Proceedings of 2004 IEEE/RSJ International Conference on
Intelligent Robots and Systems
September 28 - October 2, 2004, Sendai, Japan

Hybrid Position-Based Visual Servoing with Online Calibration for a Humanoid Robot

Geoffrey Taylor and Lindsay Kleeman
ARC Centre for Perceptive and Intelligent Machines in Complex Environments
Department of Electrical and Computer Systems Engineering
Monash University 3800, Australia
Email: {Geoffrey.Taylor;Lindsay.Kleeman}@eng.monash.edu.au

***Abstract*— This paper addresses the problem of visual servo control for a humanoid robot in an unstructured domestic environment. The important issues in this application are autonomous planning, robustness to camera and kinematic model errors, large pose errors, occlusions and reliable visual tracking. Conventional image-based or position-based visual servoing schemes do not address these issues, which motivated the proposed hybrid position-based scheme exploiting fusion of visual and kinematic measurements. Kinematic measurements provide robustness to visual distractions, and allow servoing to continue when the end-effector leaves the field of view. Visual measurements provide the complementary benefits of accurate pose tracking and online estimation of the hand-eye transformation for kinematic calibration. Furthermore, it is shown that calibration errors in the focal length and baseline can be approximated as an unknown scale of the end-effector, which can be estimated in the tracking filter to overcome camera calibration errors. The improved accuracy and robustness compared to conventional position-based servoing is demonstrated experimentally.**

I. Introduction

To perform useful tasks, a domestic humanoid robot must be able to recognize objects and accurately control the relative pose of the end-effector. Object recognition and tracking are discussed in our previous work [10]–[12], while this paper addresses the problem of end-effector control. If the joint encoders, camera parameters and kinematic model are known, control is a trivial problem in inverse kinematics [1]. However, practical humanoid robots are likely to violate these conditions. For instance, affordability requires low manufacturing tolerances and cheap sensors, while light-weight, compliant limbs are necessary for safety and low power consumption, but are difficult to model kinematically. *Visual servoing* overcomes these issues by incorporating visual measurements of the end-effector in the control loop, but faces significant challenges in an unstructured domestic environment. In particular, large pose error and obstacles may render the end-effector unobservable, while clutter confuses visual tracking.

Visual servoing schemes are generally distinguished as *image-based* or *position-based*, and *endpoint open-loop* (EOL) or *endpoint closed-loop* (ECL) [7]. An ECL controller observes both the end-effector and target, and an EOL controller observes only the target while using kinematic control. While ECL control is less sensitive to calibration errors, EOL is not affected by occlusion of the end-effector. In image-based servoing, the control error is calculated directly from image plane measurements. For humanoid control, the main drawback is the unpredictability of trajectories in Cartesian space, particularly for large initial pose error. Recent work shows that this can be avoided by decoupling orientation and translation [9]. In position-based servoing [13], [14], the control error is calculated from 3D pose parameters reconstructed from visual measurements, which facilitates Cartesian path planning. However, this approach is sensitive to camera calibration and the chosen pose estimation algorithm. Also, the gripper is not necessarily maintained within the visual field, although solutions to this issue have been proposed [2]. Other servoing schemes have been demonstrated, including 2-1/2-D visual servoing [9] and frameworks based on linear approximations [3]. A recent comparison revealed little variation in the stability, robustness and sensitivity to calibration errors of visual servoing schemes [4].

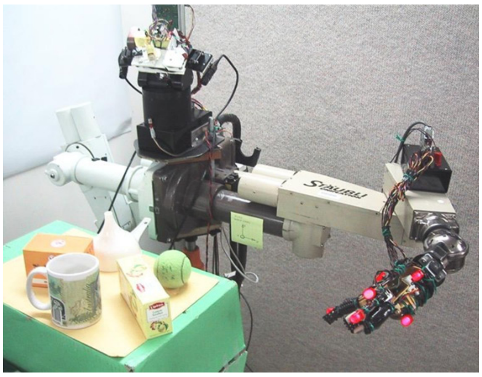


Fig. 1. Upper-torso humanoid robot for domestic tasks.

For autonomous planning, our humanoid robot must rely on internal models to calculate the control reference. Camera calibration errors thus influence positioning accuracy regardless of the approach used, and must be addressed by the controller. Many servoing schemes emphasize maintaining features within the field of view. However, the possibility of large initial pose errors and obstacles in domestic tasks may result in unavoidable loss of visual feedback. Thus, visual servoing for domestic tasks should instead be characterized by robustness to occlusions.

This work adopts position-based control, since humanoid tasks are naturally planned in Cartesian space [6]. Our

0-7803-8463-6/04/$20.00 ©2004 IEEE

proposed scheme achieves robustness to both calibration errors and occlusions by fusing kinematic and visual measurements in a Kalman filter. Kinematic measurements provide robustness to visual distractions and occlusions, while visual measurements provide accurate pose tracking. This *hybrid* approach benefits of both EOL and ECL control and offers a significant improvement over conventional schemes. However, since both camera and kinematic models are required, the influence of calibration errors is two-fold. To facilitate reliable, long-term operation, the proposed scheme therefore incorporates online calibration. Fusion of kinematic and visual measurements was previously demonstrated for control of a dextrous hand [15], and our paper improves on this work by handling occlusions and introducing online calibration.

The following section defines the configuration of our robot, and the controller is formulated in Section III. Robust end-effector tracking using visual and kinematic measurements is described in Section IV, and Section V details the experimental implementation. Finally, the results in Section VI demonstrate the improvement of the proposed scheme over similar position-based techniques.

II. Coordinate Frames and Notation

In this paper, 3D points are represented in upper-case and 2D points in lower-case. Coordinate frames are specified in superscript, such as ${}^{A}\mathbf{X}$, and the homogeneous transformation matrix ${}^{B}\mathrm{H}_{A}$ transforms points from frame A to B as ${}^{B}\mathbf{X} = {}^{B}\mathrm{H}_{A}{}^{A}\mathbf{X}$. Our experimental platform (see Figure 1) consists of a 3-axis stereo head and Puma arms, and Figure 2 shows the relevant coordinate frames. Frame C is rigidly attached to the stereo cameras and W is attached to the base of the head, while ${}^{W}\mathrm{H}_{C}$ is parameterized by pan and tilt angles. The cameras are positioned in rectilinear configuration at ${}^{C}X = \pm b$. Frame B is attached to the Puma base, E is attached to the end-effector, O locates the pose of the object, and G describes the desired relative pose of the object and gripper. The end-effector pose is equivalently represented by the transformation ${}^{W}\mathrm{H}_{E}$ or pose vector ${}^{W}\mathbf{p}_{E} = (X, Y, Z, \phi, \theta, \psi)$, where X, Y and Z are translations, and Euler angles ϕ, θ and ψ represent orientation.

III. Visual Servo Controller

The controller in this work follows the formulation in [7], with the addition of the grasp frame. The task is to control the end-effector to align the grasp frame with the object frame. The control error is the pose error between the grasp and object frames, given by the transformation:

$$ {}^{G}\mathrm{H}_{O} = ({}^{W}\mathrm{H}_{E}{}^{E}\mathrm{H}_{G})^{-1} \cdot {}^{W}\mathrm{H}_{O} \tag{1} $$

The above transformation is identity when the control task is achieved. For each new observation of ${}^{W}\mathrm{H}_{O}$ and ${}^{W}\mathrm{H}_{E}$, the controller calculates the velocity screw $(\Omega, \mathbf{V})^{\top}$ of the end-effector that drives the pose error to zero. Using proportional control with gains k_1 and k_2, the desired velocity screw in the grasp frame is:

$$ {}^{G}\Omega = k_1 {}^{G}\theta_O {}^{G}\mathbf{A}_O \tag{2} $$

$$ {}^{G}\mathbf{V} = k_2 {}^{G}\mathbf{T}_O - {}^{G}\Omega \times {}^{G}\mathbf{T}_O \tag{3} $$

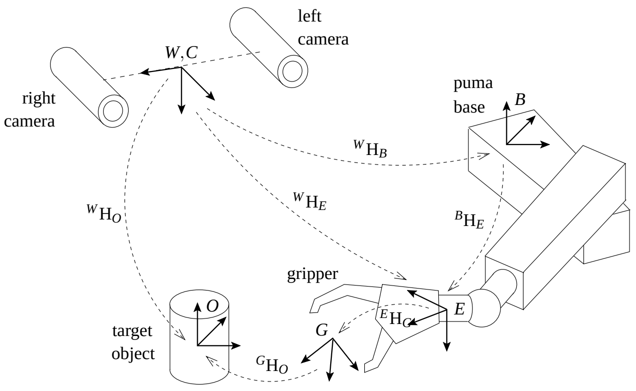


Fig. 2. Coordinate frames and transformations for visual servoing.

where ${}^{G}\mathbf{T}_{O}$ is the translational component of ${}^{G}\mathrm{H}_{O}$ and $({}^{G}\theta_O, {}^{G}\mathbf{A}_O)$ is the rotational component (angle and axis of rotation). In practice, the Puma operates in "tool-tip" mode; the velocity screw is transformed by ${}^{E}\mathrm{H}_{G}$ (see [7] for transforming velocity screw) before passing to the controller. The fundamental limitation of this formulation is the uncertainty in ${}^{W}\mathrm{H}_{E}$, which may be estimated by direct visual observation (ECL control) or kinematically through ${}^{W}\mathrm{H}_{B}$ and ${}^{B}\mathrm{H}_{E}$ (EOL control). The hybrid controller proposed in the paper fuses visual and kinematic estimates to exploit the benefits of both ECL and EOL control.

IV. Robust Gripper Tracking

The pose of the end-effector is robustly estimated by fusing visual and kinematic measurements in an *Iterated Extended Kalman filter* (IEKF), which is commonly used for visual servoing [13], [14]. In our application, the state $\mathbf{x}(k)$ consists of the end-effector pose and velocity screw, used to calculate the control error in equation (1), and additional parameters for online calibration as discussed in the following sections. Constant velocity dynamics are assumed for the pose parameters, while the calibration parameters are modelled as constants. Details of the IEKF equations can be found in [8]. The following sections describe the visual and kinematic sensor models which are used to predict the measurement vector $\mathbf{y}(\mathbf{x})$ for a given state $\mathbf{x}$. Section IV-C then summarizes the practical implementation of the IEKF.

A. Visual Measurements

Our servoing framework uses 3D model-based visual tracking with active artificial cues (LEDs), which are more readily detected than passive markers. The manually measured LED locations, ${}^{E}\mathbf{G}_i$, $i = 1 \dots 6$, form the model shown in Figure 3. The thumb and forefinger LEDs rotate about $\mathbf{P}_5$ and $\mathbf{P}_3$ through angles θ_5 and θ_3 respectively, and are recalculated whenever these angles change. The measurement model to predict the image plane projections ${}^{L,R}\mathbf{g}_i$ is formulated from the camera model as:

$$ {}^{L,R}\hat{\mathbf{g}}_i = {}^{L,R}\mathrm{P}{}^{C}\mathrm{H}_{W}{}^{W}\widehat{\mathrm{H}}_{E}{}^{E}\mathbf{G}_i \tag{4} $$

where ${}^{L,R}\hat{\mathbf{g}}_i$ are the predicted measurements, ${}^{W}\widehat{\mathrm{H}}_{E}$ is the predicted pose and ${}^{L,R}\mathrm{P}$ are the camera projection matrices.

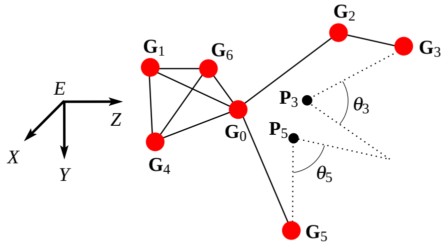


Fig. 3. Articulated model of active LED features for tracking.

In real systems with imprecise camera models, the estimated pose is an unknown transformation of the *actual* pose, which we now examine in the following error analysis. To simplify the analysis, we ignore the predicted state and kinematic measurements in the Kalman filter state estimator, and consider a state estimator that minimizes only the sum of squared image plane errors, defined by

$$D^2({}^W\mathbf{p}_E) \equiv \sum_i d^2({}^L\hat{\mathbf{g}}_i({}^W\mathbf{p}_E), {}^L\mathbf{g}_i) + d^2({}^R\hat{\mathbf{g}}_i({}^W\mathbf{p}_E), {}^R\mathbf{g}_i) \quad (5)$$

where $d(\hat{\mathbf{g}}, \mathbf{g}_i)$ is the Euclidean distance between the predicted and actual measurements. Consider $\mathbf{X} = (X, Y, Z)^\top$ observed at ${}^L\mathbf{x} = ({}^Lx, {}^Ly)^\top$ and ${}^R\mathbf{x} = ({}^Rx, {}^Ry)^\top$. Assuming pin-hole cameras with calibrated focal length f, baseline $2b$ and zero verge angle ν (rectilinear stereo), the predicted measurements in inhomogeneous coordinates are:

$${}^{L,R}\hat{\mathbf{x}} = \frac{f}{Z}(X \pm b, Y)^\top \quad (6)$$

taking the positive sign for L and negative for R. Then, the optimal reconstruction from minimization of $D^2(\mathbf{X})$ is:

$$\widehat{\mathbf{X}} = \frac{b}{{}^Lx - {}^Rx}\left({}^Lx + {}^Rx,\ {}^Ly + {}^Ry,\ 2f\right)^\top \quad (7)$$

Now, let $2b^*$, f^* and ν^* represent the *actual* baseline, focal length and verge angle. When the calibrated and actual verge angle differ (violating rectilinear stereo) the actual measurements ${}^{L,R}\mathbf{x}$ are effectively rotated by $\nu - \nu^*$:

$${}^{L,R}\mathbf{x} = \frac{f^*\left((X \pm 2b^*)\cos(\nu - \nu^*) \mp Z\sin(\nu - \nu^*),\ Y\right)^\top}{Z\cos(\nu - \nu^*) \pm (X \pm 2b^*)\sin(\nu - \nu^*)}$$

Applying a small angle approximation to the above, substituting the result into equation (7) and taking a Taylor series expansion about $f = f^*$, $b = b^*$ and $\nu = \nu^*$, the relationship between $\mathbf{X}$ and $\widehat{\mathbf{X}}$ can be approximated as:

$$\widehat{\mathbf{X}}(b, f, \nu) \approx K_1(b, \nu)\mathbf{X} + K_2(f)(0,\ 0,\ Z)^\top + \mathbf{T}(\nu) \quad (8)$$

where

$$\begin{aligned} K_1(b,\nu) &= 1 + \frac{b - b^*}{2b^*} + \frac{X^2 + Z^2}{2b^* Z}(\nu - \nu^*) && (9) \\ K_2(f) &= (f - f^*)/f^* && (10) \\ \mathbf{T}(\nu) &= (\nu - \nu^*)(2b^*X/Z,\ 0,\ 2b^*)^\top && (11) \end{aligned}$$

Since the LEDs occupy a small, distant volume, the terms $(X^2 + Z^2)/2b^*Z$ and $2b^*X/Z$ are approximately constant and equation (8) can be treated as linear.

Equation (8) reveals that focal length error scales the Z coordinate by $K_2(f)$, which can be neglected due to good calibration of f. Conversely, the verge and baseline are poorly calibrated and vary as the head moves. Verge error introduces a translation $\mathbf{T}(\nu)$, which does not effect servoing accuracy since the object and gripper are equally biased when aligned. Thus, the main contributor to reconstruction error is the uniform scale $K_1(b, \nu)$.

We now consider the gripper points $\mathbf{G}_i = (X_i, Y_i, Z_i)^\top$. To analyse the effect of $K_1(b, \nu)$, we assume $\sum \mathbf{G}_i = \mathbf{0}$ and the pose $\mathbf{T}_E = (X_E, Y_E, Z_E)^\top$ is purely translational. Let $\mathbf{G}_i^* = K_1(\mathbf{G}_i + \mathbf{T}_E)$ represent the location of LEDs observed by poorly calibrated cameras (that is, after scaling by K_1). The *actual* measurements ${}^{L,R}\mathbf{g}_i$ are the projections of $\mathbf{G}_i^*$:

$${}^{L,R}\mathbf{g}_i = \frac{f}{K_1(Z_i + Z_E)}\left(K_1(X_i + X_E) \pm b,\ K_1(Y_i + Y_E)\right)^\top \quad (12)$$

Now, let $\widehat{\mathbf{T}}_E$ represent the estimated pose. From equation (4), the predicted (unscaled) measurements are:

$${}^{L,R}\hat{\mathbf{g}}_i = \frac{f}{Z_i + \widehat{Z}_E}(X_i + \widehat{X}_E \pm b,\ Y_i + \widehat{Y}_E)^\top \quad (13)$$

Substituting equations (12) and (13) into (5) and solving the minimization analytically, the optimal pose estimate is:

$$\widehat{\mathbf{T}}_E = \frac{f(\mathbf{G}_i, \mathbf{T}_E) + 4NZ_E b^2}{K_1 f(\mathbf{G}_i, \mathbf{T}_E) + 4NZ_E b^2} K_1 \mathbf{T}_E \quad (14)$$

where $f(\mathbf{G}_i, \mathbf{T}_E) = X_E \sum X_i Z_i + Y_E \sum Y_i Z_i - Z_E(\sum X_i^2 + \sum Y_i^2)$ and N is the number of LEDs. From equation (14), the intuitive result $\widehat{\mathbf{T}}_E = K_1 \mathbf{T}_E$ (ie. the estimated pose is scaled by K_1) only occurs when $N = 1$ or $K_1 = 1$. Otherwise, the induced bias is a function of the model itself; the estimated pose for two different objects at the *same* position, and in the presence of the *same* calibration errors, will be different. For visual servoing, the pose error may not reduce to zero even when the task is achieved.

The above analysis suggests a solution to the problem of camera calibration errors, by simply estimating the unknown scale K_1 along with the pose ${}^W\mathbf{p}_E$ in the tracking filter. This is achieved by replacing equation (4) with

$${}^{L,R}\hat{\mathbf{g}}_i(K_1, {}^W\mathbf{p}_E) = {}^{L,R}\mathrm{P}\,{}^C\mathrm{H}_W\,{}^W\widehat{\mathrm{H}}_E \cdot (K_1\,{}^E\mathbf{G}_i) \quad (15)$$

To sufficiently constrain the scale, four or more measurements are required with at least one feature on each image plane. Monocular measurements do not sufficiently constrain the scale since the projection from a single camera is non-invertible. Section IV-C describes how the IEKF handles the state update when the scale is unconstrained.

B. Kinematic Measurements

The Puma controller reports the end-effector pose in the base frame, ${}^B\mathrm{H}_E$, from kinematic measurements. Thus, an appropriate measurement prediction is

$${}^B\widehat{\mathrm{H}}_E = {}^B\mathrm{H}_W\,{}^W\widehat{\mathrm{H}}_E \quad (16)$$

where ${}^B\widehat{\mathrm{H}}_E$ is the predicted measurement, ${}^W\widehat{\mathrm{H}}_E$ is the predicted pose, and ${}^B\mathrm{H}_W$ is the hand-eye transformation. Typically, hand-eye calibration is performed once and then assumed to remain constant. We propose the alternative approach of treating the hand-eye transformation as a dynamic bias between the kinematically and visually observed pose. This is implemented by adding $^{B}\mathrm{H}_W$ to the state vector, and dynamically updating the transformation through equation (16). However, solving for $^{B}\mathrm{H}_W$ requires $^{W}\mathrm{H}_E$ to be sufficiently constrained by visual measurements. When the solution is unconstrained, $^{B}\mathrm{H}_W$ does not participate the state update, as described below.

C. Kalman Filter Implementation

During servoing, the IEKF is updated at the sample rate of visual measurements. The state vector can now be summarized as $\mathbf{x}(k) = (^{W}\mathbf{p}_E(k), {}^{W}\dot{\mathbf{r}}_E(k), {}^{B}\mathbf{p}_W(k), K_1(k))^\top$ where $^{W}\mathbf{p}_E(k)$ and $^{W}\dot{\mathbf{r}}_E(k)$ are the pose and velocity screw of the end-effector, $^{B}\mathbf{p}_W(k)$ is the hand-eye transformation (expressed as a pose vector) and scale $K_1(k)$ compensates for camera calibration errors. Combining visual and kinematic measurements, the measurement vector can be written as $\mathbf{y}(k) = (^{L}\mathbf{g}_0(k), {}^{R}\mathbf{g}_0(k), \ldots, {}^{L}\mathbf{g}_6(k), {}^{R}\mathbf{g}_6(k), {}^{B}\mathbf{p}_E(k))^\top$ where $^{L,R}\mathbf{g}_i(k)$, $i = 1\ldots 6$, are the positions of the LEDs on the left and right image planes, and $^{B}\mathbf{p}_E(k)$ represents the kinematic pose of the end-effector. Measurement prediction is given by equations (15) and (16), and occluded LEDs are excluded from the state update by setting a large error variance for LEDs that are not observed.

As mentioned above, special care must be taken to constrain the estimated state. It is well known that 3D pose can be recovered from three monocular measurements, although multiple solutions may exist [5]. However, monocular measurements do not constrain K_1, while $^{B}\mathbf{p}_W(k)$ requires an estimate of visual pose. Furthermore, robust LED detection relies on global consistency (see Section V). The following hierarchy of state estimators, based on the number of visual features n_L and n_R, is adopted to ensure only constrained parameters are updated:

- $n_L < 3$ **and** $n_R < 3$**:** All visual measurements are discarded due to possible association errors and the IEKF uses only kinematic measurements.
- $n_L \geq 3$ **or** $n_R \geq 3$ **(but not both):** Three or more monocular measurements sufficiently constrain the pose and hand-eye transformation, but not scale.
- $n_L \geq 3$ **and** $n_R \geq 3$**:** Sufficient measurements exist to constrain all state variables.

When necessary, state parameters are excluded from the state update by setting the corresponding rows and columns of the measurement Jacobian to zero.

Since the IEKF solves the non-linear system equations numerically, a good initial state estimate is required. If the hand-eye transformation is already known from previous trials, the initial pose is calculated from kinematic measurements, otherwise an autonomous calibration procedure is executed. Calibration begins by scanning the workspace for the colour of the LEDs. To avoid association errors, the LEDs are then individually lit and measured in successive frames using colour filtering and image differencing. Finally, the initial pose and scale factor are estimated by minimizing equation (5) using the Levenberg-Marquardt algorithm, with the scale initially set to unity, orientation to zero, and translation estimated from the average position of LEDs. Once the pose is known, the hand-eye transformation is initialized as $^{B}\mathrm{H}_W = {}^{B}\mathrm{H}_E {}^{W}\mathrm{H}_E^{-1}$, where $^{B}\mathrm{H}_E$ is reported by the Puma controller.

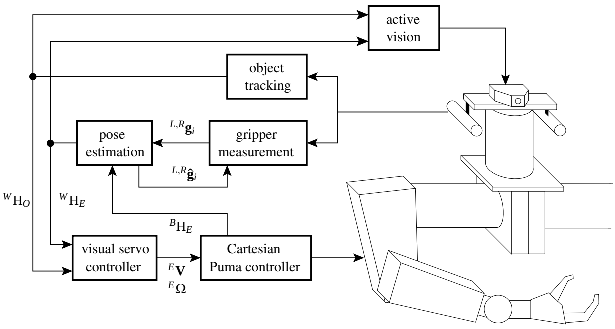


Fig. 4. Block diagram of visual servoing control loop.

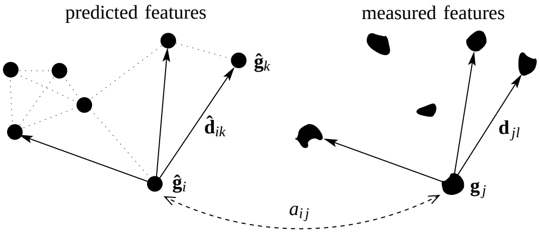
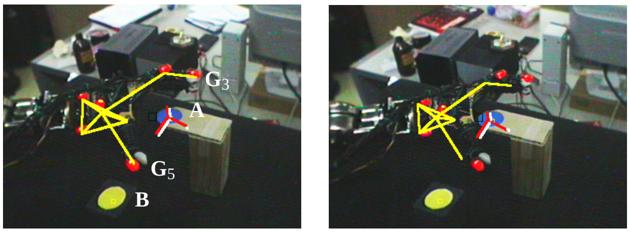


Fig. 5. Associating LEDs and candidates based on global matching.

V. Implementation

Figure 4 illustrates the components of the controller. *Active vision* controls the gaze direction to maximize visual information. The control strategy depends on the pose error: when the error is large, the cameras track only the object (with EOL visual servoing), while the mid-point between the object and gripper is tracked when the error is small. *Gripper measurement* identifies the LEDs on the gripper using the process below, which are fed along with kinematic measurements into the *pose estimation* block. *Object tracking* similarly estimates the pose of the grasping target. The *visual servo controller* calculates the pose error and velocity screw, which is finally actuated by the *Cartesian Puma controller*. The control cycle continues until the pose error is sufficiently small.

Stereo images are captured at 25 Hz and 384×288 pixel resolution and processed on a duel Xeon PC. The main problem is to measure the position of each LED in the presence of occlusions and background clutter. The background is discarded by identifying a *region of interest* (ROI) as the bounding box enclosing the predicted LED locations. A two step process is then applied to detect LEDs in the ROI. First, colour filtering produces a binary image identifying red pixels, and the centre of mass of connected blobs serves as initial LED candidates. The candidate associated with each LED (ie. the *association problem*) is identified using a global matching algorithm, which is more reliable than closest-point matching but cheaper than a full search. The process is illustrated in Figure 5; the association a_{ij} between LED $\hat{\mathbf{g}}_i$ and candidate $\mathbf{g}_j$ is



(a) Proposed controller.

(b) EOL controller.

Fig. 6. Pose of object and gripper at completion of servoing task.

supported by other prediction/candidate pairs with similar relative displacements. In this example, a_{ij} is supported by three pairs, including $\hat{\mathbf{g}}_k$ and $\mathbf{g}_l$ with matching displacement vectors $\hat{\mathbf{d}}_{ik}$ and $\mathbf{d}_{jl}$. The algorithm searches for all good associations and determines the largest self-consistent set. At least three self-consistent associations are required for sufficient confidence in the measurements.

VI. Experimental Results

The proposed method was tested using the positioning task illustrated in Figure 6(a). The target consists of two coloured markers, **A** and **B**, and the goal is to centre **A** between the thumb and index finger, while aligning all points collinearly. To achieve this goal, G is placed at the midpoint between $\mathbf{G}_3$ and $\mathbf{G}_5$ with the y-axis pointing towards $\mathbf{G}_5$, and O is centred at **A** with the y-axis pointing towards **B**. After completing each trial, the positioning accuracy is measured as the translational error e_T between $\widehat{\mathbf{A}}$ and the midpoint between the fingertips:

$$e_T = \left| \frac{1}{2}(\widehat{\mathbf{G}}_3 + \widehat{\mathbf{G}}_5) - \widehat{\mathbf{A}} \right| \tag{17}$$

where $\widehat{\mathbf{A}}$, $\widehat{\mathbf{B}}$, $\widehat{\mathbf{G}}_3$ and $\widehat{\mathbf{G}}_5$ are calculated from stereo measurements averaged over a ten frames. The rotational error e_θ is the angle formed by the lines $(\widehat{\mathbf{A}}, \widehat{\mathbf{B}})$ and $(\widehat{\mathbf{G}}_3, \widehat{\mathbf{G}}_5)$:

$$e_\theta = \cos^{-1}\left(\frac{\widehat{\mathbf{G}}_5 - \widehat{\mathbf{G}}_3}{|\widehat{\mathbf{G}}_5 - \widehat{\mathbf{G}}_3|} \cdot \frac{\widehat{\mathbf{B}} - \widehat{\mathbf{A}}}{|\widehat{\mathbf{B}} - \widehat{\mathbf{A}}|} \right) \tag{18}$$

A. Positioning Accuracy

In this experiment, the proposed controller was compared to conventional ECL and EOL schemes. The ECL controller was implemented by removing the scale and kinematic parameters and measurements from the IEKF, and the EOL controller was implemented by discarding the IEKF and using only kinematic measurements. For each case, five trials were performed with an initial pose error of about 100 mm and 0.25 rad (the camera and target both initially visible). Figure 6 shows the completion of a typical trial for the proposed and EOL controllers, with the estimated gripper overlaid in yellow and O and G indicated in white and red respectively. Table I shows the average final pose error and error variance. As expected, accuracy is lowest for the EOL controller, while the proposed controller is bounded only by the servoing termination threshold. The improvement over ECL control indicates that calibration is important for accurate visual servoing, and can be achieved using the proposed online methods.

TABLE I
Experimental positioning accuracy.

Controller	e_T (mm)	$var(e_T)$ (mm^2)	e_θ (rad)	$var(e_\theta)$ (rad^2)
proposed	5.4	0.4	0.11	0.003
ECL	29.4	0.9	0.22	0.01
EOL	54.5	1.6	0.18	0.008

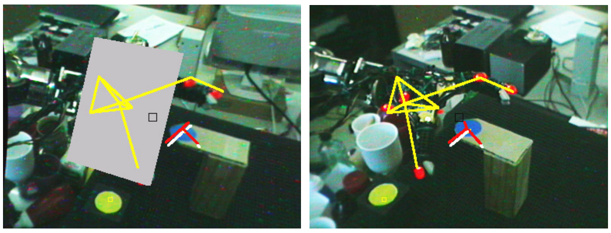


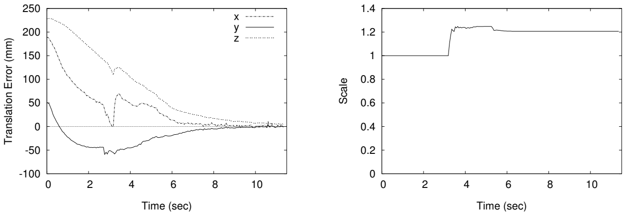
Fig. 7. Stereo image of final pose at completion of the servoing task in the presence of visual distractions and occlusions.

B. Tracking Robustness

This second experiment tests our controller in the presence of poor tracking conditions: the end-effector is initially outside the field of view and a *virtual 3D obstacle* is rendered on the left image plane. Figure 7 shows the completed task, and Figure 8 plots the translation error and estimated scale. Scale estimation only commence after the first three seconds, when the gripper enters the field of view. After six seconds, the gripper is obscured by the virtual obstacle and the scale parameter remains fixed at the most recent estimate (hand-eye transformation estimation continues based on the fixed scale). The controller achieves a final pose error of $e_T = 10.9$ mm and $e_\theta = 0.087$ radian (the virtual obstacle is removed for these measurements), which is only a small reduction in accuracy from the ideal conditions considered in the previous experiment.

C. Effect of Camera Calibration Errors

Calibration errors are now deliberately introduced to test the bounds of the error model in Section IV-A. The cali-



(a) Translation error

(b) Estimated scale

Fig. 8. Controller performance in the presence of occlusions and clutter.

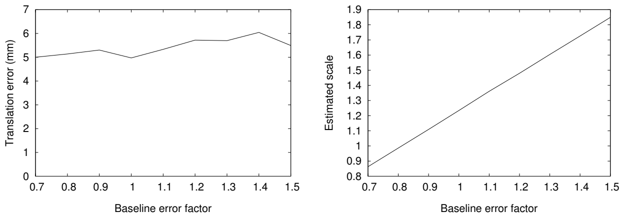


Fig. 9. Translation error and scale for baseline errors.

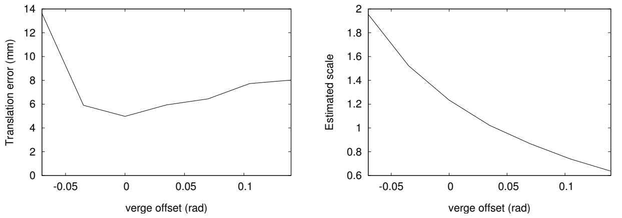


Fig. 10. Translation error and scale for verge errors.

brated baseline (*not* the mechanical baseline) was scaled by 0.7 to 1.5, and the verge angle was offset by -0.07 to 0.14 radian (-6 to +8 degrees). Figure 9 shows the translational pose error and estimated scale at the completion of trials with varying baseline, and indicate that the positioning accuracy is reasonably independent of baseline error when the scale is estimated using the proposed error model. Figures 10 shows the results for verge offset and verify that the error model is only valid within the bounds of the small angle approximations used to derive equation (8). This result indicates that a linear error model is sufficient to compensate for verge errors of about ± 0.05 rad.

VII. Discussion and Conclusions

This paper presented the development and implementation of a 3D model-based visual servoing framework for a domestic humanoid robot. The proposed framework emphasizes robustness to occlusions and online compensation for calibration errors. Sensitivity to calibration errors is usually considered the primary drawback of position-based visual servoing. However, we show that the effect of arbitrary errors in the baseline and small errors in the verge angle can be modelled as an unknown scale. Experimental results demonstrate that the accuracy of the controller is improved by estimating the unknown scale along with the pose of the end-effector. The proposed error model is appropriate for the practical camera platform, since the baseline and verge angle continuously vary as the head is actuated while the other camera parameters are fixed or easily calibrated.

Conventional position-based visual servoing schemes employ either EOL or ECL control and are therefore susceptible to hand-eye calibration errors and/or visual occlusions. The robust pose estimator proposed in this paper avoids both issues by optimally fusing kinematic and visual measurements. Kinematic measurements allow the controller to operate (with reduced accuracy) while the gripper is occluded, and visual measurements provide accurate pose control and online estimation of the hand-eye transformation. Experimental results verify both the increased accuracy gained with visual feedback and the robustness to occlusions gained with kinematic feedback. The combination of scale estimation and kinematic/visual fusion proposed in this work overcomes many of the classical problems associated with position-based visual servoing and provides a useful framework for controlling a humanoid robot in an unstructured environment.

Acknowledgment

This research was funded by the *Strategic Monash University Research Fund for the Humanoid Robotics: Perception, Intelligence and Control*, and the *Australian Research Council Centre for Perceptive and Intelligent Machines in Complex Environments*.

References

[1] M. Becker, E. Kefalea, E. Maël, C. von der Malsburg, M. Pagel, J. Triesch, J. C. Vorbrüggen, R. P. Würtz, and S. Zadel. GripSee: A gesture-controlled robot for object perception and manipulation. *Autonomous Robots*, 6:203–21, 1999.

[2] E. Cervera and P. Martinet. Visual servoing with indirect image control and a predictable camera trajectory. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 381–386, 1999.

[3] R. Cipolla and N. Hollinghurst. Visually guided grasping in unstructured environments. *Robotics and Autonomous Systems*, 19:337–346, 1997.

[4] L. Deng, W. J. Wilson, and F. Janabi-Sharifi. Characteristics of robot visual servoing methods and target model estimation. In *Proc. IEEE Int. Symposium on Intelligent Control*, pages 684–689, 2002.

[5] R. M. Haralick, C. Lee, K. Ottenberg, and M. Nölle. Analysis and solutions of the three point perspective pose estimation problem. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 592–598, 1991.

[6] A. Hauck, M. Sorg, G. Faber, and T. Schenk. What can be learned from human reach-to-grasp movements from the design of robotic hand-eye systems. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 2521–2526, 1999.

[7] S. Hutchinson, G. D. Hager, and P. I. Corke. A tutorial on visual servo control. *IEEE Trans. on Robotics and Automation*, 12(5):651–670, 1996.

[8] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Mathematics in Science and Engineering. Academic Press, New York, 1970.

[9] E. Malis, F. Chaumette, and S. Boudet. 2-1/2-D visual servoing. *IEEE Trans. on Robotics and Automation*, 15(2):238–250, 1999.

[10] G. Taylor and L. Kleeman. Fusion of multimodal visual cues for model-based object tracking. In *Proc. 2003 Australasian Conf. on Robotics and Automation*, pages 1–8, 2003.

[11] G. Taylor and L. Kleeman. Robust range data segmentation using geometric primitives for robotic applications. In *Proc. 9th IASTED International Conference on Signal and Image Processing*, pages 467–472, 2003.

[12] G. Taylor, L. Kleeman, and Å. Wernersson. Robust colour and range sensing for robotics applications using a stereoscopic light stripe scanner. In *Proc. 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 178–183, 2002.

[13] W. Wilson, C. Williams Hulls, and G. Bell. Relative end-effector control using cartesian position based visual servoing. *IEEE Trans. on Robotics and Automation*, 12(5):684–696, 1996.

[14] P. Wira and J. P. Urban. A new adaptive Kalman filter applied to visual servoing tasks. In *Fourth Int. Conf. on Knowledge-Based Intelligent Engineering Systems and Applied Technologies*, pages 267–270, 2000.

[15] Y. Yokokohji, M. Sakamoto, and T. Yoshikawa. Vision-aided object manipulation by a multifingered hand with soft fingertips. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 3201–3208, 1999.