Fast, Recurrent, Attentional Modulation Improves Saliency Representation and Scene Recognition

Xun Shi[1*], Neil D. B. Bruce[1*], and John K. Tsotsos[1]

[1]Department of Computer Science & Engineering, and
Centre for Vision Research,
York University, Toronto, Ontario, Canada
{shixun, neil, tsotsos}@cse.yorku.ca

Abstract

The human brain uses visual attention to facilitate object recognition. Traditional theories and models envision this attentional mechanism either in a pure feedforward fashion for selection of regions of interest or in a top-down task-priming fashion. To these well-known attentional mechanisms, we add here an additional novel one. The approach is inspired by studies of biological vision pertaining to the asynchronous timing of feedforward signals among different early visual areas and the role of recurrent connections from short latency areas to facilitate object recognition [7]. It is suggested that recurrence elicited from these short latency dorsal areas improves the slower feedforward processing in the early ventral areas. We therefore propose a computational model that simulates this process. To test this model, we add such fast recurrent processes to a well-known model of feedforward saliency, AIM [6] and show that those recurrent signals can modulate the output of AIM to improve its utility in recognition by later stages. We further add the proposed model to a back-propagation neural network for the task of scene recognition. Experimental results on standard video sequences show that the discriminating power of the modulated representation is significantly improved, and the implementation consistently outperforms existing work including a benchmark system that does not include recurrent refinement.

1. Introduction

Object recognition in real scenes is a hard problem. Classical computer vision algorithms rely on scanning the image to search for similar patterns that correspond to targets, a problem proved to be NP-complete [30]. In one way or another, attentive mechanisms facilitate object recognition in both biological and machine vision.

A number of computational approaches follow [22] to model visual attention in a pure bottom-up fashion. The premises of these models are a layered hierarchy that relies on a cascade of filters. The process begins by extracting image characteristics of the object (*e.g.* colour, edges, motion) into a multi-dimensional description, and gradually building up to a saliency representation. The representation is used to provide regions of interest with respect to the background that further facilitates object recognition. However, when input scenes are cluttered and noisy, implementations (*e.g.* [34,35]) that follow the above feedforward strategy fail even for the simplest target types. The main reason for this failure is that, saliency in these models is defined in a way to capture the perceptual difference between a location and its background. If such an algorithm is exposed to a cluttered scene, where the difference between a location and its background is not obvious or objects are not conspicuous in the manner defined by the algorithm, then the saliency representation will confuse a potential target with its background, such that it is difficult for the algorithm to yield the region that precisely segments the target.

Other approaches instead suggest the importance of using recurrence to model visual attention. Since an early conceptualization [24], the idea has been applied in many computational models (*e.g.* [14, 31]). They incorporate a slow and serial process starting from the top of the visual hierarchy, in which top-down information controls the bottom-up activation flow. In the case of object recognition, the nature of recurrence captured in different models (*e.g.* [2,32]) is to refine the feedforward representation to assist recognition.

Recent studies of the primate visual system provide the insight that another kind of recurrence among the early areas of the primate visual cortices could also play an im-

*Xun Shi and Neil D. B. Bruce contributed equally to this paper.

Table 1: ***Response Latencies of visual areas.*** *Earliest and median latencies are recorded from monkeys in different studies. Earliest latencies refer to the delays observed with 10% neural activations.*

Area	Earliest (msec)	Median (msec)
V1 M [23]	20-31	40
MT [27] *	25	45
MST [21]	35	45
FEF [9]	35-45	65
7a [9]	50	90
V1 P [25]	40	65
V2 P [25]	55	85

* Latency recorded uses TMS on human subjects.

portant role in attention that facilitates object recognition (see [3] for a review). This mechanism is highly dependent on timing of signal convergence in the visual hierarchy.

Bullier proposed a model that uses asynchronous information projections to facilitate early visual analysis [7]. The so-called integration model of visual processing is put forward based on the observations that the primate geniculocortical and cortico-cortical connections have different conduction speeds. He and his colleagues collected data provided by several research groups and noticed that the conduction speed of M pathway (magnocellular cells of lateral geniculate nucleus (LGN)) to dorsal areas (for motion perception, for example) is faster than the conduction speed of P pathway (parvocellular cells of LGN) to ventral areas (for object perception) [8, 25]. Higher level dorsal areas such as MT, MST and FEF, are activated more rapidly than several early ventral areas, such as V1 P and V2 P (see Table 1 for a list of feedforward latencies recorded at several visual areas). They also noted that the difference in feedforward latencies (about 20ms between MT and V1 P) permits the results of computation in higher dorsal areas to be sent back through recurrent connections to ventral neurons in time to affect the feedforward projection from the P pathway to the same group of ventral neurons. Since the computational units of the two pathways interpret different visual elements, the fast recurrent connections may naturally reinforce the ventral bottom-up object analysis. Experiments [17, 18] showed that the recurrence elicited from MT facilitates computation of ventral V1 cells in a push-pull fashion, which improves neural responses towards moving objects and suppresses background activations. However, what Bullier and his colleagues did not do is detail exactly how such a process might operate.

Motivated by the integration model, we describe in section 2 the computational components that capture the idea that the primate visual system uses results from the higher dorsal areas to modulate the computation in the early ventral areas. In section 3, the model is formalized and implemented. Several examples using static as well as dynamic images are shown to demonstrate the modulation effects.

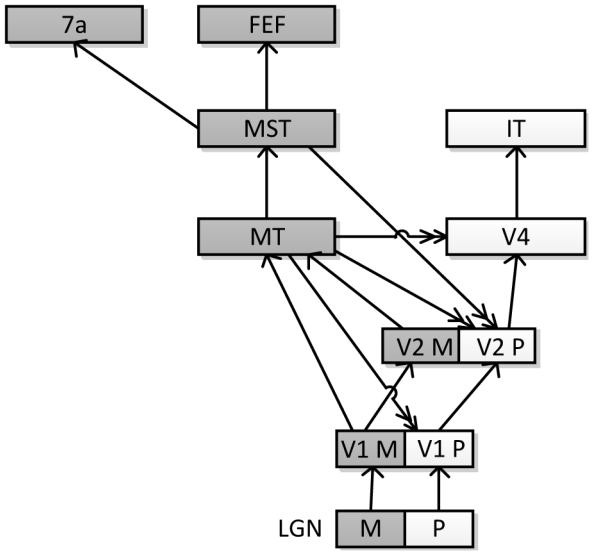

Figure 1: ***Connections between dorsal and ventral pathways.*** *Hierarchy of visual areas considered in this paper. (derived from [2, 7]) Grey blocks denote dorsal areas, and white blocks denote ventral areas. Lines denote connections. Particularly, double arrow lines denote feedback connections from higher dorsal areas to lower ventral areas that envisioned in our model.*

To test the model comparatively, we added the implementation to a back-propagation neural network for the task of scene recognition, which is presented in section 4. Results on standard video sequences of real scenes indicate that the recurrent modulation significantly augments the representation used for recognition, and the implementation consistently outperforms existing work including a benchmark system that does not include recurrent refinement. In section 5, we draw conclusions as well as discuss the implications of the proposed model.

2. Computational Model of Fast Recurrent Modulation

The representation and processes described in this section are informed by the knowledge of structures and function of the primate visual system [11, 33]. In particular, the model details the mechanism of the fast recurrent modulation [7] between neurons of higher dorsal areas and neurons of lower ventral areas.

The model assumes a multi-pathway layered hierarchy, with each layer being a retinotopic array. Layers are connected via feedforward connections; some layers have in addition, recurrent links to layers earlier in the hierarchy within the same or different pathways. Figure 1 illustrates these concepts. An element of the array is a neural assembly that includes all the units that perform the associated computations. For the rest of the paper, we use connections from MT to V1 P as an example to develop the idea.

In its most direct form, the novel aspects of our model consist of using the output of MT neurons to modulate or tune the processing of V1 P neurons in advance of the arrival of their input. The tuning is realized as a multiplicative inhibition on the inputs of the V1 P computations. The concept may be generalized to be as applicable to any other dorsal area that is connected via recurrence to a ventral area provided the timing of M and P pathway conduction is of the correct kind. Specifically, if the sum of the time required for input to arrive at a dorsal area, to be processed by that area, and to be fed back to a ventral area is less than the time required for a P pathway input to reach that ventral area, then this same mechanism may be at play.

The neurons of the two pathways compute different visual features. In macaque monkeys, it is reported that M cells are achromatic, have a higher peak sensitivity to contrast and respond to higher temporal frequencies than P cells, while P cells are color-sensitive, respond to higher spatial frequencies and show higher sensitivity at the lowest temporal frequencies. There is significant overlap in the ranges of both temporal and spatial frequency sensitivity [10]. MT neurons also have larger receptive fields (RFs) than V1 P neurons, and thus integrate input over a wider spatial context. If results from a wider context can be taken to influence the smaller V1 computations, they can act as a consistency check on those computations that informs the V1 computations of how to best respond given a wider spatial context that they, because of connectivity limitations, cannot otherwise see. The fast feedforward sweep of dorsal activations from the M channel thus underlines a mechanism of the primate visual system to provide spatiotemporal contextual guidance to the ventral analysis. In our model, context modulates the V1 ventral computations through multiplicative inhibition. The fact that these visual areas are all retinotopically organized allows such modulation to locally specific, in contrast to a global context measure (see [4, 12, 15]).

3. Formalization

The units that are essential to describe the process are formalized in the next section. Empirical comparisons are conducted to study the modulation effects as follows.

3.1. Formalization of each stage

The model will include M and P channel pathways from the retina (simplified to just represent an image with no other processing), LGN and cortical areas V1 and MT. Sets of image filters will be defined that represent the function of each. The structure of each filter lends itself naturally towards a filter bank that provides even coverage of the spectrum. Layers of filter banks are connected to form a cascade structure to extract early visual features as follows.

LGN: The center-surround receptive fields of LGN have response patterns that can be described as a Difference-of-Gaussian filter [20] given by:

$$f_{lgnS}(x,y) = \frac{1}{2\pi\sigma_c^2}\exp\{\frac{-(x^2+y^2)}{2\sigma_c^2}\} - \frac{1}{2\pi\sigma_s^2}\exp\{\frac{-(x^2+y^2)}{2\sigma_s^2}\} \quad (1)$$

where σ_c and σ_s are the bandwidth (standard deviation) for the center and surround Gaussian function respectively. In our implementation, these parameters are set to realize M cells ($\sigma_c = 3$, $\sigma_s = 4.8$) and P cells ($\sigma_c = 1$, $\sigma_s = 1.6$). LGN temporal response patterns can be described as a log-Gabor filter [13], which is defined in frequency domain as:

$$F_{lgnT}(w) = \exp\{\frac{-\log(w/w_0)^2}{2\log(\sigma_t/w_0)^2}\} \quad (2)$$

where w_0 is the center temporal frequency, and σ_t is the bandwidth. A multi-scaled temporal filter bank is realized to provide an even spectrum coverage by using different w_0 and σ_t. We configure lower temporal frequencies ($w_0 = 3, 9, 27$) for P cells and higher temporal frequencies ($w_0 = 9, 27, 81$) to represent M cells. The bandwidth $\sigma_t = 0.55w_0$, which yields approximately 2 octaves. Note that all the units in the model are spatiotemporal and most conveniently described in the frequency domain.

V1: The simple cells [16] receive feedforward projections from LGN and respond to different spatiotemporal sub-bands. The temporal profile of V1 is defined as a low-pass filter to describe the temporal selectivity. The spatial sub-band selectivity can be described as a 2D log-Gabor orientation filter that is defined in frequency domain as:

$$F_{V1S}(u,v) = \exp\{\frac{-log(u_1/u_0)^2}{2log(\sigma_u/u_0)^2}\}\cdot\exp\{\frac{-v_1^2}{2\sigma_v^2}\} \quad (3)$$

where $u_1 = u\cos(\theta)+v\sin(\theta)$, $v_1 = -u\sin(\theta)+v\cos(\theta)$, θ denotes the orientation of the filter, u_0 denotes the center spatial frequency, σ_u and σ_v denote the spatial bandwidth (standard deviation) of the simple cell along the u and v axis respectively. We implement a filter bank of 4 spatiotempoal subbands. For each, V1 P simple cells are tuned to high spatial frequencies ($u_0 = 9, 27, 81$) and V1 M simple cells are set with lower spatial frequencies ($u_0 = 3, 9, 27$), note that the ranges for M and P cells have overlaps. The bandwidths are set as $\sigma_u = 0.55u_0$, $\sigma_v = 0.55u_0$.

V1 complex cells [16] integrate energy of V1 simple cells over larger receptive fields. In this context, a quadrature pair is used to model the complex cell, which computes the square root over outputs of two simple cells given by Eq.(3) that are 90 degrees out of phase. In practice, the Hilbert transform of a V1 simple cell response gives its

quadrature pair. We follow [1] to compute V1 energy of a specific spatiotemporal orientation, which is governed by:

$$C_\theta(x,y,t) = \sqrt{(S_\theta(x,y,t))^2 + (h(S_\theta(x,y,t)))^2}; \quad (4)$$

where $S_\theta(x,y,t)$ denotes the responses of V1 simple cells in spatiotemporal domain of orientation θ, and $h(\cdot)$ denotes the computation of quadrature pair.

MT: In particular for the dorsal pathway, the computation of a MT neuron takes a simple form that integrates over V1 M complex cell responses within a larger spatiotemporal area to compute opponent energy [5] given by:

$$\begin{aligned} MT_\theta(x,y,t) = & \sum_{\Delta x,\Delta y,\Delta t} C_\theta(x,y,t) \\ & - \sum_{\Delta x,\Delta y,\Delta t} C_{\theta+\pi}(x,y,t) \end{aligned} \quad (5)$$

where $\sum$ denotes the summation of V1 complex cells over a spatialtemporal range $(\Delta x, \Delta y, \Delta t)$.

Multiplicative inhibition: The output representation of MT used to modulate V1 ventral area is generated by accumulated responses across all spatiotemporal bands. The modulation is realized as multiplicative inhibition. Since it is unclear where the MT-V1 feedback fibers terminate, we have considered three possible locations, namely, at input to the V1 ventral simple cells, at input to the V1 ventral complex cells, and at output from the V1 ventral complex cells. Out tests show similar results. Therefore, in the current paper we assume such feedback connections affect the input to the V1 ventral simple cells. The process is given by:

$$P'(x,y,t) = P(x,y,t) * Sig(\sum_\theta MT_\theta(x,y,t)) \quad (6)$$

where P denotes the output of LGN P cells, and P' is the modulated output that further connects to V1 ventral simple cells. $Sig(\cdot)$ denotes the sigmoid function. The motivation for the use of this operation is to bound the MT output in a nonlinear way.

3.2. Empirical study

We begin with a simple example that illustrates the sequence of computations. Fig. 2 shows an example of two simulated objects with plaid patterns in front of uniformly distributed pseudorandom noisy background. The intention of using a simulated scene is to provide a clear view of the computation developed along the hierarchy. In this case, both pathways compute pure spatial features. As shown, layers along the dorsal pathway have lower spatial frequency sensitivity with four orientations, and layers of the ventral pathway have higher spatial frequency sensitivity [1].

[1]In this demo, we use $u_0 = 3$ for V1 M cells and $u_0 = 27$ for V1 P cells respectively in Eq.(3).

Results of MT layer are the feedback representation that is used to modulate the input of V1 ventral neurons. Since ventral neurons have higher spatial sensitivity, they extract mostly the noisy patterns present in both targets and background (see the non-modulated V1 ventral output where one cannot tell the location of targets). Through modulation, regions of background are inhibited, leaving patterns belonging to the targets highlighted in the output.

Next, in order to enable a qualitative analysis of the impact of the model, we consider the effect of incorporating it into a model of saliency and into a model of scene recognition. The fast recurrent process fits easily in the existing saliency models. The dorsal-ventral inhibition introduced in the previous section makes an impact on computation of visual saliency especially in cluttered and noisy scenes. To test this idea, we add the proposed recurrent model into a well-known saliency model, AIM [6], to compute a saliency representation over a variety of image types to demonstrate the potential for segmentation improvement.

Since filters representing the dorsal pathway respond to both spatial and temporal information represented in the input, we conduct single image tests to evaluate modulation caused by spatial frequency variations, and use image sequences to study the motion caused modulation.

The inhibitory effect elicited from the dorsal pathway to modulate the computation of the ventral pathway is revealed using static scene images (Note: filters used in this experiment are purely spatial.). Fig. 3 demonstrates a qualitative comparison of the saliency representation computed by AIM using the proposed mechanism with the saliency representation computed by AIM alone. As can be observed, there is considerable similarity between the outlines of salient regions (reddish areas) from the modulated saliency map and real object contours shown in the original images. Finally, a substantial amount of clutter is suppressed.

We also provide examples using image sequences in Fig. 4 to illustrate how motion is involved in inhibiting ventral perception that facilitates object segregation. The dorsal feedback maps in these examples clearly highlight the regions consisting moving objects (*e.g.* cars, pedestrians). This information inhibits the ventral computation corresponding to stationary areas, such as trees, buildings and street signs. In return, moving targets are conspicuous in the saliency representation. By comparing the modulated ventral saliency maps and the non-modulated versions to human labeled data, it is obvious that salient regions depicted in the modulated ventral saliency maps more accurately reflect the boundaries of targets, facilitating figure-ground segmentation.

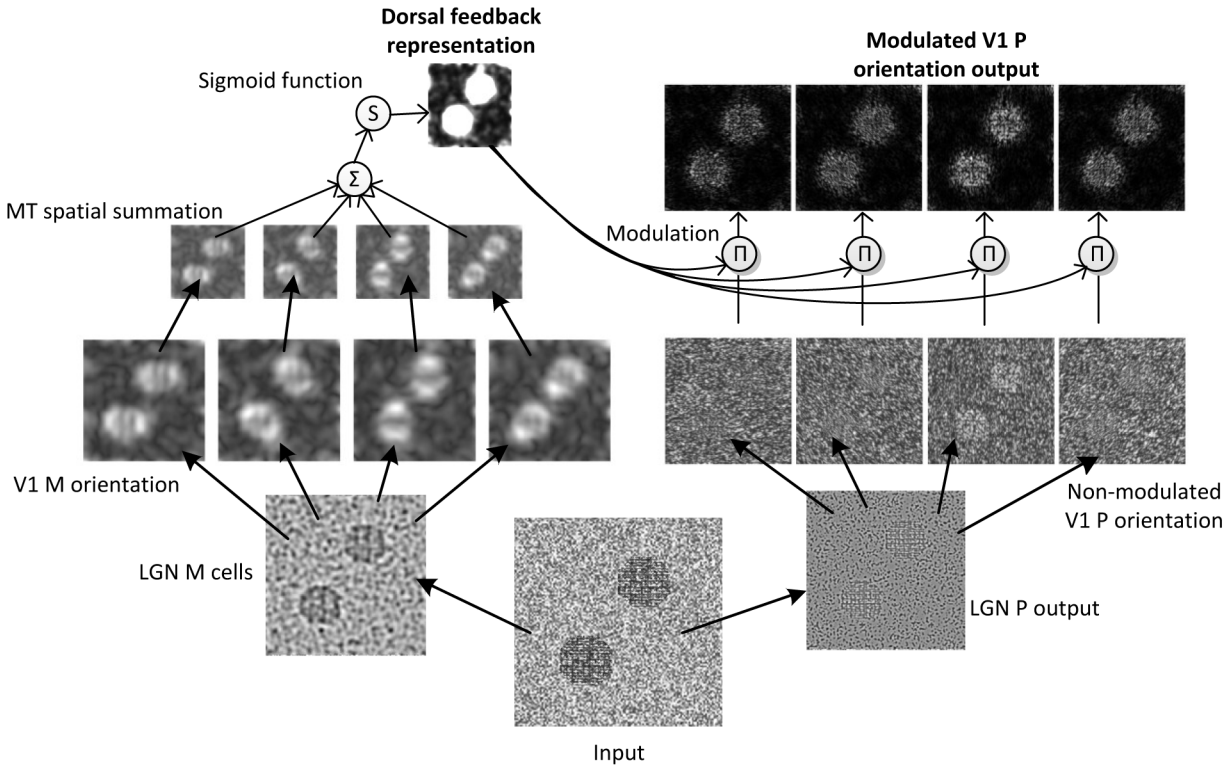

Figure 2: ***An example to demonstrate the principles of the proposed computational model.*** *Stimuli: two objects of plaid pattern within uniformly distributed pseudorandom noisy background. Both pathways compute pure spatial features. The dorsal pathway computes the lower frequency orientation features to generate a feedback representation. Higher frequency orientation features computed along the ventral pathway are then modulated through multiplication. Result of the modulated V1 P orientation output shows a clear segmentation that isolates the two objects.*

4. Improved scene recognition via fast recurrent modulation

In view of the enhanced figure-ground segmentation, we suspect that the fast recurrent modulation may deliver an improved representation for recognition. To test this idea, our implementation is added to a back-propagation neural network (as a scene classifier) to form a recognition system to recognize real scenes. Image sequences introduced in [29] are used for our testing because they include a variety of cluttered scenes. Since scenes are recorded using a hand-held camera, we want to examine how spatiotemporal information extracted by the dorsal representation may impact overall scene recognition performance. We compared our system with two existing systems that incorporate different feedforward strategies on the same set of video sequences.

We first detail the proposed recognition system, and then describe the other two existing systems.

In the proposed system that uses fast recurrent modulation (FRM), the filter cascade to simulate the dorsal pathway includes 3 temporal scales ($w_0 = 9, 27, 81$ used for Eq.(2)), 3 spatial scales ($u_0 = 3, 9, 27$ for Eq.(3)) and 4 orientations on the luminance channel. As such, there are $3 * 3 * 4 = 36$ filters providing the dorsal representation. To simplify the computation, the filter cascade for the ventral pathway does not include temporal filters, but with 3 spatial frequency sub-bands and 4 orientations on luminance channel. Compared with the spatial filters that are used to construct the dorsal representations, the ventral filters are defined in a way that covers relatively higher spatial frequency sub-bands ($u_0 = 9, 27, 81$ for Eq.(3)). The early ventral representation also includes two color opponency features. Thus there are $3 * 4 + 2 = 14$ filters that define the ventral representation. Finally, each extracted visual feature in the ventral representation is modulated using Eq.(6).

We employ a simple machine vision strategy for recognition. A holistic representation is built following [29]. Each modulated ventral feature is sliced into five-by-five non-overlapping blocks. A vector is created for each ventral feature through block averaging. Thus, each vector has 25 elements. At last, vectors of the 14 ventral features are concatenated to form the holistic representation, which is then used for recognition.

The other two existing systems use the same filters to compute visual features, but they employ different strate-

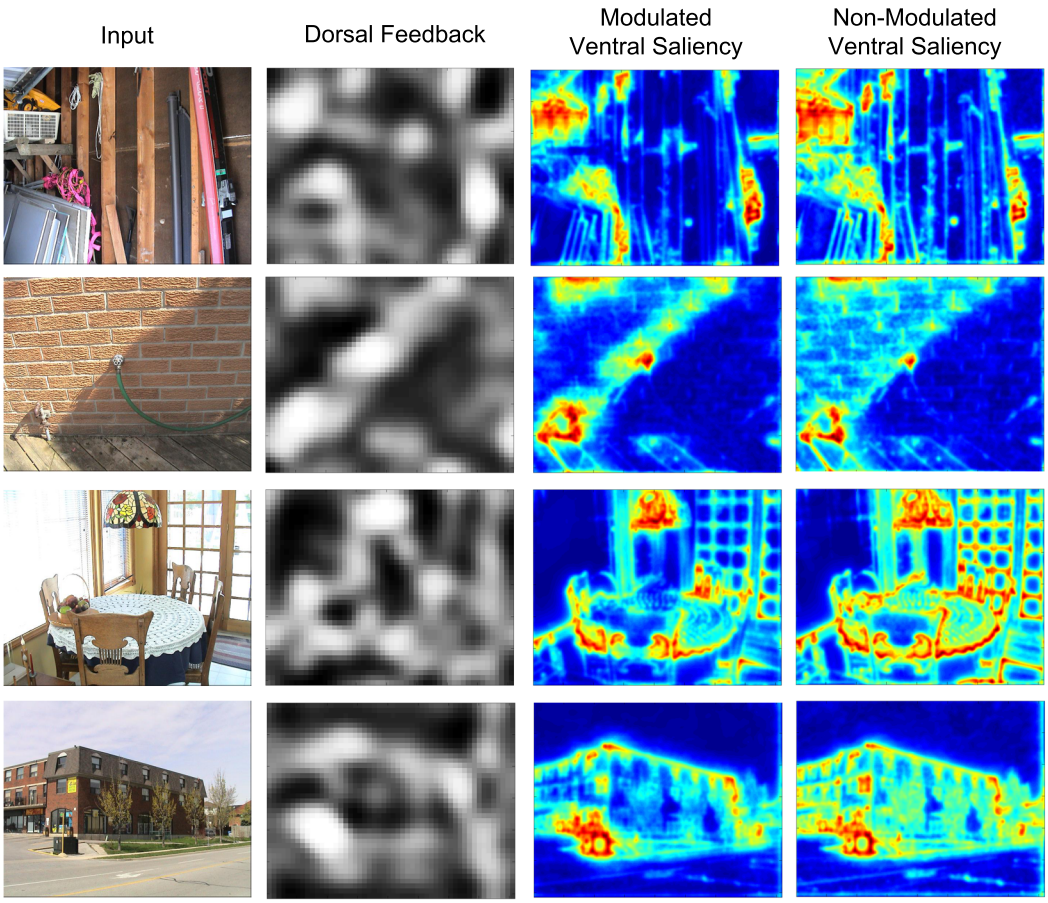

Figure 3: *An empirical single-image comparison of the ventral saliency based on the proposed fast recurrent modulation with non-modulated ventral saliency for a variety of cluttered scenes. From left to right: original images, feedback strength elicited from the fast dorsal activation, visual saliency based on modulated ventral features and visual saliency based on non-modulated ventral features.*

gies to construct the holistic representation. The first one (SI) borrows the feedforward strategy provided by Siagian and Itti [29]. The holistic representation is built based on the non-modulated ventral features. We also defined a second feedforward strategy referred to as the benchmark system (BM), which builds the holistic representation by the concatenation of dorsal features and non-modulated ventral features. The reason for using the BM is that it provides another way to handle the scene features by directly adding the results of dorsal computation in the recognition representation. That said, the recognition network of BM has direct access to all information carried in both ventral and dorsal pathways. Thus, FRM and SI both include 14 features in the holistic representation, while BM has 50 features. The length of the vector in the holistic representation for FRM and SI is $25 * 14 = 350$, and it is $50 * 25 = 1250$ for BM.

Test sequences include 3 scenarios of a university campus location, "ACB", "AnFpark" and "FDFpark" (see Fig.5). Each scenario includes 9 different scenes, with each scene under varied illuminating conditions. In our experiment, six clips of each scene are used to train the network, with another four clips (that are different from the training set) to test the performance.

Figure 5: ***Test sequences.*** *(introduced in [29]) From left to right: ACB, AnFpark and FDFPark.*

We apply the same one-hidden-layer back-propagation neural network (provided by the Neural Network Toolbox in Matlab) in all systems. To further simplify the computation and allow all the learning networks to have the same number of input nodes, we reduce the vector dimension of the holis-

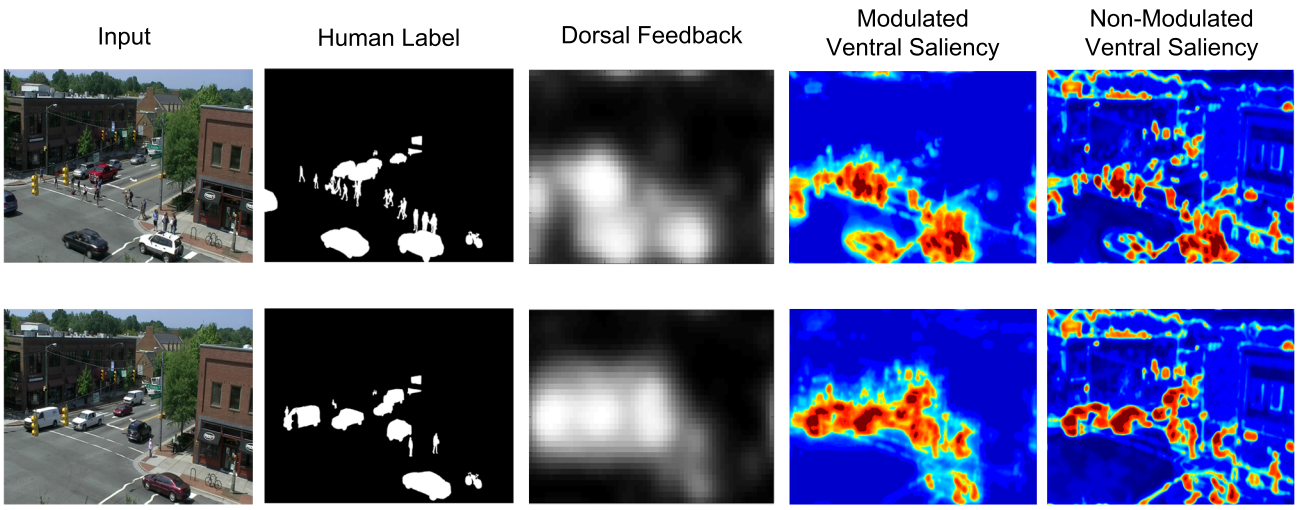

Figure 4: *An empirical study of the effects by the proposed fast recurrent modulation on image sequences. Visual saliencies based on the modulated ventral features are compared with non-modulated ventral saliency for image sequences. Also provided is the hand crafted ground truth labeling. From left to right: original image sequences, feedback strength elicited from the fast dorsal activation, visual saliency based on modulated ventral features, visual saliency based on non-modulated ventral features, and ground truth masks.*

Table 2: ***Comparison of recognition performance.*** *The percentage indicates the correctness rate, which is computed as number of correctly recognized scenes divided by number of total test scenes, as per scenario. The proposed method (FRM) consistently outperforms Siagian and Itti's method (SI) and the benchmark method (BM). Performance gain of BM over SI is also seen.*

Scene	SI	FRM	BM
ACB	90.43%	93.84%	91.25%
AnFpark	90.62%	91.45%	91.22%
FDFpark	90.26%	93.16%	92.41%

tic representation to 80 using principal component analysis available in [19]. Therefore, the input layer of the network includes 80 nodes. The output layer contains 9 nodes, with each corresponding to a scene within a scenario. The neural network contains one hidden layer of 100 nodes. For fair comparisons and to exclude the performance gain introduced outside the proposed fast recurrent modulation, all tests use the same set of network parameters. The cut-off of convergence that we use for all cases is set to 2%.

Table 2 provides a quantitative comparison of performance achieved by the three systems to correctly recognize a scene, as per scenario. Performance is measured by recognition correctness, the ratio between the number of true positives and the number of all test samples. We see from the table that FRM outperforms the other two systems for all scenarios. Therefore, the empirical conclusion drawn in the previous section that the fast recurrent modulation is able to provide a better figure-ground segmentation to facilitate object recognition is confirmed in this quantitative study.

5. Discussion

In this paper, we have proposed a computational model to describe the attentive mechanism of the fast recurrent modulation. In particular, we detailed the process by which results of computation from higher dorsal areas are used to inhibit the computation of lower ventral areas. The main role of the modulation is to improve object segmentation and further to facilitate high level visual tasks.

At first blush, it may seem that this work is simply a re-incarnation of the well-known Gist model of [28], or the Spatial Envelope of [26]. This however, is not the case. Oliva and colleagues focus on methods to improve the recognition of whole scenes, and to be sure, we do use this task as an example. Their Gist approach is motivated by the "proposition that top-down information from visual context modulates the saliency of image regions during the task of object detection". The visual context referred to is "based on a model of contextual priors (that learns the relationship between context features and the location of the target during past experience)". They integrate these priors with a simple model of image saliency. We too modulate the saliency of image regions but with local image characteristics computed independently and differently than the local image characteristics that go into saliency computation. Their Spatial Envelope approach employs a set of perceptual dimensions (naturalness, openness, roughness, etc.) that represent the dominant spatial structure of a scene. They show that these dimensions may be reliably estimated using spectral and coarsely localized information. In general, theirs is a "whole scene" approach in both cases. On

the other hand, ours is a local, image-based, approach that takes advantage of different features that are computed with different speeds in the visual system and thus can positively affect each other via fast recurrence.

We conducted empirical studies on both single images and image sequences to demonstrate the inhibitory effects of the modulation. The specific examples show the impact of this recurrent inhibition on the computation of a saliency map and for the task of scene recognition. In both bases, noticeable improvements are observed. Our direct analysis has focused only on MT-V1 recurrence, and as noted earlier, possibilities for recurrent contextual modulation may exist at other levels of the visual processing hierarchy. It will be a challenge to determine what semantic information is computed by each layer and how it may productively modulate lower layers in order to facilitate ventral visual processes.

Acknowledgments

This research was supported by the Teledyne Scientific Company and the Canada Research Chairs program. We thank Mazyar Fallah and Mario Aguilar for discussion and literature pointers.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, 1985. 52

[2] C. H. Anderson and D. C. Van Essen. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Natl. Acad. Sci. U.S.A.*, 84(17):6297–6301, 1987. 49, 50

[3] A. Angelucci and J. Bullier. Reaching beyond the classical receptive field of v1 neurons: horizontal or feedback axons? *J. Physiol. Paris*, 97(2-3):141 – 154, 2003. 50

[4] M. J. Arcaro, S. A. McMains, B. D. Singer, and S. Kastner. Retinotopic organization of human ventral visual cortex. *J. Neurosci.*, 29(34):10638–10652, 2009. 51

[5] D. C. Bradley and M. S. Goyal. Velocity computation in the primate visual system. *Nat. Rev. Neurosci.*, 9(9):686–695, 2008. 52

[6] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *J. Vis*, 9(3), 2009. 49, 52

[7] J. Bullier. Integrated model of visual processing. *Brain Res. Rev.*, 36(2-3):96 – 107, 2001. 49, 50

[8] J. Bullier. Cortical connections and functional interactions between visual cortical areas. In M. Fahle and M. Greenlee, editors, *Neuropsychol. Vis.*, pages 251–256. Oxford University Press, 2003. 50

[9] M. C. Bushnell, M. E. Goldberg, and D. L. Robinson. Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention. *J. Neuropsychol.*, 46(4):755–772, 1981. 50

[10] A. M. Derrington and P. Lennie. Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. *J. Physiol.*, 357(1):219–240, 1984. 51

[11] D. C. V. Essen and J. H. Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends Neurosci.*, 6:370 – 375, 1983. 50

[12] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47, 1991. 51

[13] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, 1987. 51

[14] K. Fukushima. A neural network model for selective attention in visual pattern recognition. *Biol. Cybern.*, 55:5–16, October 1986. 49

[15] J. D. Golomb, M. M. Chun, and J. A. Mazer. The native coordinate system of spatial attention is retinotopic. *J. Neurosci.*, 28(42):10654–10662, 2008. 51

[16] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, 148(3):574–591, 1959. 51

[17] J. M. Hupé, A. C. James, B. R. Payne, S. G. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nat.*, 394:784–787, 1998. 50

[18] J. M. Hupé, A. C. James, B. R. Payne, S. G. Lomber, P. Girard, and J. Bullier. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nat.*, 394(6695):784–787, 1998. 50

[19] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999. 55

[20] E. Kaplan, S. Marcus, and Y. T. So. Effects of dark adaptation on spatial and temporal properties of receptive fields in cat lateral geniculate nucleus. *J. Physiol.*, 294(1):561–580, 1979. 51

[21] K. Kawano, M. Shidara, Y. Watanabe, and S. Yamane. Neural activity in cortical area MST of alert monkey during ocular following responses. *J. Neurophysiol.*, 71(6):2305–2324, 1994. 50

[22] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.*, 4(4):219–227, 1985. 49

[23] J. H. Maunsell and J. R. Gibson. Visual response latencies in striate cortex of the macaque monkey. *J. Neurophysiol.*, 68(4):1332–1344, 1992. 50

[24] P. M. Milner. A model for visual shape recognition. *Psychoact. Rev..*, 81(6):521 – 535, 1974. 49

[25] L. Nowak, M. Munk, P. Girard, and J. Bullier. Visual latencies in areas v1 and v2 of the macaque monkey. *Visual Neurosci.*, 12(02):371–384, 1995. 50

[26] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001. 55

[27] A. Pascual-Leone and V. Walsh. Fast Backprojections from the Motion to the Primary Visual Area Necessary for Visual Awareness. *Science*, 292(5516):510–512, 2001. 50

[28] P. G. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. *Psychol. Sci.*, 5:195–200, 1994. 55

[29] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(2):300 –312, 2 2007. 53, 54

[30] J. K. Tsotsos. A complexity level analysis of immediate vision. *Int. J. Comput. Vision*, 1:303–320, 1988. 49

[31] J. K. Tsotsos, S. M. Culhane, W. Y. K. Winky, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artif. Intell.*, 78(1-2):507–545, October 1995. 49

[32] J. K. Tsotsos, A. J. Rodrguez-Snchez, A. L. Rothenstein, and E. Simine. The different stages of visual recognition need different attentional binding strategies. *Brain Res.*, 1225:119 – 132, 2008. 49

[33] L. G. Ungerleider and M. Mishkin. *Two Cortical Visual Systems*, chapter 18, pages 549–586. 1982. 50

[34] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition - a gentle way. *Br. Mach. Vision Conf.*, pages 472–479. Springer, 2002. 49

[35] D. Walther, U. Rutishauser, C. Koch, and P. Perona. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. In *Comput. Vision Image Understanding*, pages 41–63, 2005. 49