Natural Scene Statistics at Stereo Fixations

Yang Liu
Department of ECE
Univeristy of Texas at Austin
young76@mail.utexas.edu

Lawrence K. Cormack
Department of Psychology
University of Texas at Austin
cormack@psy.utexas.edu

Alan C. Bovik
Department of ECE
University of Texas at Austin
bovik@ece.utexas.edu

Abstract

We conducted eye tracking experiments on naturalistic stereo images presented through a haploscope, and found that fixated luminance contrast and luminance gradient were generally higher than randomly selected luminance contrast and luminance gradient, which agrees with previous literatures. However we also found that the fixated disparity contrast and disparity gradient were generally lower than randomly selected disparity contrast and disparity gradient. We discuss the implications of this remarkable result.

Keywords: fixation, stereopsis, natural scene statistics

1 Introduction / Overview

Many studies have shown that several low level luminance features at human fixations are significantly different from those at other areas. Reinagel and Zador [1999] found that the regions around human fixations tend to higher spatial contrasts and spatial entropies than random fixation regions, which suggests that the human visual system may try to select image regions that help maximize the information content transmitted to the visual cortex, by minimizing the redundancy in the image representation. By varying the patch sizes around the fixations, Parkhurst and Niebur [2003] found that the largest difference between the luminance contrast of fixated regions and that of image shuffled (pseudo-random) regions is observed when the patch size is 1 degree. In a recent study [Rajashekar et al. 2007], fixated image patches were foveated using an eccentricity-based model. Higher order statistics were analyzed than in prior studies. They found that bandpass contrast showed a notably larger difference between fixations and random patches than other higher-order statistics. Other features found to attract fixations (in decreasing order of attractiveness) included bandpass luminance, RMS contrast, and luminance. This data was subsequently used to develop an image processing "fixation predictor" [Rajashekar et al. 2008].

The body of literature on visual fixations on 2D luminance images appears to be largely consistent across the studies. However, there has been very little work done on analyzing the nature or statistics of images and scenes at the point of gaze in three dimensions. Certainly the three dimensional attributes of the world affect the way we interact with it both visually and physically. Moreover, the 3D statistics of the natural world have likely played a role in the adaptation of the visual system [Liu et al. 2008]. One reason for the lack of studies on fixations in 3D space has been the dearth of 3D ground truth natural scene data. The few databases of naturalistic stereo images do not adequately fulfill this need, since what is needed are dense disparity maps for each scene. One very recent study [Jansen et al., 2009] used natural scenes with ground truth disparity map acquired from laser scanning. They found that disparity appears to be a salient feature that affects eye movements. In particular, they found that the presence of disparity information may affect saccade length, but not duration; that disparity appears not to affect the saliency of luminance features; and that subjects tend to fixate nearer objects earlier than more distance objects. A reasonable agreement with intuition may be found in each of the results. The authors also found that in 3D noise images, subjects tended to fixate depth discontinuities more frequently than smooth depth regions. One might view this result as intuitive also; as with luminance images, such locations might be deemed interesting.

ETRA 2010, Austin, TX, March 22 – 24, 2010.
© 2010 ACM 978-1-60558-994-7/10/0003 $10.00

2 Stereo Eyetracking

2.1 Observers

Three male observers participated in this study. Their stereo ability was tested to be normal by presenting them random dot stereograms. Observer LKC has extensive experience in psychophysical studies, and knew the purpose of the experiment. Observers JSL and CHY were naïve.

2.2 Stimulus

We manually selected 48 grayscale stereoscopic outdoor scenes [Hoyer & Hyvärinen, 2000] that contained mountains, trees, water, rocks, bushes, etc., but avoided manmade objects. We didn't have the ground truth disparity data from these scenes. Instead, we relied on a local correlation method which is simple yet biological inspired to solve this issue. Models of binocular complex neurons [Anzai et al. 1999; Fleet et al. 1996; Ohzawa et al. 1990; Qian, 1994] commonly contain a cross correlation term in their response function.

The simple correspondence algorithm was defined as follows. Given a pixel (x_r,y_r) in the right image, we defined a 161×5 (3.2°×0.1°) search window centered on the same pixel location (preferring zero disparity) in the left image. Given a 1°×1° patch in the right image centered on the pixel (x_r,y_r), the algorithm computed the cross correlation between the right patch and a candidate 1°×1° left patch centered on each pixel in the 161×5 search window. The left patch yielding the largest cross correlation was deemed to be the matched patch. The location (x_l, y_l) corresponding the center of the patch was deemed the matched pixel for the right pixel (x_r,y_r), hence the horizontal disparity of (x_r,y_r) was taken to be $D(x_r,y_r) = x_r-x_l$.

It could be argued that, since we are only interested in local scene statistics, why not record the movements of both eyes, and use the disparity between the recorded left and right fixations to find the correct match? There are several reasons why this wasn't practical. First, there are fixation disparities that occur between the right eye and the left eyes. Most people have a fixation disparity that is less than 6 arcmin, but can be as large as 20 arcmin with peripheral visual targets [Wick, 1985]. When fixation disparity occurs, the image of an object point that a person is trying to fixate do not fall on exactly corresponding points. Secondly, each Purkinje eye tracker has an accuracy about 7 acrmin (the median offset from real fixations); hence the error between two eye trackers is about 14 arcmin, which corresponds to about 12 pixels. This is quite a large error, considering that the stereo images are only 800×600. Thirdly, the fixation detection algorithm (associated with eye-tracking) of the two eye paths can also introduce unwanted noise into the corresponding fixation locations; registered stereoscopic eyetracking is difficult to accomplish in practice. Lastly, since we are interested in disparity features within neighborhoods (not just points) of the fixations, a dense disparity map is required for all points in the neighborhood. Hence, local processing of the type that our disparity algorithm accomplishes would be required anyway.

2.3 Equipments

Stereo images were displayed on two 17 inch, gamma calibrated monitors. The distance between the monitors and the observer was 124 cm. Each monitor's screen resolution was set at 800×600 pixels, corresponding to about 50 pixels per degree of visual angle. The total spatial extent of each display was thus about 16◦ × 12◦ of visual angle.

A haploscope was placed between the two monitors and the observers to completely separate the displays from the left and right monitor. Eye movements were recorded by a SRI Generation V Dual Purkinje eye tracker. This eye tracker has an accuracy of < 10′ of arc, a response time of under 1 ms, and bandwidth of DC to > 400Hz. The output of the eye tracker (horizontal and vertical eye position signals) was low-pass filtered in the hardware and then sampled at 200 hHz by a National Instruments data acquisition board in a Pentium IV host computer, where the data were stored for offline data analysis. The observers used a bite bar and a forehead rest to restrict their head movements.

A viewing session was composed of 48 viewed stereo image pairs. At the beginning of each session, a 0.3°×0.3° crosshair was displayed on the centers of both monitors to help the observers to fuse by fixating on it. When the correct binocular fixation (the crosshair was perceived single) was achieved, the observer pressed a button to start the calibration. Two 3×3 calibration grids were displayed on the monitors respectively. After the observers visited all 9 dots, a linear interpolation was then done to establish the transformation between the output voltages of the eye tracker and the position of the subject's gaze on each computer display. The calibration also accounted for crosstalk between the horizontal and vertical voltage measurements. After correct calibration, a 0.3°×0.3° crosshair was displayed on the centers to force all observers to start from the same center position. The stereo images were displayed on two monitors for 10 seconds during which the eye movements were recorded. Between two consecutive image pairs, two identical Gaussian noise images were displayed for 3 seconds on both monitors to help suppress after-images corresponding to the previous stereo pairs that may otherwise have attracted fixations. Then a 0.3°×0.3° crosshair was displayed on the centers of both monitors to help the observers to fuse before the presentation of the next stereo pair.

This calibration routine was repeated compulsorily every 10 images, and a calibration test run every 5 images. This was achieved by requiring that the observer fixate for 500ms within a 5s time limit on a central square region (0.3◦ ×0.3◦) prior to progressing to the next image in the stimulus collection. If the calibration had drifted, the observer would be unable to satisfy this test, and the full calibration procedure was re-run.

Observers who became uncomfortable during the experiment were allowed to take a break of any duration they desired.

The ambient illumination in the experiment room was kept constant for all observers, with a minimum of 5 minutes luminance adaptation provided while the eye-tracker was calibrated.

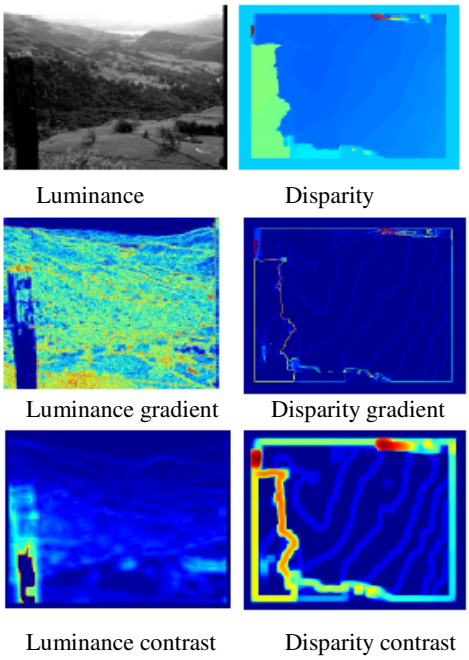

Figure 1. The luminance and disparity features

3 Analysis

3.1 Computation of scene statistics

Denote the right image as I, and the dense disparity map as D We computed the luminance gradient map:

$$G_l = \sqrt{\frac{\partial I}{\partial x}^2 + \frac{\partial I}{\partial y}^2}$$

and the disparity gradient map:

$$G_d = \sqrt{\frac{\partial D}{\partial x}^2 + \frac{\partial D}{\partial y}^2}$$

using the Matlab function gradient(X). Here we define luminance contrast as the RMS contrast of a luminance patch:

$$C_l = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\frac{I_i - \bar{I}}{\bar{I}})^2},$$

and likewise, disparity contrast as

$$C_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(D_i - \bar{D})^2}$$

which is the standard deviation of a disparity patch. All of the following analysis is based on these four scene features: luminance contrast, disparity contrast, luminance gradient, and disparity gradient. Figure 1 depicts these scene features for one example stereo pair.

3.2 Random locations

Suppose an observer made f_i fixations for the i^{th} image. Then, the total number of fixations that the observer made during a session is $\sum_i^{48} f_i$. We assume that a random observer also made the same number of fixations as the subject did. That is, for the i^{th} image, the random observer selected f_i fixations uniformly distributed on the image plane too. For each human observer, we assume that there are 100 random observers each making the same number of fixations in each image as the human observer. For example, the overall fixation that subject LKC made is 486 fixations in 48 images, so each random observer selected 486 random locations too. The total number of random locations is 48,600. We wanted to know whether or not there is a statistically significant difference between image features at fixations and those at randomly selected locations by comparing the human observer's data and the 100 random observers' data.

3.3 Fixation/Random ratios

For the i^{th} image, we computed the mean luminance contrast at the f_i fixations as:

$$C_l^i = \sum_j^{f_i} C_l(x_j, y_j)/f_i$$

where (x_j, y_j) is the location of the j^{th} fixation, and C_l is the luminance contrast map.

We also computed the mean luminance contrast at the f_i random locations as:

$$C_{rl}^i = \sum_j^{f_i} C_l(u_j, v_j)/f_i$$

where (u_j, v_j) is the location of the j^{th} random location.

We defined patch luminance gradient as the mean gradient of the patch:

$$\bar{G_l} = \sum_1^N G_l(x, y)/N,$$

where G_l is the luminance gradient map. Similarly, we computed the mean patch luminance gradient of the f_i fixations for the i^{th} image as:

$$G_l^i = \sum_j^{f_i} \bar{G_l}(x_j, y_j)/f_i$$

where (x_j, y_j) is the location of the j^{th} fixation.

We also computed the mean patch gradient of the f_i random locations the i^{th} image as:

$$G_{rl}^i = \sum_j^{f_i} \bar{G_l}(u_j, v_j)/f_i$$

where (u_j, v_j) is the location of the j^{th} random location.

For each image, we then defined the fixation-to-random luminance contrast ratio $RC_l = C_l^i/C_{rl}^i$ and the fixation-to-random luminance gradient ratio $RG_l = G_l^i/G_{rl}^i$. If $RC_l > 1$, it means that the fixated patches generally have a larger luminance contrast than randomly selected patches on the image being considered. If $RC_l < 1$, then the meaning is reversed. The same meaning applies to the luminance gradient ratio.

The same analysis method that was used on the luminance contrast and luminance gradient was also applied to for the analysis of disparity. We calculated the mean disparity contrast C_d^i on the fixated patches, and the same quantity C_{rd}^i on the randomly selected patches. The ratio of disparity contrast between the fixated patches and the randomly selected patches is defined as $RC_d = C_d^i/C_{rd}^i$.

The mean patch disparity gradient at the fixated patches (G_d^i) and the randomly selected patches (G_{rd}^i) was also calculated. The ratio of the disparity gradient between the fixated patches and the random patches is defined as $RG_d = G_d^i/G_{rd}^i$. If the ratios are significantly greater than 1, then fixated patches tend to have a larger disparity contrast and gradient than randomly picked locations.

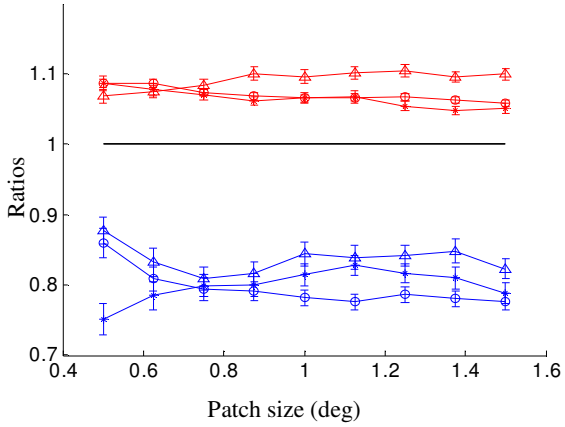

Figure 2. Mean luminance gradient ratios (red), and mean disparity gradient ratios (blue) of three observers.

We ran 100 simulations for each image, and plotted the mean luminance gradient ratio $\overline{RG_l}$ with 95% confidence intervals (CI), and the mean disparity gradient ratio $\overline{RG_d}$ with 95% CIs, for all subjects as shown in Figure 2. For better comparison, we plotted 1 as a straight horizontal line across all patch sizes. The red curves show the ratios of luminance gradient, and the blue curves showed the ratios of disparity gradients. Different markers were used to represent the observers: LKC (*), CHY (o), JSL (Δ). We made a

similar ensemble comparison plot for the mean luminance contrast ratio and mean disparity contrast ratio, as displayed in Figure 3. It is clear that all the mean ratios of luminance gradient/contrast and their 95% CIs fall well above 1, while the mean ratios of disparity gradient/contrast fell well below 1. All of the results we obtained lead to the conclusion that humans tend to fixate at regions of higher luminance variation, and lower disparity variation.

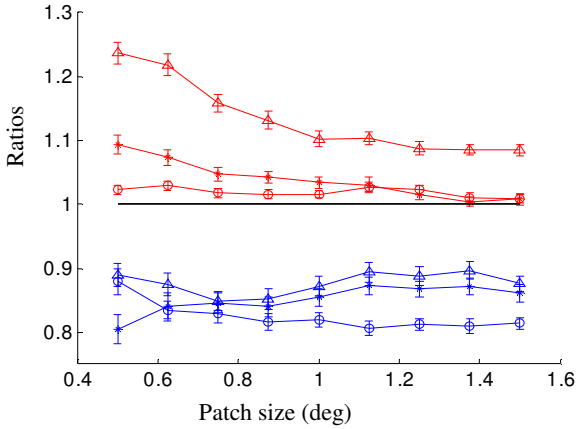

Figure 3. Mean luminance contrast ratios (red), and mean disparity contrast ratios (blue) of three observers.

4 Discussion

The most immediate explanation for this behavioral phenomenon is that the binocular vision system actively seeks to fixate scene points that will simplify or enhance stereoscopic perception. The nature of this enhancement is less clear. We suggest that the binocular visual system, unless directed otherwise by a higher-level mediator, seeks fixations that simplify the computational process of disparity and depth calculation. In particular, we propose that the binocular vision system seeks to avoid, when possible, regions where the disparity computations are complicated by missing information, such as occlusions, or rapid changes in disparity, which may be harder to resolve.

Many fMRI studies have reported that the V1 activity increases with disparity variation [Tootell et al. 1988; Backus et al. 2001]. Georgieva et al. [2009] showed that activity in occipital cortex and ventral IPS, at the edges of the V3A complex was correlated with the amplitude of the disparity variation that subjects perceived in the stimuli. All of these studies show that the processing of large, complex disparity shapes involves the allocation of more neuronal resources and energy than do smooth disparity fields.

Another (related) possibility is that when viewing an object, observers tend to fixate towards the center of the object, leaving the object boundaries (often associated with depth discontinuities) to peripheral processing involving larger receptive fields both in space and disparity. Such a strategy would allow the fovea to operate under better-posed stereo viewing conditions, and to process 3D surface detail, while the periphery could simultaneously encode the (statistically) large disparities associated with object boundaries.

References

ANZAI, A., OHZAWA, I., AND FREEMAN, R. D. 1999. Neural Mechanisms for Processing Binocular Information II. Complex Cells. *J Neurophysiol*, *82*, 2, 909–924.

BACKUS, B. T., FLEET, D. J., PARKER, A. J., & HEEGER, D. J. 2001. Human Cortical Activity Correlates With Stereoscopic Depth Perception. *J Neurophysiol*, *86*, 4, 2054-2068.

FLEET D. J., WAGNER H., AND HEEGER D. J. 1996. Neural Encoding of Binocular Disparity: Energy Models, Position Shifts and Phase Shifts. *Vision Research*, *36*, 1839-1857.

GEORGIEVA, S., PEETERS, R., KOLSTER, H., TODD, J. T., & ORBAN, G. A. 2009. The Processing of Three-Dimensional Shape from Disparity in the Human Brain. *J. Neurosci.*, *29*, 3, 727-742.

HOYER, P. O., AND HYVÄRINEN, A. 2000. Independent Component Analysis Applied to Feature Extraction from Colour and Stereo Images . *Network Computation in Neural Systems, 11*, 191--210.

JANSEN, L., ONAT, S., AND KÖNIG, P. 2009. Influence of Disparity on Fixation and Saccades in Free Viewing of Natural Scenes. *Journal of Vision 9*, 1, 1-19.

LIU, Y., BOVIK, A. C., AND CORMACK, L. K. 2008. Disparity Statistics in Natural Scenes. *Journal of Vision 8*, 11, 1-14.

OHZAWA, I., DEANGELIS, G., AND FREEMAN, R. 1990. Stereoscopic Depth Discrimination in The Visual Cortex: Neurons Ideally Suited as Disparity Detectors. *Science*, *249*, 4972, 1037-1041.

PARKHURST, D. J., AND NIEBUR, E. 2003. Scene Content selected by active vision. *Spatial Vision 16*, 2, 125-154.

QIAN, N. 1994. Computing Stereo Disparity and Motion with Known Binocular Cell Properties. *Neural Computation*, *6*, 3, 390-404.

RAJASHEKAR, U., VAN DER LINDE, I., BOVIK, AND CORMACK, L. K.. 2007. Foveated Analysis of Image Features at Fixations. *Vision Research 47*, 25, 3160-3172.

RAJASHEKAR, U., VAN DER LINDE, I., BOVIK, A. C., AND CORMACK, L. K. 2008. GAFFE: A Gaze-attentive Fixation Finding Engine. *IEEE Transactions on Image Processing 17*, 4, 564-573.

REINAGEL, P., AND ZADOR, A. M. 1999. Natural Scene Statistics at the Center of Gaze. *Networks 10,* 4, 341-350.

TOOTELL, R. B., HAMILTON, S. L., SILVERMAN, M. S., & SWITKES, E. 1988. Functional Anatomy of Macaque Striate Cortex. I. Ocular Dominance, Binocular Interactions, and Baseline Conditions. *J. Neurosci.*, *8*, 5, 1500-1530.

WICK, B. 1985. Forced Vergence Fixation Disparity Curves at Distance and Near in an Asymptomatic Young Adult Population. *American Journal of Optometry and Physiological Optics*, *62*, 9, 591-599.