Real-Time Tracking Using Stereo and Motion:
Visual Perception for Space Robotics

H. Keith Nishihara*   **Hans Thomas***   **Eric Huber**†   **C. Ann Reid***

*Teleos Research
576 Middlefield Road
Palo Alto, CA 94301 USA
Tel (415) 328-8800   Fax (415) 328-8880
E-mail {keith, hans, car}@teleos.com

†The Mitre Corporation
1120 Nasa Road One, Suite 600
Houston, TX 77058
Tel (713) 483-9939
E-mail huber@aio.jsc.nasa.gov

KEY WORDS AND PHRASES

Computer vision, EVAHR, robotic manipulation, visual perception.

INTRODUCTION

The state-of-the-art in computing technology is rapidly attaining the performance necessary to implement many early vision algorithms at real-time rates. This new capability is helping to accelerate progress in vision research by improving our ability to evaluate the performance of algorithms in dynamic environments. In particular, we are becoming much more aware of the relative stability of various visual measurements in the presence of camera motion and system noise. This new processing speed is also allowing us to raise our sights toward accomplishing much higher-level processing tasks, such as figure-ground separation and active object tracking, in real-time. This paper describes a methodology for using early visual measurements to accomplish higher-level tasks; it then presents an overview of the high-speed accelerators developed at Teleos to support early visual measurements. The final section describes the successful deployment of a real-time vision system to provide visual perception for the Extravehicular Activity Helper/Retriever robotic system in tests aboard NASA's KC135 reduced gravity aircraft.

LOW-LEVEL MEASUREMENTS FOR HIGH-LEVEL VISION TASKS

Computer vision systems typically exist as a primary input to some higher-level process. Although many systems have been constructed where there is limited or no feedback from the high-level process to the vision system, there is an emerging belief in the vision community that incorporating powerful feedback mechanisms will greatly increase the capability and durability of various vision algorithms; this new area of vision research has been termed active vision.

Many new issues are raised when we start to think about visual perception as an active, dynamic process interacting closely with higher-level goal directed behavior. For example, what makes a good measurement in this context? Clearly, a perceptual aid for machine vision ought to recover some basic useful information [1]. Furthermore, it should have an easy-to-model behavior that allows its user to employ it intelligently in new situations.

Two particularly important qualities of a visual measurement are meaningfulness and minimality.

Meaningful. A visual measurement device should derive useful information from the visual scene. This usually means recovering something about the physical surfaces that gave rise to the visual images. Range from stereo, surface orientation, and local image velocity are examples. In addition, there is considerable latitude in how information can be presented as an output, and this can significantly influence the effectiveness of the device for solving perception problems. As far as possible, output from the measurement device should exhibit a consistent, dynamic behavior that encourages the learning of strategies for making more specialized measurements. For example, in the case of a stereo correlator, static estimates of range would be enhanced by information about the shape of the correlation peak used to derive that range and the stability of that information across time and spatial position.

Minimal. A user's ability to exploit a measurement device effectively in a wide range of sensing environments depends to a large extent on how well that user is able to anticipate what the device will do in a new situation. This

is easier to do with devices that have consistent, easy-to-model behaviors, and this, in turn, tends to be easier to achieve with simpler measurements. For example, a sensing device that tries to do a lot in one shot, (e.g., a sophisticated but monolithic face recognition system) typically operates on a restricted range of inputs and exhibits extremely non-linear behavior. This makes it difficult to apply in novel imaging environments because one does not have a good model of what it would do, for example, on non-face images. As a side effect, this minimality criterion encourages the use of computations that consume fewer resources and this boosts overall performance.

The combination of these two criteria leads to the question: What is the minimal measurement that produces meaningful information? In the stereo and motion sensing domains, this has led us to some new perspectives on how to define these computational problems. For example, instead of attempting to compute a dense stereo range map, we are focusing on the problem of computing and communicating the results of a single range measurement over a patch of surface. This distinction can be significant when issues of interaction with higher-level knowledge and control are considered.

In stereo matching, for example, a measurement over a small sensing area may fail due to the absence of matchable features. To recover, the calling agent can try switching to a larger measurement window, or it could move the original measurement patch to a slightly different position, or it could decide to move the sensor head to a better vantage point. In any case, the calling agent is aware of the changes made and their implications for the measurement. It is in possession of knowledge of the task to be accomplished, and it is aware of the measurement difficulty and the character of the possibly degraded information obtained. At the same time this agent does not have to know much about the detailed workings of the measurement algorithm itself. As long as it exhibits a consistent and predictable behavior, it can be effectively treated as a black box.

Sign-Correlation Algorithm

The first class of computations studied extensively in this context has been image matching algorithms applicable to stereo range finding and optical flow field measurement. We have developed a computational theory for measuring stereo and motion disparity that is consistent with the measurement-tool objectives and we have had some success at demonstrating the validity of that model for biological systems.

Binocular stereo, the measurement of optical flow, and many alignment tasks involve the measurement of local translation disparities between images. Marr and Poggio's zero-crossing theory made an important contribution towards solving this disparity measurement problem. The zero-crossing theory, however, does not perform well in the presence of moderately large noise levels as has been illustrated by the inability of zero-crossing-based approaches to solve transparent random-dot stereograms—which, interestingly, can be perceived correctly by the human visual system. The sign-correlation algorithm builds on Marr and Poggio's ideas, addressing many of the weaknesses of the original work.

The sign-correlation algorithm continues to use the zero-crossing primitive for matching, but the matching rule is changed. Instead of matching zero contours, we correlate the signal's *sign* in an area. This subtle change makes a significant difference in the behavior of the matcher. Sign-correlation continues to provide useful disparity measurements in high-noise situations long after the zero-crossing boundaries surrounding the signed regions cease to have any similarity. An intuitive explanation of why the two approaches perform so differently follows from the fact that the sign of the convolution signal is preserved near its peaks and valleys long after increasing noise has caused the zero contours to be fully scrambled. Thus, area correlation of the sign representation yields significant correlation peaks even with signal-to-noise ratios of 1 to 1. Since sign-correlation still operates off the zero crossing representation, the key strengths of Marr and Poggio's theory are preserved.

PRISM-3

The sign correlation algorithm has been implemented in the PRISM-3 real-time vision system. A pair of stereo cameras has been mounted on an active pan-tilt-vergence mechanism. The cameras have a stereo baseline of 22.2 cm and the camera vergence angle is computer controlled. The head can move through a 180 degree rotation in under a second and exhibits a positioning repeatability on the order of 50 arc seconds standard deviation in pan, 20 arc seconds in tilt, and 6 arc seconds in

vergence.

The two video cameras share the same pixel clock in order to minimize timing skew between the cameras that would result from using only horizontal and vertical video synchronization signals. The left and right camera video is digitized using commercial (DataCube) digitizer hardware, and parallel digital video streams are fed to two dedicated Laplacian-of-Gaussian convolvers (developed by Teleos). These convolvers allow video rate convolution with operator center diameters ranging from 1.6 pixels to 16.6 pixels.

The convolved video signals are fed from the two convolvers to a binary correlator board (also developed at Teleos) which carries out high-speed correlations on the sign bits of the input video streams.

The PRISM-3 correlator board performs 36 correlations in parallel on rectangular windows of adjustable size. The correlator board is operated by an external control processor (currently a 68040 single board computer). At the start of a measurement cycle, this processor writes the pixel coordinates of the next measurement to be made into registers on the correlator along with information about the disparities at which correlation measurements are to be made. A set of correlations with 32 by 32 pixel windows at 36 different disparities takes 100 microseconds to complete. The correlation results are then read into the control processor. If a well formed peak is identified in the data, quadratic interpolation is used to refine the peak disparity. These steps on the CPU take an additional 200 microseconds.

With correlations taken at even pixel disparities at a single vertical disparity, the above 300 microsecond cycle allows a disparity peak to be located in a 72 pixel disparity search range with a third to a tenth of a pixel resolution. Vertical disparity errors between 1 and 2 pixels are well tolerated.

The correlator hardware is also configured to allow correlations to be computed between successive frames from a single camera, allowing optical flow measurements to be made. In the tracking application described below, the system has been programmed to handle image velocities as large as 50 pixels per frame in any direction with subpixel measurement resolution.

The dedicated hardware incorporates standard off-the-shelf TTL components and makes extensive use of field-programmable gate arrays (FPGAs) to achieve high performance while maximizing flexibility in reconfiguring the hardware design.

Tracker Module

Tracking and control applications require fast, low-latency response from the sensor to be of value. A natural limit on speed is the frame rate of the camera system; for most commercially available cameras this is either 30 or 60 frames per second.

At 30 Hz, a person three meters from a camera walking across the field of view at 1 meter per second will traverse about 38 arc minutes per frame. With a 50mm lens the interframe motion disparity will be on the order of 30 pixels. This estimate is for one set of parameters—disparity magnitude varies approximately linearly with lens focal length, subject distance, subject speed, and frame rate—but it gives an indication of the kind of matching performance that will be required to follow human scale motions.

Similarly, the head position control must be responsive to velocity commands at the 30Hz rate with maximum acceleration and velocity limits set sufficiently high to allow smooth pursuit tracking motions.

A tracking system designed to meet these performance specifications was implemented on the PRISM-3 architecture as three subsystems, a low-level electronic tracking system, a mechanical servoing system, and a figure stabilization system. These individual mechanisms operate as loosely coupled parallel process threads. The electronic tracker makes high performance image-based measurements of optical flow and stereo range and attempts to follow electronically an externally designated patch of surface so long as it remains within the camera field of view. The mechanical tracker operates the active camera head in velocity mode using a PID control algorithm. This system attempts to keep the head pointed so that the coordinates of the surface patch tracked by the electronic tracker are kept close to the center of the camera field of view. The figure stabilization submodule uses stereo measurements to assess the extent of the figure associated with the tracked patch. If the tracked patch is not centered on that figure, this module sends an error bias signal to the electronic tracker in an attempt to push it back to the center of the figure. This helps to maintain tracking on figures undergoing rotation that would otherwise lead an optical-flow-based tracking scheme astray.

VISUAL PERCEPTION FOR SPACE ROBOTICS

The Automation and Robotics Division in the Engineering Directorate at the Johnson Space Center recently used PRISM-3 in a successful demonstration of autonomous, vision-guided grasping of a simple target. Testing took place during a flight on NASA's KC135 Reduced Gravity Aircraft as part of Phase 3A of the Extravehicular Activity Retriever/Helper Project (EVAHR). These tests are the first to prove that autonomous robots can use computer vision to guide robotic manipulation and grasp of moving objects in microgravity.

The EVAHR is equipped with a 7-degree-of-freedom robot arm and a dextrous hand consisting of three active and two passive fingers. The PRISM-3 vision system provides the EVAHR's control system with continuous measurements of the position and velocity of a given object, enabling the arm to move to intercept the object. During tests aboard the KC135, a four-inch ball was released to move freely in space during the brief periods of microgravity induced on the aircraft. PRISM located and tracked the ball, enabling the EVAHR to catch it seven times in a number of tries.

Vision-guided grasping of moving objects is a basic skill both in space helper [2] and retrieval tasks and in making the transition from flying to attachment to a spacecraft. Making this transition is particularly demanding as the spacecraft is moving relative to the robot even if the robot is station-keeping with the spacecraft.

Plans are under development to use PRISM-3 in a follow-on EVAHR grasping experiment using more complex targets.

Additional space-related applications are under consideration in two areas: in-space assembly (for example, for operations involving the Shuttle Remote Manipulation System), and in the use of visually-guided Rover navigation for autonomous and/or supervised planetary exploration.

REFERENCES

[1] Nishihara, H. K., 1991. *Minimal Meaningful Measurement Tools*. Technical Report, TR-91-01, Teleos Research

[2] Erickson, J. D.; Grimm, K. A.; and Pendleton, T. W., 1994. An Intelligent Robot for Helping Astronauts. In *Proceedings of the Conference on Intelligent Robotics in Field, Factory, Service & Space*, NASA Conference Publication 3251, vol. 1, 421-433.

Planning and Scheduling Workshop

The Planning and Scheduling Workshop is a single track within the overall i-SAIRAS 94 meeting. It focuses on planning and scheduling as they apply to space exploration, with specific attention to practical, working systems. The workshop includes papers of particular technical interest because they describe fielded planning or scheduling systems and emphasize the reasons for a particular system's success or failure.

The workshop combines formal presentations with opportunities for questions, discussion, and debate among speakers and workshop participants. A number of panels throughout the workshop allow participants to air their views and to exchange ideas about important topics in the area of planning and scheduling.

The theme of the workshop is technology transfer, with specific attention to possible "dual uses" of technology. The workshop attempts to establish connections between technology developed for space and that developed for nonspace (often private industry) markets — especially the manufacturing and airline industries, since they have many characteristics in common with space applications. Presentations in this track include discussions of technology developed in government research labs for particular space applications that can apply to nonspace applications, as well as technology developed for nonspace applications that can sometimes work perfectly for space.

The Planning and Scheduling Workshop comprises the following sessions:
- Session PS-AT   *Astronomy Planning and Scheduling*
- Session PS-DS   *Decision Support Aspects*
- Session PS-MS   *Mission Support*
- Session PS-NT   *New Techniques*

Planning and Scheduling Workshop:
Astronomy Planning and Scheduling

Astronomy Planning and Scheduling