

Optimal sensor design for estimating local velocity in natural environments

Tal Tversky, Wilson S. Geisler

Center for Perceptual Systems, University of Texas at Austin, Austin, TX, USA

ABSTRACT

Motion coding in the brain undoubtedly reflects the statistics of retinal image motion occurring in the natural environment. To characterize these statistics it is useful to measure motion in artificial movies derived from simulated environments where the “ground truth” is known precisely. Here we consider the problem of coding retinal image motion when an observer moves through an environment. Simulated environments were created by combining the range statistics of natural scenes with the spatial statistics of natural images. Artificial movies were then created by moving along a known trajectory at a constant speed through the simulated environments. We find that across a range of environments the optimal integration area of local motion sensors increases logarithmically with the speed to which the sensor is tuned. This result makes predictions for cortical neurons involved in heading perception and may find use in robotics applications.

Keywords: natural scene statistics, motion, speed tuning, heading perception

1. Introduction

Coding in the brain is undoubtedly based on the properties of the environment and the tasks that are necessary for survival and reproduction. Therefore, in order to understand how the brain performs a given perceptual task, it is important to understand the statistics of the environment relevant to that task. Motion provides a great deal of useful information for many perceptual tasks including heading estimation, image segmentation, distance estimation and shape estimation. The heading task is particularly relevant because, in general, most of the visual motion experienced by a human is due to self-motion. All visual information available for heading estimation is contained in the spatiotemporal image produced on the retina by self motion through the 3D environment. In the mammalian visual system, and in many machine vision systems, motion estimation starts with local motion estimates in the image plane. Thus, accurate local image motion estimation is presumably critical for making inferences about physical motion in the environment. The question we are concerned with here is the following: How should local motion detectors be designed in order to make accurate estimates when locomoting through the world?

To answer this question two sets of statistics are needed. First, we need to know the spatiotemporal pattern of light falling on the image plane. Second, we need to know the true projected local motion at each point in the image plane (i.e., the “ground truth”). To know the ground truth image motion, it is necessary to know the actual translation and rotation of the image plane through the environment as well as the three dimensional structure of the environment.

There has been a significant amount of research done measuring and analyzing image motion statistics [1-3, for a review see 4]. None of this work, however, ties the image motion with its real world sources. Zanker and Zeil’s [5] work comes the closest to making the relevant measurements; they gathered image statistics by moving a camera on a robotic gantry through known insect flight paths. However, their work still lacks ground truth information about the true projected motion on the image plane and is specific to the motion and visual environment of insects.

Clearly, measuring the statistics of image motion is a much simpler task than measuring the sources of that motion—the statistics of the structure of the environment and the statistics of the motion through it. Huang et al [6] and a number of other studies [7-9] provide some of the relevant scene statistics by measuring ranges as a function of location in the image plane. Calow et al [7] and Roth and Black [8] use Huang et al.’s range database combined with a collection of estimated motions through the world to create a set of instantaneous ground truth local image velocities at every point in an image. Roth and Black then use these velocities as prior probabilities that inform a machine vision optic flow

estimation algorithm. This is a promising approach, but still lacks a few key pieces of information. Foremost, there is no way to connect the ground truth velocities (estimated from the range data) with image information. Also, occlusions in the world cause gaps in the range information available in the database. This means that instantaneous velocities are accurate, but it is impossible to know how these velocities change over time.

In order to measure statistics that reflect complete information about the spatiotemporal image, the motion of the image plane through the environment, and the structure of the environment, we generate artificial movies using a ray tracer with a world model that is based on measured scene statistics. The main advantage we gain with this approach is that we know the ground truth image motion and its source at every point in the image. Also, we can generate a much larger set of statistics in a short period of time since we are only limited by computational time rather than by the labor intensive process of physical measurement. The disadvantage to this approach is obvious: Our measured statistics can only be as accurate as the models that we use. However, for the particular issue we consider here this disadvantage is unlikely to have a substantial effect (see Discussion). Also, in another ongoing project we are testing our world model by collecting and analyzing video sequences taken over known trajectories with a calibrated camera.

Our world model is based in part on the measurements of Huang et al [6]. They found that the statistics of range are consistent with a world of piecewise smooth regions. More specifically, they found that the range statistics of forest scenes can be modeled well with a world consisting of a flat ground plane populated with a Poisson distribution of cylinders (trees). To match measured image statistics, the reflectance of every surface in our model world is given a frequency spectrum that falls with $1/f$ frequency [9]. The choice of texture with $1/f$ noise is partly motivated by Dong and Atick's [1] observation that the spatiotemporal power spectra of video clips can be modeled as a collection of translating $1/f$ surface patches. Using our world model, we can generate statistics that can help us understand how local motion detectors should be designed in order to make accurate estimates when locomoting through the world.

Motion detectors in image processing and biological models of motion selective neurons commonly assume that local motion is well approximated by translational motion in the image plane. For such detectors the accuracy of local velocity estimates depends on several factors including sensor noise, the extent to which the motion breaks the translational assumption, stimulus ambiguity (e.g. the aperture problem), and the spatial and temporal area over which information is pooled. Larger areas of integration typically contain more complex, non-translational motions. On the other hand, smaller areas of integration are more susceptible to the aperture problem and to sensor noise. For locomotion through an environment, there should be an ideal integration area for local motion detectors that will balance these constraints in order to maximize average accuracy. Furthermore, it is plausible that this optimal area depends upon the local image speed (magnitude of velocity). In this paper, we consider how the optimal integration area of local image velocity detectors varies with local image speed.

2. Methods

2.1. Ray Tracer

Using a ray tracer [10, 11] we generated movies for an observer translating through a model world. Each movie consists of 6 frames. At a 30 hertz sampling rate, this corresponds to movie duration of 200 ms, which is approximately the minimum time for a human fixation. Each movie frame was 316 by 252 pixels corresponding to a visual angle of 39 by 33 degrees. We defined a valid location in the scene as one which, for the largest aperture size, there is no clipping. Valid locations span 28 by 20 degrees. Four different kinds of scenes were generated at two different human walking speeds (3 and 5 feet per second) making eight different scene conditions. The four different kinds of scenes are shown in Fig. 1: two "forest" scenes (a,b), one flat-wall scene (c) and one ground-plane only scene (d). The two forest scenes were based on our world model of scene statistics with two different densities of cylinders. The cylinders had a radius of 2 feet and were distributed with a density of either 5×10^{-3} cylinders/ft² (low density) or 2×10^{-3} cylinders/ft² (high density). The reflectance of all the surfaces in all the scenes was generated using a $1/f$ procedural texture. This texture is the weighted sum of Perlin noise [12] of different frequencies. In each movie, the model observer moved forward and gazed in the direction of translation, 10 degrees down from the horizon. The observer's eyes were located 5 feet

above the ground. In the forest movies, if a cylinder was within 1 foot of the observer, it was not drawn. Image movies and range movies were both rendered simultaneously using a feature of the ray tracer that allows 16 bit accuracy for range values spanning 1 foot to 1000 feet. The range images and the known observer motion are all that is necessary to compute the ground truth local projected motion at each image location.

2.2. Sampling

For each of the eight conditions, 100 movies were generated, each with a new random seed. The random seed determines both the random texture on all the objects and the random placement of the cylinders within the scene. For each of the eight conditions, 20,000 image locations were sampled. Each sample was obtained by randomly picking a movie and a valid location within that movie. At each sampled location, motion estimates were computed at a range of different aperture size. The estimated motion vector was then subtracted from the ground truth motion vector and the length of this difference vector was used as the error. Errors were binned by the magnitude of the ground truth retinal speed using equal quantile binning with 15 bins. In all conditions, the distribution of velocity amplitudes was found to be approximately log normal, resulting in an approximately logarithmic speed binning.

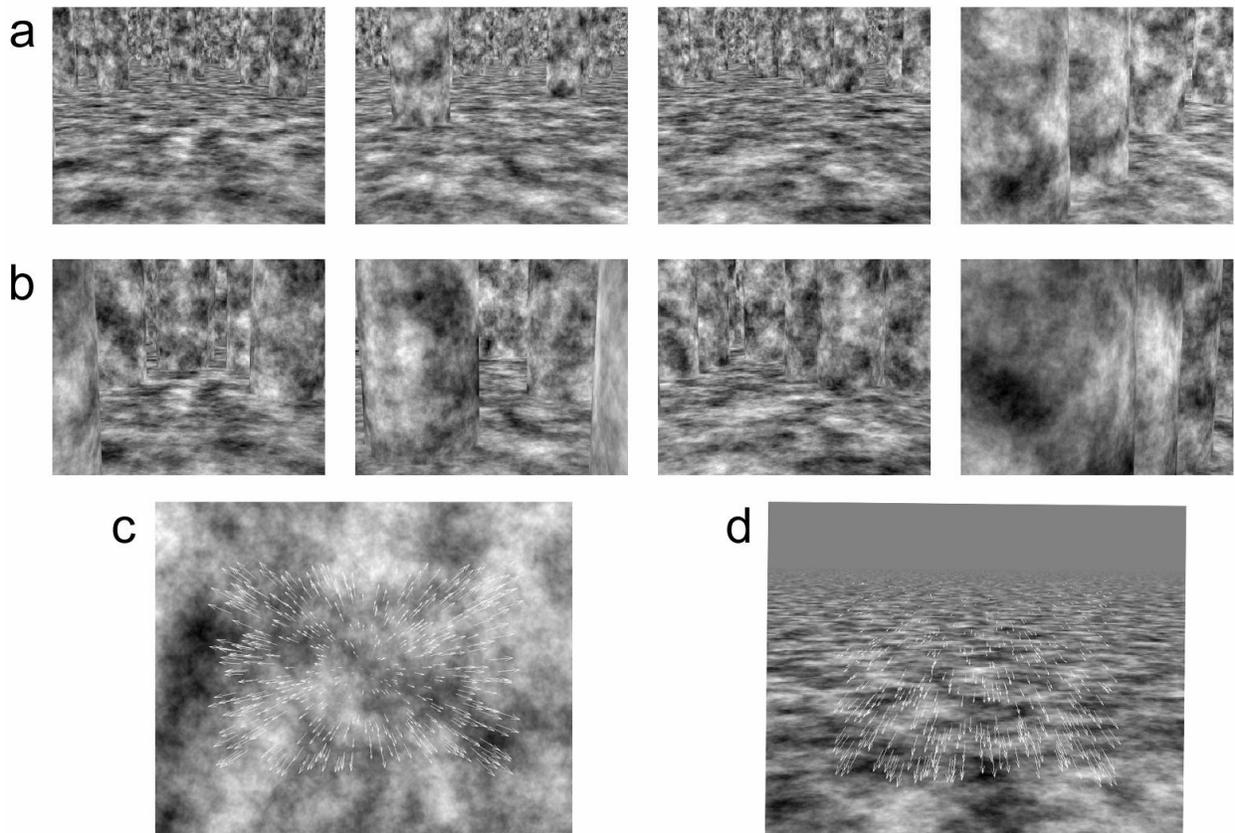


Fig. 1. Examples of frames from movies generated using each of the four different world models. (a) Four frames from different movies using the forest model with a low density of cylinders. (b) Four frames from different movies using the forest model with a high density of cylinders. (c) Flat vertical plane with superimposed sample of estimated motions (white arrows). (d) Flat ground plane with superimposed sample of estimated motions.

2.3. Motion estimation algorithm

For the current study, our aim is not to design an efficient or accurate computer motion detection algorithm, but rather to understand the distributions of velocities in natural environments and their effects on local motion processing. Therefore we wanted a simple motion estimation algorithm whose errors will reflect the environmental constraints. We use a coarse to fine iterative solution to the gradient constraint equation [13]. In separate tests (results not shown) we find the algorithm to be a good, unbiased estimate of the two-dimensional motion energy in a movie. We believe that the measured errors in the motion estimates are overwhelmingly due to the complexities of the motion in the scene, or are due to aperture problems in the scene. At each sampled location the images are windowed with a Gaussian before they are given to the motion estimation algorithm. The reported ‘aperture size’ is twice the standard deviation of the Gaussian window function.

2.4. Best aperture

For each speed bin, at each aperture size, we averaged the error in the estimated velocity over all the samples. Figure 2 plots the average error as a function of aperture size for a particular speed bin. Notice that the function is U-shaped, showing us that for this condition and speed bin the environmental constraints balance against each other to give an ideal aperture size. Too small an aperture leads to large errors because of a lack of information (aperture problem). Too large an aperture leads to large errors because there are too many different motions in the aperture. Figure 3 shows the same plot (error as a function of aperture size) for all 15 speed bins. Notice again that there is a well-defined optimal aperture size. In subsequent figures we plot the best aperture size for each speed bin.

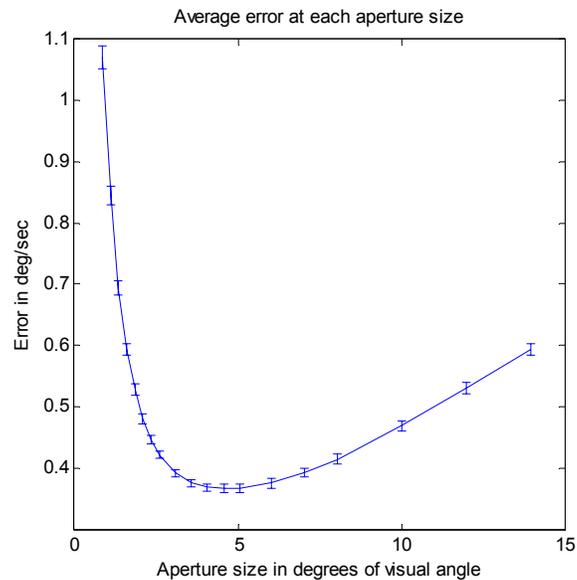


Fig. 2. If we look at the average error for each aperture size across all samples, we see that the function is U-shaped and has a clear minimum. Error bars show standard error. The plot is for a single velocity bin for the low density cylinder scenes with a walking speed of 3 ft/sec. (The error amplitude is the length of the difference vector obtained by subtracting the estimated velocity vector from the ground truth velocity vector.)

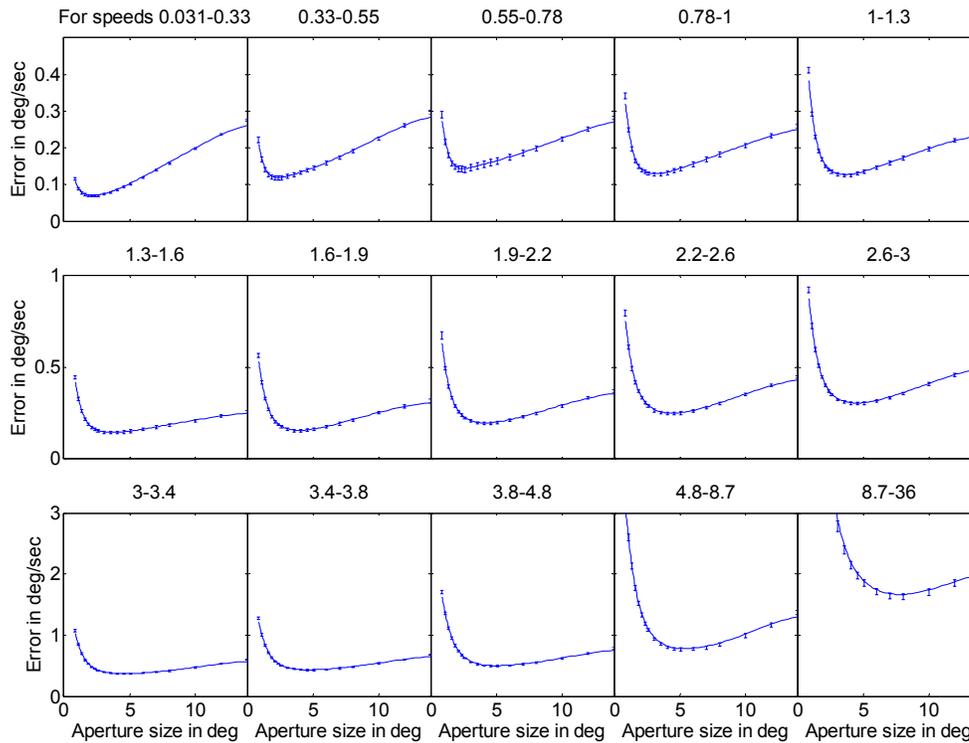


Fig. 3. Average error as a function of aperture size for different speed bins. The speed range for each bin in deg/sec is indicated at the top of each plot. These plots are from the low density cylinder scenes with a walking speed of 3 ft/sec (the plot in Fig. 2 is the first plot in the third row). The minimum of each curve is the optimal aperture size for that speed bin.

3. Results

3.1. Forest scenes

As can be seen in Figure 4, ideal aperture size increases monotonically with local image speed. This is true for all four forest conditions: both walking speeds and both cylinder densities. When ideal aperture size is plotted as a function of log speed (Fig. 4c), it becomes apparent that ideal aperture size increases approximately linearly with log speed. The solid line in Figure 4c shows the best fitting straight line to the combined results: $aperture = 3.3 \cdot \log(speed) + 2.6$, where $aperture$ is in units of deg, $speed$ is in units of deg/sec.

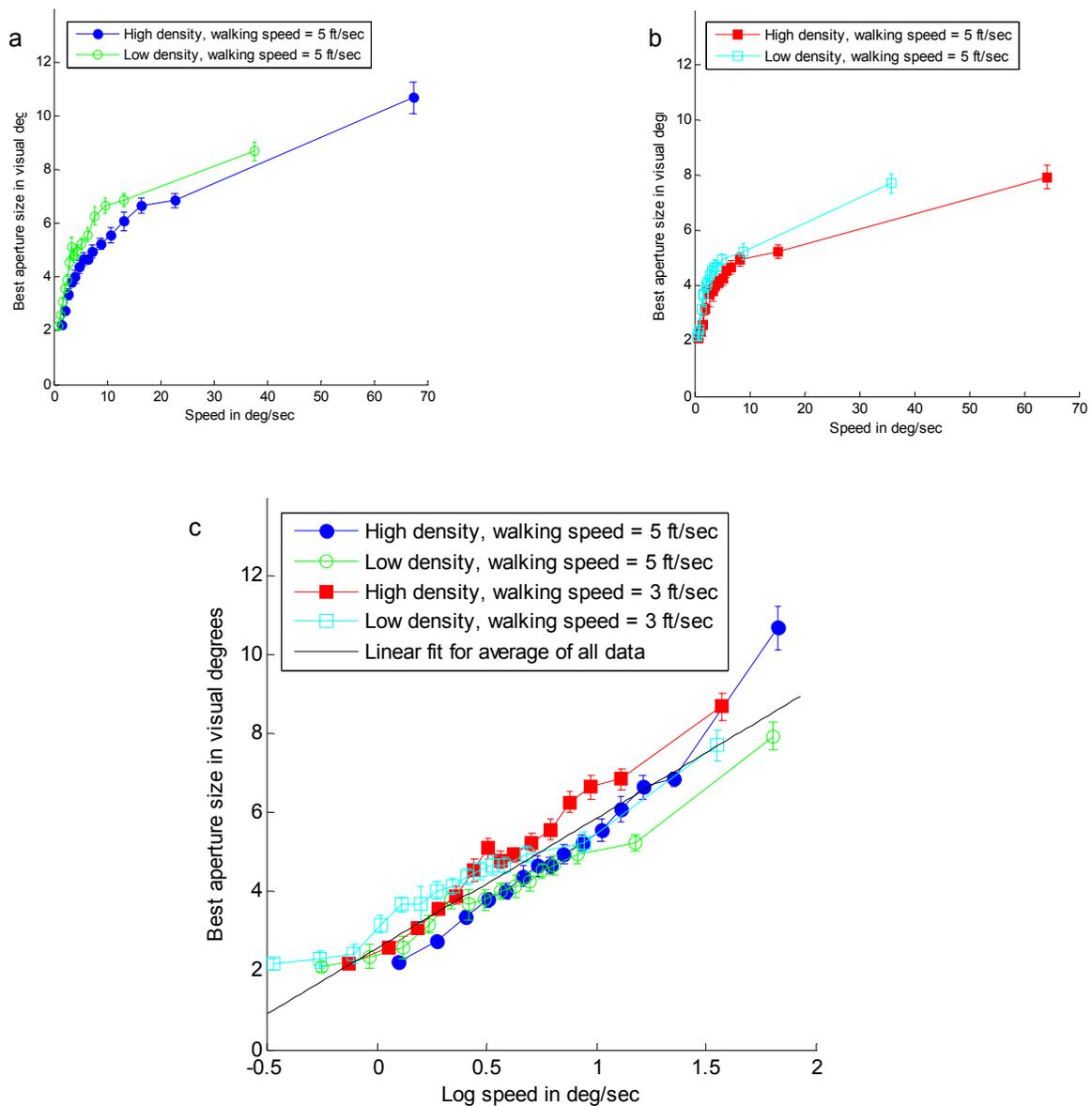


Fig. 4. Best aperture at each velocity at two different walking speeds. For forest scenes with a ground plane and (a) low density of cylinders, (b) high density of cylinders and (c) both low and high density as a function of log velocity. Error bars on all plots are the estimated standard deviation of the interpolated minima from plots of the type shown in Figure 3.

3.2. Flat wall scenes

For the flat wall scenes, the best aperture size is noisy and the optimal aperture sizes are large (Fig. 5a). The reason becomes obvious when one looks at the average error as a function of aperture size (Fig. 5b). For the flat wall scenes, the error function is not U-shaped, except perhaps for the slowest speed bins. Above a certain minimum aperture size, there is no significant decrease in accuracy. To understand this, consider the nature of optic flow in this environment. There are no motion borders in the scene (objects) and the optic flow changes slowly across the plane. Thus, even for a large aperture the average of the velocities will remain close to the velocity in the center of the aperture. The only

exception to this is when the aperture overlaps the focus of expansion. Motion tends to be slow near the focus of expansion and hence at slow speeds somewhat more U-shaped error functions are observed, with minima at small aperture sizes.

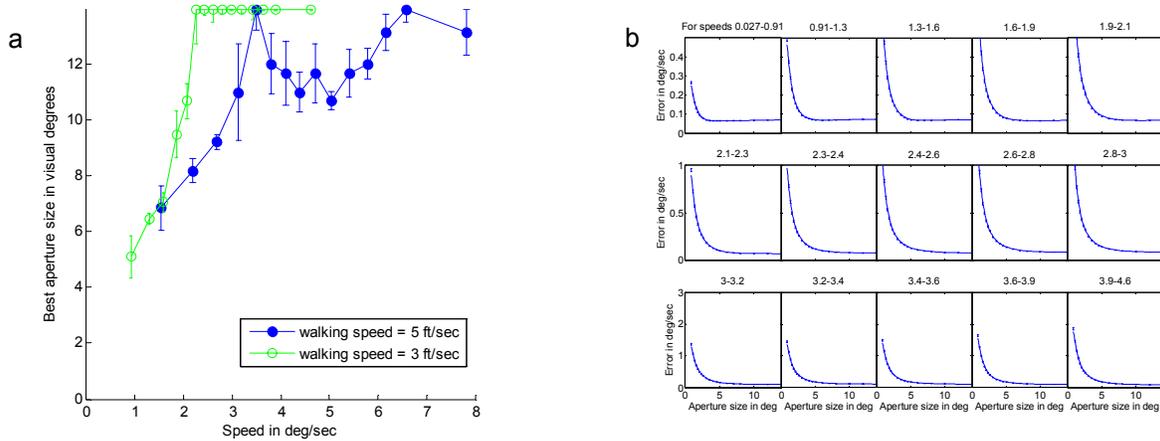


Fig. 5. (a) Best aperture size at each speed, for two different walking speeds toward a flat wall. Notice that in the 5 ft/sec condition the aperture size quickly reaches the maximum of 14 degrees. (b) Average error as a function of aperture size for different speed bins for the 3 ft/sec condition. For the flat wall condition the error function is flat above a certain aperture size. Plots in (b) were qualitatively similar for a walking speed of 5 ft/sec.

3.3. Ground plane scenes

For an environment consisting of only a flat ground plane (Fig. 6a), the ideal aperture sizes have an unusual peak at low speeds. There is no obvious explanation for why these particular slow velocities would require a larger area of summation, and hence we analyzed the velocity estimation errors in more detail. Our first analysis determined optimum aperture size separately for minimizing the difference in vector length and for minimizing the difference in vector angle (Fig. 6b). Our second analysis determined optimum aperture size separately for minimizing the difference in the horizontal (x) vector components and for minimizing the difference in the vertical (y) vector components (Fig. 6c). In all cases, there was a reliable peak at low speeds. Notice that these alternative error metrics are impractical because they each ignore a whole dimension of motion information in the environment. We used these error metrics solely to see if the peak at low speeds was isolated to a particular dimension of velocity.

Our third analysis was to bin errors across spatial locations in the scene, instead of across speed (Fig. 7a). For each spatial bin, we calculated the error as a function of aperture size (as in Fig. 2) and picked the aperture size that minimizes error. As can be seen, the source of the peak in aperture size at low speeds is due to a particular region in the environment (the bright pixels in Fig. 7a). Note, that for the ground plane environment, each image location has a specific velocity, at a particular walking speed. We verified that, in fact, the peaks at low speeds in Fig. 6a correspond to the speeds at the locations of the bright pixels.

Although the ray tracer is configured to perform anti-aliasing, it is possible that the peaks are due to some kind of aliasing artifact created in the rendering of the scene. To test this, we rendered movies at double the spatial and temporal resolution of the originals, and then smoothed and down-sampled the movies to ensure that there was no aliasing. The result is shown in Fig. 7b. Rather than removing the peak, it is accentuated. Apparently, additional smoothing, which removes high frequencies, expands the region requiring large aperture sizes.

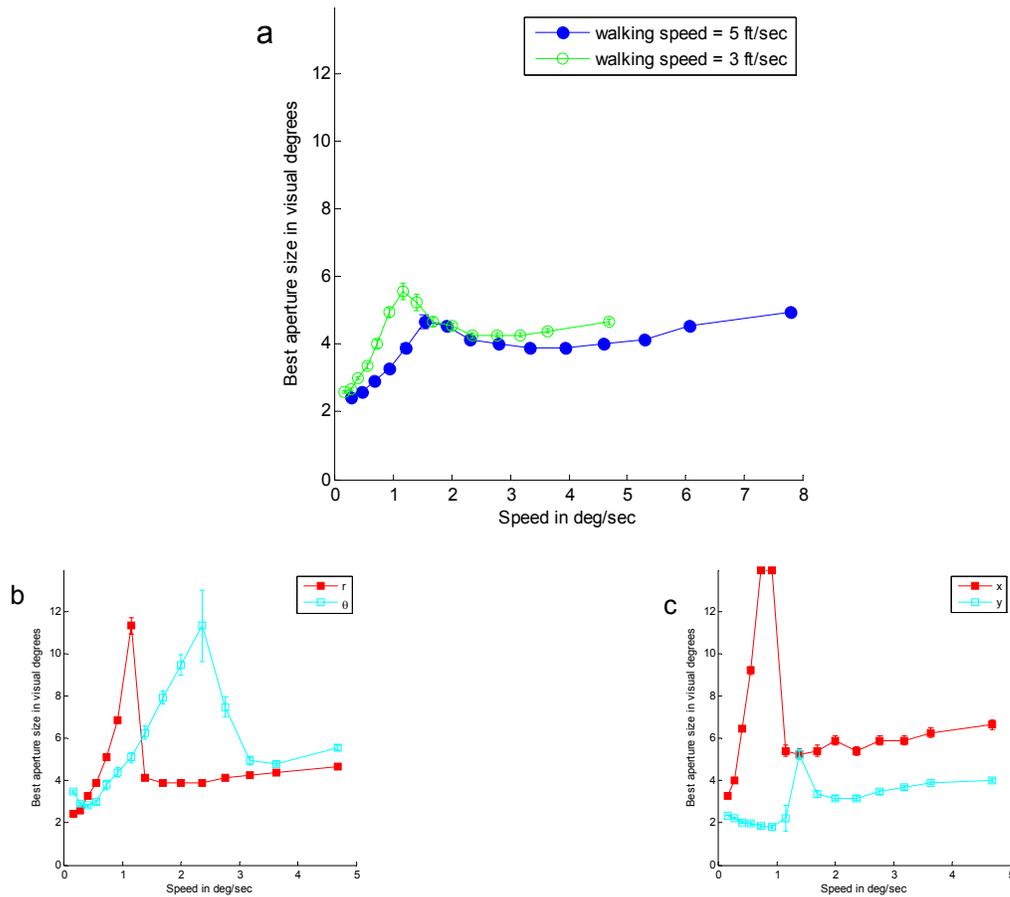


Fig. 6. (a) Best aperture size at each velocity for ground plane scenes. There is a distinctive rise and then fall in aperture size at low velocities that we do not see in our other scenes. (b) Best aperture size for minimizing average error in velocity magnitude (r) or orientation (θ) for a walking speed of 3 ft/sec. (c) Best aperture size for minimizing error in x-component and y-component of velocity for a walking speed of 3 ft/sec. The peak at low speeds is particularly pronounced when minimizing error in the horizontal vector component. Plots in (b) and (c) were qualitatively similar for a walking speed of 5 ft/sec.

Given this result and the fact that the locations of the bright pixels are far from the observer suggests that the unusual peaks in Fig. 6 are explained by spatial compression of the ground plane due to perspective. Such spatial compression in the real world and in a simulated world (if anti-aliased and sampled properly) shifts frequencies into higher bands where the limited spatial resolution of the visual systems filters them out. It is this lost spatial frequency content that allows accurate motion estimation within a small aperture. The regions in the image near the horizon do not require an increase in aperture size because at those distances there is almost no motion at all. The fact that there is a larger low-speed peak when minimizing error for the horizontal component of velocity (Fig. 6c) is likely due to the larger velocity gradient in the vertical direction than in the horizontal direction.

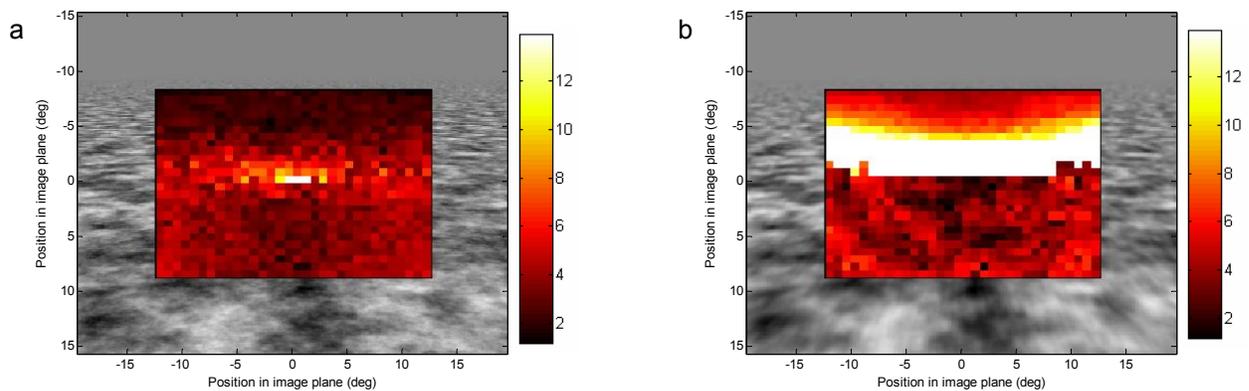


Fig. 7. (a) Best aperture size for minimizing error in estimated velocity at each spatial location for ground plane scenes. Brightness represents the best aperture size in degrees. The gray scale image surrounding the central plot is included to give a sense of the location of each bin in the image. The bright areas in the spatial map correspond to the peaks at low speeds in Fig. 6. (b) Best aperture size for minimizing error in estimated velocity at each location for scenes generated at twice the resolution and then filtered and down sampled to remove any possible aliasing. The gray scale image surrounding the central plot is from the filtered movie, and therefore has visibly less high frequency information. The increase in the size of the bright area indicates that the peaks are not due to aliasing.

4. Discussion

The aim of this study was to determine how local motion detectors should be designed to optimally code the local image motion that occurs during translation through the natural environment. More specifically, we asked how the integration area (aperture size) should vary as a function of image speed. To do this we used a ray tracer to generate movies from the perspective of an observer translating at two different walking speeds through four different classes of model world that were textured with $1/f$ noise. Two of the classes of model world were based on the measured statistics of natural forest scenes. The other two classes of model world were simple textured planes lying either below the observer, as a ground surface, or frontal-parallel, as if the observer were walking towards a textured wall.

Across almost all scenes and walking speeds, we find that the ideal aperture size increases monotonically with speed. For the ground plane scenes there is a small peak in the optimal aperture size at slow speeds that is likely an effect of spatial compression of the ground plane due to perspective. For the frontal-parallel plane scenes there is no ideal aperture size except at the slowest speeds; rather there is a minimum aperture size beyond which there are negligible changes in accuracy. Most importantly, for the forest scenes we find that the ideal aperture size increases approximately linearly with log speed for all four scene conditions.

How general are these results? Consider, in comparison to our artificial statistics the best possible real world measurements. Ideal measurements would include the positions and orientations of the eye of observers as they walk through real forest environments, as well as the calibrated image information and range data from the observers' point of view. Thus, there are two obvious inaccuracies in our model. First, it ignores the complex eye movements humans make when walking. Second, real forests have leaves, branches, grass and other objects of varying size and complexity that create additional motion boundaries. However, given that our findings are robust to changes in walking speed and density of objects in the scene, it seems unlikely that more detailed scene data or tracking of eye movements would change our qualitative finding that optimal integration area increases monotonically with speed.

Given our findings one might expect that the receptive field sizes of speed-tuned neurons in cortex should increase with the speed to which the neurons are tuned. There is evidence that the preferred speed of neurons in area MT increases with retinal eccentricity [14]. Given that receptive field size tends to increase with retinal eccentricity this result is consistent with our findings. However, a more definitive test would be to determine whether there is a positive

correlation between preferred speed and receptive field size at the same eccentricity, but we do not know of such a study.

Our results may also have implications for computer and robotic vision. Machine vision algorithms that use local translational motion detectors, VLSI vision systems for example, might benefit from altering the integration area of their motion detectors based on their speed tuning. Also, robots that move through man-made flat environments may benefit from having their motion detectors designed in order to compensate for the effects of spatial compression of the ground plane due to perspective.

The question we have addressed in this paper is quite specific. There are two basic routes for further research: expanding and building on the world model, or exploring further questions about the optimal design of sensors. These two research paths are interdependent since studying additional aspects of optimal sensor design might necessitate adding detail to the world model. For example, instead of studying the ideal area of spatial integration, we might investigate the optimal temporal integration time of motion detectors. Just as the ideal spatial area of integration depends largely on the statistics of the spatial derivatives of motion in the scene, we might expect that the ideal temporal integration time depends on the temporal derivatives of motion experienced by the observer. To study this question we would need detailed information about eye gaze durations during locomotion.

This paper shows that the statistics of the environment clearly do have an effect on the design of ideal motion sensors. In retrospect it might seem obvious that the ideal integration area of a motion sensor should increase with speed. However, it is impossible to know for certain if this is true without examining environmental statistics in detail. This is one example of why it is important to study the statistics of the natural environment relevant to a particular perceptual task. This paper also points out that, because of the difficulty of making the necessary statistical measurements in the natural world, computer simulation based on measured statistics can be a powerful tool. Computing power has reached the point where ray tracers can create virtually photo-realistic images with ease. The incredible computing power at our disposal allows us to simulate statistics that might be very difficult or impossible to measure in the natural world.

5. Conclusion

We found that over a range of simulated environments, the ideal aperture size for local motion detection increases monotonically with speed. For simulated forest environments, the ideal aperture size of motion detectors increases linearly with log speed. This result makes predictions for cortical neurons involved in heading perception. Also, these results have implications for the design of computer and robotic motion detectors. This work demonstrates that the statistics of the environment have an influence on the optimal design of motion sensors. Finally, simulating scene statistics is a promising approach for generating detailed statistics that might otherwise be difficult or impossible to gather.

ACKNOWLEDGEMENTS

Supported by NIH grant EY11747. Direct correspondence to TT (tal@cs.utexas.edu) or WSG (geisler@psy.utexas.edu).

REFERENCES

1. D. W. Dong, and J. J. Atick, "Statistics of natural time-varying images," *Network: Computation in Neural Systems* 6(3), 345-358 (1995).
2. R. O. Dror, D. C. O'Carroll, and S. B. Laughlin, "The Role of Natural Image Statistics in Biological Motion Estimation," *Proceedings of the First IEEE International Workshop on Biologically Motivated Computer Vision*, 492-501 (2000).
3. J. H. van Hateren, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proceedings of the Royal Society B: Biological Sciences* 265(1412), 2315-2315 (1998).

4. E. P. Simoncelli, and B. A. Olshausen, "Natural Image Statistics and Neural Representation," *Annual Review of Neuroscience* 24, 1193-1216 (2001).
5. J. M. Zanker, and J. Zeil, "An Analysis of the Motion Signal Distributions Emerging from Locomotion through a Natural Environment," *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, 146-156 (2002).
6. J. Huang, A. B. Lee, and D. Mumford, "Statistics of range images," in *IEEE Conference on Computer Vision and Pattern Recognition*, (2000), pp. 324-331.
7. D. Calow, N. Kruger, F. Worgotter, and M. Lappe, "Statistics of optic flow for self-motion through natural scenes," in *Dynamic Perception*, U. J. Ilg, H. H. Bülthoff, and H. A. Mallot, eds. (IOS Press, 2004), pp. 133–138.
8. S. Roth, and M. J. Black, "On the Spatial Statistics of Optical Flow," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, (2005), pp. 42-49.
9. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A: Optics, Image Science, and Vision* 4(12), 2379-2394 (1987).
10. MegaPOV, <http://megapov.inetart.net/>
11. POV-ray, <http://www.povray.org/>
12. K. Perlin, "Improving noise," *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 681-682 (2002).
13. J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *European Conference on Computer Vision*, (1992), pp. 237-252.
14. J. H. Maunsell, and D. C. Van Essen, "Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation," *Journal of Neurophysiology* 49(5), 1127-1147 (1983).