values using a convolution of a mask of the type shown in Figure 3. Each image is convolved with the mask to generate a new image. This process is repeated N times to generate a defocused final image. The parameters (f_0, f_1, f_2) and this number N are the parameters for the defocus.

f_2	f_1	f_2
f_1	f_0	f_1
f_2	f_1	f_2

Figure 3: Mask for discrete convolution in defocus simulation

Finally, the intensity of an image pixel is subject to noise in the image formation process. We model image noise as an additive term with a Gaussian distribution and a standard deviation of σ gray levels.

3 Analysis and simulations

To present our results a concise manner, we have divided the possible disturbances into two groups. The first group, which includes blur, defocus, camera sensitivity and noise, can be assessed in broad qualitative terms. Therefore, in Section 3.1 we present a discussion of these distortions and their effects on state estimation, as observed through a number of simulation experiments. In Section 3.2 we give several quantitative results that illustrate the effects of lens distortion

on state estimation. In all simulations, we utilize a 512 x 512 pixel image with 256 gray levels. The vision algorithm utilized is a simple centroid tracking scheme, where the state to be determined is the location of the centroid of a thresholded feature in the image plane. For details of the implementation, see [2].

3.1 Blur, defocus, camera sensitivity and noise

Blur will present a problem for any camera system viewing a moving object. It is the hope of the designer that the sample rate of the camera system will be high enough compared to the speed of the moving features to render the distortion from blur negligible. Nonetheless, it can be noted that the centroids to be determined in the visual state estimation algorithm defined above will represent the centroids of the features at approximately the midpoint of the sample interval, assuming the motion is relatively constant over the interval. For small sample periods, this represents a delay in the data of approximately one-half of the sample period, in addition to any delays which arise from calculations or are inherent in the feedback controller utilized. Thus we see the need for high frequency in order to avoid increasing the delays in the feedback control law, which could result in instability of the system. When blur is present, care must be taken to consider its effects when choosing system parameters, such as a threshold level. Blur can be partially negated by the use of asynchronous shutters (which have an exposure time much smaller than the sample interval) or certain image processing steps [6].

With regard to defocus, our experiments have shown that typical values for the defocus parameters (f_0, f_1, f_2), (i.e. parameters that reflect a system that has been brought into focus by a human operator or by an autofocus mechanism) produce negligible effects on the state estimates. This matches expectations, since many computer vision algorithms exploit Gaussian blurring as a preprocessing step to help cope with image noise.

Camera sensitivity is of particular concern when the vision algorithms rely on intensity thresholding for segmentation of image features. When camera sensitivity varies greatly across the image plane, it is possible that there will not exist a suitable global intensity threshold for feature segmentation. In such cases, methods such as local histogramming can be used to compensate (see, for example [10]). For most cameras, the sensitivity issue is not significant. Therefore, in the following simulations, we will assume this distortion to be negligible.

Noise, as mentioned above, is an omni-present problem. Assuming that the underlying noise process is Gaussian, we expect that the effects of additive noise as described above will be averaged out over a large area. In fact, since the image is thresholded, we will only be concerned with points which are disturbed from their initial intensity enough to cross the threshold, either downward or upward. This can pose problems for feature segmentation and the estimation of feature locations. Analytically, we can see that over a large region of continuous intensity, the effects of noise with zero mean will average out. The same sort of result is seen when the image is thresholded. The Gaussian noise in the gray scale image translates to salt-and-pepper noise in the thresholded image. Over a large region of pixels with value one, such as an extended feature in a thresholded image, the centroid is relatively undisturbed by the noise.

In the following simulations, the state to be estimated is the image plane location (in (u, v)) of the centroid of a stationary, uniform circle. Again, the image plane is assumed to be $[-1, 1] \times [-1, 1]$. The simulations utilize a circle with a radius, in the image plane, of 0.02 units and a brightness of 255 gray levels on a background with a brightness of zero. The threshold of the vision algorithm is set at 200 gray levels. Estimation error data was taken with the circle placed such that the ideal centroid was at the vertices of a regular grid over the image plane.

Errors in the state estimates in the case of no lens distortions were on the order of 10^{-4} units over the entire image plane, due to quantization noise and thresholding. The estimation errors for noise of $\sigma = 75$ gray levels show negligible effect (errors on order of 10^{-3} units).

3.2 Lens distortion

Equations 5 – 8 demonstrate that the effects of p_1 and p_2 are similar, as are those of s_1 and s_2. Since we are concerned only with a circularly symmetric feature in the image plane, we can analyze only p_1 and s_1, knowing that the effects of p_2 and s_2 will mimic those results under proper coordinate transformations. Thus we are concerned only with the distortions k_1, p_1 and s_1. By noting that the distortion equations are linear in the distortion coefficients, we avoid presenting a family of plots for various values of each distortion.

In Figures 4, 5 and 6, we see the error in location of the centroid of the circle defined above, over a portion

of the image plane, for distortion parameter values of 0.05 (which represent distortions readily apparent to the naked eye). The error on the remainder of the image plane can be generated by noting, from the distortion equations (3)–(8), the symmetries of the distortion fields for each parameter. A negative value for k_1 yields distortions which are approximately equivalent to those shown, but with reversed signs. Negative values for p_1 and s_1 reflect the error graphs about the v axis, with appropriate sign changes.

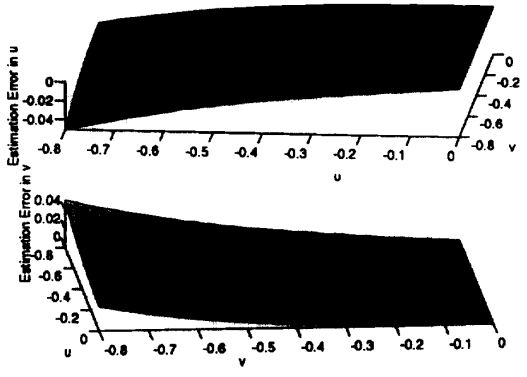

Figure 4: State estimation error for k1 = 0.05 units^{-2} (all values in image plane units)

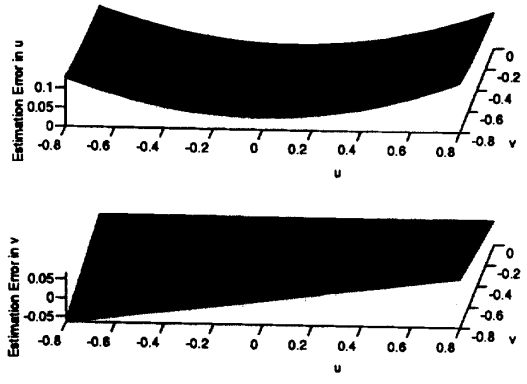

Figure 5: State estimation error for p1 = 0.05 units^{-1} (all values in image plane units)

Considering these state estimation error graphs and the associated distortion equations, we can generate a set of qualitative guidelines for use of this type of visual state estimation under these distortions. Clearly,

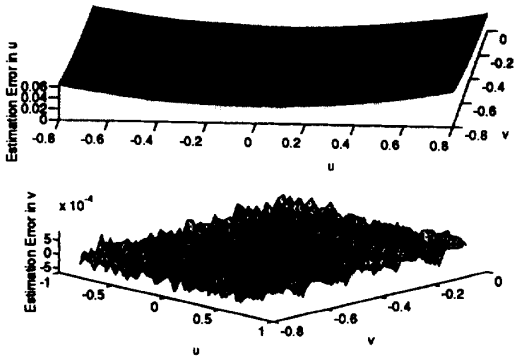

Figure 6: State estimation error for s1 = 0.05 units^{-1} (all values in image plane units)

in any case where lens distortions are present, best results can be obtained when the features of interest are near the optical axis. In most other cases, it is the type of state information to be derived that determines the steps which should be taken when each distortion is present. As an example, consider a state which consists of the planar orientation and location of a single, rigid link. This state could be generated by inscribing one circle on each end of the link and locating their centroids using the scheme detailed above, then calculating the orientation and position from those endpoints. By maintaining one end of the link near the optical axis, the orientation measurement would not be disturbed badly by radial distortions such as k_1. If the orientation was limited to some arc of $[\theta_1, \theta_2]$ in the world reference frame, and distortions like p_1 or s_1 were present in the system, it might be best to orient the camera so that the direction of maximum distortion lies along the midpoint of this arc, minimizing the angular distortion. One might also take the vertex closest to the optical axis as the best measure of the location.

When features of interest are moving between sample frames, and image plane velocities are to be calculated as part of the state estimation, lens distortions also impose an additional velocity distortion as the feature moves across the image plane. This velocity distortion arises from the variation of the magnitude of position distortion over the image plane. Thus, even when calculations are based on differences of image values (in this case, positions), the distortion still presents a problem.

Cases of state estimation for objects which are to

be recognized by relative position of two or more features demonstrate another difficulty presented by lens distortions. If the image of the object extends over a large area of the image plane, or rests far from the optical axis, the features of interest may depart from their ideal locations to a degree which causes a loss of identification and a failure of the state estimation routine. Careful bounds on measurement error must be set in such cases to avoid a loss of data. In the simulated case of simple centroid tracking, bounds were set on the minimum and maximum area of a contiguous, thresholded region of value one which would be labeled as a feature.

We see from the simulations performed that, when the distortions in the lens array are small, such as to be difficult to notice with the naked eye, the error in state estimation for our simple scheme is quite acceptable. Further, with the use of the data accumulated and the distortion fields derived, suitable guidelines for selecting a proper camera setup and appropriate operating goals can be obtained for most visual state estimation schemes.

4 Conclusions

Herein we have demonstrated the quantitative and qualitative effects of various disturbances and aberrations on visual state estimation for direct visual servo of robotic systems. This information is useful in selecting the proper camera system for applications which could be very sensitive to errors in the state feedback, and offers guidelines for dealing with distortions in existing equipment. This research has been applied to simulations of a real–time robot control system utilizing visual feedback for robotic state estimation of a balancing robot in [2].

Acknowledgements : This research is partially supported by The National Science Foundation and the Electric Power Research Institute under the joint NSF/EPRI Intelligent Control Initiative, Grant number ECS-9216428.

References

[1] P. K. Allen, A. Timcenko, B. Yoshimi, and P. Michelman. Trajectory filtering and prediction for automated tracking and grasping of a moving object. In *Proc. IEEE Int'l Conference on Robotics and Automation*, pages 1850–1856, April 1992.

[2] B. E. Bishop. Real time visual control with application to the acrobot. Master's thesis, University of Illinois at Urbana–Champaign, Coordinated Science Laboratory, May 1994.

[3] A. Castano and S. A. Hutchinson. Hybrid vision/position servo control of a robotic manipulator. In *Proc. IEEE Int'l Conference on Robotics and Automation*, pages 1264–1269, Nice, France, May 1992.

[4] B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing in robotics. *IEEE Trans. on Robotics and Automation*, 8(3):313–326, June 1992.

[5] J. T. Feddema and O. R. Mitchell. Vision-guided servoing with feature-based trajectory generation. *IEEE Trans. on Robotics and Automation*, 5(5):691–700, October 1989.

[6] Berthold Klaus Paul Horn. *Robot Vision.* McGraw Hill Book Company, New York, 1986.

[7] O. Khatib. A unified approach for motion and force control of robot manipulators: The operational space formulation. *IEEE Journal of Robotics and Automation*, RA-3:43–53, 1987.

[8] N. Papanikolopoulos, P. K. Khosla, and T. Kanade. Vision and control techniques for robotic visual tracking. In *Proc. IEEE Int'l Conference on Robotics and Automation*, pages 857–864, April 1991.

[9] Azriel Rosenfeld and Avi Kak. *Digital Picture Processing.* Academic Press, New York, 1982.

[10] T. J. Tarn, A. Bejczy, A. Isidori, and Y. Chen. Nonlinear feedback in robot arm control. In *IEEE Conf. on Decision and Control*, 1984.

[11] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, August 1987.

[12] L. E. Weiss, A. C. Sanderson, and C. P. Neuman. Dynamic sensor-based control of robots with visual feedback. *IEEE Journal of Robotics and Automation*, RA-3(5):404–417, October 1987.

[13] J. Weng, P. Cohen, and M. Heriou. Calibration of stereo cameras using a non-linear distortion model. In *Int'l Conf. on Pattern Recognition*, pages 246–253, 1990.