Motion Segmentation:

A Biologically Inspired Bayesian Approach

Greg Corrado
gcorrado@stanford.edu

Motion Segmentation is the attribution of motion information to elements in a visual scene. Here we propose a motion segmentation algorithm that uses the properties of motion sensitive neurons as inspiration for its fundamental units. Each unit is treated as a local Bayesian estimator embedded in a larger belief network. The heuristics that we employ to drive segmentation in this network is that there should be at most one motion at each point in space.

Introduction:

Motion is one of the best understood aspects of visual processing in the brain. We know a lot about how motion is perceived by biological visual systems [Adelson & Movshon 1982]. We know a surprising amount about how motion information is represented in the brain [Albright 1993]. We even have some good ideas about how biological systems might extract motion information from the visual scene [Adelson & Bergen 1985, Simoncelli 1993]. But we know shamefully little about how motion information is combined across space and time to construct a "motion scene."

To interpret the visual world, any system must parse the continuous stream of sensory input into distinct salient elements. This is true in the simplest case of foreground background segregation all the way through the difficult task of object recognition. It is no different in the realm of motion vision. To estimate the motion of objects well, information must be combined across space and time - but not across sources. Motion elements must be segmented from each other and from the motion of the background to allow a system to make intelligent judgments about its environment. But how? Can we use what we know about biological vision to construct an algorithm?

A particularly compelling example of the need for this algorithm is the case of motion transparency. Motion transparency occurs whenever two motions appear in the same region of space. This might happen in a reflection on a pane of glass, or in the translating specularities of a moving reflective object, or in the jungle as the shadows of leave shimmer across the form of an approaching predator. In these cases it is the motion itself

that allows us to separate one object from another. Because this is such a crisp example the challenges of motion segmentation, we will use a synthetic example of motion transparency to explore our algorithm. In addition we will use human perception of this example as inspiration for how to construct our algorithm.

Broadly, our approach will be to construct an interconnected network of local motion processors. The characteristics of each processing unit will be taken very directly from known properties of neurons in are MT of the primate visual cortex – the apparent seat of motion processing in the human brain. We will interpret the output of these unit as a Bayesian estimate of the probability of local motion given the images, and use belief propagation to reason about the motion scene.

The goal of this approach is really two fold. Not only might we construct a algorithm which successfully segments a motion scene, we might generate testable hypothesis for future neurophysiological studies of motion processing.

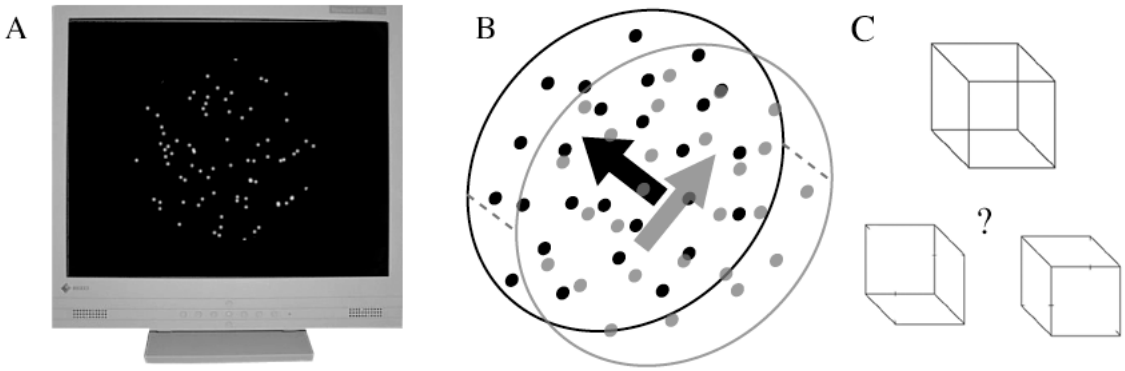
Introduction:

Consider the following synthetic motion transparency scene. We sparsely populate a region of an otherwise blank screen with high contrast Gaussian dots. [See Fig 1A.] At random we select half of the dots to drift with an arbitrary velocity – meaning that on each frame we displace each dot within this population by the same horizontal and vertical position. The dots move in this direction on each subsequent frame until they reach the end of the boundary of our region, at which time they are redrawn to appear at the opposite boundary. The other half of the dots, not yet accounted for drift in an different arbitrary velocity. [See Fig 1B.] We will call this motion presentation Transparent Random Moving Dots, or TRMD.

What is the percept human observers experience when viewing a TRMD? It is not a collection of dots, each moving independently. Rather, all the dots moving in a particular direction appear to grouped - intuitively belong to the same object. The two species of dots seem to coalesce into two distinct transparent textured sheets sliding across each other.

Interestingly, many observers report an apparent difference in the depth of these surfaces. The display is ambiguous however – there is no information in the visual scene that gives a cue as to which sheet of dots is in the foreground, and which is in the background. The inherent ambiguity of this display induces a bistabilty to this percept. In particular, though the apparent depth ordering of the dot planes may be stable for many seconds, it varies from presentation to presentation of an identical stimulus reminiscent of the famed Necker cube. [See Fig 1C.] We use this apparent crosstalk between motion processing and depth perception when viewing the TRMD as inspiration on how to segment transparent motion.

Figure 1



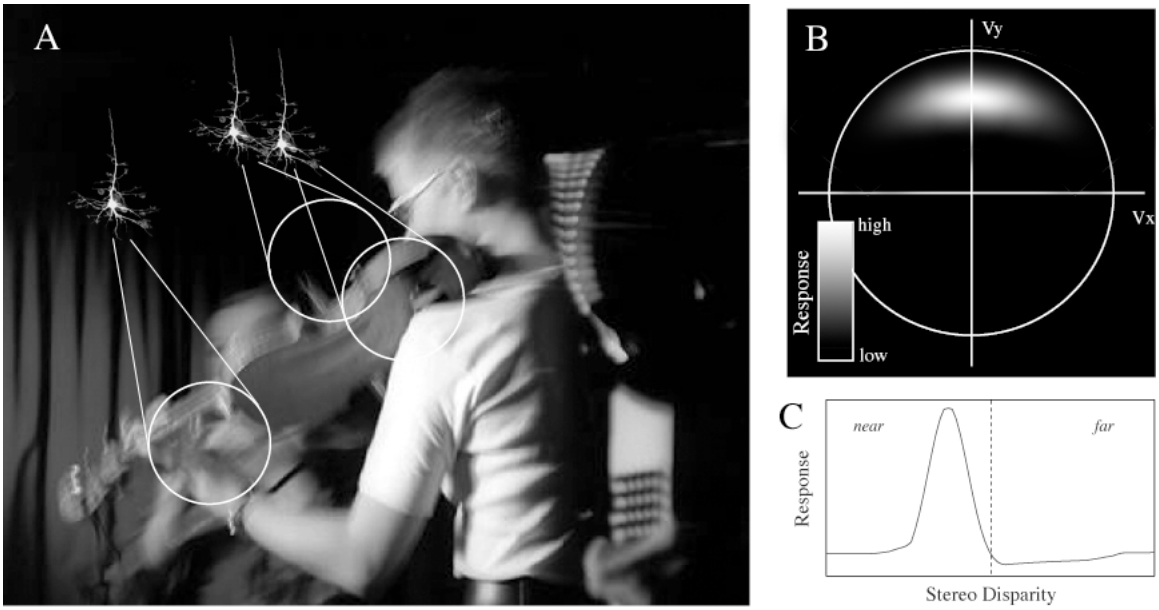
Now we turn our attention to some observations about motion processing in the brain in area MT. The neurons in area MT are some of the most well studied in the primate visual system, if not the entire brain. We can sketch some of their basic properties here to guide our thinking. Like many neurons in the visual system, they respond only to part of image, called the receptive field. These receptive fields tile the image in largely overlapping fashion, with neurons responding to adjacent receptive fields adjacent and most densely connected in the brain. [See Fig 2A.] Receptive fields tend not to have crisp boundaries, or perfectly symmetric shapes, but they can be well approximated as Gaussian masks through which the neuron views the world.

As a whole, we said that area MT processes visual motion - but what does that mean in terms of individual neurons? Each neuron appears to be a tuned nonlinear filter, responding most vigorously to a particular type of motion within its receptive field. In

simplest terms, each neuron has a preferred direction and speed of motion to which it responds regardless of spatial scale, image contrast, etc. Typically MT neurons give little or no response to motions that differ significantly from this preferred ideal, and so are often thought of as having a velocity tuning that is the product of Gaussians in radial coordinates. [See Fig 2B.]

A third interesting property of MT neurons is that a great many of them are sensitive to stereo vision cues. Most MT neurons respond strongly to motion with zero binocular disparity – that is to motion in the plane containing the point to which the eyes are verged. But many neurons respond better to motion in a band beyond (negative disparity) or before (positive disparity) this plane. [See Fig 2C.] While this intermingling of stereo and motion information is at first puzzling, we will make use of this to help solve the motion segmentation problem.

Figure 2



Approach:

If it is indeed the task of the motion vision system to estimate object motion, then the properties of physical objects (the ultimate source of visual input) provide important constraints to resolve the ambiguity in the visual scene. In the case of transparency, we offer the following supposition as a constraint: *there is at most one object, and thus one object motion, at any point in space*. Rarely, if ever, are there truly two object motions at a single location in a natural scene. Therefore, a motion system concerned with objects need only represent one motion at each point in space. A visual system with the capacity to represent several well-defined object motions per point would not only be wasteful, but critically, would tend to proliferate uncertainty into the most unlikely of scene interpretations.

If the *one-point-one-motion* conjecture is correct, we would predict that when a visual system *is* confronted with an artificial stimulus where there *are* two motions a single point, such as the TRMD, it must make a compromise. It must either ignore one of the motions entirely, or it must represent the two motions at different spatial locations. Given that the visual system can in principle extract motion and localize elements in angular space more confidently than it can localize elements in depth, the parsimonious choice would be to attribute the inconsistency to an error in depth judgment. The phenomenology of the TRMD depth illusion may be an example of just this. Were the human visual system subject to the one-point-one-motion constraint, we would expect it to represent the two motions of the TRMD at different depths despite there being no other evidence to support this. Thus, the one-point-one-motion conjecture offers a convenient explanation for the TRMD depth illusion, and will serve as the central principle directing our algothim.

To be specific, we propose to create an ensemble of artificial neurons with response properties similar to that of real MT neurons. We will construct these units so that as a population, these neurons compute a probability distribution over motion vectors given the local information available in their receptive fields. These Bayesian motions estimates will be based on a Gaussian model of image noise, combined with a prior for

smooth and slow motions. We will in turn use these local estimates as nodes in a Markov Random Field that combines information across space and attempts to relax conflict among the nodes. This process amounts to an implicit solution of the motion segmentation problem, which groups coherent motion elements and segregates disparate ones. The representational trick is to have incompatible (non-common-fate) motions repel each other in the 3rd dimension of binocular disparity, thus implementing the *one-point-one-motion* constraint. This can be done by constructing a single compatibility capturing this notion and then employing Loopy Belief Propagation to iterate toward a single stable segmentation of the motion scene.

As an input for our model we construct simple synthetic motion stimuli, like those used to study MT neurons. In particular, we made movies of the transparent random moving dot pattern described above. Each dot is white Gaussian blob on a black background, which drifts in one of two directions at a speed less than one standard deviation per frame (so that correspondence can be established trivial). Call each frame of the movie $I(x,y)$ and the resulting image sequence $I(x,y,t)$. We won't consider any depth information being present at the input – as if you were verging on the screen the movie was playing on and there was nothing else in the field of view. This means the disparity map associated with each frame is zero for all locations and all times, so we can simply write: $I(x,y,t,d) = I(x,y,t)$ for $d = 0$, and $I(x,y,t,d) = 0$ otherwise.

Each model neuron has five parameters: $\hat{x}$ and $\hat{y}$ the location of its receptive field in image coordinates, the two parameters of its preferred motion vector $\hat{v}$, and $\hat{d}$ its preferred stereo disparity. We can capture the receptive field sensitivity profile and the disparity tuning of this neuron with a three dimensional Gaussian mask,

$$I_{\hat{x}\hat{y}\hat{d}}(x,y,t,d) = \frac{N(\vec{\mu},\vec{\sigma})}{\max N(\vec{\mu},\vec{\sigma})} I(x,y,t,d) \quad \text{where} \quad \vec{\mu} = [x-\hat{x}, y-\hat{y}, d-\hat{d}].$$

If the neuron is a Bayesian estimator, then its response can be written as

$$R_{\hat{x}\hat{y}\hat{d}\hat{v}}(t) = P(\hat{v}) \prod_i P(I_{\hat{x}\hat{y}\hat{d}}(x_i, y_i, t, d_i) \mid \hat{v}),$$

where the product is over all points *i* where the mask is non-zero. Common assumptions are that the prior $P(\hat{v})$ is a symmetric Gaussian centered at zero – biasing us toward small velocity estimates, and that the likelihood function $P(I(x_i, y_i, t, d_i) \mid v)$ is of the form

$$P(I(x_i, y_i, t, d_i) \mid v) \propto \exp\left[-\frac{1}{2\sigma^2} \int w_i(x,y)\left(I_x v_x + I_y v_y + I_t\right)^2 dxdy\right]$$

where $w_i(x,y)$ is a small window centered around (x_i, y_i) and $I_k \equiv \frac{\partial}{\partial k} I(x,y,t,d)$. This likelihood function can be derived assuming smooth motion within $w_i(x,y)$, intensity constancy, and independent Gaussian image noise – though other likelihood functions can be arrived at with similarly reasonable assumptions. (See Weiss and Fleet 2002),

Up until this point, there is nothing new in what we are suggesting. It is almost completely a recapitulation of Weiss, Simoncelli, and Adelson 2002 – a paper which uses a model of Bayesian motion estimation to explain a number of puzzling human perceptual illusions. The novel part here is the *one-point-one-motion* constraint and the use of Markov Random Fields to propagate information across space.

Results:

Well of course I'm nothing close to being finished – probably not even half way to finished, but I'll describe what I've done so far.

The first step was generating the TRMD movies, and other simple motion stimuli to test algorithms on. Dots are placed uniformly and randomly in a circular aperture. Each dot is a small Gaussian blob with a peak at full range of the intensity scale. To prevent the motion from being jerky or having irregular 'shimmering' it is necessary to store dot locations at a 10x subpixel resolution. Displacements between frames are kept small enough that the I don't have to confront the correspondence problem or motion aliasing – the corresponding dot in the next frame is much closer than other dots. Ultimately it turned out that the best way to compute frames was to: (1) track dot locations as real numbers, (2) create an upsample matrix of zeros, (3) place ones (effectively delta functions) in the correct locations by rounding the real values, (4) convolve this sparse

matrix with a Gaussian kernel in x and y, (5) subsample back down to pixels, and then (6) finally renormalize. Other methods I tried were either vastly slower, produced motion shimmer, or had weird boundary effects if dots were randomly placed too close together.

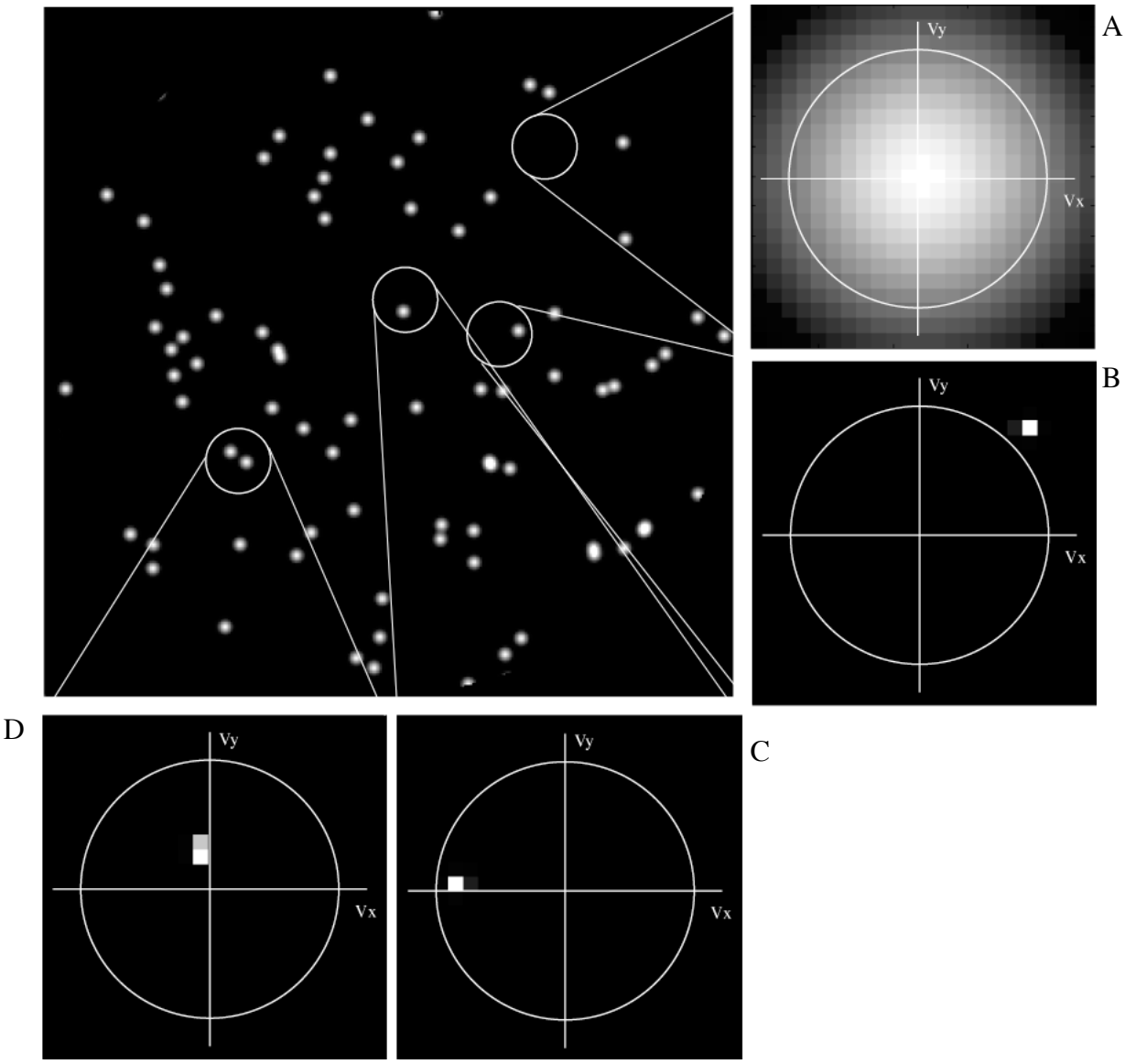
Using this method I have produced TRMD movies containing two and three motions. I have also produced movies with a single constant motion, a single motion that changed over time (spirals, jerks, and acceleration), and other motion displays such as drifting Gaussian bands which allowed be to explore susceptibility to the aperture problem.

The next major step has been to re-implement what others have done to generate Bayesian estimates of local motion. This has proven to be surprisingly computationally taxing – particularly in terms of storage. At this point I have debugged implementation of all the calculations described above up to the point of embedding them in the Markov Random Field.

A few notes on design decisions. I chose the sizes of the receptive fields such that they tend to contain 0, 1, occasionally 2, and only very rarely 3 dots. This decision will have a huge impact on how well my implementation ultimately works. I suspect that this parameter will need to be tuned carefully later in the process. I chose the size of the micro neighborhoods in the calculation of the likelihood function to be about the size of the dots so that they almost always contained at most one part of one dot. This is necessary if we are to meet the assumption that there is strictly one smooth motion within each micro neighborhood – which allowed us to write the likelihood in that form. The size of the image noise term was chosen arbitrarily and changing it does not seem to affect performance. I suspect this is because the synthetic images I'm are essentially noiseless. I'm planning on adding image noise in the future, and hope nothing breaks. The width of the Gaussian prior for slow motions was chosen to have a width comprable to the variance of the image velocities – is seems that it could be argued that this should be the optimal choice of prior which would naturally arise from long integrated experience in a world with my chose artificial image statistics. Also I chose to build in a characteristic scale rather than construct a full image pyramid – the computations are

slow as it is. Instead, to compute derivatives I used precomputed filter taps derivatives of derivatives of Gaussians, which seems to work well. Results are shown in Fig 3.

Figure 3



Next steps:

The next broad goal is obviously for me to marry these units in a Markov Random Field. But first I've been experimenting with just two nodes and compatibility functions relating the two of them. The aim is to identify the class of compatibility functions which allows units with unambiguous image views to propagate information to regions with

ambiguous information, and induce those notes to split their representation between disparities.

In particular what I plan on doing is using a unit with only one dot [e.g. Fig 3B] in its receptive field as an inducer node. Using messages passed by this node, I would like a node with two dots [e.g. Fig 3C] to re-evaluate its evidence as being most consistent with two planes of motion, one consistent with the inducer, and another in a different disparity plane. For testing purposes I'm trying to do this without feedback from the ambiguous node onto the inducer. I've tried a few things at this point, one of which almost works, but it'll take some more playing to really see.

The next miniature experiment would be to consider what compatibility function would allow a now disparity separated two dot node, to induce a matching disparity in our original inducer. My hope is that by conceptually breaking the problem down into these two distinct processes, I can sidestep a lot of the confusing and counterintuitive pitfalls of recurrent feedback.

If this approach fails to yield compelling results, there are a number of ready backup plans. One plan which we could fall back on, would be to directly use spectral analysis combined with k-means to clustering, or other similar analysis to segment the velocity measures. Another possibility I might consider would be to try using some form of anisotropic diffusion in velocity space, but that would require a lot of additional thought.

Summary:

The main thrust of the work so far has been first in implementing and understanding Bayesian motion estimation in a local field, and second in arriving at a plausible direction to take the algorithm in terms of segmenting the scene. The first step in segmenting must be to generate meaningful tokens, and we have done that.

References:

Adelson EH & Bergen JR. Spatiotemporal Energy Models for the Perception of Motion, Journal Optical Society of America A, 2(2):284-299 (1985).

Adelson EH & Movshon JA. Phenomenal coherence of moving visual patterns. Nature. 1982 Dec 9;300(5892):523-5.

Albright TD. Cortical processing of visual motion. Rev Oculomot Res. 1993;5:177-201.

Simoncelli EP. Distributed analysis and representation of visual motion. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge MA, January 1993.

Weiss Y & Fleet DJ. Velocity likelihoods in biological and machine vision in R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki (eds) Probabilistic Models of the Brain: Perception and Neural Function, MIT Press, 2002. pages 77-96

Weiss Y, Simoncelli EP & Adelson EH. Motion Illusions as Optimal Percepts. Nature Neuroscience June 2002 Volume 5 Number 6 pp 598 - 604