In Proceedings of the Sixth International Conference on Cognitive Modeling, 326-331.
Mahwah, NJ: Lawrence Earlbaum.

Vergence Control in Fixation with Minimal Disparity Information

Weilie Yi (wyi@cs.rochester.edu)
Department of Computer Science, University of Rochester
Rochester, NY 14627 USA

Dana H. Ballard (dana@cs.rochester.edu)
Department of Computer Science, University of Rochester
Rochester, NY 14627 USA

Abstract

Vergence control is an important part of biologically plausible models of active vision, which plays an important role in cognition. Our vergence system is based on the output of three disparity-selective neurons corresponding to zero, near, and far disparities. A neuron's output is defined as the distance between the feature vectors of two points in the left and right images. Vergence control is generated by using these three disparity values, in search for global minimum of disparity. To escape from local minima, the image is subsampled, and gradually expanded to the original scale. Empirical results shows that the method is effective and robust when applied to targets in natural scences.

Introduction

In the human binocular visual system, because of the different viewpoints of the eyes, two retinas receive similar, but slightly different images of the physical world. To ensure the object of interest is in the fovea, where the majority of optical sensors gather (Yarbus, 1967), the brain sends motor control signals to muscles in the eyes and orient the eyeballs, according to the images perceived by the visual cortex. When the intersection of the lines of sights falls on the object of interest, the object is fixated on, and the vergence control is completed.

A central component of any visual system is the ability to perform figure-ground segmentation. Vergence control ability can greatly facilitate this process by isolating the parts of the world that have small disparities near the point of gaze. The small disparity constraint can be as powerful as color and shape cues(Coombs, 1992) in isolating useful objects. Vergence is also special because it provides depth information needed for reaching.

Vergence control has been implemented in various artificial binocular active vision systems (Hansen and Sommer, 1996; Manzotti et al., 2001; Sturzl et al., 2002). The goal of active control is to minimize the distance between the central pixel/patches of the two views. For the sake of simplicity, one camera could be designated as the reference camera, and computation of vergence control parameters only occurs on the other camera.

There are basically two methods to determine the control parameters to verge the second camera. One is to solve the Stereo Correspondence Problem. This problem asks which pixel in the first image corresponds to which pixel in the second one, or, in this case, which pixel in the verging camera's image corresponds to the central pixel of the dominant camera's image. Solving this problem involves heavy computation, because every possible pixel in the vergence view has to be compared with the central pixel in the reference view, and the one with maximal similarity is labeled as the corresponding pixel(Manzotti et al., 2001).

The other method is to use a close loop control module, which takes error as input and uses PID control to generate vergence parameters(Hansen and Sommer, 1996). This method does not guarantee convergence because the error curve could have multiple minima with various values due to the change of image characteristics in an nonlinear and unpredictable fashion(Manzotti et al., 2001).

In a symmetrical binocular vision system, which doesn't have a dominant camera, the same problem persists(Sturzl et al., 2002).

To make the vergence control computationally efficient and biologically more plausible, we incorporated the idea of disparity selective cells(Ohzawa et al., 1996) and minimized the amount of information needed for the vergence control problem. The motivation is that, in a foveated visual system, the majority of information comes from the fovea. So the computation has to focus on the centers of retinal images. In our approach, three neurons are responsible for detecting zero, tuned near and tuned far disparities respectively. According to the relative values of these three neuron's output, control command is generated and the two views are updated, entering the next loop of execution. Inspired by psychopysiological studies, we also applied Scale Space Theory to ensure a better convergence.

The rest of this paper is arranged as follows. The next section introduces the simple cell model and our basic algorithm for vergence control. Section 3 provides a modified version of this algorithm. Experimental results are presented in Section 4. Finally, conclusive remarks are given in Section 5.

Vergence Control Algorithm

In our binocular vision system, only horizontal disparities exist. Further more, since the fovea lies in the center of the image, we only consider the distance between the central point in one image and points on the middle horizontal line of the other image, neglecting vertical varieties. As shown in Figure 1, an error curve has all the information we need to compute the vergence control

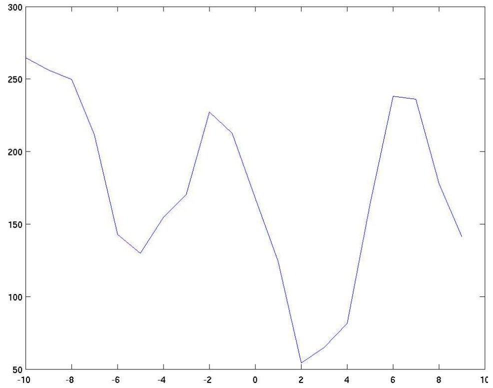


Figure 1: Error curve for the right eye. The X axis is the relative horizontal position from the center of the retinal image. The Y axis is the Euclidean distance, between the image patch at position X in the right image and the one at position -X in the left image. For example, the patches in the center of the two images have an error of about 150, while the patch centered at (4, 0) in the right image and the patch centered at (-4, 0) in the left image have an error of around 80. To fixate on any particular point, the vergence system must reduce its error to zero.

parameters. The minimum of this curve corresponds to the optimal verge angle. In our algorithm, we only use a minimal subset of it to direct the vergence, instead of searching for the minimum in a brute force manner.

Simple Cell Model

On the visual pathway of the brain, visual stimuli are sent from the retina to the primary visual cortex via LGN (lateral geniculate nucleus) (Hubel, 1988). The majority of the orientation sensitive neurons are called *simple cells*, whose receptive fields can be modeled by Gabor filters (Petkov and Kruizinga, 1997). In the classical view, *binocular neurons* in the visual cortex receive inputs from simple cells and they are tuned to specific disparity values (Hubel, 1988) (Qian, 1994) (Qian and Zhu, 1997). Disparity selective binocular neurons can be classified to five categories, according to their tuning curve: tuned zero disparity cells, tuned near/far disparity cells and untuned near/far disparity cells.

Receptive Fields

The base representation, or the receptive fields of the first layer image processing units, could be modeled by Gabor filters (Daugman, 1980). The general form of a 2-D Gabor filter is:

$$F(x,y) = \frac{1}{2\pi\alpha\beta} e^{-\pi[\frac{(x-x_0)^2}{\alpha^2} + \frac{(y-y_0)^2}{\beta^2}]} e^{i(\xi_0 x + \nu_0 y)} \qquad (1)$$

We will use Gaussian derivatives to approximate Gabor filters (Rao and Ballard, 1995):

$$G(x,y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (2)$$

$$G_1^0 = \frac{\partial}{\partial x} G(x,y) = -2xe^{-(x^2+y^2)} \qquad (3)$$

$$G_1^{\pi/2} = \frac{\partial}{\partial y} G(x,y) = -2ye^{-(x^2+y^2)} \qquad (4)$$

Basing on Eqs. 3 and 4, orientation selective receptive fields with the same spatial frequency could be synthesized as:

$$G_1^\theta = k_{11}(\theta) G_1^0 + k_{21}(\theta) G_1^{\pi/2} \qquad (5)$$

where θ is the preferred orientation and

$$k_{11}(\theta) = cos(\theta)$$
$$k_{21}(\theta) = sin(\theta)$$

Similar steerability exists in other orders. For each pixel on the image, a feature vector **V** is computed basing on a set of selective fields, typically 0th to 3th Gaussian derivatives, by doing convolution with discrete filters on the image.

$$\boldsymbol{V}(x,y) = \sum_{|i-x|<R, |j-y|<R} I(i,j) \boldsymbol{G}(i-x, j-y) \qquad (6)$$

where **G** is a vector of discretized Gaussian derivative receptive fields. A screen shot of the computation of base representation is in Fig. 2.

Minimal Information Model

Of the five categories of disparity selective neurons, we only used one neuron from the first 3 categories (tuned zero, tuned near and tuned far), whose receptive field is easy to implement. The zero disparity neuron's output is defined as the Euclidean distance between the feature vectors of the central pixels of two images.

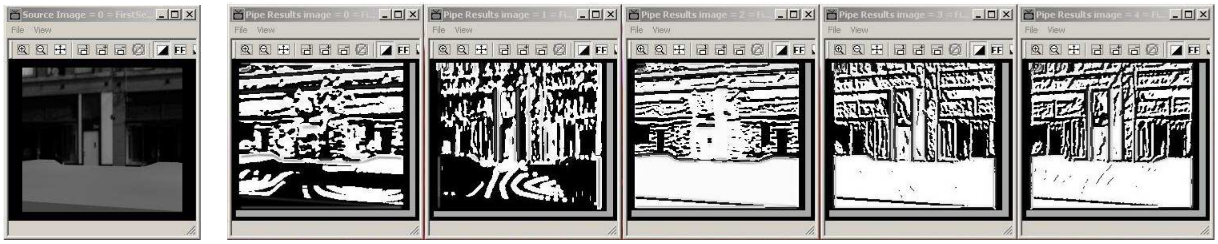
$$d^{zero} = (\boldsymbol{V}^l(\hat{x},\hat{y}) - \boldsymbol{V}^r(\hat{x},\hat{y})) \cdot (\boldsymbol{V}^l(\hat{x},\hat{y}) - \boldsymbol{V}^r(\hat{x},\hat{y}))^T \qquad (7)$$

where $< \hat{x}, \hat{y} >$ is the central pixel of an image, and $\boldsymbol{V}^l$ and $\boldsymbol{V}^r$ stand for the feature vectors of left image and right image respectively. The near and far disparities are defined as

$$d^{near} = (\boldsymbol{V}^l(\hat{x}+1,\hat{y}) - \boldsymbol{V}^r(\hat{x},\hat{y})) \cdot (\boldsymbol{V}^l(\hat{x}+1,\hat{y}) - \boldsymbol{V}^r(\hat{x},\hat{y}))^T \qquad (8)$$

$$d^{far} = (\boldsymbol{V}^l(\hat{x}-1,\hat{y}) - \boldsymbol{V}^r(\hat{x},\hat{y})) \cdot (\boldsymbol{V}^l(\hat{x}-1,\hat{y}) - \boldsymbol{V}^r(\hat{x},\hat{y}))^T \qquad (9)$$

The vergence control algorithm only uses the output of these three disparity tuned neurons.


Source Image = 0 = FirstSc...
Pipe Results image = 0 = Fi...
Pipe Results image = 1 = Fi...
Pipe Results image = 2 = Fi...
Pipe Results image = 3 = Fi...
Pipe Results image = 4 = Fi...
File View
FF

(a) (b)

Figure 2: Filter Outputs. (a) shows the raw image sent over network, and (b) shows five filtered images used in a vergence control operation. These filters, from left to right, are 1st order and 2nd order Gaussian derivatives. The sizes of these filters are all 15 by 15

The Basic Vergence Control Algorithm

Our basic vergence control algorithm, which doesn't work well, is like follows.

Vergence Control Algorithm 1

```
calculate errors;
WHILE (d_near < d_zero OR d_far < d_zero)
{
  IF (d_near < d_zero)
     converge the eyes;
  ELSE
     diverge the eyes;
  re-calculate errors;
}
```

The vergence control here is symmetric: Both cameras adjust their orientation symmetrically basing on the output of the three disparity neurons. In the loop, outputs of the three disparity neurons are compared. We tried to used quadratic function to model the local error curve according to the three values, but failed due to the unpredictable nature of the views. This algorithm terminates when the minimum is found, e.g. the zero disparity value is smaller than both near and far disparities.

A problem with this algorithm is that it is always looking for a minimum which is next to the starting point. The minimum it finds is not guaranteed to be a global one. We will address this problem in the next section.

Saccadic Suppression and Scale Space Theory

The solution of the local minimum problem is motivated by a psychological observation called saccadic suppression and the Scale Space theory in computer vision.

Psychological studies discovered that when the eye is performing a saccadic movement, the resolution of the images sent to the visual cortex is reduced to keep the stability of the view(Ross and Ma-Wyatt, 2004; Thilo et al., 2004). This is mainly because of the prohibitive lateral retinal interconnections(Hubel, 1988). Computationally speaking, retinal images are subsampled before being sent over the optic nerve. Since every saccade is generally followed by a fixation, fixation can take advantage of saccadic suppression by processing the shrunken images.

In our revised algorithm, vergence control begins with a highly shrunken image pair. Suppression is implemented as another loop which gradually decreases the suppression factor so that the images become finer and more details are unveiled. As the original sized images are restored, the accuracy of vergence comes to a maximum. With this suppression loop, the pixel pair with global minimum of distance is always close to the centers of the images.

Closely related to the saccadic suppression observation, Scale Space Theory was introduced to study the properties of images in different scale levels(Lindeberg and Romeny, 1994; Lindeberg, 1994). For any image I, a continues family of images $I(x, y, \sigma)$ is a series of blurred versions of $I(x, y)$ in which $I(x, y, \sigma)$ is the original image $I(x, y)$ when $\sigma = 0$:

$$I(\sigma) = I \otimes G(\sigma) \qquad (10)$$

where $G(\sigma)$ is a Gaussian kernel $G(x, y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2+y^2}{2\sigma^2}}$ and $\otimes$ is a convolution operator. The space (x, y, σ) is called a scale space, and σ the scale parameter.

One of the basic properties of the scale space is "non-creation of extrema"(Lindeberg, 1994), which means, as the scale parameter increases, local extrema would merge together, and when σ is above a certain value, only one extremum exists.

The Enhanced Vergence Control Algorithm

Inspired by the theories described above, we revised our algorithm to simulate the suppression process. Intuitively when the scale parameter, or equivalently the suppression factor, increases, local minimum tend to disappear and only the global one persists, at a slightly different location, because of the blurring effect. Our enhanced vergence control algorithm is as follows.

Vergence Control Algorithm 2

```
initialize scale
WHILE(scale > 0)
{
  calculate errors;
  WHILE (d_near < d_zero OR d_far < d_zero)
  {
    IF (d_near < d_zero)
       converge the eyes;
    ELSE
       diverge the eyes;
    re-calculate errors;
  }
  lower scale;
}
```

Note that this algorithm doesn't guarantee convergence to the global minimum of error, which in some extreme cases, e.g. wall paper illusion, is impossible(McKee and Mitchison, 1988).

Experiments

We implemented this algorithm in the Virtual Reality Laboratory at University of Rochester. A virtual human running on an SGI station walks in a town, and the images it sees are sent over network to a PC where the algorithm runs. This PC has a Datacube$^{\text{TM}}$ MaxRevolution image processing board, which is dedicated to the computation of image convolutions. Vergence control commands are sent back to the virtual human and the view ports of its two eyes are updated.

The changes of disparity curves are illustrated in Figure 3. Note only the three disparity values, d^{near}, d^{zero} and d^{far}, as shown in Figure 3(b), were actually used in the experiments, while Figure 3(a) gives the complete disparity curves for reference.

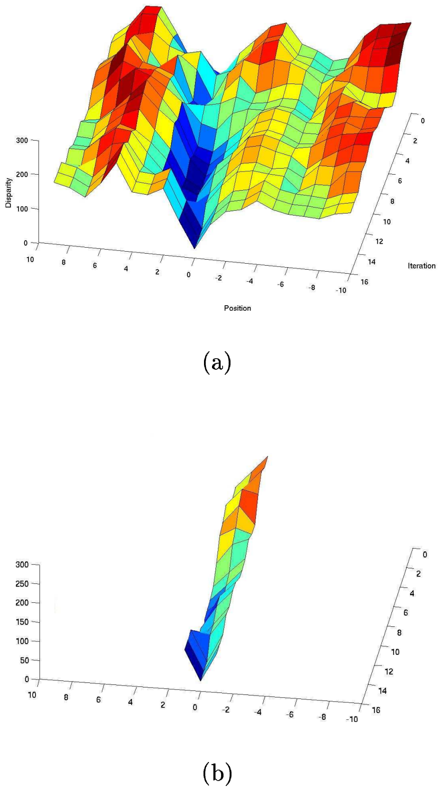
Since vergence facilitates depth perception. We used the deictic information, the distance between two eyes, to compute how far the fixated object is from the viewer. In our experiment, the object gradually moves away and so the vergence angle decreases. We calculated the distance by solving the following equation

$$e/2 = d\tan(\alpha/2) \tag{11}$$

where e is the distance between two eyes, d is the distance between the fixated object and the viewer, and α is the vergence angle. In our experiments, $e = 0.05m$. We used linear regression to extract the relation between the computed distance and the true distance which is available from the simulator program. The result is shown in Figure 4.

Conclusions

We presented a biologically and psychologically plausible model for vergence control. In this single cell based model, minimal disparity was used to determine the vergence command. Inspired by Scale Space Theory and saccadic suppression, a coarse-to-fine loop was introduced to help the process converge.



(a)

(b)

Figure 3: Vergence Control and Error Curves. This plot shows how the error curve (the curve on an error-position plane) changes over the iterations. Only the iterations with the starting scale parameter are illustrated. Later ones with scale parameters down to zero are not displayed for the sake of clarity. Subfigure (a) shows the complete error surface, while (b) only shows the output of the three disparity tuned neurons, which correspond to positions -1, 0 and 1. Only these three values are used in our algorithm, which successfully reduce the error at the point x=0 to zero.

We cannot compare this model to others at this time because we do not have a proper criteria. For instance, it is not reasonable to compare the computational complexity of various models without considering the physiological mechanism, which is the object of these models, and which is unknown. Our work suggests, it is likely that the brain can go around the Correspondence Problem and do vergence control with visual information which is very local to the fovea.

Acknowledgments

This work was supported by NIH grants MH60624 and P41 RR 09283.

References

Coombs, D. J. (1992). *Real-Time Gaze Holding in Binocular Robot Vision.* PhD thesis, University of

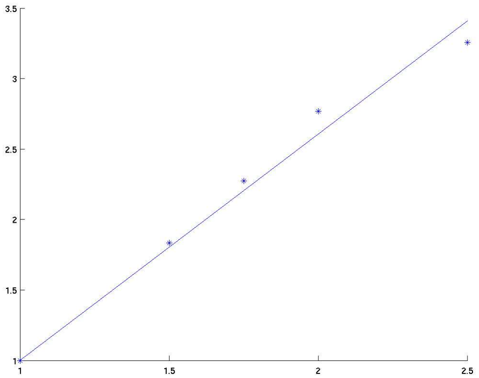


Figure 4: Computation of distance. The X axis is the computed distance between the fixated object and the viewer, while the Y axis is the true distance known from the simulating software.

Rochester.

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profile. *Vision Research*, 20:847–856.

Hansen, M. and Sommer, G. (1996). Active depth estimation with gaze and vergence control using gabor filters. In *Proc., 13th Int. Conf. Pattern Recognition*, volume A, pages 287–291, Vienna, Austria.

Hubel, D. H. (1988). *Eye, brain, and vision.* Scientific American Library.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision.* Kluwer Academic Publishers.

Lindeberg, T. and Romeny, B. M. T. H. (1994). *Linear scale-space*, chapter 1-2. Series in Mathematical Imaging and Vision. Kluwer Academic Publishers.

Manzotti, R., Gasteratos, A., and Metta, G. (2001). Disparity estimation in log polar images and vergence control. *Computer Vision and Image Understanding*, 83:97–117.

McKee, S. P. and Mitchison, G. J. (1988). The role of retinal correspondence in stereoscopic matching. *Vision Research*, 8:1001–1012.

Ohzawa, I., DeAngelis, G. C., and Freeman, R. D. (1996). Encoding of binocular disparity by simple cells in the cat's visual cortex. *J. Neurophysiol.*, 75:1779–1805.

Petkov, N. and Kruizinga, P. (1997). Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. *Biological Cybernetics*, 76(2):83–96.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation*, 6(3):390–404.

Qian, N. and Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research*, 37:1811–1827.

Rao, R. P. and Ballard, D. H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505.

Ross, J. and Ma-Wyatt, A. (2004). Saccades actively maintain perceptual continuity. *Nature Neuroscience*, 7(1):65–69.

Sturzl, W., Hoffmann, U., and Mallot, H. A. (2002). Vergence control and disparity estimation with energy neurons: Theory and implementation. In Dorronsoro, J. R., editor, *Proc., Int'l. Conf. on Artificial Neural Networks*, pages 1255–1260.

Thilo, K. V., Santoro, L., Walsh, V., and Blakemore, C. (2004). The site of saccadic suppression. *Nature Neuroscience*, 7(1):13–14.

Yarbus, A. L. (1967). *Eye movements and vision.* New York, Plenum Press.