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

# Integration of Robust Visual Perception and Control for a Domestic 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 describes a complete vision-based framework that enables a humanoid robot to perform simple manipulations in a domestic environment. Our system emphasizes autonomous operation with minimal *a priori* knowledge in an unstructured environment, with robustness to visual distractions and calibration errors. For each new task, the robot first acquires a dense 3D image of the scene using our novel stereoscopic light stripe scanner that rejects secondary reflections and cross-talk. A data-driven analysis of the range map identifies and models simple objects using geometric primitives. Objects are reliably tracked through clutter and occlusions by exploiting multimodal cues (colour, texture and edges). Finally, manipulations are performed by controlling the end-effector using a hybrid position-based visual servoing scheme that fuses visual and kinematic measurements and compensates for calibration errors. Two domestic tasks are implemented to evaluate the performance of the framework: identifying and grasping a yellow box without any prior knowledge of the object, and pouring rice from an interactively selected cup into a bowl.**

## I. Introduction

A practical humanoid robot in domestic and industrial applications will be expected to operate with the flexibility, skill and intelligence of a human. In recent years, humanoid robotics research has made important steps towards this goal by addressing problems such as locomotion [11], interaction [9] and learning [2]. Our work develops the perception and control skills necessary to perform simple manipulations (such as pouring a drink) with minimum *a priori* knowledge in an unstructured domestic environment, which forms the basis of important applications such as assisting the elderly and disabled. The proposed framework is based on visual sensing, and integrates our previous work in 3D range scanning [17], data-driven object classification and modelling [16], multimodal tracking [15], and visual servoing [13]. To perform tasks with maximum flexibility, our novel methods address the following key challenges:

*1) Ad hoc tasks with unknown objects:* As a universal aid, almost every task performed by a domestic robot will involve different target objects and surroundings. Many visual perception frameworks for robotic applications require the robot to learn a separate model for each unique object [1], [12], which reduces the flexibility to perform *ad hoc* tasks. In contrast, our work emphasizes a *data-driven* approach to perception, and allows the robot to identify cups, bowls, boxes and other simple objects without explicit models. Lack of prior information about the environment also affects sensing, since vision cannot rely on the presence of particular cues, and task planning.

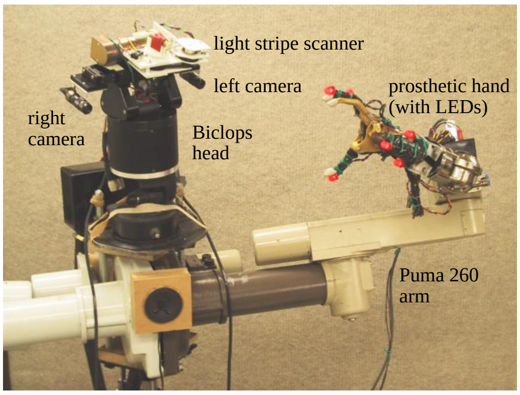


Fig. 1. Metalman, an experimental upper-torso humanoid robot.

*2) Robust sensing in uncontrolled conditions:* The opportunity for association errors in visual sensing is particularly acute in the cluttered, dynamic, unpredictable environment of a domestic robot. Visual sensing is susceptible to background distractions, lighting variation and occlusions. Our work exploits active sensing techniques and fusion of multimodal measurements to develop visual perception and control algorithms with a high tolerance to interference.

*3) Robustness to calibration errors:* Many humanoid robots rely on accurate camera and kinematic calibration to locate and grasp objects [1], [6]. However, poor system models can result from operational wear and the complexity of calibration. Considerations of safety lead to light, compliant structures that are difficult to model kinematically. Furthermore, cheap construction and low maintenance allow robots to achieve widespread application, but make accurate manual calibration impractical. Reliance on manual calibration should be minimized in practical humanoid systems, and our methods instead emphasize active online calibration that can be performed autonomously and whenever necessary.

The following Section provides an overview of our experimental upper-torso humanoid platform (Figure 1), and summarizes the integrated perception and control methods. Two domestic tasks are implemented to evaluate the performance of the proposed framework: identifying and retrieving a yellow box from a cluttered scene without

0-7803-8463-6/04/$20.00 ©2004 IEEE

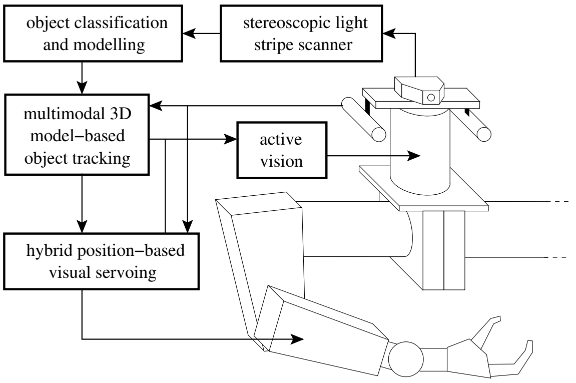

object classification and modelling
stereoscopic light stripe scanner
multimodal 3D model−based object tracking
active vision
hybrid position−based visual servoing

Fig. 2. Block diagram of system components.

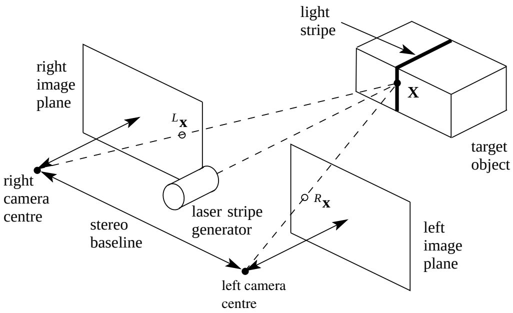


Fig. 3. Stereoscopic stripe scanner.

any prior knowledge of the object, and pouring rice from an interactively selected cup into a bowl. The implementation and results are presented in Sections III and IV.

## II. Proposed Framework

Figure 1 shows our experimental upper-torso humanoid robot, also known as Metalman. The platform consists of hardware-synchronized stereo PAL cameras mounted on a three-axis active head, and two Puma arms with prosthetic hands. A 5 mW laser stripe scanner is mounted above the cameras for active 3D sensing, and LED markers are attached to the hands to facilitate tracking. All image processing is performed on a dual 2.4 GHz dual Xeon PC, with extensive use of parallelization through MMX/SSE instructions and POSIX threads.

Our complete perception and control framework is illustrated in Figure 2. The *stereoscopic light stripe scanner* and *object classification and modelling* blocks work together to provide Metalman with data-driven, textured, polygonal models of objects in the workspace. Extracted objects are classified as cups, bowls, boxes or other simple types based on geometry, which provides Metalman with the flexibility to perform *ad hoc* tasks with previously unknown objects without a large database of learned models. The world model is continuously updated by applying *multimodal 3D model-based object tracking* to captured stereo frames, which exploits both texture and geometric information in the polygon models. Finally, the desired manipulation is performed by controlling the end-effector using visual feedback in a *hybrid position-based visual servoing* framework. During servoing, *active vision* controls the gaze direction to maximize observation of the object and end-effector. The following sections describe the main components of the proposed framework in greater detail.

### A. Stereoscopic Light Stripe Scanning

Reliable object classification and 3D modelling requires dense and accurate colour/range measurements. Passive stereo is usually associated with human 3D sensing but relies on the presence of suitable textures, which is difficult to ensure in uncontrolled conditions. Alternatively, light stripe ranging offers good accuracy but presents other unique challenges: the sensor must cope with secondary reflections, cross talk and other spurious measurements while operating in ambient indoor light. Single-camera scanners typically identify the stripe as the brightest observed feature, without any mechanism to detect and reject interference. Robust scanners have been proposed in previous work to address this problem [5], [8], [19], but remaining issues include assumed scene structure, sensor noise modelling, error recovery and capturing colour.

To overcome these issues, we proposed the stereoscopic light stripe scanner shown in Figure 3 [17]. On corresponding stereo scanlines, measurements of a point $\mathbf{X}$ on the light plane satisfy a linear relationship (in projective space) of the form $^L\mathbf{x} = \mathbf{H}^R\mathbf{x}$, where the transformation H is related to the stereo geometry and position of the light plane. The primary reflection can therefore be identified from a set of noisy candidate measurements by finding the candidate stereo pair ($^L\mathbf{x}$,$^R\mathbf{x}$) that minimizes the squared image plane error $E$ with respect to the ideal projection $^R\hat{\mathbf{x}}$:

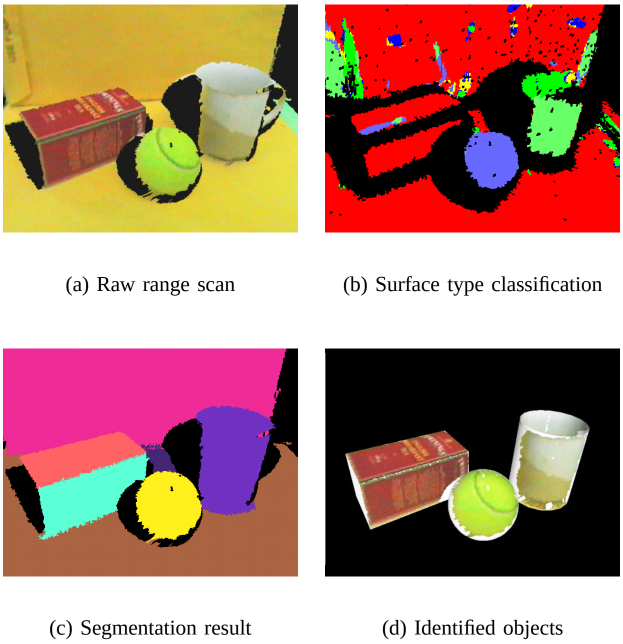
$$E(^R\hat{\mathbf{x}}) = d^2(^L\mathbf{x}, \mathrm{H}^R\hat{\mathbf{x}}) + d^2(^R\mathbf{x}, ^R\hat{\mathbf{x}})$$

where $d^2(\mathbf{x}_1, \mathbf{x}_2)$ is the Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$. The 3D reconstruction is recovered by back-projecting the ideal projection $^R\hat{\mathbf{x}}$ onto the light plane.

Unlike other robust light stripe methods, our validation and reconstruction algorithm is optimal with respect to sensor noise, and achieves greater precision than a single camera scanner. We also demonstrate a simple active method to calibrate the scanner using an arbitrary non-planar target, which allows the sensor to be calibrated during normal operation. Finally, interference rejection allows the scanner to operate in ambient indoor light and thus capture implicitly registered colour and range.

### B. Object Classification and Modelling

Object modelling and classification provides the link between low-level range sensing and high-level planning. For maximum flexibility, the robot must be capable of recognizing cups and other common objects without having seen them previously. This problem is solved by modelling classes of objects as collections of geometric primitives (for example, a cup is a cylinder open at one end) and applying split-and-merge segmentation to extract data-driven primitives from the range map. The extracted primitives are matched with the generic models to locate and classify

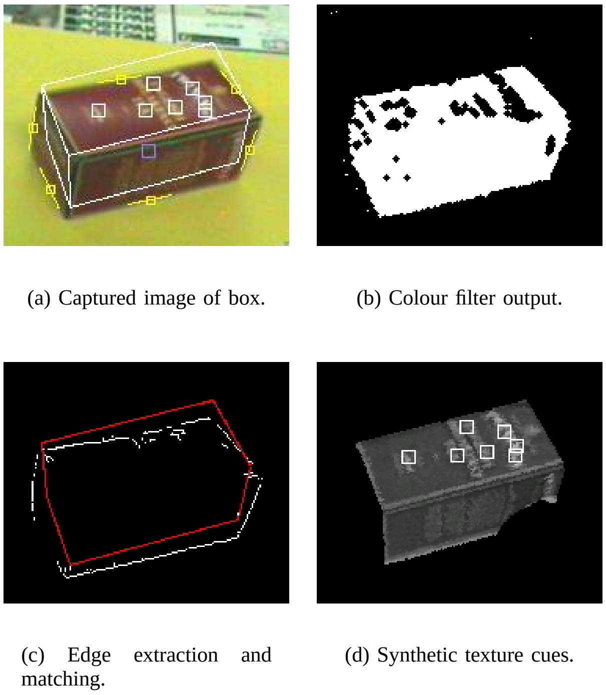

(a) Raw range scan

(b) Surface type classification

(c) Segmentation result

(d) Identified objects

Fig. 4. Object modelling and classification.


(a) Captured image of box.

(b) Colour filter output.

(c) Edge extraction and matching.

(d) Synthetic texture cues.

Fig. 5. Scene analysis for grasping task.

objects in the scene [14]. At the core of the segmentation algorithm is our novel *surface type classifier*, first proposed in [16]. Unlike conventional methods [3], our classifier identifies the local shape at each point on a surface by analyzing principal curvatures and convexity, without the need for fitting arbitrary approximating surface functions. This approach achieves greater noise robustness and ultimately simplifies segmentation. Figures 4(a) and 4(b) show the colour/range scan and surface type classification result for a typical scene, with regions of different shape shown in homogeneous colour. The final segmentation is shown in Figure 4(c) and the textured, polygonal models for identified objects are shown in Figure 4(d).

### C. Multimodal 3D Model-Based Tracking

Once objects have been identified, the world model is updated in real-time by tracking objects in captured stereo frames. Tracking is necessary even for static objects (typical of domestic scenes) to compensate for camera motion and detect collisions or unstable grasps. Conventional model-based tracking algorithms are based on the detection of a single cue such as edges [18] or image templates [7]. However, arbitrary objects may contain too many or too few single-modal features for reliable matching, while lighting variations and background clutter make detection unreliable. Thus, tracking of any single cue is likely to eventually fail in an uncontrolled domestic environment. To overcome this problem, we proposed a multimodal tracking algorithm based on the notion that different cues exhibit independent, complementary failure modes and are therefore unlikely to fail simultaneously [15]. The tracker fuses edge, texture and colour cues in a Kalman filter framework, which allows arbitrary objects to be tracked in visual conditions that cause single-cue trackers to fail, including low contrast, lighting variations, and occlusions.

To illustrate the algorithm, Figure 5(a) shows a subimage captured while tracking the box. The approximate pose (predicted from the previous frame) is overlaid in white. Figure 5(b) shows the output of a colour filter constructed from texture data, and the pixel centroid (blue square in Figure 5(a)) serves as the colour cue. Figure 5(c) shows the output of edge detection, which uses the colour filter to identify silhouette edges. Edge pixels (white) are matched to the predicted edges (red outline), and the resulting measurements are shown in yellow in Figure 5(a). Finally, texture cues are identified by applying a texture quality measure [10] to the synthetic rendered image of the object in Figure 5(d). The extracted templates (outlined in white) are matched to the captured image using sum-of-squared intensity difference, and the matched locations are shown in Figure 5(a). The multimodal measurements are passed to the Kalman filter equations, which minimize the observation error using appropriate measurement prediction models to estimate the new pose.

### D. Hybrid Position-Based Visual Servoing

If the camera parameters and kinematic model of the robot are accurately known, manipulations can be performed using open-loop *look-then-move* kinematic control (for example, see [1]). However, reliance on accurate modelling and calibration may be unrealistic for practical humanoid robots, due to considerations of cost, safety and long-term operation. The effect of kinematic errors can be reduced by closing the loop using visual measurements of the end-effector, in the framework of *visual servoing* [4].

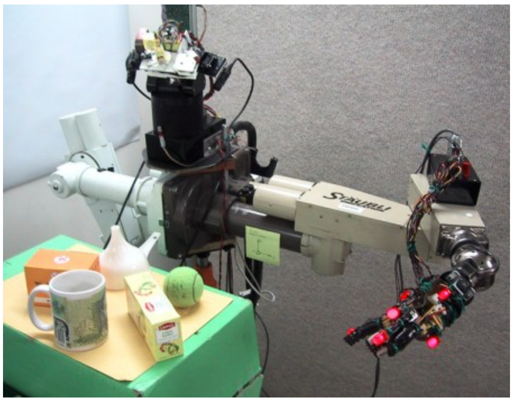


Fig. 6. Arrangement of objects for Experiment 1.

The basic task in visual servoing is to control pose of the end effector to achieve a desired pose relative to the object. In conventional endpoint closed-loop position-based visual servoing, the control error is calculated from the 3D pose of the object and end-effector, reconstructed from visual measurements. However, conventional methods are sensitive to camera calibration errors and fail when the end-effector is obscured (for example, due to a large pose error). We overcome this problem by proposing a self-calibrated hybrid position-based framework [13] that fuses visual and kinematic measurements. Kinematic measurements allow servoing to continue when the end-effector is obscured and improves tracking robustness, while visual measurements provide accurate pose estimation. The hand is tracked using active LED markers to further improve the robustness of visual sensing. The hand-eye transformation is continuously estimated as the bias between the kinematic and visual pose, while camera calibration errors are compensated by introducing a *visual scale* parameter. The proposed method offers greatly improved performance compared to conventional position-based servoing.

## III. Experiment 1: Grasping an Unknown Object

In this first experiment, Metalman is given the simple, common task of finding and retrieving an unknown object specified only as a *yellow box*. The initial experimental arrangement is shown in Figure 6. User interaction is required to establish the initial gaze direction, but the remainder of the task is performed autonomously and requires all of our robust perception and control techniques.

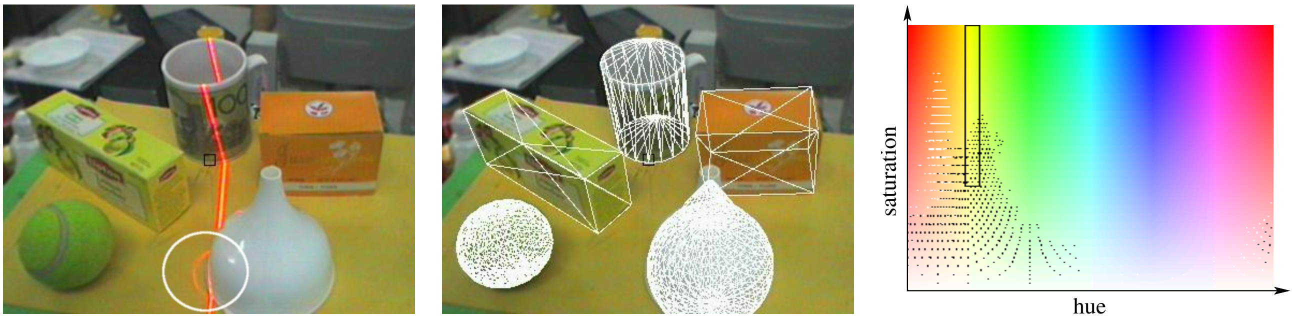
Following the framework outlined above, the robot initially obtains a 3D colour/range map of the scene and applies segmentation to construct a list of candidate objects. Figure 7(a) shows that light stripe scanning causes significant secondary reflections even in this simple scene, but are successfully rejected by our robust scanner. The wireframe overlay in Figure 7(b) shows the extracted objects. Using the classification results, the cup, ball and funnel are immediately rejected as objects of interest. For the remaining two candidates, the *yellow box* (on the left) is identified by tallying the number of texture pixels from the associated polygon model within manually predefined ranges of hue, saturation and intensity. Figure 7(c) plots the hue and saturation of texture pixels for the yellow box (black points) and orange box (white points), with the colours of interest bounded by the black rectangle. The yellow box is successfully identified as the candidate with the highest proportion of yellow pixels.

After identifying the target, a collision-free grasp and approach set-point are planned (see [14]) and passed to the visual servo controller. Figure 8 shows selected frames from the right camera during execution of the grasp. Since the end-effector is initially outside the field of view, servoing commences using kinematic control at 25 Hz, as shown in Figure 8(a) (the estimated pose is indicated by the yellow wireframe overlay). As the end-effector enters the field of view, the controller switches to kinematic and visual fusion with online calibration, as shown in Figure 8(b). During servoing, the box (indicated by the white overlay) is simultaneously tracked using our multimodal algorithm at a reduced rate of 2 Hz, so that the computational load does not compromise the stability of the controller. Figure 8(c) shows the box successfully grasped and lifted.

## IV. Experiment 2: Pouring Task

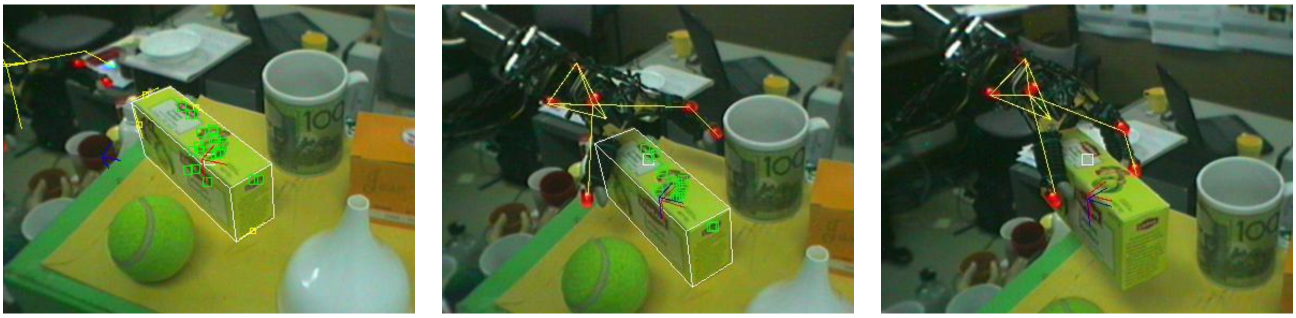
Using a similar experimental arrangement as above, this experiment requires Metalman to grasp a cup of rice and pour the contents into a bowl, both of which are initially unknown. The cup is manually selected to simulate the type of interaction necessary when task specifications are ambiguous. As before, light stripe scanning and segmentation produce the set of candidate objects shown in Figure 9(a). The target bowl is identified as the cylinder with the largest radius. However, two remaining cylinders are identified as possible cups of rice. To resolve the ambiguity, Metalman asks the user for additional information by clicking on the desired cup. While a graphical user interface is appropriate for tele-operated robots, a practical humanoid may alternatively use verbal or gestural interaction [9], [12].

Once the cup and bowl have been identified, task planning generates a series of set-points to carry out the grasping and pouring motion, which are passed to the visual servo controller. Selected frames from the right camera during servoing are shown in Figure 9 (see accompanying video for the complete sequence). As before, the end-effector is initially outside the field of view, as shown in Figure 9(b). Thus, servoing initially employs kinematic control and switches to kinematic/visual fusion when the end-effector becomes visible. Figure 9(c) shows the end-effector just before the grasp, and Figure 9(d) shows the cup successfully grasped and lifted. The cup is visually tracked (at 2 Hz) until grasped, and then assumed to remain fixed relative to the hand. The pouring motion is implemented by positioning the cup above the bowl, and rotating the wrist while lowering the cup towards the bowl, as shown in Figures 9(e) and 9(f). The bowl is successfully tracked throughout the pouring manoeuvre, despite occlusion and a low contrasting background, to ensure accurate placement of the cup.



(a) Secondary reflection (circled). (b) Extracted models (wireframe overlay) (c) Colour histogram to identify yellow box.

Fig. 7. Scene analysis for grasping task.



(a) Box tracking and kinematic servoing. (b) Visual servoing to planned grasp. (c) Cup grasped and lifted.

Fig. 8. Selected frames from grasping task (see http://www.irrc.monash.edu.au/gtaylor/chapter7/chapter7.html for complete sequence).

## V. Summary and Future Work

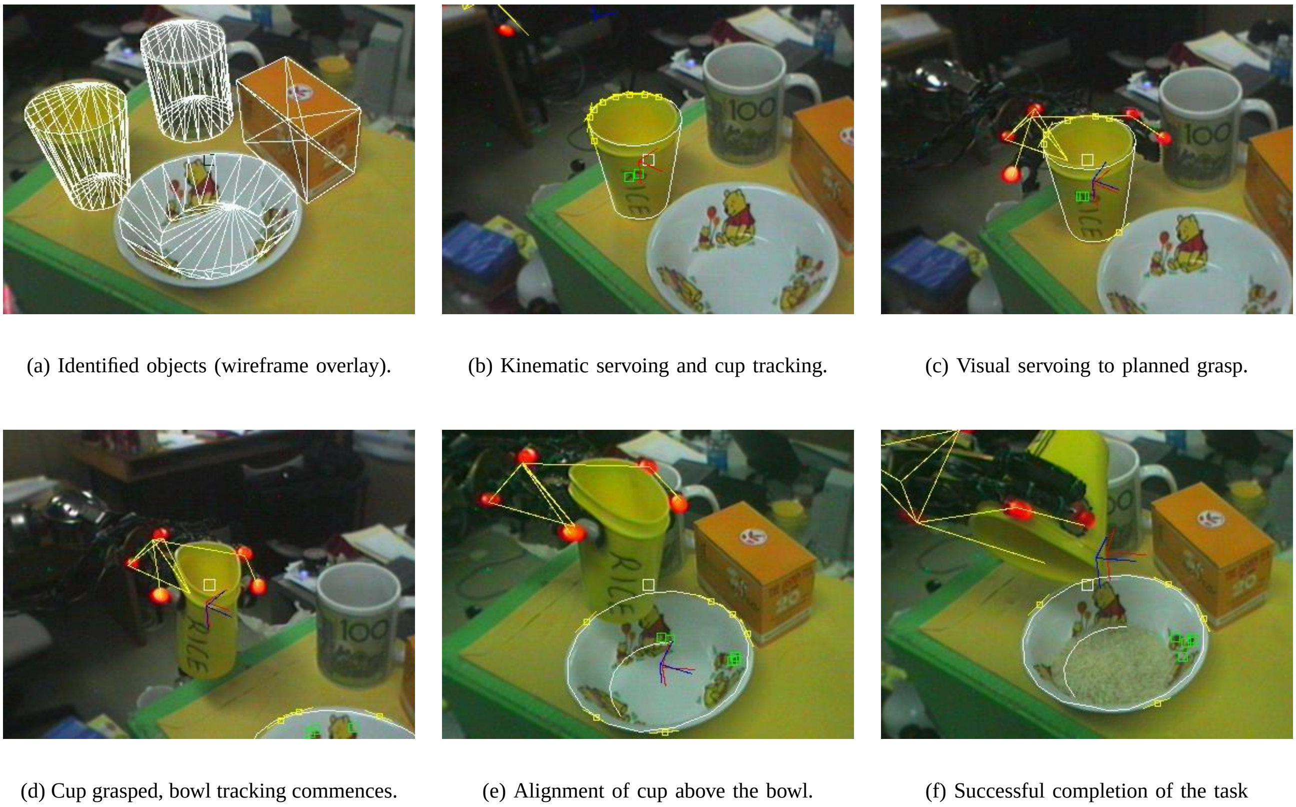
This paper described the integration of robust vision-based perception and control techniques to enable a humanoid robot to perform manipulation tasks in an unstructured domestic environment with minimal *a priori* knowledge. Stereoscopic light stripe scanning provides the robot with an accurate 3D range map of the workspace, despite the presence of secondary reflections, cross-talk and other sources of noise. Data-driven segmentation and object modelling allow the robot to classify previously unknown objects, which provides the level of flexibility necessary for *ad hoc* tasks. Similarly, multimodal 3D model-based tracking enables the robot to track objects with varying appearance while rejecting visual distractions. Finally, manipulations are performed using a self-calibrated position-based visual servoing scheme that fuses visual and kinematic measurements to robustly track the end-effector despite clutter, occlusions and calibration errors.

Two real-world tasks involving classification and manipulation of previously unknown objects were experimentally implemented to demonstrate the effectiveness of the proposed framework. The successful completion of both tasks confirms that our robust perception and control techniques provide a suitable framework for a practical humanoid robot. Nevertheless, the experiments were contrived to simplify scene analysis and task planning. Grasp stability analysis is important when operating in an unpredictable environment, but was neglected in the current implementation. This could have been implemented by tracking objects during grasping, or by integrating tactile and force sensors. Similarly, grasp and trajectory planning could account for obstacles using more sophisticated algorithm.

This work offers a number of interesting directions for future research. Cooperative servoing of both arms provides the opportunity for greater flexibility and more sophisticated manipulations. Learning techniques may allow the robot to recognize new classes of objects by associating collections of geometric primitives with textural or vocal tokens introduced by the user. By interacting with objects (for example, viewing hidden surfaces), models could be refined to improve classification and tracking. Ultimately, the skills developed in this work could be integrated with existing results in verbal and gestural interaction, dextrous manipulation, locomotion and navigation to create a practical robotic assistant for immediate applications such as helping the elderly and disabled with domestic chores.

## Acknowledgment

This research was funded by the *Strategic Monash University Research Fund for the Humanoid Robotics:*



(a) Identified objects (wireframe overlay).

(b) Kinematic servoing and cup tracking.

(c) Visual servoing to planned grasp.

(d) Cup grasped, bowl tracking commences.

(e) Alignment of cup above the bowl.

(f) Successful completion of the task

Fig. 9. Selected frames from pouring task (see http://www.irrc.monash.edu.au/gtaylor/chapter7/chapter7.html for complete sequence).

*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] D.C. Bentivegna, A. Ude, C.G. Atkeson, and G. Cheng. Humanoid robot learning and game playing using PC-based vision. In *Proc. IEEE/RSJ 2002 Int. Conf. on Intelligent Robots and Systems*, volume 3, pages 2449–2454, 2002.

[3] P. J. Besl and R. C. Jain. Segmentation through variable-order surface fitting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(2):167–192, March 1988.

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

[5] M. Magee, R. Weniger, and E. A. Franke. Location of features of known height in the presence of reflective and refractive noise using a stereoscopic light-striping approach. *Optical Engineering*, 33(4):1092–1098, April 1994.

[6] A. Morales, E. Chinellato, A. H. Fagg, and A. P. del Pobil. Experimental prediction of the performance of grasp tasks from visual features. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3423–3428, 2003.

[7] B.J. Nelson and P.K. Khosla. An extendable framework for expectation-based visual servoing using environment models. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 184–189, 1995.

[8] J. Nygards and Å. Wernersson. Specular objects in range cameras: Reducing ambiguities by motion. In *Proc. IEEE Int. Conf. on Multisensor Fusion & Integration for Intelligent Sys.*, pages 320–328, 1994.

[9] O. Rogalla, M. Ehrenmann, R. Zollner, R. Becher, and R. Dillman. Using gesture and speech control for commanding a robot assistant. In *Proc. 11th IEEE Int. Workshop on Robot and Human Interactive Communication*, pages 454–459, 2002.

[10] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[11] A. Takanishi, M. Ishida, Y. Yamazaki, and I. Kato. The realization of dynamic walking by the biped walking robot wl-10rd. In *Proc. '85 Int. Conf. on Advanced Robotics*, pages 459–466, 1985.

[12] M. Takizawa, Y. Makihara, N. Shimada, J. Miura, and Y. Shirai. A service robot with interactive vision – object recognition using dialog with user –. In *Fist Int. Workshop on Language Understanding and Agents for Real World Interaction*, pages 16–23, 2003.

[13] G. Taylor and L. Kleeman. Hybrid position-based visual servoing with online calibration for a humanoid robot. In *IROS 2004, submitted*.

[14] G. Taylor and L. Kleeman. Grasping unknown objects with a humanoid robot. In *Proc. 2002 Australiasian Conference on Robotics and Automation*, pages 191–196, 2002.

[15] 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.

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

[17] 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 Int. Conf. on Intelligent Robots and Systems*, pages 178–183, 2002.

[18] M. Tonko, K. Schäfer, F. Heimes, and H.-H. Nagel. Towards visually servoed manipulation of car engine parts. In *Proc. IEEE Int. Conf. on Robotics and Automation*, pages 1366–1371, 1997.

[19] E. Trucco, R. B. Fisher, A. W. Fitzgibbon, and D. K. Naidu. Calibration, data consistency and model acquisition with a 3-D laser striper. *Int. Journal of Computer Integrated Manufacturing*, 11(4):292–310, 1998.