

Visual Tracking with Subpixel Resolution using an Analog VLSI Computational Sensor

Ziyi Lu

Bell Laboratories
Lucent Technologies (China) Co., Ltd.
Shanghai 200030, P. R. CHINA

Bertram E. Shi

Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, HONG KONG

Abstract

This paper describes the application of a computational vision sensor to active binocular tracking. The sensor outputs are used to control the vergence angles of the two cameras and the tilt angle of the head so that the center pixels of the sensor arrays image the same point in the environment. One distinguishing feature of the sensor used here is the possibility to resolve target motions with sub-pixel resolution. This is due to the use of a phase based algorithm which integrates information over multiple pixels.

1 Introduction

Work in computational sensors has been motivated by the increasing need for low-latency vision systems for applications such as visual servoing and human computer interfaces. Computational sensors combine both sensing and image processing on the same silicon substrate. Both local and global image processing operations are often implemented using analog VLSI circuits, due to their high speed, low power and low area[7][8][9].

Visual tracking has been a popular application for computational sensors[1][6]. Most of these systems have been based upon extracting the most salient point in the image using a winner-take-all network. Saliency has been measured by image intensity, temporal or spatial gradients, or combinations of these features. If the sensor is fixed, these sensors continuously report the location of the target in the image plane as it changes over time. These sensors can also be used in active vision systems, where their output is used to move the sensors so that the target is stabilized in the image. Note that since the winner-take-all network identifies the pixel corresponding to the most salient point, these systems can only resolve target motions with pixel level accuracy.

This work introduces the use of an analog VLSI computational sensor for visual tracking on an active binocular vision platform. The sensor outputs are used to control the vergence angles of the two cameras and the tilt angle of the head to keep a target centered in the two sensor arrays. The trajectory of the target in environmental coordinates can be recovered using triangulation.

One feature which distinguishes the sensor used in this work is its ability to resolve target motions with sub-pixel accuracy. This translates into a corresponding increase in the resolution with which the target trajectory in the environment can be recovered. Subpixel resolution may be especially advantageous for analog VLSI computational sensors, since the resolution and pixel count of these sensors is lower than pure CCD or CMOS imagers due to the additional circuitry required at every pixel. For example, the computational sensors reported for visual tracking, including the one here, have typically had on the order of 30 by 30 pixels or less.

Sub-pixel resolution is achieved via a phase-based technique for disparity estimation using the outputs of filters similar to Gabor filters. Disparity between a two images is measured by the phase differences between the complex valued Gabor filter outputs of pixels in the two images[10]. Fleet, etc. [4] suggested that the phase-based disparity is stable under scale perturbation and smooth contrast variations. Theimer and Mallot [12] pointed out that the phase method was especially suited for the concept of active vision.

2 The computational sensor

The vision sensor consists of a 2D array of 25 x 25 phototransistors converting light into current. An array of continuous time analog processing circuits first subtracts the DC offset from the input image and then filters it with a filter whose impulse response can be approximated by

3 Phase-based disparity estimation and tracking

The filter outputs corresponding to the center pixel were used in a phase-based approach to estimate the disparity between the actual and desired target position. This approach exploits the fact translations in the input image result in phase shifts in the complex valued outputs of Gabor or similar filters. The phase shift is approximately linear in the amount of translation. Phase based approaches have been shown to be robust in the presence of contrast and luminosity imbalances[2]. The DC offset subtraction performed by the sensor further decreases this sensitivity. In addition, they do not require a search for correspondences based on feature or region based similarity measures. Although phase based approaches usually use the output of Gabor filters, they are insensitive to the exact form of the filter transfer function[13]. In particular, Crespi et. al. studied the effect of replacing the Gaussian modulating function with the 1D version of our filter and found only slight degradation[3].

If an input image $u_0(m, n)$ is convolved with the kernel $h(x)$ in (1), the output of the filter is

$$v_o(m, n) = \sum_k \sum_l u_0(m-k, n-l) f(k, l) e^{j(\omega_{x_o}k + \omega_{y_o}l)}$$

If the image is translated by $(\Delta x, \Delta y)$, $u_\Delta(m, n) = u_0(m - \Delta x, n - \Delta y)$, the output becomes

$$v_\Delta(m, n) = \tilde{v}_0(m, n) e^{-j(\omega_{x_o}\Delta x + \omega_{y_o}\Delta y)}$$

where

$$\begin{aligned} \tilde{v}_0(m, n) &= \sum_k \sum_l u_0(m-k, n-l) f(k - \Delta x, l - \Delta y) e^{j(\omega_{x_o}k + \omega_{y_o}l)} \\ &\approx v_0(m, n) \end{aligned} \quad (2)$$

The approximation holds for translations $(\Delta x, \Delta y)$ which are small compared with the width of $f(m, n)$ and the periods of ω_{x_o} and ω_{y_o} . The range of validity also depends upon the input image. However, the wider the convolution kernel, the less accurately the estimated translation reflects the local characteristics of the translation.

The phase difference between the outputs at pixel (m, n) , is given by

$$\Delta\psi(m, n) = \psi_0(m, n) - \psi_\Delta(m, n) \approx \omega_{x_o}\Delta x + \omega_{y_o}\Delta y$$

where

$$\psi_0(m, n) = \arg\{v_0(m, n)\}$$

$$\psi_\Delta(m, n) = \arg\{v_\Delta(m, n)\}$$

If we set $\omega_{x_o} > 0$ and $\omega_{y_o} = 0$, the horizontal displacement Δx can be estimated by [10]

$$\Delta x \approx \Delta\psi(m, n) / \omega_{x_o} \quad (3)$$

or more accurately by [4]

$$d(m, n) \approx \frac{2\Delta\psi(m, n)}{\psi_0'(m, n) + \psi_\Delta'(m, n)}$$

$\psi_0'(m, n)$ and $\psi_\Delta'(m, n)$ denote the rate of change of the phase in the horizontal direction. In this work, we adopt the first measure since it is simpler and for the vergence control, we are primarily concerned with zeroing the disparity rather than estimating it. Similarly, Δy can be estimated if we set $\omega_{x_o} = 0$ and $\omega_{y_o} > 0$.

Tracking of a target in the center of the sensor array is achieved by updating the tilt and vergence angles of the cameras to zero the disparity between the actual and desired target position. The sensor is configured so that the filter outputs of the center pixel in the array are permanently connected to the output. With the target in the center of the array, the sensor outputs are recorded with the sensor tuned to horizontal and then to vertical orientations. The orientation selectivity is adjusted by changing the values of bias voltages which change the gains of the transconductance amplifiers G_{2x} and G_{2y} in Figure 1. During tracking, the sensor outputs with horizontal and vertical orientation tunings are sampled every 4ms. The phase differences between the current and reference outputs are used to estimate the disparity. Assuming that the tilt and vergence axes are aligned with the camera axis, the angular movement required to zero the disparity is approximately proportional to the disparity. This movement is achieved using closed loop position control with blended motion moves updated every sample period.

By choosing equal reference outputs for the left and right sensors, the estimated horizontal disparity between the left and right images is zeroed during tracking. Assuming the disparity estimation near zero is accurate, both sensors track approximately the same point in the environment. The location of that point can be estimated via triangulation. Figure 2 shows the geometry of the binocular head looking from the top down and from the side. The X , Y and Z coordinates of the tracked point can be recovered from the vergence and tilt angles via the equations

$$\begin{aligned}
X &= \frac{Z_{\text{verg}}}{2} (\tan \theta_r - \tan \theta_l) \\
Y &= Z_{\text{verg}} \sin \theta_{\text{tilt}} \\
Z &= Z_{\text{verg}} \cos \theta_{\text{tilt}}
\end{aligned}
\tag{4}$$

where $B = 25\text{cm}$ is the baseline length and

$$Z_{\text{verg}} = \frac{B}{\tan \theta_l + \tan \theta_r}$$

Although we assume here that the vergence axes are aligned with the camera center and that the optical axis intersects the center pixel in the array, the actual model used in the estimation takes into account position and angular offsets.

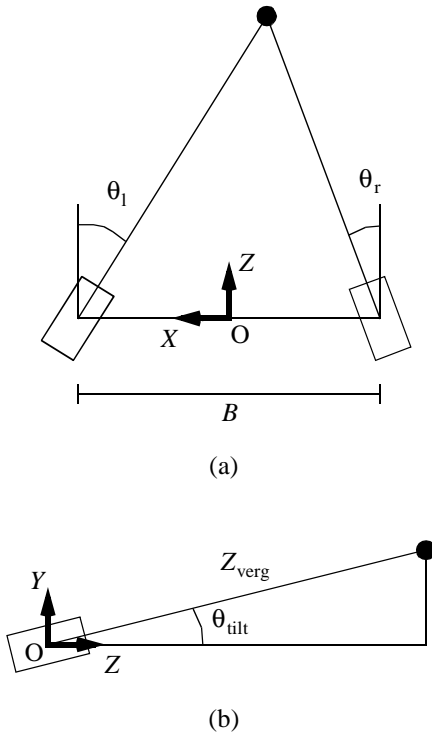


Fig. 2: Coordinate frames associated with the binocular vision head looking (a) from the top and (b) from the left side.

4 Experimental results

4.1 Disparity estimation

To examine the disparity estimated by the sensor, a vertically oriented step edge target was used. We manually configured the cameras so that the step edge transition was

imaged by the center pixel of one camera and recorded the outputs of the center pixel. The vergence angle of the camera was swept over the range -5° to 5° relative to the starting position with increments of 0.2° . At each position, the output was digitized and stored. The filter parameters were $\lambda = 0.28$, $\alpha = 1.04$, $\omega_{x0} = 0.34$ and $\omega_{y0} = 0$.

The disparity Δx for each location was estimated from the phase difference between the current and reference output using (3). The ideal and estimated disparities are plotted in Figure 3. Assuming that the offset between the vergence axis and the lens center is small, the disparity should be proportional to $\tan \alpha \approx \alpha$ where α is the vergence angle. From the figure, we observe that this is true for small α . The degradation for larger α is primarily because the approximation in (2) is less valid. Note however, that the sensor can distinguish the disparities much smaller than a single pixel.

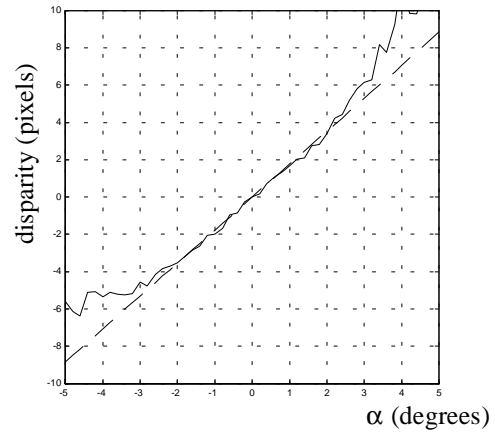


Fig. 3: The disparity estimated from the sensor (solid line) as the camera vergence angle is swept over 10 degrees. The dashed line shows the theoretical prediction.

4.2 XZ Tracking Performance

In this experiment, we kept the tilt axis stationary and controlled only the vergence angles to move to track the same vertical edge target as used in the previous section. An X table provided controlled translations of the target both from left to right and directly towards the binocular vision head at a linear speed of 25.4 cm/s . Figure e4 shows trajectories recovered with the target moving 25.4 cm from left to right. Figure 5 shows the estimated trajectory for the target moving 38.1 cm towards the binocular head.

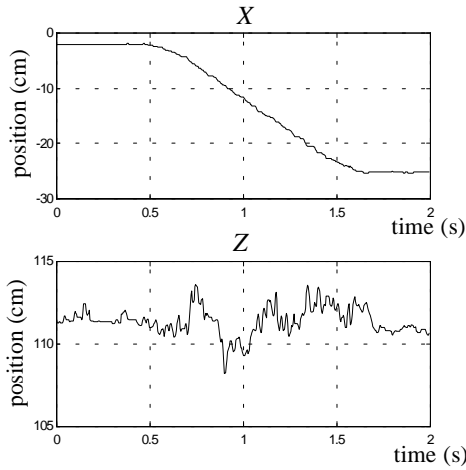


Fig. 4: Recovered trajectory for target translating from left to right.

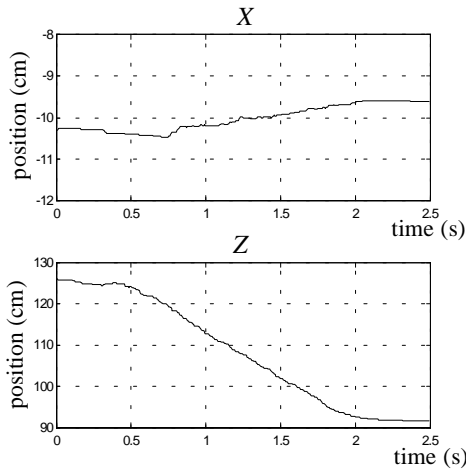


Fig. 5: Recovered trajectory for target translating towards binocular system.

Temporal noise in the output is introduced by sources such as noise from the analog processing circuits, lighting variations detected by the photosensors and electromagnetic interference picked up by signal lines and quantization noise during A/D conversion. This noise was measured by digitizing the chip output 1000 times at the reference position. The measured variance of the estimated disparity was $\sigma_{\Delta x}^2 = (0.10 \text{ pixel})^2$.

Even when the system tracks a stationary target, we can observe temporal variations in the vergence angles leading to variations in the estimated location. These variations are

predominantly due to this sensor noise. Noise in the disparity estimated by the sensor leads to fluctuations in the vergence angles with variance

$$\sigma_{\theta_l}^2 = \sigma_{\theta_r}^2 = \left(\frac{K}{f}\right)^2 \sigma_{\Delta x}^2 \quad (5)$$

where $f = 14\text{mm}$ is the focal length of the lens and $K = 0.146 \text{ mm/pixel}$ is a conversion factor. If we assume that $\theta_{\text{tilt}} = 0$ in (4), we find that small variations $\delta\theta_l$ and $\delta\theta_r$ in the vergence angles lead to variations in the estimated position of

$$\begin{aligned} \delta X &= \frac{B}{(\tan\theta_l + \tan\theta_r)^2} \left(\frac{\tan\theta_r}{\cos^2\theta_l} \delta\theta_l + \frac{\tan\theta_l}{\cos^2\theta_r} \delta\theta_r \right) \\ \delta Z &= \frac{B}{(\tan\theta_l + \tan\theta_r)^2} \left(\frac{\delta\theta_l}{\cos^2\theta_l} + \frac{\delta\theta_r}{\cos^2\theta_r} \right) \end{aligned} \quad (6)$$

Assuming that the variations $\delta\theta_l$ and $\delta\theta_r$ are uncorrelated, we can use (5) and (6) to predict the variance of the fluctuations in the recovered X and Y positions due to the sensor noise. Figure 6 shows the predicted variance and measured variance match quite closely. The variation in the recovered X coordinate is much smaller than that for Z .

4.3 XYZ Tracking

In this experiment, tilt and vergence angles were controlled to track a 5cm black square on a white background. The target rotated at 64 rpm where the center of the square was displaced by 6.35cm from the center of rotation. At the same time, the target moved backwards by 25.4 cm at a speed of 20.3 cm/s.

Acknowledgements

This work was supported in part by the Hong Kong Research Grants Council under Grants HKUST675/95E and HKUST782/96E.

References

- [1] V. Brajovic and T. Kanade, "Computational sensor for visual tracking with attention," *IEEE J. of Solid State Circuits*, vol. 33, no. 8, pp. 1199-1207, Aug. 1998.
- [2] A. Cozzi, B. Crespi, F. Valentinotti and F. Worgotter, "Performance of phase-based algorithms for disparity estimation," *Machine Vision and Applications*, vol. 9, pp. 334-340, 1997.

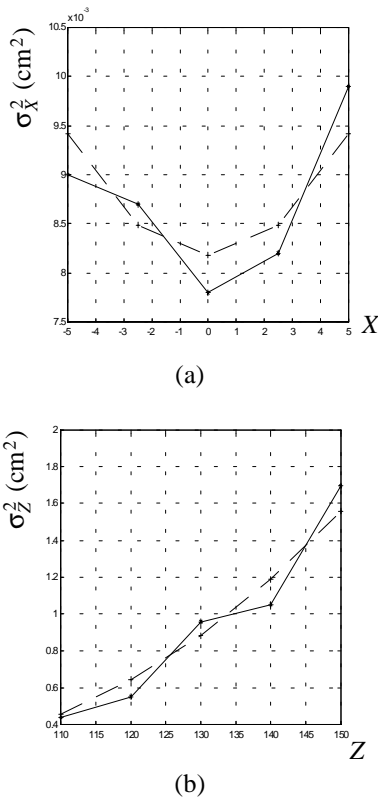


Fig. 6: (a) Comparison of the measured variance of the estimated X coordinate (solid line) with that estimated assuming the variation is due to sensor noise (dotted line) for a target at $Z = 120\text{cm}$ and various X coordinates. (b) Comparison of the measured variance of the estimated Z coordinate with the estimate for a target located at $X = 0$ and varying distances Z .

[3] B. Crespi, A. G. Cozzi, L. Raffo and S. Sabatini, "Analog computation for phase-based disparity estimation: continuous and discrete models," *Machine Vision and Applications*, vol. 11, pp. 83-95, 1998.

[4] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understanding*, Vol. 53, pp. 198-210, 1991.

[5] HelpMate Robotics Inc., Shelter Rock Lane., Danbury, CT 06810-8159.

[6] T. K. Horiuchi, T. G. Morris, C. Koch, and S. P. DeWeerth. "Analog VLSI circuits for attention-based, visual tracking," in M. C. Mozer, M. I. Jordan and T. Petsche, eds., *Advances in Neural Information Processing Systems 9*, Cambridge, MA: MIT Press, pp. 706-712, 1997.

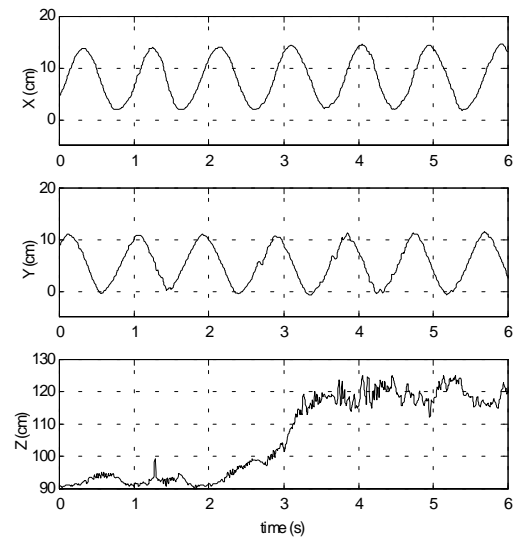


Fig. 7: Recovered trajectories for square target moving simultaneously in X , Y and Z .

[7] T. Kanade and R. Bajcsy, "Computational Sensors: A Report from the DARPA Workshop, *IUS Proceedings*, 1993.

[8] C. Koch and H. Li, eds., *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, Los Alamitos, CA: IEEE Computer Society Press, 1995.

[9] T. Roska and L. O. Chua, "The CNN Universal Machine: An Analogic Array Computer," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 40, pp. 163-173, Mar. 1993.

[10] T. D. Sanger, "Stereo disparity computation using Gabor filters," *Biological Cybernetics*, Vol. 59, pp. 405-418, 1988.

[11] B. E. Shi, "Focal Plane Implementation of 2D Steerable and Scalable Cortical Filters," *Journal of VLSI Signal Processing*, vol. 23, no. 2/3, pp. 319-334, Nov./Dec. 1999.

[12] W. M. Theimer and H. A. Mallot, "Phase-based binocular vergence control and depth reconstruction using active vision," *CVGIP: Image Understanding*, Vol. 60, pp. 343-358, 1994.

[13] C.-J. Westelius, H. Knutsson, J. Wiklund and C.-F. Westin, "Phase-based disparity estimation," in *Vision as Process*, J. L. Crowley and H. I. Christensen, eds., Berlin, Germany: Springer-Verlag, 1995, Ch. 11, pp. 157-178.