Disambiguating Visual Motion Through Contextual Feedback Modulation

Pierre Bayerl
pierre@neuro.informatik.uni-ulm.de
Heiko Neumann
hneumann@neuro.informatik.uni-ulm.de
Department of Neural Information Processing, University of Ulm,
D-89069 Ulm, Germany

Motion of an extended boundary can be measured locally by neurons only orthogonal to its orientation (aperture problem) while this ambiguity is resolved for localized image features, such as corners or nonocclusion junctions. The integration of local motion signals sampled along the outline of a moving form reveals the object velocity. We propose a new model of V1-MT feedforward and feedback processing in which localized V1 motion signals are integrated along the feedforward path by model MT cells. Top-down feedback from MT cells in turn emphasizes model V1 motion activities of matching velocity by excitatory modulation and thus realizes an attentional gating mechanism. The model dynamics implement a guided filling-in process to disambiguate motion signals through biased on-center, off-surround competition. Our model makes predictions concerning the time course of cells in area MT and V1 and the disambiguation process of activity patterns in these areas and serves as a means to link physiological mechanisms with perceptual behavior. We further demonstrate that our model also successfully processes natural image sequences.

1 Introduction

The analysis and interpretation of moving shapes and objects based on motion estimations is a major task in everyday vision. However, motion can locally be measured only orthogonal to an extended contrast (aperture problem), while this ambiguity can be resolved at localized image features, such as corners or junctions from nonoccluding geometrical configurations. The models previously suggested to solve the aperture problem, which integrate localized motion signals to generate coherent form motion. For example, the vector sum approach averages movement vectors measured for a coherent shape (Wilson, Ferrera, & Yo, 1992). Local motion signals of an object define possible solutions of the motion constraint equation. (Horn & Schunk, 1981). If several distinct measures are combined, their associated constraint
lines in the velocity space intersect and thus yield the velocity common to the individual measures (intersection of constraints, IOC) (Adelson & Movshon, 1982). Bayesian models combine different probabilities for velocities with a preference for, e.g., slower motions (Weiss & Fleet, 2001; Weiss, Simoncelli, & Adelson, 2002). Such preferences were encoded as statistical priors in the Bayesian estimation process. Like for the IOC, Bayesian models mostly assume that motion estimates belonging to distinct objects were already grouped together. Unambiguous motion signals can be measured at locations of significant 2D image structure such that curvature maxima, corners or junctions. These sparse features can be tracked over several frames to yield robust movement estimates and predictions (feature tracking) (Del Viva & Morrone, 1998). Coherent motion is often computed by smoothing and interpolating sparse measures estimated at localized image contrasts. Thus, the inverse problem of motion estimation is regularized by minimizing a smoothing constraint over surface regions (Horn & Schunk, 1981; Nagel, 1987) or along boundaries (Hildreth, 1984).

In this letter, we investigate the mechanisms of the primate visual system to process fields of visual motion induced by moving objects or selfmotion. Motion information is primarily processed along the dorsal pathway in the visual system, but mutual interactions exist at different stages between the dorsal and ventral path (Van Essen & Galant, 1994). Our modeling of mechanisms of cortical motion processing focuses on the integration and segregation of visual motion in reciprocally connected areas V1 and MT. Our model dynamics are defined by spatially local isotropic interactions of velocity tuned cells. Additional signals indicating luminance boundaries or explicitly detected motion discontinuities could easily be used to improve our results, but would make it impossible to analyze the basic functionality of our system, which we claim to represent a fundamental mechanism of motion processing in the visual system. Despite its simplicity, the model is able to explain experimental data and, without parameter changes, to successfully process real-world data used for model benchmarking (Barron, Fleet, & Beauchemin, 1994). The new contribution of this work is a unified mechanism of motion disambiguation that deals with localized features as well as elongated boundaries. It is demonstrated how recurrent feedback processing between two model areas with similar dynamics operating on two different scales stabilizes local feature measurements. In addition, we show how such features trigger a filling-in process along boundaries to resolve the aperture problem and therefore arrive at globally consistent motion estimates through local lateral interactions. Such local lateral processes are biased and controlled by feedforward-feedback interactions that

\footnote{In case of spatial object occlusions, junctions may signal incorrect velocities, since their boundaries belong to different object shapes (McDermott, Weiss, & Adelson, 2001). We do not particularly consider the corresponding segmentation problem, but briefly discuss consequences in section 4.}
can be viewed as scale-invariant mechanisms of hypothesis testing and correction.

The letter is structured as follows. In section 2, we describe the model and the model dynamics. In section 3, we present computational simulations, followed by a discussion of the model and the results in section 4.

2 The Model

Motion analysis in visual cortex starts with primary visual area V1 and is subsequently followed by parietal areas such as MT/MST and beyond. These areas communicate with a bidirectional flow of information via feed-forward and feedback connections. The mechanisms of this feedforward and feedback processing between model areas V1 and MT can be described by a unified architecture of lateral inhibition and modulatory feedback whose elements have been developed in the context of shape processing and boundary completion (Neumann & Sepp, 1999). In this section, we present the model dynamics within and between model cortical areas involved to realize the integration and segregation of inherently ambiguous input patterns.

In a nutshell, the model consists of two areas with similar architecture that implement the following mechanisms.

1. Feedback modulation. Cells in area V1 are modulated by cell activations from model area MT. Cells in MT can, in principle, also be modulated by higher areas such as MST or attention. Since we focus here on the two stages of V1-MT interaction, the feedback to model area MT is set to zero.

2. Feedforward integration.

3. Lateral shunting inhibition enhancing unambiguous signals (see Figure 1).

2.1 Input Stage of Model V1 (Initial Motion Detection). The input stage in V1 consists of a set of modified elaborated Reichardt detectors (ERD) (Adelson & Bergen, 1985) to measure local motion for a specific range of velocities at each location (population code) (Pouget, Zemel, & Dayan, 2000). The functionality of the correlation-based detector is described in appendix A. The activities of this input stage \( c^{(3)}_{(x, Δx)} \) for different velocities (encoded by \( Δx \)) at different locations \( x \) (see Figure 1 and appendix A) indicate unambiguous motion at corners and line endings, ambiguous motion along contrasts, and no motion for homogeneous regions. Physiological findings suggest that motion-sensitive cells in V1 respond to orientation as well (component motion) and that only MT cells get (more) invariant to orientation indicating “pattern motion” (Movshon, Adelson, Gizzi, & Newsome, 1985; Movshon & Newsome, 1996). We focus our investigation on the
role of feedback in motion disambiguation and therefore employ motion-sensitive cells in model V1 that are not orientation selective (see the discussion in section 4.1). The input stage is divided in two steps. The first concerns cells selective to static-oriented contrasts at a fixed spatial frequency and independent of contrast polarity. Consistent with Movshon and Newsome (1996), these cells also respond to very weak contrasts, which is realized through shunting normalization (see appendix A). The second concerns direction-selective cells, pooling over all orientation-selective cells at different time steps, which yields a representation of visual motion independent of contrast orientation. This simplification does not contradict basic properties of motion-selective cells in V1, since immediate response properties of model cells concerning visual motion basically indicate component motion except at very few locations in the image with intrinsic two-dimensional structures such as corners or line endings (see sections 3 and 4.1).

We claim that the important difference between V1 and MT is the spatial size of receptive fields and that the proposed mechanisms within each area and between each area reveal properties consistent with physiological and psychophysical observations and yield highly accurate visual motion estimations (see section 3). In the following, we describe how different parts of the model contribute to the disambiguation of ambiguous motion signals in a biased competition process, while motion discontinuities remain preserved.
2.2 Motion Processing in Model Area V1 and MT (Motion Integration).

Two areas, model V1 and model MT, with similar dynamics subsequently process the initial motion signal in a recurrent loop (an outline is given in Figure 1). Here we outline the details of the architecture of both model areas, which consists of three steps: feedback modulation, feedforward integration, and lateral inhibition (see Table 1 for the parameterizations of processing stages in model V1 and MT).

The computational logic of recurrent feedback processing between two areas is that the higher area integrates localized measures from the lower area over a larger spatial range, or neighborhood, and thus evaluates these signals in a much larger context (Hupé et al., 2001; Friston & Bülchel, 2000). Consider the higher area, MT, as one in which cells evaluate their input via their feedforward input connection strength. Then the resulting activity could be interpreted as a signature of the degree of match between expected input (encoded by the input weights of receptive field [RF] kernels) and the current input signal. Feedback in turn functions as a predictor that enhances those signals in the lower area that are compatible with respect to feature specificity by way of top-down modulation (Grossberg, 1980; Ullman, 1995; Mumford, 1994). Such excitatory modulatory feedback interaction has the effect of providing activations in V1 that match the “expectations” of MT, a competitive advantage in subsequent mutual inhibition. Modulatory feedback has the further advantage that only compatible patterns get emphasized in the lower area and no activity is produced where no signal is provided by the input. Thus, our model adopts the “no strong loop” hypothesis developed by Crick and Koch (1998). Also, modulatory feedback can be compared with the intersection of constraints (IOC) analysis of motion integration, since the input signal constrains the feedback signal and the intersection of both gets emphasized.\(^2\) Before integrating the modified input

\(^2\) In our model, we particularly focus on the feedforward-feedback interaction between areas V1 and MT. This does not deny, however, the existence of (modulatory) feedback connections from, for example, area MST also. Such an extension can be incorporated by adopting the model mechanisms of MT-V1 feedback modulation also at the stage of model MT to include MST-MT feedback as well.
signal, it is squared to sharpen the distribution. Then the signal is processed by cells with isotropic spatial and isotropic directional gaussian RFs (see equation 2.2). This feedforward integration can be compared with vector average if the population code is interpreted as the sum of its components. Cell activities are normalized subsequently by lateral shunting inhibition (see equation 2.3). The sum of activations of cells sensitive to any velocity at a specific location normalizes the total energy (Simoncelli & Heeger, 1998). Therefore, unambiguous signals (indicating the presence of only one or only a few velocities) get emphasized, while ambiguous signals (indicating many possible velocities) lead to flat population responses. This step is similar to a feature tracking mechanism, whose behavior in our model emerges from the dynamic behavior of the system without employment of any explicit feature detection mechanism.

Each model area consists of three stages of processing whose dynamics are defined by the following equations (compare Figure 1; see appendix B for the steady-state solutions used for the simulations):

\[ \partial_t v^{(1)} = -v^{(1)} + net_{IN} \cdot (1 + C \cdot net_{FB}) \]  
\[ \partial_t v^{(2)} = -v^{(2)} + (v^{(1)})^2 \ast G_{\sigma_1}^{(x,space)} \ast G_{\sigma_2}^{(\Delta x,velocity)} \]  
\[ \partial_t v^{(3)} = -0.01 \cdot v^{(3)} + v^{(2)} - \left( \frac{1}{2n} + v^{(3)} \right) \cdot \sum_{\Delta x} v^{(2)} \],  

where \( n \) denotes the number of cells tuned to different velocities at any specific location. \( net_{IN} \) is the input of the model area (e.g., the output of the correlation detector for model V1), and \( net_{FB} \) is the feedback signal (e.g., the output of model MT for model V1). \( \ast \) denotes the convolution operation and \( G_{\sigma_i} \) are gaussian kernels in space(\( G_{\sigma_1} \)) and in the velocity domain (\( G_{\sigma_2} \)). Velocity is coded by a spatial shift \( \Delta x \) per frame.

Compared to model V1, the enlarged RF sizes in model MT (RF ratio V1:MT = 1:5; see Table 1) lead to less ambiguous signals, since the influence of unambiguous motion features generated by, e.g., line endings, is larger due to a larger aperture. We propose a solution to the aperture problem by feedback of activities from a higher stage, which provide a larger context of motion integration. Thus, less ambiguous MT responses from large spatial regions are combined with ambiguous but spatially localized V1 cell responses such that these signals in turn become less ambiguous. The temporal evolution of this disambiguation process in turn spreads, or fills in, unambiguous motion information along the moving shape outline like a traveling wave triggered by localized features. Again, we emphasize that this disambiguation is achieved by the network dynamics without the employment of specialized feature detectors and explicit mechanisms for decision making.
To handle continuous sequences of images (instead of iterating just over the correlation results of one image pair), the feedback signal has to "follow" the detected motion. This is realized by shifted synaptic feedback connections corresponding to the velocity (represented by $\Delta x$) in accordance with the cells' maximum sensitivity and thus represents a kind of prediction mechanism.

3 Results

The model dynamics described in the previous section emerges from a layered architecture of cells, or groups of cells, in which each area is represented by three layers. Model cells were considered as single compartment entities whose gradual activation dynamics follow shunting, or membrane, properties. The network can be shown to obey bounded input-output stability. Since the equations equilibrate fast, we simulate the dynamics using steady-state equations in order to save computing time. Computational simulations demonstrate the ability of the presented neural architecture to cope with synthetic image sequences as well as with natural sequences. The same set of parameters (see Table 1) was used for all simulations. The number of iterations varied, depending on how many frames were available. The number of cells to represent the velocity space was chosen to be 225 ($\Delta x \in \{-7, \ldots, 7\}^2$, including $\Delta x = 0$).

We investigated the model by probing it with artificially generated image sequences utilized in physiological and psychophysical experiments. Furthermore, we demonstrate that the same model using the same parameter settings is able to process realistic sequences. These test cases were also used in benchmark tests for technical motion detection approaches.

3.1 Empirical Data. Simulations of model V1 and MT qualitatively confirm the results from neurophysiological recordings of the time course of MT cells (Pack & Born, 2001). Model MT cell responses along the extended contrasts initially signal normal flow (perpendicular to the contrast orientation and corresponding to component motion) as in experimental observations. In the psychophysical experiment, the cells’ selectivity gradually switches after approximately 150 ms to signal the true velocity. In our model, this change is facilitated by filling in correct motion interpretations derived from intrinsic feature signals at line endings or corners (see Figure 2). As a consequence, our model suggestions are twofold: (1) the spatial distance of MT cells from unambiguous motion features directly influences the time until the signal is completely disambiguated and (2) scale invariance is achieved through temporal integration of disambiguated motion estimates along boundaries. Note that the filling-in process preserves spatially localized responses due to modulatory feedback. Our model generates disambiguated motion cell responses in MT as well as in V1. This predicts that recordings of motion-sensitive cells in V1 should result in similar time pat-
terns as for recordings in MT. There is no clear evidence that V1 cells behave similar to MT cells. This may be a consequence of the RF size of V1 cells, which makes it extremely difficult to obtain spatially accurate physiological measurements over time for moving objects. However, recent physiological studies show that subpopulations of V1 cells near line endings actually encode pattern motion and show similar dynamics as cells in MT (Pack, Livingstone, Duffy, & Born, 2003) (see section 4.1).

The following experiment shows how feedback processing causes the model to generate neural hysteresis and emphasizes the role of feedback for the disambiguation process. Figure 3 illustrates how neural hysteresis is generated by processing a random dot cinematogram, gradually changing from rightward to leftward motion or vice versa. Such displays induce perceptual hysteresis, indicating interactions with a short-term memory (Williams & Phillips, 1987). Such an interaction may help to lock in and keep the initially detected velocity (e.g., leftward motion) until the stimu-
lus has changed enough to switch to another alternative (e.g., rightward motion). The feedback signal in our model can be interpreted as short-term memory since it contains information from preceding time steps. Model simulations show that hysteresis is generated as a consequence of feedback processing. Without feedback, our model generates no hysteresis behavior, and the motion signal never gets completely disambiguated (maximum of approximately 80% correct flow information for purely left- or rightward motion; see Figure 3B). The latter effect occurs since the correlation detector used in the input stage cannot distinguish between the different stimulus dots, which leads to a high ambiguity in the correlation signal (similar to most other motion detectors). If feedback is applied, motion gets completely disambiguated and 100% right- or leftward motion is indicated (even in the presence of dots with switched direction) until the signal represented by MT cells switches nearly immediately to 100% of the other direction. This behavior can be interpreted as neural decision making.

3.2 Real-World Data

3.2.1 Artificial Sequences. Figures 4 and 6 illustrate the ability of the model to segregate different regions of visual motion. The artificially gener-

Figure 2: *Facing page.* Example of processing an artificial test sequence (100 × 100 pixel). (A) Temporal development (t = 1, . . . , 9) of the direction indicated by local model MT cell populations at three different image locations (indicated by simulated “electrodes”). The direction of motion measured at the corner (electrode 1, filled circle) points toward the true direction of motion (45 degrees) from the beginning, while on the edge (electrodes 2 and 3, open circles and triangles) flow initially measured orthogonal to the orientation of local contrast (90 degrees) gradually changes to the correct direction. Cell populations spatially more distant from locations with initially unambiguous estimations take longer to get disambiguated. (B) Real data showing the temporal course of direction indicated by a cell population in macaque area MT located on a moving bar pattern, redrawn and adapted from Pack and Born (2001): (1) 60 ms after stimulus onset, motion is orthogonal to the orientation of the bar. (2) 150 ms after stimulus onset, correct motion is indicated. The temporal activity pattern of model cell populations (A) qualitatively matches the temporal pattern of activity in macaque MT (B) by rescaling the temporal axis by an appropriate factor. (C) Temporal development of V1 cell population responses. Populations of motion-sensitive cells representing the local velocity space are illustrated at given locations indicated by small circles (dark = low activity, light = high activity). t = 1: Initial flow estimations (V1) at different (subsampled) cell locations and activity patterns of selected cell populations (large arrows: true direction of motion). t = 2, . . . , 6: Demonstration of the disambiguation process.
ated sequence of a flight over Yosemite National Park simulates self-motion over a static environment, which leads to an expanding flow pattern, with the exception of clouds moving horizontally to the right. The resulting flow estimation demonstrates that the model is able to extract motion induced by gray-level sequences of different frequencies, although the correlation detector is based on only one small scale (the clouds are represented on a much coarser scale than the ground).
Figure 4 shows the quantitative analysis of the processing results for the Yosemite sequence. We obtain a flow field of 100% density with an average angular error of 6.20 degrees after 10 iterations. This result compares well with technical solutions that produce less accurate estimations in most cases (Barron et al., 1994; Nicolescu & Medioni, 2003). The median error of 2.95 degrees after 10 iterations illustrates the robust model performance when outliers near region boundaries are excluded that occur exclusively at the horizon. Note that Barron et al. (1994) did not use iterations and multiresolution to refine the results of the computational models they have compared, specifically in those approaches, where iterations and multiresolution processing would be optional. However iterations were used, for example, to propagate smoothness constraints (Horn & Schunk, 1981), which can be compared to our relaxation process. Again, we point out that our approach relies on a single-scale estimation of visual motion, which is subsequently integrated and in turn updated by context information.

3.2.2 Real Camera Images. Natural image sequences recorded by cameras are noisy, and there is a high probability of encountering complex

---

Figure 3: Facing page. Perceptual hysteresis effect and model feedback for motion disambiguation. The proportion of MT cell activities indicating rightward motion ($\sum \text{activities}_{right} / \sum \text{activities}_{left+right}$) is plotted for each frame, processing two random dot cinematograms (each sequence shows 60 moving dots and consists of 60 frames, 40 $\times$ 40 pixel). In the first sequence (sequence a, the solid line), the dots are initialized with a random position and a velocity of three pixels per frame to the right. In each frame, one right-moving dot changes its direction by moving to the left. In the second sequence (sequence b, the dashed line), the dots have an initial direction to the left, switching one after the other to the right. (A) Feedback processing disambiguates the signal and generates a directional hysteresis effect that indicates the inertia generated by locking in the prediction from top-down feedback of a motion direction measured over time. Both sequences show an initial ambiguity in estimated motion (relative activity is 80% and 20% for correct and incorrect motion, respectively), because the correlation detector confounds corresponding dots within a certain neighborhood. Therefore, wrong velocity cues are detected in addition to the correct ones. After a few iterations of feedback processing, the MT cells are disambiguated and indicate a perfectly coherent motion signal (100% or 0% rightward). The response for sequence a (solid line) switches from rightward to leftward motion between 60% and 75% of changed dot directions, whereas for sequence b (dashed line), it switches from leftward to rightward motion between 40% and 25% of changed dot directions (hysteresis). (B) Without feedback ($C = 0$; see Table 1), no hysteresis is generated. The sum of cell activities indicating rightward motion is proportional to the relative number of dots moving to the right. The initial ambiguity of 80% versus 20% correct and incorrect motion responses is not resolved.
imaging conditions like occlusions and nonrigid motion. The proposed architecture successfully deals with a large set of natural sequences, including traffic sequences and animals in motion. Figure 5 illustrates the results for two examples: a traffic sequence with the moving taxi and a moving zebra. Feedback processing combines temporal information in order to segregate motion and eliminate outliers and noise. The spreading of unambiguous flow is visible especially where the aperture problem arises (along the stripes of the zebra) and where occlusion falsifies initial motion estimates (the border of the cars). Object segregation capabilities of the model are sketched in Figure 6. Concurrent interaction (shunting inhibition) in velocity space (including zero velocity, $\Delta x = 0$) segregates regions of different motion from each other as well as from regions with stationary contrast configurations.

4 Discussion

In this letter, we presented a new computational framework of recurrent motion processing to integrate and segregate visual motion signals. In the following, we discuss the model’s biological plausibility, compare it with existing models of motion processing, and summarize the major contributions.

4.1 Relevance and Biological Plausibility. There is both structural and functional evidence for the mechanisms and layered organization of our model. Anatomical and physiological studies suggest inter- and intra-areal connections also used in our model (Van Essen & Galant, 1994; Maunsell, 1995). Motion-sensitive cells can be found in MT as well as in V1 (Maunsell & Van Essen, 1983). Physiological studies (Movshon et al., 1985) have shown that cells in V1 are sensitive to component motion (motion along oriented contrasts) while cells in MT are less sensitive to oriented components but signal pattern motion. Since our focus is on the investigation of feedback in the processing of motion stimuli, we kept the input stage to the model as simple as possible. As a consequence, we assumed motion-sensitive cells in V1 to be independent of contrast orientation. We claim that this does not constrain the capability of our model to resolve ambiguities among motion-sensitive V1 cells if they were sensitive to orientation as well, because motion signals indicated by such cells inherently contain ambiguities by the definition of component motion. However, such an approach would obscure the disambiguation process described by our model, which is a consequence of lateral and feedback interactions between both model areas, differing in the spatial scale of integration. We claim that model V1 could be adapted by incorporating further details to match physiological properties of V1, such as spatial frequency tuning, contrast polarity, and orientation tuning (Hubel & Wiesel, 1968).

There is recent physiological evidence that V1 cells can partially encode pattern motion (thus, motion independent of the orientation). Pack et al.
Figure 4: (A) Detected motion using an artificially generated sequence of four frames simulating a flight over the Yosemite national park (316 × 252 pixel resolution); shown: motion indicated by V1 cell populations. Expanding flow is caused by self-motion through a static scene (the ground) and rightward translational flow by moving clouds (upper region of the horizon). Motion is detected even though the underlying gray-level structure has strongly varying spatial frequency content, or scale. Both regions of motion (ground and clouds) are clearly segregated from each other after a few iterations of V1-MT feedback processing. (B,C) Angular error of the optical flow direction indicated by MT cell populations processing the first two frames of the Yosemite sequence. Since there is no ground-truth available for the movement of the clouds, we assumed a horizontal motion to the right (according to Barron et al., 1994). (B) Spatial configuration of the angular error at different time steps (dark = large errors, light = small errors). Error peaks occur at flow discontinuities (clouds and ground). Note that the regions containing these outliers are much smaller than the RF size of MT cells (indicated at the top left of each image). (C) Temporal development of the average and the median angular error (of the entire flow field). The average error almost converges after five iterations and slightly improves to 6.20 degrees after 10 iterations. The median error of 2.95 degrees (after 10 iterations) illustrates the accuracy of the model excluding outliers, such as errors of approximately 180 degrees, which occur at region boundaries along the horizon.
Figure 5: Detected motion using real-world sequences (four frames). (A) The Hamburg taxi sequence (256 × 190 pixel resolution). (B) A walking zebra (http://www.junglewalk.com, 320 × 240 pixel). Shown: motion indicated by MT cell populations. Initially, erroneous flow information is eliminated where no motion occurs, and slight directional corrections can be observed where flow information is affected by the aperture problem (at the border of the cars and the stripes of the zebra).

(2003) showed that the time course of a subpopulation of V1 cells is similar to the time course of cells in MT solving the aperture problem near line endings. This is consistent with the prediction of our model that cells in V1 and MT are disambiguated simultaneously as a consequence of feedback and local competitive interaction. Lateral inhibition (normalization) is accomplished in our model by an isotropic center-surround interaction, which represents the simplest way to realize that functionality. The shunt-
Figure 6: Object segregation capabilities of the model. Display of maximal MT/V1 activities that were normalized against global minima and maxima for the sequences in Figures 4 and 5A (dark = low activation, light = high activation). Low activation appears due to strong competition at locations of high motion contrast. (A) In the taxi sequence, the cars are clearly segregated from the background and against each other. (B, C) The Yosemite sequence reveals two major regions of motion: ground and sky/clouds. It shows that the motion signal in V1 (C) contains some structural information that is averaged in MT (B).
ing equation (see equation 2.3) reflects the neuronal activity saturation. As a result of our modeling investigations, we provide evidence that experimentally observed behavior is generated by network dynamics that emerges from several elementary operations and functional principles of layering and connectivity. With the detail of employed model components, it was not our primary focus to quantitatively fit physiological data. Instead, our model serves as a link between physiology and perceptual behavior, which also enables processing real-world image sequences for benchmarking. The model suggests key principles of motion disambiguation and integration.

The proposed modulatory feedback mechanism is supported by recent physiological investigations of feedback connections between early visual areas (V1, V2, and V3) and MT (Hupé et al., 2001; Friston & Büchel, 2000). For example, Hupé et al. (2001) show that cell activities in V1 are highly affected by feedback from MT in an excitatory manner shortly after stimulus onset. This is consistent with our model, in which only excitatory feedback modulation is used ($1 + net_{FB} \geq 1$; see equation 2.1). As a result of recurrent processing, and consistent with physiological recordings of the time course of MT neurons (Pack & Born, 2001), our model disambiguates the motion signal shortly after stimulus onset. Here, the time to establish the final percept is influenced by the strength of the feedback connections, the RF field size ratio between V1 and MT, and, as a prediction of our model, the spatial extension of the region of ambiguous motion (see Figure 2). The time course of MT cell populations was also investigated by Pack and Born (2001) for different bar lengths (2–8 degrees). Consistent with our results, the time required to disambiguate such stimuli was roughly proportional to the bar length (Pack, personal communication, December 2003).

4.2 Comparison with Existing Models of Motion Processing. Different stages and computational mechanisms utilized by our model were also used in other biologically inspired as well as computational models. For the purpose of better readability, we organize our discussion according to categories of existing models of motion estimation; pure feedforward and recurrent models.

4.2.1 Feedforward Models. Simoncelli and Heeger (1998) proposed a model of detecting motion energy in areas V1 and MT using linear spatiotemporal filters. Individual motion estimates are normalized by dividing individual responses through the average response of activity in a spatial neighborhood. This center-surround mechanism has also been employed in our model. We achieve such a normalization by an antagonistic mechanism that involves shunting inhibition. The net effect leads to a divisive inhibition at individual locations by average activities integrated over a neighborhood in the space-velocity domain. Unlike Simoncelli and Heeger, we have incorporated a mechanism of modulatory feedback that disambiguates the motion signal and spreads activities over longer distances, solving the aperture
problem. It is worth mentioning that their filters in V1 and MT differ from our mechanisms and model the fact that motion-sensitive cells in V1 have less or no speed tuning. Such filters could also be used in our model, but in order to focus on the influence of feedback processing, we omitted any additional parameters. Nowlan and Sejnowski (1994, 1995) described a model of motion integration that utilized an explicit selection signal that is computed to determine the regions in the visual field where velocity estimations are most reliable. Motion-sensitive cells are then gated by this signal to produce the final estimate. Note that the way they learn how to compute the selection signal is an elegant method that may be applied to learn a normalization process like the one described by Simoncelli and Heeger (1998). Our model differs in several ways from Nowlan’s approach. While their approach utilizes a feedforward scheme, our model combines feedforward estimates with feedback integration and prediction. As a consequence, initial rough estimates are integrated and evaluated over time within a recurrent loop of matching velocities and motion predictions generated in area MT. As a by-product, the determination of reliable motion estimates is computed implicitly in our model instead of explicitly generating a decision-like selection signal. Weiss and Fleet (2001) and Weiss, Simoncelli, and Adelson (2002) solved the problem to determine the velocity of a moving object using a Bayesian estimation approach. They estimate coherent motion of a moving shape by maximizing the posterior probability of velocity votes given the detected image motion. This formulation leads to a probability representation in velocity space for all measures of single moving objects. In the spirit of IOC computation, all probability distributions are multiplicatively combined, including a given prior of expected velocities in the scene. The aperture problem is solved implicitly by maximizing the posterior from all individual measures. In our model, we do not directly combine all initial estimates, since this requires a priori knowledge about which moving parts in the stimulus belong together. Instead, we let initial motion signals being modulated by a predictive signal from the higher processing stage of area MT, which serves as a local prior. In order to achieve a global consistent estimate, this process is iterated to allow propagation of disambiguated motion signals along extended shape boundaries.

4.2.2 Recurrent Models. Grossberg, Mingolla and Viswanathan (2001) and Mingolla (2003) presented a model of motion integration and segmentation in MT and MST based on inputs from the form pathway (modeled as the FACADE framework—Form-And-Color-And-DEpth) (Grossberg, 1994). They studied how motion signals from partly occluded patterns can be integrated and segregated in a recurrent fashion. In contrast to our approach, their feedback signal (from MST) inhibits MT activities and has a more global character due to the RF size of MST cells (depending on the stimulus, these RFs cover 50% up to 100% of the entire stimulus). The authors suggest that such a mechanism of feedback inhibition and selection
also helps to solve the aperture problem (Mingolla, 2003). Unlike our model, this is realized by a decision-like mechanism through the inhibitory influence of global context information delivered by large-spanning kernels. We predict, therefore, that any resolution of uncertainty from the aperture problem should be independent of the length of the bar stimulus. Instead, our model propagates salient motion information along extended boundaries through recurrent interaction of MT and V1 cells with different RF sizes. This filling-in mechanism achieves size-invariance properties. Also, its temporal properties concur with experimental observations as the time needed for disambiguation increases with distance from locations of unambiguous motion. Lidén and Pack (1999) proposed a model of recurrent lateral motion interactions, which is able to produce a traveling wave of motion activation to solve the aperture problem. Like our model, they use the normalization similar to the mechanism described by Simoncelli and Heeger (1998) to emphasize salient motion estimates. In contrast to our model, their normalization mechanism is not isotropic in the velocity space.

The propagation is done by recurrent lateral excitation leading to an unbounded filling-in process, which has to be constrained by long-range inhibition of motion cells of different directional selectivity and by a separately processed motion boundary signal. In the absence of concurrent motion signals from multiple objects, their model leads to completely filled-in motion fields, which must be gated by multiplying the input signal in order to display only relevant motion patterns. Conversely, our model implements a kind of “soft gating” by biasing the input signal during feedback processing (see equation 2.1) and therefore produces spatially restricted motion estimates at all time steps without an explicit computation of motion or form boundaries. Koechelin, Anton, and Burnod (1999) describe a model of motion integration along the V1-MT pathway that utilizes mechanisms of recurrent lateral interactions. Their model utilizes a multiplicative combination of feedforward input and the result of lateral integration, which leads the authors to claim that their approach implements a neural mechanism of Bayesian estimation. Salient motion features are emphasized through normalization, and the results of recurrent lateral modulation (gating) are used to propagate these features. Though these mechanisms seem to be rather similar compared to those proposed in our model, the realization and behavior differ in many respects. For example, their gating process leads to strong inhibition of the input signal once the model has focused on one specific velocity while the stimulus changes to another velocity. Such lateral multiplication intensifies the winner-takes-all characteristic of their model (Koechelin et al., 1999) and makes it more vulnerable to errors. Our model follows a gradual prediction-and-correction philosophy realized by an exclusively excitatory modulation of the feedforward input through the feedback signal, which is followed by a center-surround competitive mechanism to realize a biased competition. Essential to our model is the decoupling into different areas with different RF sizes. That provides a larger context to the
higher area and thus the ability to correct (bias) and disambiguate cell activities in earlier areas with higher spatial accuracy. Another important point is that Koechelin et al. did not include cells sensitive to zero velocity or include an interaction with static form information in their model. This renders it impossible to segregate moving objects from a static background in a spatially localized fashion. Combined with the winner-takes-all property, their model may even form some regions of motion in a static image sequence with spatiotemporal noise that actually are not present. Finally, the results published in Koechelin et al. (1999) show that they fail to solve the aperture problem for moving bars in cases when they are longer than the size of their RFs. Contrary to this behavior, in our model a traveling wave of activation emerges, which helps to disambiguate motion signals along extended bars and shape outlines independent of the ratio between shape size and RF size. This concurs with the temporal evolution of MT cell activities investigated by (Pack & Born, 2001). Our proposed architecture is demonstrated to deal with large varieties of shape or object size as to provide a mechanism of size invariant motion integration.

5 Conclusion

In sum, we presented a model of motion processing in area V1 and MT capable of handling synthetic as well as artificial image sequences. The model shows the following key properties: initial detection of raw flow information, temporal spreading of reliable motion signals to gradually correct uncertain flow estimates, and the ability to sharply segregate regions of individual visual motion. We showed how to solve the aperture problem by contextual modulation, how feedback acts as short-term memory to account for hysteresis effects in motion disambiguation, and how global consistency is achieved by local interactions. Our model is unique in the sense that it combines mechanisms of local lateral interaction with modulatory and purely excitatory feedback to solve ambiguities of detected visual motion.

Our approach makes several new contributions. First, we propose a model of cortical feedforward and feedback processing in the dorsal pathway of motion integration implementing a neural hypothesis-test cycle of computation. Most important, the feedback mechanism is part of top-down modulatory enhancement of initial activities that match signal properties at a higher processing stage. Second, the disambiguation of initial estimates is solved by the interplay between top-down modulation and subsequent lateral competition. Consequently, the network dynamics propagate disambiguated motion signals along shape boundaries, thus realizing a guided filling-in process (Neumann, 2003). This mechanism is important in that it provides a means to process objects of different sizes in an invariant fashion. Third, the model serves as a link between physiological recordings (e.g., Pack & Born, 2001) and psychophysical investigations of perceptual motion integration (Williams & Phillips, 1987). Beyond this, the model is
able to process real-world stimulus sequences to yield accurate motion estimations. We believe that this further justifies the explanatory competence of key computational elements of the model, as most other biologically inspired models do not compare the quality of their results against other technical or nontechnical models.

Based on model simulations, we make several predictions concerning the computational mechanisms involved in early motion perception:

1. The disambiguation process observed for macaque MT cells (Pack & Born, 2001) should also be observed in V1 cells as a direct consequence of feedback interactions. This model prediction is partly confirmed by the findings of Pack et al. (2003).

2. The time to disambiguate regions of ambiguous motion depends on the distance of such regions from unambiguous motion features (e.g., induced by corners or line ends). This time is consumed by the increased number of feedforward-feedback cycles necessary to bias responses that cohere with the apparent motion direction.

3. We predict that without feedback, no perceptual hysteresis is generated, and motion activity patterns remain ambiguous.

At the current stage of modeling, the process of motion grouping is solely based on proximity in the spatial and the velocity domain. Interactions with the form pathway could enhance the model performance by grouping motion information preferably along contours. Ownership cues arising from occlusion can also be deduced from the ventral form pathway. Psychophysical investigations suggest that motion features are integrated only when they are intrinsic to the moving boundary, while extrinsic signals should be suppressed (Shimojo, Silverman, & Nakayama, 1989). This topic needs further investigation since it is yet unclear how signals from the form and motion pathway are integrated utilizing mainly excitatory interactions between cortical areas. However, even without all these extensions, the model already yields psychophysically and physiologically consistent results for a broad range of motion stimuli. In addition, the quality of estimated motion direction in real-world sequences compares well with technical solutions without explicit parameter tuning for each type of sequence.

In all, the proposed model provides further evidence for key computational principles that are involved in the cortical computation of sensory stimuli, their integration and segregation. Neumann and Sepp (1999) have proposed the basic mechanisms of feedforward feature detection, subsequent integration and matching, and subsequent modulation through feedback to implement a neural hypothesis-testing paradigm for boundary integration in static form perception. Model simulations of model V1-V2 interaction demonstrated context-dependent changes of orientation selectivity, texture density suppression, and subjective contour completion as observed in physiological and psychophysical experiments. An exten-
sion of this architecture by Thielscher and Neumann (2003) incorporates a model area V4 to investigate the segregation of textures generated from bars of different orientations. Whereas these models focus on the inter-areal feedforward-feedback interaction, we also proposed a model of intra-areal recurrent processing of V1 contour processing (Hansen & Neumann, 2004). These neurodynamical models of static form perception utilize the same core mechanisms of layered processing in cortical architecture.

In this letter, we now propose the same core mechanisms to account for the processing of temporally varying stimuli in the cortical motion pathway. Given the evidence gathered from our computational experiments, we claim that the early processing stages in visual cortex along the ventral and the parietal pathway are organized in a homologous fashion. Modulatory feedback and subsequent divisive inhibition realize a mechanism of biased competition already at an early stage with a similar behavior as the one proposed by Desimone and Duncan (1995) for attention mechanisms to filter out irrelevant information. We have proposed a concept of feedback as part of a layered structure and representation and presented an implementation of multiple loops of recurrent interaction whose dynamics realize multilevel cortical hypothesis testing cycles.

Appendix A: Equations for Initial Motion Detection (Input Stage)

Here we describe the equations used to generate an initial raw motion estimation, which is used as input stage to V1. We use (oriented) model complex cells to compute a spatiotemporal correlation (modified elaborated Reichardt detectors [ERD], similar to Adelson and Bergen (1985)) to measure local motion for a specific range of velocities at each location. Equations A.1 to A.4 (illustrated in Figure 7A) generate a raw motion signal (encoded as population code in \( c(x, 1) \)), which is computed from a pair of images and is further processed by the main model equations, B.1 to B.3.

\[
c(1)_{i,x,\alpha} = I_{i,x} \ast \partial^2_{x(\alpha)} G_\sigma \ast \frac{1}{0.01 + \sum_\beta |I_{i,x} \ast \partial^2_{x(\beta)} G_\sigma | \ast G_\sigma}.
\]

\( c_{i,x,\alpha} \) represents oriented complex cells responses normalized through shunting inhibition. Eight orientations \( \alpha \) were used for the simulations (* denotes the convolution operator, \( G_\sigma \) is a gaussian function, and \( \partial^2_{x(\alpha)} \) the second directional derivative in direction \( \alpha \)). The convolution with the second derivative of an isotropic gaussian filter (with approximately \( \sigma = 1 \)) was computed by applying two times a sampled first derivative of an isotropic gaussian filter (\( \sigma = 0.75 \)) in order to preserve (numerically) a dc-component of zero. Figure 7B illustrates some examples for sampled first derivatives of a gaussian. Figure 7E shows a zoomed part of a frame from an example sequence to compare the size of the kernels (printed at the same scale) with typical image structures processed by the model.
Figure 7: (A) Schematic overview of the correlation detector used as the input stage of V1 (for details, see the text). (B) First derivative gaussian kernels for different orientations. (C) Isotropic gaussian kernel with $\sigma = 1$ used in model V1 and the correlation detector (see the text). (D) Gaussian kernel ($\sigma = 7$) used in model MT (see the text). (E) Zoomed part of a frame from an example sequence, printed at the same scale as the kernels in B–D.
\( c_{t,x,\Delta x}^{(2+)} \) and \( c_{t,x,\Delta x}^{(2-)} \) represent half-detectors for a specific velocity defined by a shift \( \Delta x = (\Delta x, \Delta y) \) between two successive frames and can be interpreted as a raw correlation of model complex cell responses. These results are obtained by summing over all orientations (as a consequence, these results are no longer orientation specific) and pooling over a small spatial neighborhood with an isotropic gaussian kernel \( (\sigma = 1) \), illustrated in Figure 7C:

\[
\begin{align*}
\hat{c}_{t,x,\Delta x}^{(2+)} &= \left( \sum_{\alpha} c_{t,x,\alpha}^{(1)} \cdot c_{t+1,x+\Delta x,\alpha}^{(1)} \right) \ast G_{\sigma} \\
\hat{c}_{t,x,\Delta x}^{(2-)} &= \left( \sum_{\alpha} c_{t+1,x,\alpha}^{(1)} \cdot c_{t,x+\Delta x,\alpha}^{(1)} \right) \ast G_{\sigma}.
\end{align*}
\]

\( \hat{c}_{t,x,\Delta x}^{(3)} \) builds the final population code at each location indicating raw motion estimations \( ([\cdot]_+ \) is a rectification operator realizing \( \max(\cdot, 0) \)). A shunting inhibition normalized these activities, which differs from the standard implementation of the full Reichardt detector, which uses subtractive inhibition:

\[
\hat{c}_{t,x,\Delta x}^{(3)} = \frac{[\hat{c}_{t,x,\Delta x}^{(2+) - 0.5 \cdot [\hat{c}_{t,x,\Delta x}^{(2-)}]}]}{1 + [\hat{c}_{t,x,\Delta x}^{(2+) + 1}]}.
\]

The key features of the mechanism are that (1) the responses of normalized complex cells of all orientations are involved for motion estimation and that (2) even stationary patterns can produce (weak) responses in a spatial velocity code (e.g., induced by a periodic gray-level pattern).

The resulting activities \( \hat{c}_{t,x,\Delta x}^{(3)} \) for different velocities (encoded by \( \Delta x \)) at different locations \( x \) indicate unambiguous motion at corners and line endings, ambiguous motion along contrasts, and no motion for homogeneous regions.

**Appendix B: Equations for Motion Integration**

Here we present the steady-state equations corresponding to equations 2.1 through 2.3.

\( v^{(1)} \) samples the input signal \( \text{net}_{\text{IN}} \), depending on the model area; see Table 1) and multiplicatively enhances incoming signals matching to the feedback signal \( \text{net}_{\text{FB}}, \) depending on the model area; see Table 1).

\[
v^{(1)} = \text{net}_{\text{IN}} \cdot (1 + C \cdot \text{net}_{\text{FB}}).
\]

\( v^{(2)} \) realizes the integration in space and the velocity domain by isotropic gaussian filters. In the velocity domain, the filter size is identical in both model areas \( (\sigma = 0.75) \). Note that the resolution of the velocity space was set to \( 15 \times 15 \) for all simulations. The spatial kernel is a dirac-impulse (a
gaussian with $\sigma = 0$) in model V1 and a gaussian filter with $\sigma = 7$ in model MT (illustrated in Figure 7D):

$$v^{(2)} = (v^{(1)})^2 \ast C_{\sigma_1}^{(x,\text{space})} \ast C_{\sigma_2}^{(\Delta x,\text{velocity})}$$  \hspace{2cm} (B.2)

$v^{(3)}$ represents normalized velocity estimations in both model areas, which is realized by a shunting inhibition with the sum over all velocities. As a consequence, ambiguous signals are weakened, while unambiguous signals are emphasized:

$$v^{(3)} = \frac{v^{(2)} - \frac{1}{2\pi} \cdot \sum \Delta x v^{(2)}}{0.01 + \sum \Delta x v^{(2)}}.$$  \hspace{2cm} (B.3)

References


Received August 8, 2003; accepted April 1, 2004.