Biologically Motivated Local Contextual Modulation Improves Low-Level Visual Feature Representations

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

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

Abstract. This paper describes a biologically motivated local context operator to improve low-level visual feature representations. The computation borrows the idea from the primate visual system that different visual features are computed with different speeds in the visual system and thus they can positively affect each other via early recurrent modulations. The modulation improves visual representation by suppressing responses with respect to background pixels, cluttered scene parts and image noise. The proposed local contextual computation is fundamentally different from exiting approaches that involve "whole scene" perspectives. Context-modulated visual feature representations are tested in a variety of existing saliency algorithms. Using real images and videos, we quantitatively compare output saliency representations between modulated and non-modulated architectures with respect to human experimental data. Results clearly demonstrate that local contextual modulation has a positive and consistent impact on the saliency computation.

Keywords: local context, visual saliency, recurrent modulation.

1 Introduction

A biologically motivated local context operator is proposed to improve low-level visual feature representations. Specifically, the computation is inspired by knowledge of the primate visual system that visual features are calculated through visual pathways with different speeds and thus they can positively affect each other via recurrent modulations [6].

It has been proposed in the literature that visual context influences visual perception to a great extent [10]. For example, knowing the context of a workbench improves performance in searching for tools. However, the optimal representation of visual context is still unclear. Most existing theories view context from a "whole scene" perspective. In [16–18], context is a low-dimensional representation of the whole image, termed as "gist" or spatial envelope. The purpose of employing such a holistic representation is to identify scene types and to use

* Indicates equal contribution.

A. Campilho and M. Kamel (Eds.): ICIAR 2012, Part I, LNCS 7324, pp. 79–88, 2012.

the identified scene knowledge as a prior to predict spatial locations of specific objects. In [2], context is defined as scenery structure, and it is used to predict the likelihood of an object being present. Although these works have been shown to benefit object processing, there are other biologically motivated routes to use context in vision, specifically at less abstract representational levels.

In this paper, visual context is considered from a local feature representational perspective, which is fundamentally different from existing models. The work is motivated from recent studies of the asynchronous visual feature processing of the primate visual system [6,11,15]. Specifically, it is noted that: 1) visual signals are processed via the two main visual pathways at different speeds, with signals projecting through the dorsal pathway significantly faster than those through the ventral pathway, and 2) there exist recurrent connections that cross the two pathways, namely, from higher-level dorsal regions to lower-level ventral regions. It is thus highly likely that our brain utilizes these speed differences to apply dorsal percepts to positively improve ventral processing via recurrent mechanisms, which consequently impact later computations in the ventral pathway.

In a previous study [19], a computational model of local context is developed to improve visual feature representations. The computation first extracts visual features of different types, which are consistent with the primate dorsal and ventral pathway characteristics. The model then uses "dorsal" features to modulate "ventral" features via multiplicative inhibition. The modulation improves "ventral" representation by suppressing responses with respect to background pixels, cluttered scene parts and image noise, all of which could negatively impact processing of a real target.

To investigate whether the proposed local context operator can improve visual feature representations, it is convenient to test our work on several existing visual saliency models. The influence of local contextual modulation can be directly reflected through saliency performance. Experiments using real world images and videos have been conducted. Comparison to human experimental data indicates that contextually modulated feature representations have positive and consistent impact on the saliency performance.

The rest of the paper is organized as the follows. Section 2 introduces the general model of local contextual modulation. Second 3 formalizes the computational components to capture the idea. Section 4 describes the experimental procedure and test results. Section 5 concludes the work.

2 Modeling Local Contextual Modulation

The model of local contextual modulation is informed by our knowledge of the structure and functions of the two-pathway visual hierarchy [1, 20]. The dorsal pathway receives input from magnocellular layer of the lateral geniculate nucleus (LGN) and continues via dorsal layers of the primary visual cortex (V1), middle temporal cortex (MT) to the posterior parietal cortex for motion and action perception. The ventral pathway starts from parvocellular layers of the LGN, through ventral layers of V1, V2, V4 to the inferior temporal cortex for high

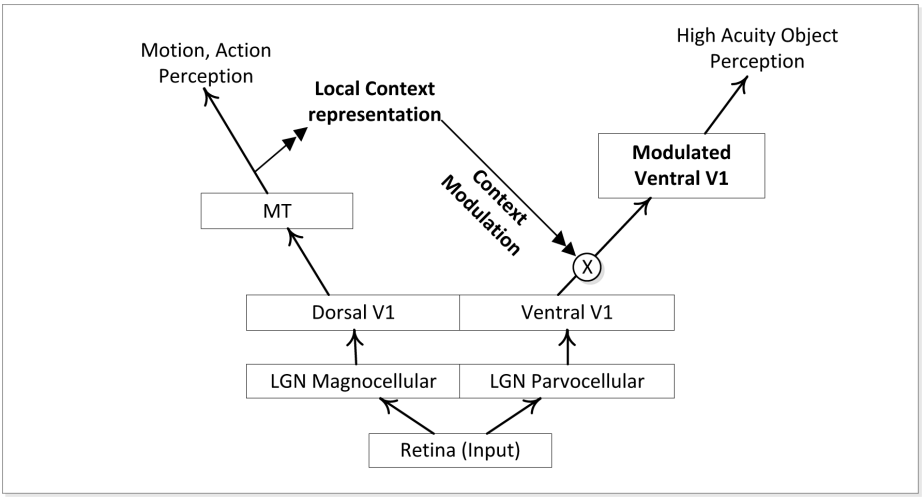

Fig. 1. A hierarchical view of the proposed model. Visual input is computed along the dorsal and ventral pathway. The double arrow lines depict the process of local context is created from MT and is used to modulate ventral V1 output.

acuity object perception. Magcellular cells are achromatic, have a higher peak sensitivity to contrast and respond to higher temporal frequencies than parvocellular cells, while parvocellular cells are color-sensitive, respond to higher spatial frequencies and show higher sensitivity at the lowest temporal frequencies. The model further relies on the facts that: 1) the two pathways compute different visual features, 2) pathway connections have different conduction speeds, with the dorsal pathway conducting signals faster than the ventral pathway, and 3) timing of the process allows that fast dorsal percepts can influence ventral processing via recurrent modulation.

Figure 1 illustrates the general flowchart of the proposed model. Two sets of visual features are separately calculated, corresponding to the dorsal features and the ventral features. In our case, they are defined differently in spatiotemporal scale. Specifically, our dorsal representation employs high-temporal-low-spatial scales and the ventral representation includes low-temporal-high-spatial scales. This is consistent with the neurophysiological properties of the primate visual system [7].

It is important to note that the ventral computation represents the usual kind of feature processing seen in other models. On the other hand, the local context representation is formed based on the dorsal features, and is used to modulate ventral computation. Therefore later stages of (ventral) visual processing reflects inhibited signals as opposed to the usual (non-inhibited) outputs.

3 Formalization

This section discuss the computational components needed to formalize the model. Before defining the modulation, computations in the two visual pathways

are detailed. The modeled hierarchy starts from the retina[1], via LGN to V1 and MT. Computation at each layer is represented as a bank of image filters.

The center-surround receptive fields (RFs) of the LGN have spatial response patterns consistent with Difference-of-Gaussians filter [13] given by:

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

where σ_c and σ_s are the bandwidth for the center and surround Gaussian respectively. In our implementation, they are set for magnocellular ($\sigma_c = 3$, $\sigma_s = 4.8$) and parvocellular ($\sigma_c = 1$, $\sigma_s = 1.6$) respectively. The LGN temporal response pattern is described as a log-Gabor filter [8] defined in the frequency domain as:

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

where w_0 is the center temporal frequency, σ_t is the bandwidth. We employ a multi-scale temporal filter bank to provide an even spectrum coverage by using different w_0 and σ_t. Specifically, for parvocellular $w_0 = 3, 9, 27$, and for magnocellular $w_0 = 9, 27, 81$. $\sigma_t = 0.55w_0$ for both cases.

Area V1 receives feed-forward projections from LGN and integrates energy along different spatiotemporal orientations. The spatial selectivity is described as a 2D orientated log-Gabor filter defined in frequency domain as:

$$F_{V1spatial}(u,v) = \exp\{\frac{-log(u_1/u_0)^2}{2log(\sigma_u/u_0)^2}\} \cdot \exp\{\frac{-v_1^2}{2\sigma_v^2}\} \quad (3)$$

where $u_1 = u\cos(\theta) + v\sin(\theta)$, $v_1 = -u\sin(\theta) + v\cos(\theta)$, θ denotes the orientation. Our implementation has 6 orientations. u_0 denotes the center spatial frequency, σ_u and σ_v denote the spatial bandwidth along u and v axis respectively. In our case, ventral V1 uses $u_0 = 9, 27, 81$ and dorsal V1 uses $u_0 = 3, 9, 27$. The bandwidths in both cases are set to $\sigma_u = 0.55u_0$ and $\sigma_v = 0.55u_0$. The temporal profile of a V1 neuron is defined as a lowpass filter in time domain as:

$$f_{V1temporal}(t) = \exp\{\frac{-t^2}{2\sigma_t^2}\} \quad (4)$$

where σ_t denotes the temporal bandwidth. In our implementation, $\sigma_t = 1$.

Area MT integrates opponent energy [3] provided by dorsal V1 as:

$$MT_\theta(x,y,t) = \sum_{\Delta x,\Delta y,\Delta t} V1_\theta(x,y,t) - \sum_{\Delta x,\Delta y,\Delta t} V1_{\theta+\pi}(x,y,t) \quad (5)$$

where $\sum$ denotes the summation of dorsal V1 energy over range $(\Delta x, \Delta y, \Delta t)$.

The output of MT is further integrated across spatiotemporal orientations to produce the **local context representation**, which is then used to modulate ventral V1 representations. Although several aspects of computation at the

[1] To simplify the computation, the retina is represented directly by the input image without any processing.

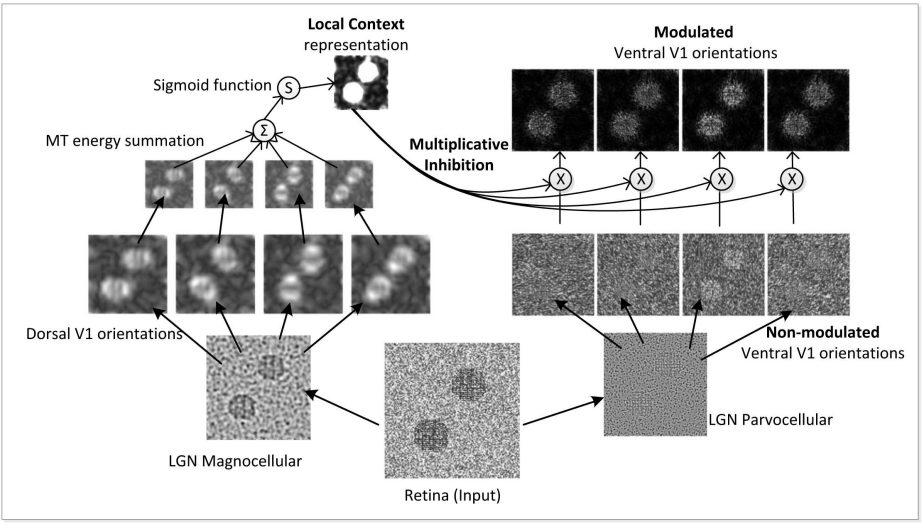

Fig. 2. A demonstration of the proposed computational model of local contextual modulation. Stimuli: two circles of plaid pattern in front of uniformly distributed pseudorandom noisy background. Both pathways compute pure spatial features. The dorsal pathway is sensitive to lower frequency (coarse scales) features to generate an early recurrent representation. Higher frequency orientation (fine scales) features being calculated by the ventral pathway are then multiplicatively inhibited by the context representation. Result: The noisy background is suppressed while targets remain.

neuronal level remain to be characterized, experimental results of orientation-selective neurons [9, 14] have suggested that the biophysical underpinning of such an operation can be described mechanistically by multiplication. Further, the goal of modulation is to improve feature representation by suppressing inconsistent responses. Thus, we propose the inhibition process between MT and ventral V1 as **multiplicative inhibition**. Since it is unclear where the MT-ventral V1 feedback fibers terminate, we have considered two possible locations: input and output of ventral V1. Our tests show similar results. Therefore, in what follows we assume the modulation is at the output of ventral V1 as:

$$V1'_V(x,y,t) = V1_V(x,y,t) * Sig(\sum_{\theta} MT_\theta(x,y,t)) \tag{6}$$

where $V1_V$ denotes output of ventral V1, $V1'_V$ is modulated output, and $Sig()$ is a sigmoid function used to rectify the MT output in nonlinearly between 0..1.

To briefly illustrate the principles, Figure 2 shows an example of selecting two simulated circles with plaid patterns in front of uniformly distributed noise. In the example, V1 has 4 orientations. It clearly shows that dorsal V1 highlights object regions (but without spatial details). Non-modulated ventral V1 extracts details but of both objects and background, by which boundaries of the circles are unclear. MT integrates the output of dorsal V1 to produce a context representation. The representation then inhibits output of ventral V1, after which

background activations are suppressed, leaving the two circles standing out. The system is then able to process the targets without background interference.

4 Experiments

In order to enable a quantitative evaluation of the impact of local context, it is convenient to incorporate the modulated feature representations into a number of feed-forward models of visual saliency [5,12,21]. The purpose of the evaluation is to determine the effect of contextual modulation *as a general and intermediate process* to the existing works. Feed-forward models used in the evaluation capture the computation of visual saliency from different perspectives. In particular, saliency in [12] is defined as strength of summed visual feature activations, while in the other two proposals [5, 21], visual saliency arises from measuring self-information (but differently) based on natural image statistics. It is thus natural to deem these original models as starting points that provide baseline performance. The proposed computation fits itself easily into these models by applying the context representation to modulate feature maps provided by the models. This is such that in the revised models, saliency representations are calculated based on the modulated feature maps. One can then evaluate to what extent context improves performance over baseline scores.

Saliency performance is measured using receiver operating characteristic (ROC) curves, which have been widely used in related works. For a given ground truth (G) and a saliency map (S), saliency values are normalized between [0...1]. By varying the threshold $\delta \in [0...1]$, a smooth curve is generated as true positive ($G(x,y) >= \delta$ and $S(x,y) >= \delta$) rate versus false positive ($G(x,y) < \delta$ and $S(x,y) >= \delta$) rate, where (x,y) is pixel coordinate.

The implementations are tested with cluttered images [5] to evaluate spatial context, and tested with surveillance videos spatiotemporal context.

Figure 3 compares spatial modulation. Saliency maps produced by the original models and the modulated versions are paired in groups. Reddish pixels indicate salient regions. It is clearly shown that there are more similarities between real objects in the input images and the reddish regions in the *modulated* saliency maps (right image of each pair) than the reddish regions in the *original* saliency maps (left image of each pair). The main difference between a pair of saliency maps lies in the fact that a substantial amount of cluttered background is inhibited. For each input image, salient regions produced by the three original models are different. However, in the modulated versions, salient regions are all confined to the context representation, leaving the remaining regions mostly in blue (not salient).

Mean ROC curves are generated based on human fixation densities [5]. From the lower charts of Figure 3, it is obvious that curves produced from modulated saliency maps (solid lines) augment their original works (dashed lines) significantly. Areas under mean ROC curves are calculated. As concluded in Table 1, modulation raises areas under curves in all cases, which further confirms that local contextual facilitation is generally effective in improving different saliency measurements.

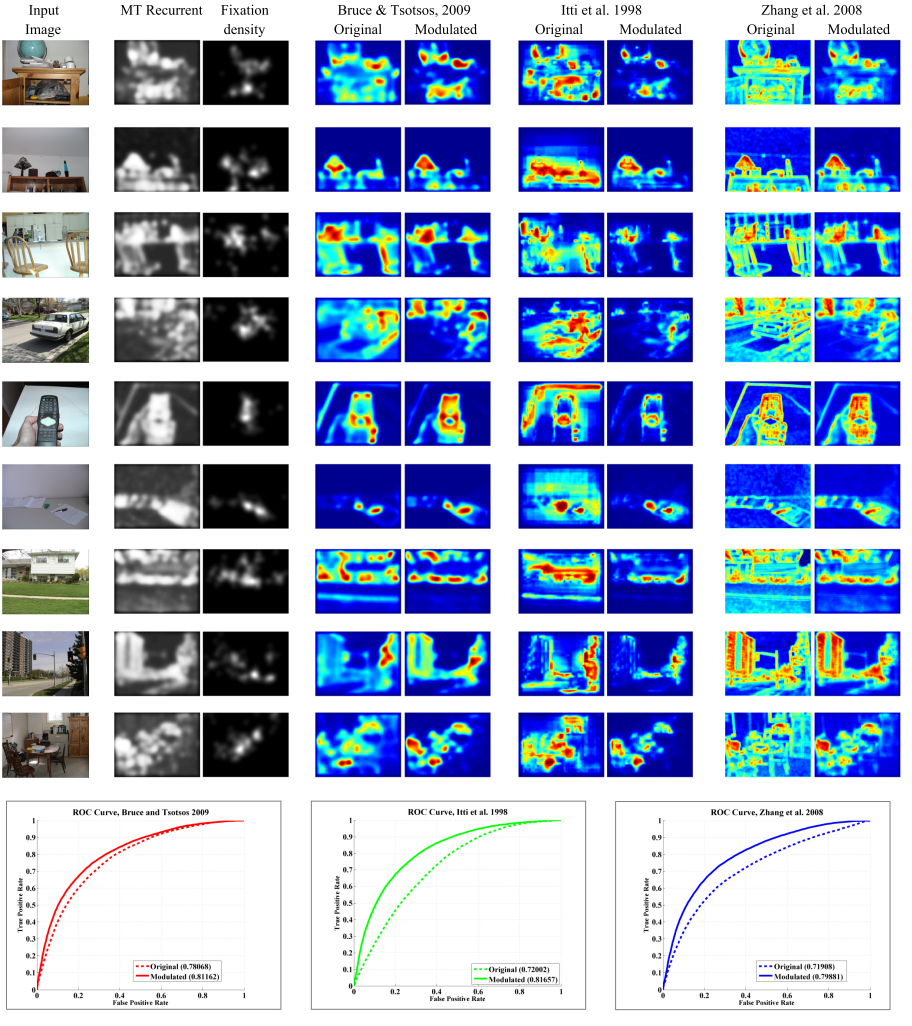

Fig. 3. Use early recurrence to improve visual saliency. Up: Impact of modulation for three different saliency algorithms [5, 12, 21]. **Down:** Mean ROC to access saliency improvements between modulated saliency maps (solid lines) and non-modulated saliency maps (dashed lines). Numbers in brackets indicate areas under curves.

Table 1. Comparison of areas under mean ROC curves

	Original	**Modulated**	Improvements
Bruce & Tsotsos 2009 [5]	0.781	**0.812**	+3.97%
Itti et al. 1998 [12]	0.720	**0.817**	+13.47%
Zhang et al. 2008 [21]	0.719	**0.799**	+11.13%

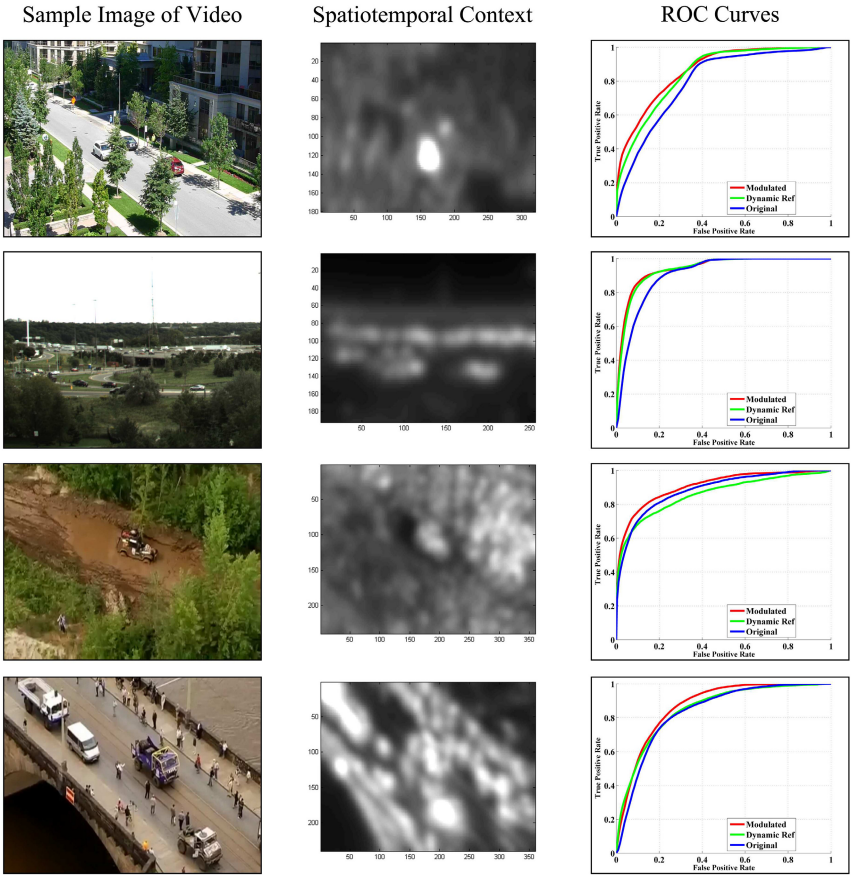

Fig. 4. Use early spatiotemporal context to improve visual saliency. Left: image from test videos. Middle: early recurrent representation that highlight regions consisting with moving objects. Right: Mean ROC curves of original AIM [5] (blue), a spatiotemporal alternative [4] (green) and modulated AIM (red).

Figure 4 illustrates how spatiotemporal (motion) context is involved in improving saliency maps. Videos are shot from various viewing angles and under different lightening conditions. Targets (i.e. vehicles and pedestrians) are manually labeled and saved as ground truth files for evaluation. As shown, context clearly highlights regions corresponding to moving stimuli in most cases, leaving stationary and cluttered parts suppressed. Improvement over original saliency model [5] is obvious by comparing the mean ROC curves (red lines versus blue lines). Also compared is a reference model similar to [4], where saliency is computed using output of spatiotemporal filters (green lines).

5 Conclusion

In this paper, we have proposed a novel computational approach to visual information processing, which is inspired by research of the primate visual system. In its most simplified form, the model applies local context (activations of dorsal regions) to improve visual feature representations (computed in the ventral pathway). The modulation improves visual representation by suppressing responses with respect to background pixels, cluttered scene parts and image noise.

The proposed contextual modulation is a local and image-based operation, which is different from existing context models involving scene *Gist* [16–18]. Although both approaches are proposed with the goal of having contextual representation affect visual perception, their motivations and biological foundations are different. The main focus of Oliva and colleagues is to use context to prime the input image with regions that are most likely to contain targets. Such a process is based on a model that learns target features and locations from past experience. Context described in our work, on the contrary, captures the characteristics of early visual cortical structures and computational principles, that context exerts its influence at a very early stage.

The work has been applied to three saliency models to investigate the influence of local context. Quantitative analysis is provided. Saliency maps generated based on the modulated feature representation significantly augment their non-modulated versions. The results clearly demonstrate that our proposed local contextual modulation is a robust and generally applicable process. In this paper, contextual modulation has been focused primarily to simulate MT-ventral V1 recurrent processing, possibilities for similar modulation process may exist between other form of visual computation. This presents additional fruitful avenue for further work.

Acknowledgments. This research was supported by the Teledyne Scientific Company.

References

1. Anderson, C.H., Van Essen, D.C.: Shifter circuits: a computational strategy for dynamic aspects of visual processing. PNAS 84(17), 6297–6301 (1987)
2. Bar, M.: Visual objects in context. Nat Rev Neurosci 5(8), 617–29 (Aug 2004)
3. Bradley, D.C., Goyal, M.S.: Velocity computation in the primate visual system. Nature Reviews Neuroscience 9(9), 686–695 (August 2008)
4. Bruce, N.D., Tsotsos, J.K.: Spatiotemporal saliency: Towards a hierarchical representation of visual saliency. In: Paletta, L., Tsotsos, J.K. (eds.) Attention in Cognitive Systems, pp. 98–111. Springer-Verlag, Berlin, Heidelberg (2009)
5. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9(3) (2009)
6. Bullier, J.: Integrated model of visual processing. Brain Research Reviews 36(2-3), 96 – 107 (2001)
7. Derrington, A.M., Lennie, P.: Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque. The Journal of Physiology 357(1), 219–240 (1984)

8. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A 4(12), 2379–2394 (Dec 1987)
9. Gabbiani, F., Krapp, H.G., Koch, C., Laurent, G.: Multiplicative computation in a visual neuron sensitive to looming. Nature 420(6913), 320–4+ (2002)
10. Henderson, J.M., Hollingworth, A.: High-level scene perception. Annual Review of Psychology 50(1), 243–271 (1999)
11. Hupé, J.M., James, A.C., Payne, B.R., Lomber, S.G., Girard, P., Bullier, J.: Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature 394(6695), 784–787 (August 1998)
12. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (Nov 1998)
13. Kaplan, E., Marcus, S., So, Y.T.: Effects of dark adaptation on spatial and temporal properties of receptive fields in cat lateral geniculate nucleus. The Journal of Physiology 294(1), 561–580 (1979)
14. McAdams, C.J., Maunsell, J.H.R.: Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. The Journal of Neuroscience 19(1), 431–441 (1999)
15. Nowak, L., Munk, M., Girard, P., Bullier, J.: Visual latencies in areas v1 and v2 of the macaque monkey. Visual Neuroscience 12(02), 371–384 (1995)
16. Oliva, A.: Gist of the scene. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) The Encyclopedia of Neurobiology of Attention, pp. 251–256. Elsevier, CA (2005)
17. Oliva, A., Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Int. J. Comput. Vision 42(3), 145–175 (May 2001)
18. Schyns, P.G., Oliva, A.: From blobs to boundary edges: Evidence for time and spatial scale dependent scene recognition. Psychological Science 5, 195–200 (1994)
19. Shi, X., Bruce, N., Tsotsos, J.: Fast, recurrent, attentional modulation improves saliency representation and scene recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. pp. 1–8 (june 2011)
20. Ungerleider, L.G., Mishkin, M.: Two Cortical Visual Systems, chap. 18, pp. 549–586 (1982)
21. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision 8(7) (2008)