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 [1]. We know a surprising amount about how motion information is represented in the brain [2]. We even have some good ideas about how biological systems might extract motion information from the visual scene [3, 4]. 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, and integrate information within an element but not across 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?

Current work [5, 6, 7] has focused on two stage approaches. In the first stage, motion elements are grouped in a process of assignment. Then within each group object motion is estimated, often from a fixed set of rigid transformations. Various techniques based on masking, motion subtraction, or motion layer assignment appear in the literature with various degrees of success. While these algorithms can be computationally efficient, they are both biologically implausible and by their very nature fail to make the best use of available data. To make the best use of the data, we suggest and integrated probabilistic estimation of the motion of at each point in space. This approach does not explicitly limit the motion transformations (even allowing non-rigid motion extraction), or the number of separate object motions in the scene. Instead we will use priors of likely motions in real scenes (such as fast motions are less common than slow motions) and natural constraints on interrelationship between motion elements (such as there is rarely truly two different motions at exactly the same point in space).

Broadly, our approach has been to construct an interconnected network of local motion processors. Each of these units make a Bayesian estimate of the probability of local motion given the image data and our priors. Moreover, 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. Using belief propagation in a larger Markov Random Field (MRF) interconnecting these unit, we reason about the motion scene to group, segment, and integrate information. Thus our niche is a single integrated probabilistic graph approach to motion vision.

We find that this approach is able to recover a number of interesting human perceptual experiences, which more traditional techniques fail to capture. We find that our network is able to improve motion estimates by integrating information across space for elements with a common fate. Replicating earlier Bayesian motion work, we find our network estimates the motion of an ellipse as either rotation or deformation depending on its shape. Moreover, much like humans this percept can be drastically altered by the presence and relative motion of spatially remote cues. In an expended 2.5D MRF we find that this approach can successful extract multiple transparent motions and induce illusory separation in depth, much like human percept. Ultimately, this raises hope that a full 3D Markov Random Volume might be able to recover complex volume motion from sparse data much like humans percept.

Approach:

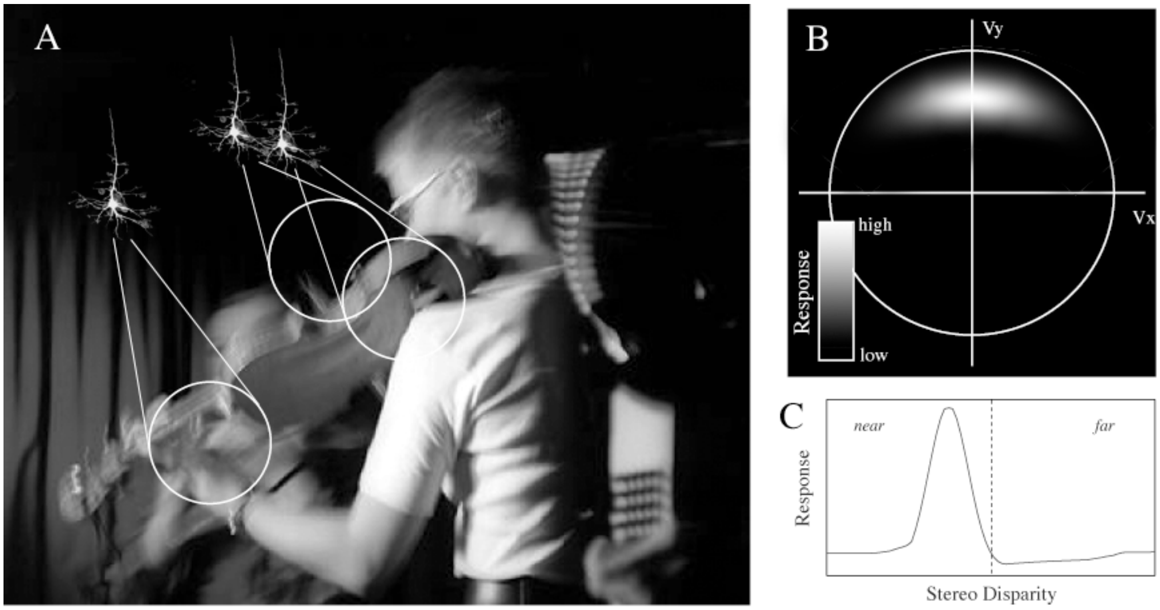
How is it that biological systems extract motion in a unified approach? We can answer this question by making some observations about motion processing in the brain in area MT. The neurons in area MT are some of the most well studied in the primate visual system, if not the entire brain. We sketch some 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 1A.] 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. The main point is that the first stage of processing is massively parallel local motion extraction.

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 as a population they encode the relative likelihood of any possible motion at a particular location. [See Fig 1B for an example of what a response might look like if an upward motion were presented.]

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 1C.] 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.

All of this is underscored by the massive interconnection between processing Thus, it seems that the brain single integrated network approach to motion vision.

Figure 1



Constraints:

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.

For one thing, because moving things require energy to accelerate and to maintain their speed in our friction filled world, most things move slowly if at all. This will be our first constraint, namely that *slower motions should be more likely than fast ones*.

Another major idea comes from the fact that objects in the world have spatial extent. That means that motion in adjacent regions of the image tends to be correlated. This will be our second constraint, namely that *motion information should propagate constructively in space*.

A third major constraint is that objects rarely experience infinite accelerations. Thus, motion tends to be smooth from one point in time to the next. This leads to our third major constraint, namely that *motion estimates should maintain continuity and vary slowly over time*.

A fourth constraint arises to handle the case of transparent motion, the apparent appearance of two motions in a single location, *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.

Design:

To be specific, we propose the following single integrated probabilistic graph approach to motion vision. We will 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. 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.

Results:

We constructed a set of local Bayesian motion estimators as follows. Each estimator examined only a small image patch. The sensitivity of the estimator was very local, operating through a Gaussian mask 11 pixels in diameter. This soft bounded window gave very good local motion estimates and avoided boundary problems.

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 window is non-zero. We assume that the prior $P(\hat{v})$ is a symmetric Gaussian centered at zero – biasing us toward small velocity estimates. This implements our first constraint, namely that *slower motions should be more likely than fast ones.*

A second common assumption is that the likelihood function $P(I(x_i,y_i,t,d_i)\,|\,v)$ is of the form

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

where $w_i(x,y)$ is the small Gaussian window centered around (x_i,y_i) and $I_k \equiv \frac{\partial}{\partial k}I\left(x,y,t,d\right)$. 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 [7].

Given the structure of the images, many estimators suffered from the aperture problem, and have very broad probability distributions along the axis consistent with their observed window [9]. Given their very local view the motion estimates from these detectors were often noisy, or the result of local image variation not representative of global object motion.

The Markov Random Field (MRF) in which these units were imbedded was constructed specifically to allow the pooling of appropriate information for more precise motion extraction. Specifically, our local estimators were arranged in a hexagonally packed grid with non-overlapping windows, making estimator relatively independent. The graph defining the MRF was complete, connecting each node to each other node. Interestingly, early implementations only having local connectivity failed to produce any satisfactory results. Evidentially, these computations demand direct non-local interaction of estimators to work properly, making this null result interesting in its own right.

Inference in the MRF was carried our using Loopy Belief Propagation [10]. Nodes were selected in a random order to message to the rest of the network. Let the *B* be the beliefs of the messaging node. Messages we constructed as 1 + *k*(*B*-mean(*B*)). Belief which are favored in *B* with have a positive (*B*-mean(*B*)) and so a message value greater than 1, while beliefs that are contrary will have a negative *k*(*B*-mean(*B*)) and so a message value less than 1. Receiving nodes multiply their own beliefs by the message and renormalize. This implements a simple compatibility function in which the beliefs in *B* reinforce compatible beliefs in other nodes, and suppress incompatible beliefs. The critical aspect of the messaging is actually the factor *k*, which varies inversely with the absolute distance between any two nodes on the graph. Thus, although information is broadcast globally its relevance in inversely proportional to spatial separation. This implemented our second constraint, namely that *motion information should propagate constructively in space* and simultaneously our fourth constraint that, *there is at most one object, and thus one object motion, at any point in space*.

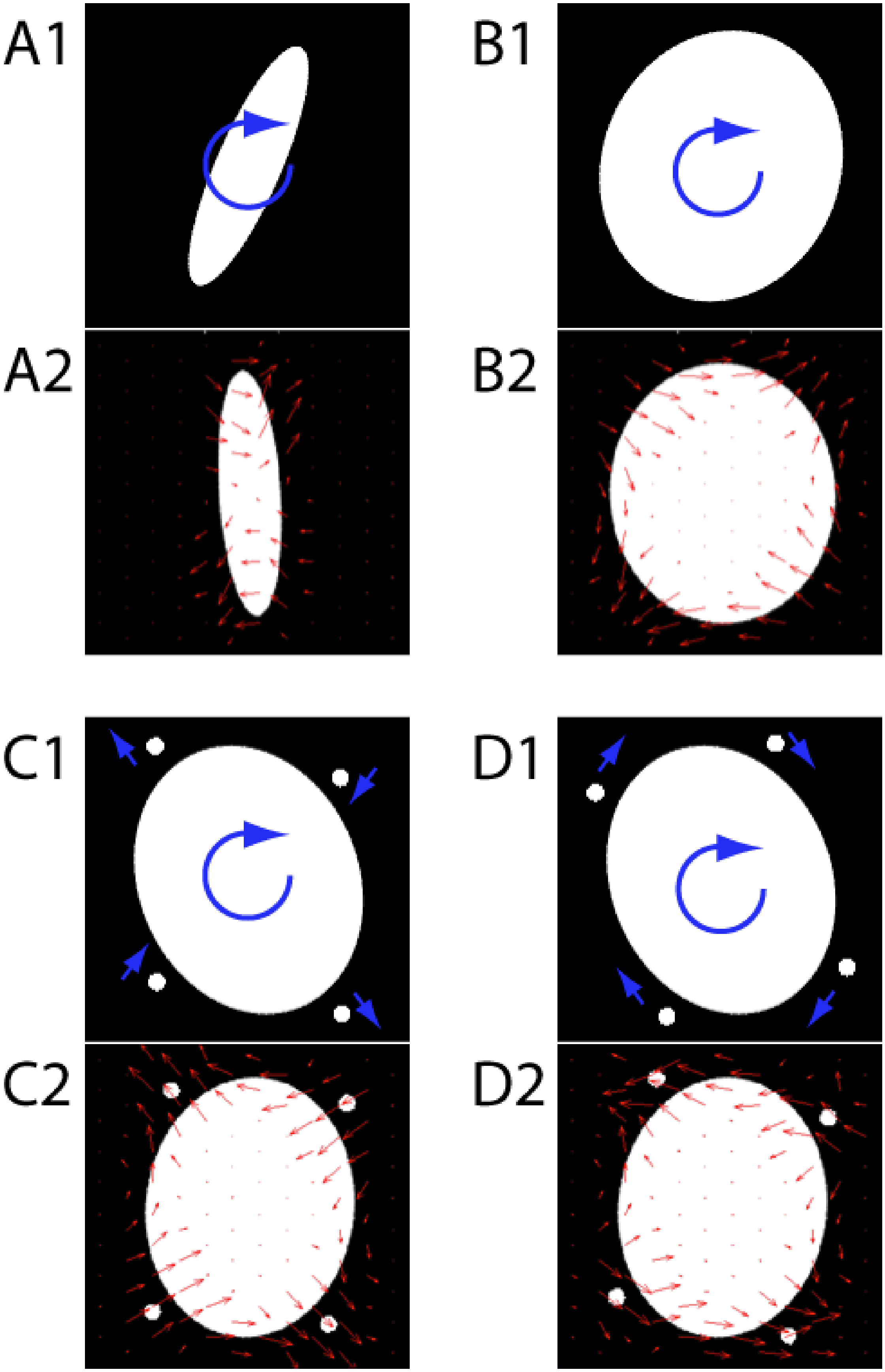
To implement or third constraint, namely that *motion estimates should maintain continuity and vary slowly over time*, we let our belief carry over from one time step to the next. Following the ideas of dynamic Bayesian Networks and Kalman filtering, the beliefs for the next frame of video were degraded forms of the beliefs given the previous frame. If we assume that motions change by adding a small random Gaussian acceleration, then we can propagate our beliefs forward simply by convolving our current beliefs with a Gaussian.

Testing:

When humans view a narrow ellipse that is rotating angularly (Fig 2A1), we normally perceive it as a rigidly rotating shape. If however we view the same motion applied to a fat ellipse (Fig2B1), we perceive non-rigid deformation. This is surprising given they both movies can be constructed by rigid rotation, and most motion extraction algorithms recover the same rigid motion. This dependence of perceived motion on object shape has been previously explained by global Bayesian motion estimation [8]. Essentially a prior in favor of smooth and slow motions biases motion estimates in favor of deformation along the gentle curves of the fat ellipse, if we assume Gaussian image noise.

We wanted to see whether our local Bayesian estimators, communicating only through the Markov Random Field, could arrive at the same conclusions as the global Bayesian approach. Can we computer this with a single integrated probabilistic graph approach. In fact this approach is very successful, even if only given two frames of video from which to extract the motions. (Fig 2A2 and Fig 2B2).

Figure 2

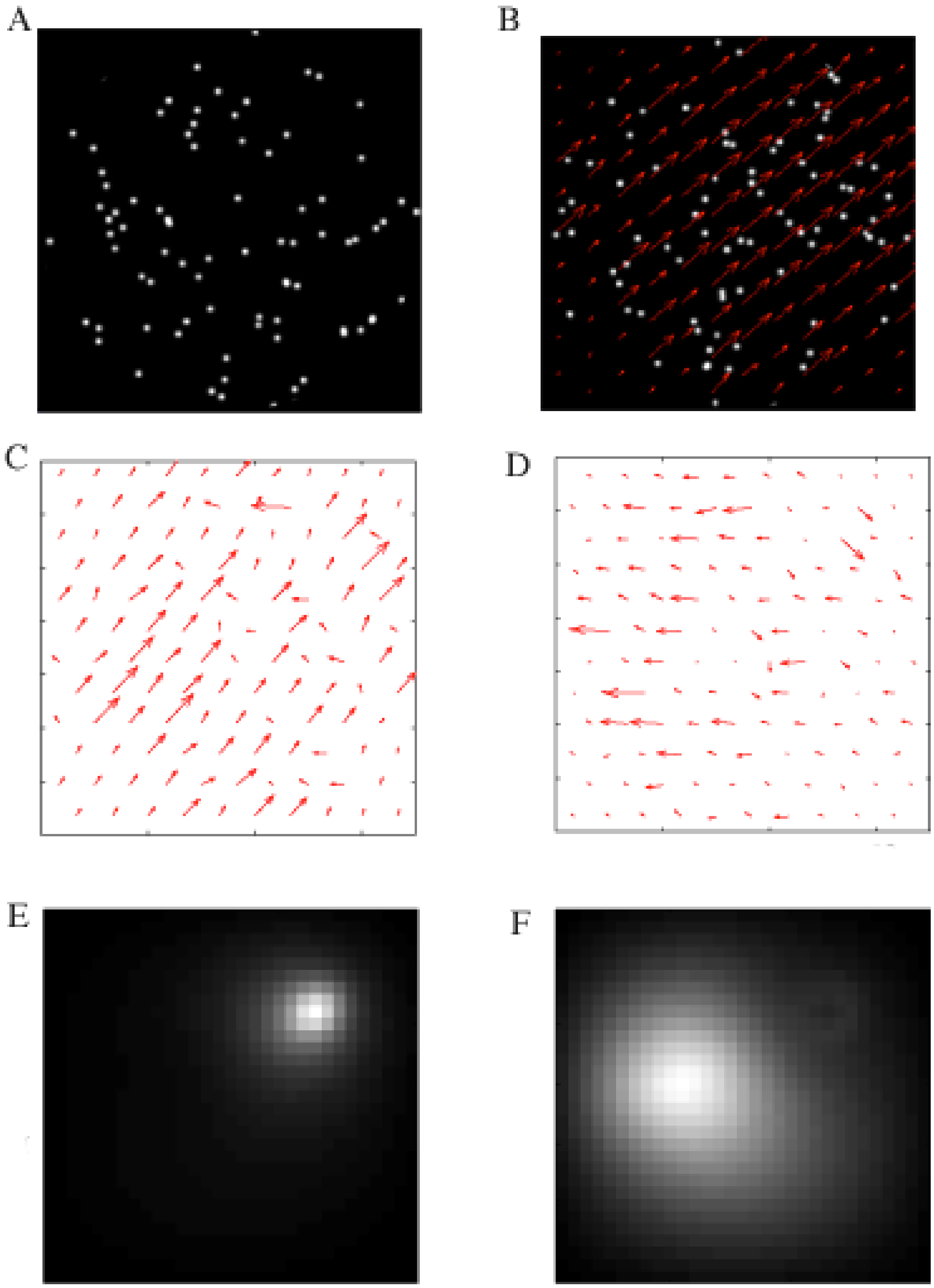


To really test how information is being propagated, we can make use of another interesting perceptual trick. If we flank the fat ellipse with dots we can affect human perception. Specifically, if the dots move only away from and toward the ellipse [Fig 2C1], they strengthen the percept of non-rigid deformation. If on the other hand, the dots, rotate along with the ellipse [Fig 2D1] they abolish the percept of non-rigid deformation, and the fat ellipse now beings to rotate rigidly like the narrow ellipse did before. This would not be surprising at all if the dots were in *contact* with the ellipse. Then they would create corners where the aperture problem would be solved and induce the above motion. The interesting thing is that for human vision this effect persists even if the dots are not connected by instead float some distance away from the edge of the ellipse. This dependence non-local image information has also been previously explained by global Bayesian motion estimation using a layering approach [8]. But again we wanted to see if we could accomplish this with a single integrated probabilistic approach.

Again we find that this integrated probabilistic graph approach does well. Figure 2C2 and 2D2 show the responses our network gives to each of the two cases. As we would expect, in the case where the dots move radial, the impression of deformation is only enhanced. In the second case where the dots rotated angularly, the precept of deformation is largely disturbed, tending more toward rotational motion as in Fig 1A2.

We can do more extensive testing by using a different class of example stimuli. Figure 3A shows a field of well separated Gaussian blobs. Each blob drifts in a particular direction and velocity. If the blobs all drift together, the motion information should be grouped across them. This allows much better estimation of the motion than would have been afforded by considering any single dot in isolation. Figure 3B, shows that our approach does the integration properly, getting good estimates of motion at points where there is little data, or in fact points where a dot has been but no longer is

Figure 3



This first result is not very impressive, because this could have been accomplished simply by integrating *all* the information across the scene. We can push the model much further, if we divide the dots into two populations, each moving in a different direction. Figure 3C and 3D show that our graphical model (now with two nodes at each point in the image) can simultaneously extract both motions, only integrating information across data that share a common fate. To see better how this happens we can look at the raw probability distributions over speed at one such pair of nodes Figure 3E and 3F show the beliefs at a location where there is currently motion up and to the right, but recently there was motion to the left. The belief in 3E is sharp and well formed, supported by its neighbors in the MRF and reinforced by the imaged data at that particular moment. The belief in 3F however, has no data to directly drive it, but lives on though temporal persistence in the network and the support of its neighbors, some of whom do have data presently consistent with its belief. This double extraction without an explicit masking or subtraction is very impressive and displays the strength of this graphical approach.

Discussion:

We have shown that a single unified probabilistic graph approach to motion grouping segmentation and integration is tractable. We can in fact follow the lead of the biological vision approach of are MT, where simple local motion processing is made rich by dense interconnections of massively parallel processors.

The primary drawback to this approach is the intensity of computation, and the resulting slowness on current serial processing architectures. This is approach is completely out of the question for real time systems, or even non-real time systems with large amounts of data.

Where it not for computational boundaries, we are hopeful that this work could be taken much further in the form of a 3D Markov Random Volume, capable of extracting three-dimensional motion, on three-dimensional surfaces. This would allow this approach to come into its own, extracting motions not readily accessible given current algorithms. It may be that there are some optimizations that can be made in this type of computation to facilitate implementation in the near future, but presently this goal remains out of reach.

References:

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

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

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

[4] 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.

[7] 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

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

[8] Weiss Y. and Adelson E.H. *Perception* , volume 29 pages 543-566 (2000)