Local Analysis of Visual Motion - Review

Omer Cohen Sidon
omercs123@gmail.com

Abstract

This world is ever changing. Everything is constantly moving, and so to survive in this world, biological life forms, including humans, had to develop a way to estimate the motion that is all around them so that they can survive in this world. Motion is perceived by us humans using our eyes: When we move, or when objects in the world move, the visual image projected onto our retinas changes accordingly. This movement is processed by the brain in a way that allows motion estimation. In this work, we review a paper called "Local Analysis of Visual Motion" (Simoncelli, 2003) that presents a model for the estimation of movement based on a few basic assumptions. We first explain the mathematical model and then make comparisons between the model and biological mechanisms that reside in the mammalian brain, allowing it to process and estimate motion. In the end of this article we also criticize the model and in the appendix we add a link to a MATLAB implementation of the original paper's model.

1 Introduction

Images are basically two-dimensional projections of the three-dimensional world. The surface of the image could be several things, including an array of light sensors in the television camera, or the photo-receptors in the back of our eyes. The amount of light that fell on a spatial position in a particular time is measured by the image brightness. Movement in the three-dimensional world is projected into the two-dimensional image. The movement of the projected position of each point in the world is referred to as the motion field.

Machine vision systems' first goal is usually to estimate the motion field. Evidently, such estimation is also done by biological systems. The motion field must be estimated from the spatiotemporal pattern of image brightness. This is usually done by assuming that the brightness of objects in the real world remains constant over time. In this case, the estimation of constant-brightness points, called the "optical flow" is the same as estimation of motion field.

When estimating an optical flow, we cannot talk about the motion of just one point without considering its environment. However, by doing so, we might not get a unique solution for the motion. To demonstrate that, in figure 1 we can see a cube moving to the right. The cube has three visible sides, each with a different symmetry, and the graphs near the cube show us the best estimation we can get from looking through a small aperture. For example, in the case of the upper side we can see a constant brightness region (a "blank wall"). If we look through the aperture while the cube is moving slightly to the right, we have no way of measuring the motion of the cube (the velocity is under constrained) and so we have maximal ambiguity. Another case is the case of symmetry along one axis (the "aperture problem"). In this case, motion can be witnessed only in the direction perpendicular to the symmetry, and so if the cube moves to the right, there is an ambiguity regarding the velocity of the cube, as can be seen in the corresponding graph in figure 1. Lastly, in the third case, which can be seen on the side of the cube that does not have symmetry along any axis, a cube movement to the right can be easily detected and perfectly estimated with no ambiguity.

In addition to the cases of a blank wall and the aperture problem, there are other cases in which the velocity cannot be determined. For example, there could be locations in the image where there is more then one velocity. In par-

ticular, this can occur at occlusion boundaries of objects, where the object and the background are both present in any spatial neighborhood that we consider and they have different velocities typically. We will address these issues later, however for the most of this paper, we will restrict our attention to the simple case of one translational motion over a local patch of the image.

2 Related Work

The psychologist J.J. Gibson was the first to note that important environmental information was embedded in the pattern of local retinal image velocities (Gibson, 1950). He initialized a scientific quest to understand the mechanisms that might serve to estimate and represent these velocities. Years later, (Marr and Poggio, 1976) developed a philosophy according to which the key to understanding visual motion processing in the brain, is understanding the nature of motion information embedded in the visual world, and the problems that may arise when one attempts to extract this information.

The works of (Horn et al., 1986; Verri and Poggio, 1989) discussed the problem of using the optical flow as an estimation of the motion field. The subject of ambiguity in estimation of motion due to image symmetries was covered in the works of (Wallach, 1935; Fennema and Thompson, 1979; Marr and Ullman, 1981).

Solutions to the motion estimation problem that include more than one velocity have focused on decomposing the image into consistently moving layers of brightness content, and estimating the motion within those layers (Wang and Adelson, 1994; Darrell and Pentland, 1995). Some authors even suggested that this solution might be implemented biologically (Darrell and Simoncelli, 1993; Nowlan and Sejnowski, 1995).

The model that is presented in this paper should be treated as a building block, which could be further combined with a more complete solution for the analysis of motion. We shall now present the computational model and try to interpret its components physiologically.

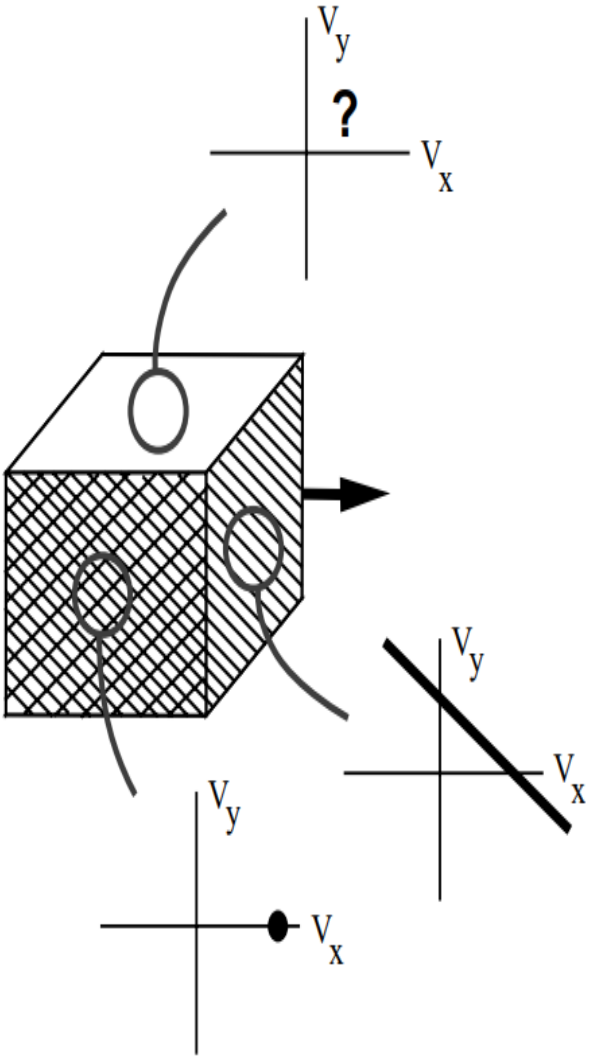

Figure 1: Conceptual illustration of motion estimation in three different regions of an image of a horizontally translating cube. In a region of constant brightness (top face of the cube), the local velocity is completely unconstrained. We refer to this as the blank wall problem. In a region where the brightness varies only along a unique spatial direction (striped side of the cube), the brightness changes are consistent with a one-dimensional set of velocities: one can determine the motion perpendicular to the stripes, but not the motion along the stripes. This is known as the aperture problem. Finally, in a region where the brightness changes in all spatial directions (hatched side of the cube), a unique velocity is consistent with the observed brightness changes.

3 Model

The problem of motion estimation is formulated here as an estimation problem. More specifically, we use Bayesian analysis to estimate the velocities in both axes (x and y). This means that given $I(x, y, t)$, which are the brightness measurements, we try to estimate $v = (u, w)$ which is the velocity vector. Using the Bayesian framework, we actually try to estimate $P(u, w|I)$. To do so, we use Bayes' theorem for the posterior distribution:

$$P(u, w|I) \propto P(I|u, w)P(u, w) \tag{1}$$

3.1 Calculating the prior distribution

Calculation of the prior distribution, $P(u, w)$, requires us to make the following assumption: Lower velocities are preferred to higher velocities. For example, using this assumption, if we enter a dark room, we will assume no movement at all. More generally, we solve ambiguities, i.e. situations when more than one velocity is possible, by choosing the velocity that has the lowest magnitude. To model this, we use normal distribution centered at zero:

$$P(u, w) = \exp\left[-\frac{u^2 + w^2}{2\sigma_v^2}\right] \tag{2}$$

Where σ_v indicates the standard deviation of the velocity distribution.

3.2 Calculating $P(I|u, w)$

Calculation of $P(I|u, w)$ requires us to use the brightness constancy assumption:

$$I(x + u\Delta x, y + w\Delta y, t + \Delta t) = I(x, y, t) \tag{3}$$

Which, using Taylor series expansion up to first order, can be replaced with a differential version:

$$I_x u + I_y w + I_t = 0$$

Where (I_x, I_y) are the spatial derivatives and I_t the temporal derivative of the image brightness. Next, we assume that the derivative measurements are corrupted by a small amount of noise, which is definitely true. A choice of computational convenience would be to assume that the noise is additive Gaussian, denoted by n, Which means that:

$$I_x u + I_y w + I_t = n$$

and so:

$$P(I|u, w) = \exp\left[-\frac{(I_x u + I_y w + I_t)^2}{2\sigma_n^2}\right]$$

Where σ_n indicates the standard deviation of the noise variable n. Another thing to consider here is that using this model we only try to estimate one velocity for the entire image, so we should sum over the gradient measurements. Assuming also that the noise variables are independent we get:

$$P(I|u, w) = \exp\left[-\sum \frac{(I_x u + I_y w + I_t)^2}{2\sigma_n^2}\right] \tag{4}$$

3.3 Calculating the posterior

Now that we calculated both $P(I|u, w)$ and $P(u, w)$ (the prior), we can use Bayes' theorem and calculate the posterior probability, which gives us the distribution over the velocities (u, w), given the input image brightness gradients (I_x, I_y, I_t):

$$P(I|u, w) \propto \exp\left[-\frac{u^2 + w^2}{2\sigma_v^2} - \sum \frac{(I_x u + I_y w + I_t)^2}{2\sigma_n^2}\right] \tag{5}$$

4 Experiments

Using MATLAB, we can examine the behavior of our model. To do so, we create several images, each with a different symmetry, and simulate a movement of the image using a specific velocity (magnitude and direction). After that, we run the Bayesian analysis to see if the suggested model can successfully estimate the true velocity. We examine a few types of images:

- A "blank wall" image (all-0 image)
- Image that contains only vertical lines
- Image that contains only horizontal lines
- Image that contains only diagonal lines
- A random image

We also examine two types of velocities:

- $v = (u, w) = (30, 0)\frac{pixels}{second}$
- $v = (u, w) = (0, 30)\frac{pixels}{second}$

Looking at the results of these experiments, which are shown in figure 2, we can see that the estimation of the motion depends critically on the input image, or more specifically, on the symmetry in the input image.

In the case of the "blank wall" image, it does not matter if the motion is pointed right or down, the estimation has no preferred direction. It is centered at zero and decays as the magnitude of the velocity rises. This can be understood, considering the fact that a "blank wall" moving is still a "blank wall", so the movement does not give us any information.

In the case of the "vertical" image, due to symmetry along the vertical axis, it was impossible to detect any vertical motion, so estimation is only successful when the image moved to the right, and anyway the estimation along the vertical axis preferred velocities with smaller absolute value. Analogously, in the image with the horizontal symmetry we were able to estimate correctly only vertical motion.

The case of the "diagonal" image is actually very similar to the horizontal and vertical cases, as it has symmetry along one axis, and so estimation was done correctly only for motion in the axis perpendicular to the symmetry. In this case, motion to the right or down was estimated in the same way, i.e., we could only determine that the motion was going more right-down than left-up.

Finally, **the case of a "random" image**, meaning an image with no symmetry, is rather convenient for motion estimation and as can be clearly seen in figure 2, the motion was estimated quite successfully.

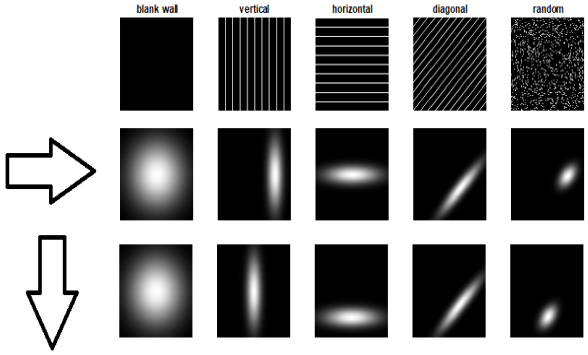

Figure 2: Results of motion estimation. Original images appear in the first row, the second row shows the motion estimation for the case of $v = (u, w) = (30, 0)$ pixels/second, which indicates movement to the right. Horizontal axis from left to right is an estimation of u and vertical axis from up to down is an estimation of w (both range from minus 50 to plus 50 pixels per second). The third row is like the second row, only this time the image moves in the direction of $v = (0, 30)$ pixels/second

5 Discussion

In this section, we will describe the relationship between the elements of the model's framework and the functional properties of neurons that lie in the so-called motion pathway of mammals such as cats or monkeys. This should not be though of as a quantitative physiological model of motion processing, but as more qualitative assignment of computational function to neural populations along the motion pathway.

Starting with equation 5, if we hold the velocity (u, w) fixed, then this function is tuned for the velocity of the input image. Meaning, it is maximized when the underlying image structure is consistent with the velocity (u, w), and it decreases as the image motion deviates from that velocity. We may identify this property with those neurons in the visual area MT that are known as "pattern selective" (Adelson and Movshon, 1982). These neurons respond vigorously to a visual stimulus moving with a particular speed and direction, and are relatively indifferent to the stimulus' spatial pattern.

We can make qualitative comparison of this model to the typical neural response in area MT. The lower left panel of figure 3 shows polar plots of the response of a model MT neuron with preferred velocity $u = 0.3, w = 0$ pixels per frame, as a function of the normal direction of a drifting sinusoidal grating. As with MT pattern cells, the response is tuned for the direction of motion, and reaches its maximum when this direction matches the preferred direction. Figure 4 shows the posterior function plotted as a function of stimulus speed, for a drifting sinusoidal grating. The maximum response occurs at 0.3 pixels per frame, the speed corresponding to the chosen u.

Having identified the posterior distributed with the MT pattern responses, we now wish to identify the components from which those responses are generated. The nominator of the exponent of

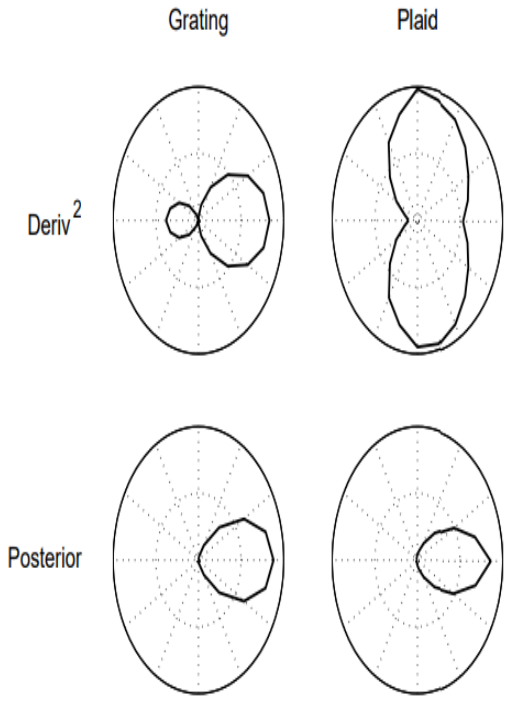

Figure 3: Direction tuning curves for model components. Upper left: Model V1 neuron (squared directional derivative) response to drifting sinusoidal grating; Upper right: Model V1 neuron response to drifting sinusoidal plaid; Lower left: Model MT neuron (posterior probability) response to grating; Lower right: Model MT neuron response to plaid.

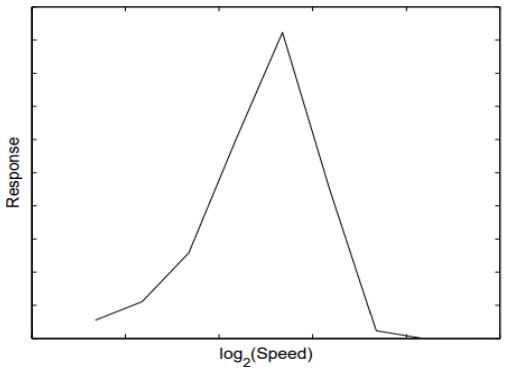

Figure 4: Speed tuning curve for posterior distribution (model MT neuron) to a drifting sinusoidal grating. The absolute scale of the horizontal axis is arbitrary, but the tick marks correspond to increments of speed by multiples of 2.

equation 5 may be expanded, up to a factor of two, as follows:

$$f(I|u,w) = \sum \left[u^2 I_x^2 + 2uwI_xI_y + w^2 I_y^2 + 2uI_xI_t + 2wI_yI_t + It^2\right] \quad (6)$$

where the sum is taken over a small spatial neighborhood. Many of the terms contain squared image derivatives, and those that contain products of image derivatives may be written as a difference of two squared derivatives. For example:

$$\begin{aligned} I_xI_y &= \left[(I_x + I_y)^2 - (I_x - I_y)^2\right]/4 \\ &= \left[I_{d1}^2 - I_{d2}^2\right]/4 \end{aligned} \quad (7)$$

where I_{d1} and I_{d2} are derivatives at angles ± 45 degrees. An analogous transformation allows us to write the term I_xI_t as a difference of two derivatives. Since one of the axes is now time, the result is difference of space-time oriented derivatives, $\{I_r, I_l\}$ which are most responsive to vertically oriented structures that are moving horizontally. Similarly, the product I_yI_t results in a difference of squared upward and downward derivatives, $\{I_u, I_d\}$. Combining all of this gives us:

$$f(I|u,w) = \sum \left[u^2(I_x^2 + \sigma_n^2/\sigma_v^2) + uw(I_{d1}^2 - I_{d2}^2)/2 + w^2(I_y^2 + \sigma_n^2/\sigma_v^2) + u(I_r^2 - I_l^2)/2 + w(I_u^2 - I_d^2)/2 + I_t^2\right] \quad (8)$$

The purpose of this representation is to show that our model contains terms that could arise in the responses of primary visual cortical neurons. The receptive field of so-called simple cells in primary visual cortex (area V1) of cats and monkeys are selective for stimulus position and orientation (Hubel and Wiesel, 1962). Many simple cells are also direction-selective: they give stronger responses for stimuli moving in one direction than the opposite direction. This type of model can explain the primary properties of these cells, including selectivity for stimulus orientation, spatial frequency, temporal frequency and direction.

Another point of comparison between our model and the bioligocal system, are the linear derivative filters used in our Bayesian model.

They bear some similarities to the response properties of simple cells: they are tuned for spatial orientation, spatial frequency, temporal frequency and direction. Example direction tuning curves of squared derivative operators for both gratings and plaids are shown in the top row of figure 3. Note that, as in V1 neurons, the response to the plaid is bimodal: unlike the posterior responses shown on the bottom row, the operator responds to each of the components of the plaid rather than the pattern as a whole.

6 Conclusion

Simoncelli in his model provided a Bayesian analysis for local motion based on a minimal set of assumptions:

- brightness conservation.
- a simple model of measurement noise.
- a prior preference for slower speeds.

Given these premises, the components of the optimal solution can be seen to have properties matching those of neurons in the pathways of mammals (cats and monkeys), at least qualitatively.

7 Future Work and criticism

In my opinion, the main drawback of Simoncelli's paper is that he does not try hard enough to justify his assumptions. For example, The Bayesian model is probably to simplistic to describe the motion estimation. A more correct model should include a more realistic model of uncertainty in photoreceptors, as well as in the subsequent neural responses. The prior assumption is also too simplistic, maybe instead of just assuming that slower velocities are more common, he could try to develop a prior distribution based on actual experiments that show which velocities are more common in different environments. Obviously, common velocities in a jungle would be different than common velocities in the dessert.

Finally, as Simoncelli mentioned himself in his paper, the formulation of motion estimation problem using brightness constancy assumption is simplistic in assuming that a single velocity accounts for the motion in each local region. The problem of many velocities in one place (e.g. near occlusion boundaries and in presence of semi-transparent surfaces) is too common to overlook. Maybe some notion of segmentation can be introduced to deal with this problem. For example, in the case of an object moving behind a window: The mammalian eye sees the object moves behind the partly-transparent window and the window does not move, however, the mammalian mind can understand that the window and the object behind it are completely different items (segmentation), and so perhaps it can process their velocities separately.

References

Edward H Adelson and J Anthony Movshon. 1982. Phenomenal coherence of moving visual patterns. Nature, *300(5892):523–525.*

Trevor Darrell and Alex P Pentland. 1995. Cooperative robust estimation using layers of support. IEEE Transactions on Pattern Analysis and Machine Intelligence, *17(5):474–487.*

Trevor Darrell and Eero Peter Simoncelli. 1993. Separation of transparent motion into layers using velocity-tuned mechanisms. *Citeseer.*

Claude L Fennema and William B Thompson. 1979. Velocity determination in scenes containing several moving objects. Computer graphics and image processing, *9(4):301–315.*

James J Gibson. 1950. The perception of the visual world.

Berthold Horn, Berthold Klaus, and Paul Horn. 1986. Robot vision. *MIT press.*

David H Hubel and Torsten N Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, *160(1):106–154.*

David Marr and Tomaso Poggio. 1976. From understanding computation to understanding neural circuitry.

David Marr and Shimon Ullman. 1981. Directional selectivity and its use in early visual processing. Proceedings of the Royal Society of London. Series B. Biological Sciences, *211(1183):151–180.*

Steven J Nowlan and Terrence J Sejnowski. 1995. A selection model for motion processing in area mt of primates. Journal of Neuroscience, *15(2):1195–1214.*

Eero P Simoncelli. 2003. Local analysis of visual motion. The visual neurosciences, *109:1616–1623.*

Alessandro Verri and Tomaso Poggio. 1989. Motion field and optical flow: Qualitative properties. IEEE Transactions on pattern analysis and machine intelligence*, 11(5):490–498.*

Hans Wallach. 1935. Über visuell wahrgenommene bewegungsrichtung. Psychologische Forschung*, 20(1):325–380.*

John YA Wang and Edward H Adelson. 1994. Representing moving images with layers. IEEE transactions on image processing*, 3(5):625–638.*

A Appendices

A MATLAB simulation of the model presented in this paper was submitted through e-mail but can also be found in the github repository: https://github.com/omercs123/simoncelli