2010 12th International Workshop on Cellular Nanoscale
Networks and their Applications (CNNA)

Binocular Disparity Calculation on a Massively-Parallel Analog Vision Processor

Soumyajit Mandal
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Email: soumya@mit.edu

Bertram Shi
Hong Kong Univ. of Science & Technology
Clear Water Bay, Kowloon, Hong Kong
Email: eebert@ee.ust.hk

Piotr Dudek
The University of Manchester
Manchester M60 1QD, UK
Email: p.dudek@manchester.ac.uk

***Abstract*—We studied neuromorphic models of binocular disparity processing and mapped them onto a vision chip containing a massively parallel analog processor array. Our goal was to make efficient use of the available hardware while preserving the fundamental computations performed by the models. We also developed an optical fixture that used mirrors to simultaneously focus two images onto the vision chip. This fixture simulates two horizontally-separated virtual cameras, thereby allowing us to run our binocular disparity estimation algorithms using a single image sensor in real time.**

I. Introduction

One of the challenges in neuromorphic engineering is to take often elaborate neural computation models and implement them in silicon, within tight constraints of size and power. Similarly, a skillful "interpretation" of a model is often necessary in order to implement it in real-time on various embedded processors or low-power devices. In this paper we investigate the problem of stereopsis, or depth perception based on binocular cues, using a massively-parallel analog vision processor chip.

II. Theory

Binocular disparity is defined as the distance, in image coordinates, between the locations of a given object as seen by the two eyes. It is an important cue for depth perception. Some cells in the primary visual cortex (V1) of various mammals have receptive fields that are sensitive to binocular disparity. A biologically-plausible model of the response of such disparity-selective cells is the binocular energy model [1]. The model, which spatially filters images obtained from the left and right eyes, generates approximations of such complex receptive fields and can be used in binocular disparity estimation algorithms. Combinations of Gabor functions are often used for spatial filtering, because receptive fields tuned to various disparities can be easily generated by varying the relative spatial position and/or phase of the functions. These strategies are known as the position-shift and phase-shift models, respectively [2]. Binocular disparity maps of images can be produced by using a heterogeneous population of such receptive fields and selecting the best-fitting receptive field at every point.

The spatial impulse response of a Gabor filter consists of a complex sinusoidal carrier that is modulated by a Gaussian kernel. Gabor filters are bandpass in nature and select a range of spatial frequencies. The center frequency is equal to ω, the frequency of the carrier, while the bandwidth is inversely proportional to σ, the width of the Gaussian kernel. A generic one-dimensional Gabor filter may be written as

$$g(x, \sigma, \omega, \phi) = e^{-x^2/2\sigma^2} e^{j(\omega x + \phi)} \quad (1)$$

where ϕ denotes the phase of the filter. The basic phase-shift model for generating disparity-selective receptive fields is shown in Figure 1. Images from the left and right eyes, denoted by I_l and I_r, respectively, are filtered by the complex Gabor filters $g(x, \sigma, \omega, \phi_r)$ and $g(x, \sigma, \omega, \phi_l)$. The figure shows one-dimensional filters for simplicity, but in general two-dimensional filters may be used. The outputs of the filters are denoted by $Y_l(d, \sigma, \omega) \exp(j\phi_l)$ and $Y_r(d, \sigma, \omega) \exp(j\phi_r)$, where Y_l and Y_r are independent of ϕ_l and ϕ_r, respectively, and the stereo disparity is denoted by d. These signals correspond to responses of pairs of binocular simple cells in the visual cortex.

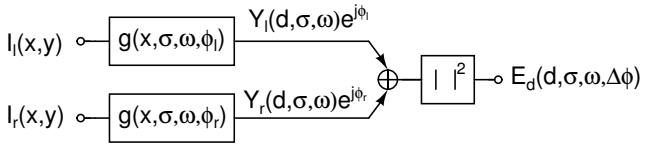

Fig. 1. The phase-shift model for generating disparity-selective receptive fields that is described in this paper.

The disparity energy E_d is found by summing the two simple cell outputs and calculating the magnitude of this complex variable, as shown in Figure 1 and explicitly defined below:

$$E_d(d, \sigma, \omega, \Delta\phi) = \left| Y_l(d, \sigma, \omega) e^{j\phi_l} + Y_r(d, \sigma, \omega) e^{j\phi_r} \right|^2 \quad (2)$$

where the relative phase shift between the filters applied to the two images is $\Delta\phi = \phi_l - \phi_r$. The output of the model is the disparity energy E_d at each pixel, as shown in Figure 1. It corresponds to the output of a particular binocular complex cell in the primary visual cortex for a particular stereogram, i.e., set of left and right images I_l and I_r. The dependence of E_d on stereo disparity d is known as the neural tuning curve and can be estimated by presenting stereograms with

978-1-4244-6678-8/10/$26.00 ©2010 IEEE

different disparities to the neuron. If $d \ll \sigma$, it can be shown that $E_d \propto \cos^2\left(\left(\omega d + \Delta\phi\right)/2\right)$ [3]. Thus the peak value of E_d occurs for a disparity of $d_{pref} \approx -\Delta\phi/\omega$ pixels. In other words, the neuron is 'tuned' to be maximally responsive to a preferred disparity of d_{pref}. It can be shown that position-shift models give rise to tuning curves with similar shapes.

The disparity map of a given stereogram can be estimated by using a population of neurons with different values of preferred disparity d_{pref}. The neurons can be sensitive to disparity either via phase-shift mechanisms, position-shift mechanisms, or a combination of both. The dependence of E_d on d_{pref}, or equivalently $\Delta\phi$ for phase-shift neurons, is known as the population response curve. The simplest way to estimate the disparity map is to pick, at each pixel, the neuron that has the largest response within the population. The preferred disparity of this neuron is equal to the estimated disparity, which we denote as d_{est}. The population of cells should contain enough distinct values of d_{pref} to allow d_{est} to be estimated to the desired accuracy.

We shall concentrate on phase-shift models in this paper because they are particularly suitable for efficient hardware implementations. It has also been shown that the values of d_{est} obtained from phase-shift models are more accurate than from position-shift models when disparities are small relative to σ, the spatial scale of the Gabor filter kernels [3]. However, phase-shift models underestimate disparities when they become comparable to σ. In addition, they cover a more limited range of preferred disparities than position-shift models for given values of σ and ω: The phase-shift population tuning curve only has a unique maximum for $-\pi < \Delta\phi < \pi$, limiting the range of disparities that can be unambiguously estimated to $-\pi/\omega < d < \pi/\omega$. Fortunately, several algorithms have been proposed to alleviate this problem [3], [4].

III. Hardware

We implemented the disparity model shown in Figure 1 on a SCAMP (SIMD Current-Mode Analog Matrix Processor) imaging system. The system consists of a SCAMP-3 vision chip, an FPGA-based microcontroller, optics, and user interface software. The SCAMP-3 chip, which is a programmable 128×128 image sensor, was fabricated in a 0.35μm CMOS technology and has been described elsewhere [5], [6]. Each pixel is 50μm×50μm in size, has a fill factor of 6% and contains a simple analog switched-current processor that is connected to neighboring pixels. This locally-connected architecture was inspired by the biological retina, and makes SCAMP well suited for massively-parallel execution of low-level image processing algorithms, particularly those that rely heavily on local operations such as spatial filtering. Other advantages of the SCAMP architecture include the absence of any data transfer bottlenecks between the sensors and the processors, thus allowing high frame rates. In addition, the analog architecture results in low power consumption, and high dynamic range can be obtained by locally adaptive sensing. The chip can execute approximately 20×10^9 operations per second (i.e., 20GOPS) while consuming 250mW of power. The resultant power efficiency of 80GOPS/W is much higher than both general purpose processors ($\sim$ 0.1GOPS/W), and digital signal processors ($\sim$ 5GOPS/W) [6].

The entire SCAMP processor array operates using a SIMD (Single Instruction Multiple Data) paradigm. The processors operate on continuous analog values but run in discrete time at a typical clock frequency of 1MHz. Each processor contains eight general-purpose analog storage registers and can execute a limited set of instructions, including inversion, addition, division by two and three, comparison, and loading external data. A specialized register allow each pixel to access values stored within pixels in its local von Neumann neighborhood. These are useful for operations that require pixels to share information, such as spatial filtering. A multiplier is not available, and multiplication must be carried out using repeated additions. Each processor receives the same set of instructions because of the SIMD architecture. However, branches in the control flow are possible because conditional statements are supported via a blanking mechanism. In this mechanism a given set of instructions is executed only by the pixels where the result of a comparison operation evaluates as true.

The limited nature of the SCAMP instruction set and the small amount of local storage available within each processor impose constraints on the algorithms that can be implemented on it. In addition, programs must be written in assembly language, since no compilers for high-level languages are currently available. Finally, the analog nature of SCAMP computations requires special strategies, such as correlated double sampling and dynamic range scaling, to be used in order to maintain acceptable accuracy during the computation [7]. Division is a particularly error-prone operation, with a typical rms error of $\epsilon = 5\%$. However, an error-compensated division subroutine, which reduces the error to approximately $\epsilon^2 = 0.25\%$, has been developed for the special case of division by two. We restricted divisions in our algorithms to factors of two in order to take advantage of this fact.

IV. Implementation

We implemented the disparity algorithm for filters tuned to three different preferred disparities: negative, zero and positive. The outputs of these filters, while insufficient for generating a complete disparity map of the image, are able to provide a feedback or error signal of the right sign for tracking algorithms that try to keep an object at a fixed distance from the image sensor. Such auto-focusing mechanisms may be useful for a wide range of applications, such as image stabilization and object tracking in mobile robots.

We shall assume that our imaging system generates two virtual cameras that are separated along the horizontal, or x axis. As a result, most disparity information will lie along this axis. For simplicity we shall therefore only use horizontally-oriented disparity detectors that use Gabor filters. However, the algorithm can be easily extended to detectors with other orientations if necessary [3].

In order to estimate the disparity energy we begin by filtering the images from the left and right eyes using Gabor

filters with phase shifts ϕ_l and ϕ_r, respectively, as shown in Figure 1. Phase-shifted Gabor filters can be written in a particularly simple form in the case when we restrict phase shifts to the set $\phi \in \{0, \pm\pi/2\}$. In fact, we can use the facts that $\sin(x) = \cos(x - \pi/2)$ and $\cos(x - \pi) = -\cos(x)$ to show that

$$\begin{aligned} g(x, \sigma, \omega, 0) =& e^{-x^2/2\sigma^2} \left[\cos(\omega x) + j \sin(\omega x)\right] \\ g\left(x, \sigma, \omega, \frac{\pm\pi}{2}\right) =& \pm\, e^{-x^2/2\sigma^2} \left[\sin(\omega x) - j \cos(\omega x)\right] \end{aligned} \tag{3}$$

Equation 3 can be rewritten as

$$\begin{aligned} g(x, \sigma, \omega, 0) =& g_1 + j g_2 \\ g\left(x, \sigma, \omega, \frac{\pm\pi}{2}\right) =& \pm\,(g_2 - j g_1) \end{aligned} \tag{4}$$

where $g_1 = \exp\left(-x^2/2\sigma^2\right)\cos(\omega x)$ and $g_2 = \exp\left(-x^2/2\sigma^2\right)\sin(\omega x)$. Thus all three complex Gabor filters can be formed by combining the outputs of only two real quadrature filters, i.e., g_1 and g_2. This result allows us to use SCAMP to estimate $E_d(d, \sigma, \omega, \Delta\phi)$ for three values of $\Delta\phi$ in real time. We used a spatial frequency of $\omega = 0.8 \approx 2\pi/8$ radians/pixel in our filters, while the standard deviation was $\sigma = 2.6$ pixels, corresponding to a spatial bandwidth of 1.9 octaves [8]. As a result, the three preferred phase shifts, i.e., $\Delta\phi \in \{0, \pm\pi/2\}$ within our neural population correspond to preferred disparities of $d_{pref} \in \{0, \pm 2\}$ pixels, respectively.

We had to approximate the shapes of the spatial filters $g_1(x)$ and $g_2(x)$ in order to implement them in a computationally efficient way on SCAMP. We denote the approximate filter shapes by $\widetilde{g_1}(x)$ and $\widetilde{g_2}(x)$, respectively. Firstly, in order to keep the processing relatively local we set $\widetilde{g}(x) = 0$ for $|x| > 4$ pixels, i.e., $|x| > 1.5\sigma$. In addition, we assumed that $\widetilde{g}(x) \in \{0, 1, \pm 1/2\}$ for integer values of x, i.e., at each pixel. This approximation allows us to apply both filters by using only the two accurate operations of shifting and dividing by two. The best-fitting values of $\widetilde{g_1}(x)$ and $\widetilde{g_2}(x)$ subject to these constraints for $|x| \leq 4$ are given by $[0, -0.5, 0, 0.5, 1, 0.5, 0, -0.5, 0]$ and $[0, -0.5, -1, -0.5, 0, 0.5, 1, 0.5, 0]$, respectively. Figure 2 compares the exact and approximate filter shapes. We see that they match fairly well. The rms difference between the exact and approximate shapes is 0.129 for g_1 and 0.127 for g_2.

The final steps in finding the disparity energy are to sum the outputs of the Gabor filters and then calculate the squared magnitude of this complex quantity, as shown in Figure 1. However, the squaring operation is difficult to implement directly on SCAMP since the processors don't contain multipliers. We therefore used a two-segment piecewise linear approximation to the required square law. Let us assume that the SCAMP registers can accurately store (as currents) real numbers between ± 100. Numbers should remain within this range after squaring in order to prevent saturation. In our scheme the value x to be squared is first rectified and then passed through the following function:

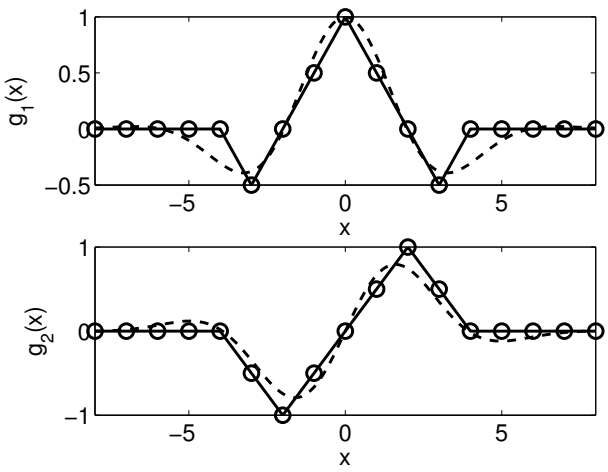

Fig. 2. The one-dimensional real filters used to generate disparity-sensitive responses. Ideal functions $g_1(x)$ and $g_2(x)$ and their approximate versions $\widetilde{g_1}(x)$ and $\widetilde{g_2}(x)$ are drawn with dashed and solid lines, respectively.

$$y = \begin{cases} x/2 & \text{if } x \leq 50 \\ 3x/2 - 50 & \text{if } x > 50 \end{cases} \tag{5}$$

Equation 5 was chosen to approximate the ideal function, i.e., $y = x^2/100$, by dividing the range of $|x|$ (0 to 100) into two regions and assigning linear functions with different slopes to each region. The straight lines were designed to intersect on the $y = x^2/100$ curve at the boundary between the regions, thus ensuring that the overall function remains continuous. We decided to use $x = 50$ as the boundary, as shown in (5), because this choice results in regions of equal size and also leads to particularly convenient linear functions: the only operations required to compute $x/2$ and $3x/2 = x + x/2$ are division by two and/or addition. Figure 3 compares (5) with $y = x^2/100$. The two functions are reasonably close to each other, with an rms separation of 4.54.

A test image containing a range of horizontal disparities is shown in Figure 4. The left half of the image corresponds to one virtual camera, while the right half corresponds to the other. The two halves of the image are shifted using the nearest neighbor connections of the SCAMP processors so that they overlap in the center of the array. Disparity calculations can now be performed using only local operations. As an example, Figure 4 also shows the simulated output of the filter tuned to $\Delta\phi = 0$, i.e., zero disparity, when run on the test image. The simulations were run on a SCAMP-specific hardware simulator that includes the effects of noise and mismatch between the pixels. As expected, the output of the filter is large only within the small region in the center of the test image where the disparity between the left and right halves is close to zero.

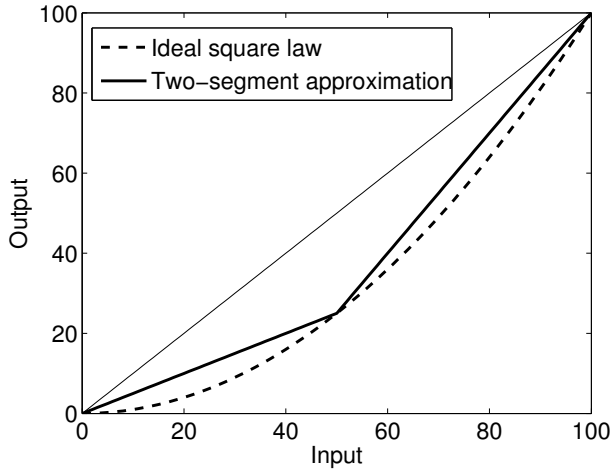

Fig. 3. An ideal square law and the two-segment approximation implemented on-chip.

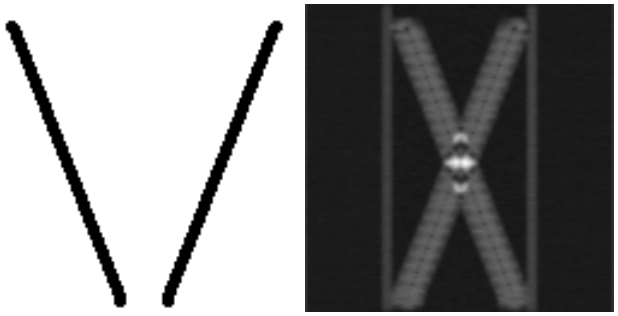

Fig. 4. The output of the filter tuned to zero horizontal disparity (right) when run on a test image (left).

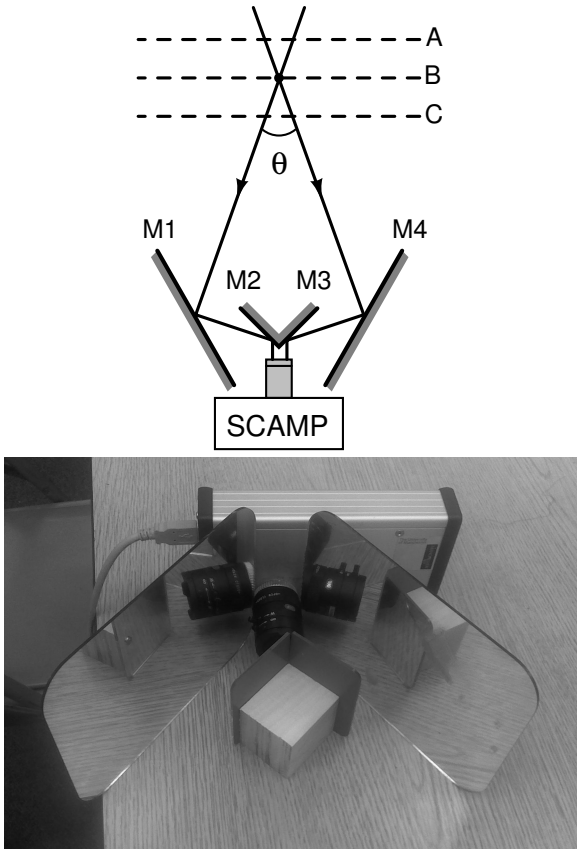

Fig. 5. Experimental setup, (top) schematic and (bottom) photograph.

V. Results

The experimental setup is shown in Figure 5. We used a combination of four planar mirrors (M1 – M4) to simultaneously focus two different views of the scene onto a single SCAMP camera, thus simulating the presence of two horizontally separated virtual cameras, or 'eyes'. The inner mirrors (M2 and M3) are always at right angles to each other. The angle between the inner and outer mirrors can be varied to select the vergence angle θ, i.e,, the range of distances over which stereo vision is obtained [9]. For example, in Figure 5 objects present at the planes labeled A, B and C will produce positive, zero and negative disparity, respectively. In addition, the camera contains an adjustable-focus lens. Thus our optical system is *catadioptric*, i.e., contains both mirrors (reflecting elements) and lenses (refracting elements) [10].

Captured images were prefiltered with a simple first-order high-pass filter, which has an impulse response of $[-0.5, 1, -0.5]$. The high-pass filter whitens the spatial frequency content of the images I_l and I_r in Figure 1. Without prefiltering, the disparity energy calculated by the algorithm would depend strongly on background illumination level, which is undesirable. In addition, we low-pass filtered the image in the vertical direction before applying the disparity-selection algorithm.

A typical scene captured by our optical setup at a frame rate of 10Hz is shown in Figure 6. The frame rate was limited by exposure time, not the SCAMP processor. The dark region at the center of the image was due to the non-reflecting hinge between the inner mirrors, which was unavoidable with the current setup. We usually tested our algorithm by moving the bright object (a white Styrofoam cup) towards and away from the sensor. As a result the object should excite the three disparity filters at three specific distances from the sensor. In other words, the filter outputs should become bright ('high') for brief periods of time, and remain relatively dark ('low') otherwise.

Experimental disparity-detection results are shown in Figure 7. The figure shows the output of the filter tuned to a disparity of $\Delta\phi = -\pi/2$ as the cup shown in Figure 6 was moved towards the camera, with the rest of the image kept approximately constant. The cup is furthest in the leftmost frame, and closest in the frame on the right. We denote this distance by d. The center of the frame in the middle contains two bright spots because at this intermediate value of d the images of the cup and its reflection (see Figure 6) have the right amount of binocular disparity to excite the filter. We verified that the three filters display maximal responses at different values of d, as predicted theoretically. These responses should have unique maxima in order to provide

Fig. 6. Typical stereogram obtained from our imaging setup. Each virtual camera occupies approximately half the pixels. The bright object at the center of each half-image is a Styrofoam cup that was moved to test the disparity algorithm.

meaningful feedback on the location of objects. We found that this condition was satisfied over a range of at least 60cm, which is sufficient for many applications.

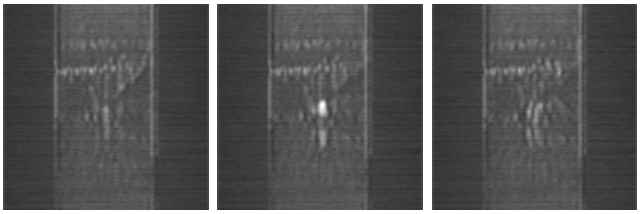

Fig. 7. Experimentally measured outputs of the disparity-sensitive filter tuned to $\Delta\phi = -\pi/2$.

VI. Conclusion

In the current implementation we used disparity filters that operated on a single spatial scale, because the spatial frequency ω was fixed. In addition, all the filters were oriented along the x-axis. We would like to extend our work to include spatial filters at several scales and orientations. However, the amount of local storage available in cellular processor array systems such as SCAMP is typically quite limited. Therefore an interesting direction for future research would be to adaptively determine the required range of scales and orientations based on the characteristics of the current scene.

Acknowledgment

The authors would like to thank the organizers of the annual Telluride Neuromorphic Cognition Engineering Workshop, where the work described in this paper was carried out.

References

[1] I. Ohzawa, G. DeAngelis, and R. Freeman, "Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors," *Science*, vol. 249, no. 4972, pp. 1037–1041, 1990.

[2] D. J. Fleet, H. Wagner, and D. J. Heeger, "Neural encoding of binocular disparity: Energy models, position shifts and phase shifts," *Vision Research*, vol. 36, no. 12, pp. 1839–1857, Jun. 1996.

[3] Y. Chen and N. Qian, "A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms," *Neural Comput.*, vol. 16, no. 8, pp. 1545–1577, 2004.

[4] E. K.-C. Tsang and B. Shi, "Estimating disparity with confidence from energy neurons," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 1537–1544.

[5] P. Dudek and P. Hicks, "A general-purpose processor-per-pixel analog SIMD vision chip," *IEEE Transactions on Circuits and Systems I*, vol. 52, no. 1, pp. 13–20, 2005.

[6] P. Dudek and S. Carey, "General-purpose 128×128 SIMD processor array with integrated image sensor," *Electronics Letters*, vol. 42, no. 12, pp. 678–679, 2006.

[7] P. Dudek and P. Hicks, "A CMOS general-purpose sampled-data analog processing element," *IEEE Transactions on Circuits and Systems II*, vol. 47, no. 5, pp. 467–473, 2000.

[8] J. R. Movellan, "Tutorial on Gabor filters," 2008. [Online]. Available: http://mplab.ucsd.edu/tutorials/gabor.pdf

[9] M. Inaba, T. Hara, and H. Inoue, "A stereo viewer based on a single camera with view-control mechanisms," in *Proceedings of the 1993 IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 3, 1993, pp. 1857–1865.

[10] J. Gluckman and S. Nayar, "Rectified catadioptric stereo sensors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 224–236, 2002.