An Attentional Framework for Stereo Vision

Neil D. B. Bruce and John K. Tsotsos
Department of Computer Science, York University, Toronto, Ontario
Centre for Vision Research, York University, Toronto, Ontario
{neil, tsotsos}@cs.yorku.ca

Abstract

The necessity and utility of visual attention are discussed in the context of stereo vision in machines and primates. Specific problems that arise in this domain including binocular rivalry, and the deployment of attention in three-dimensional space are considered. Necessary conditions are outlined for achieving appropriate attentional behaviour in both the aforementioned domains. In this light, we outline classes of existing computational models of attention and discuss their applicability for realizing binocular attention. Finally, a stereo attention framework is presented by considering the tenets of an existing attentional architecture that extends naturally to the binocular domain, in conjunction with the connectivity of units involved in achieving stereo vision.

1. Introduction

Stereo vision is an important mechanism in animal and machine vision, allowing judgements to be made based on disparity between the images captured by each eye. However, the consideration of two eyes introduces a variety of issues pertaining to the nature of the (neural) circuitry involved that warrant a closer look. The following work considers specifically the relationship between stereo vision and visual attention with consideration to biological plausibility. We propose a stereo model of visual attention based on an existing attention model, the Selective Tuning model, that extends naturally to the binocular domain. The formulation of Selective Tuning is based largely on theoretical issues relating to computational complexity, and resolving issues of signal interference inherent in a visual pyramid processing context.

The sections that follow take a close look at issues that arise in combining attention with binocular vision. Inspired by the primate visual system, we demonstrate that a distributed pyramid architecture with appropriate disparity selective neural hardware may afford a means of generating appropriate shifts of attention in three dimensional space and allow resolution of interference that arises when there exists conflict between the inputs received by the two eyes. No previous efforts towards achieving attentional processing in a biologically plausible stereo vision architecture are found in the literature. In this light, we consider the capacity of other attentional architectures to be extended to stereo visual processing revealing potential confounds for a variety of attention models with regard to their ability to explain primate visual attention, and to initiate appropriate shifts of attention in three-dimensional space.

2. The Need for Attention

Numerous previous works have addressed the necessity of visual attention in animals and machines (See for example [1,2,3]). As such, the following section does not endeavor to present a novel argument for *why* attention should be considered in the context of machine vision. Instead, the discussion is limited to highlighting a set of important principles central to the issue of why attention is needed, with a focus on those that impact on the derivation of a stereo attention framework for machine vision.

Arguments in favor of attentive behaviour most often focus on the issue of computational complexity. There exist some rather compelling arguments demonstrating that the task of visual search, in the absence of an explicit target, is NP complete [2,3]. On the basis of this observation, one might conclude that the general form of this problem is not solved by the human brain. There exists a variety of rather convincing psychophysical results in favor of this hypothesis [2], and implicating attention as a mechanism for resolving the dilemma posed by the complexity of the problem.

A second issue highlighted in [4] that receives relatively less attention, involves issues of signal interference inherent in a pyramid processing architecture. This second issue is of particular importance in the context of stereo vision. Signal interference in a pyramid processing architecture may be seen as arising from one of four situations, depicted in figure 1: (For a more detailed description see [4])

a. Context Effect: The response of a given unit at the top layer of the pyramid is derived from a large number of units at the input layer. As such, the context of the stimulus at the input affects its representation at the highest layer.

b. Blurring: Each input unit ultimately projects to a large number of units at the output layer. This introduces significant difficulty in localizing the source of a given response.

c. Cross-talk: Because multiple units at a given layer converge on a single unit at the next layer, the response represents interference between distinct events within those multiple units. This issue is of particular importance in the context of stereo vision since typically, there exist many neurons in the human visual system that respond to input from the two eyes. Often the input from each eye is in agreement in which case interference is not an issue. However, in the event that the two eyes receive disparate input, an appropriate mechanism is required to resolve such difficulties. One might claim that such a situation is highly artificial, and not a concern in practice. This is far from the truth; interference will arise sporadically as a result of monocular occlusion. Further, in primates instances arise in which disparity is sufficiently large that binocular fusion does not occur. Interference in this context is a worthwhile consideration and may prove to be a useful component for complex machine vision systems. These considerations are further elaborated on in the sections that follow.

d. Boundary Effect: Input level activation corresponding to interpretive units at the side of the pyramid receive a lower weight among units at the top of the pyramid by virtue of their position.

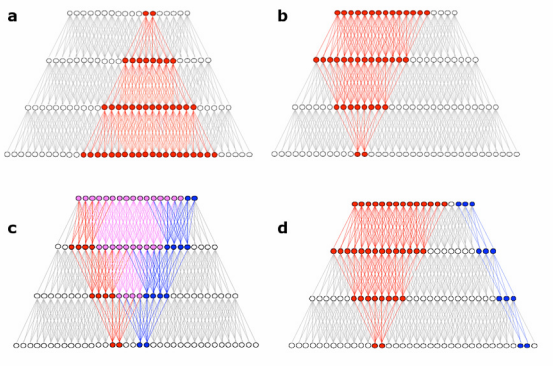

Figure 1. Various signal interference issues that arise in the context of pyramid processing.

Another role of attention in primates, is the preparation of various types of eye movements. Rapid shifts in gaze, known as saccades, allow fast deployment of fixation to a target of interest allowing expeditious sampling of the surrounding environment. The utility of such a mechanism in a machine vision context is also apparent. There also exist a variety of other types of eye movements involved in tracking, exploring the environment, and stabilizing the eyes, all of which require the consideration of binocular control. In the presence or absence of eye movements, intuitively we would expect the visual system to have the ability to attend to a single *world* event even in the case that such an event falls on non-corresponding retinal coordinates. Consider the situation where an observer is fixating on a close object and far in the background an event occurs to which attention is shifted. Such an event will fall on a very different position in each retinal image, and will require a very different movement of each eye taking into account the depth of the target. This raises an important consideration in realizing attention in stereo vision. The mechanism must allow for selection of non-corresponding regions in each eye, so that a single scene event may be selected and processed without interference and so that appropriate shifts in fixation might be initiated. This last issue is an important consideration in discussion that appears in the sections that follow.

3. An Attentional Framework for Stereo Vision

The formulation of a framework for attentive processing in stereo vision is achieved in the context of an existing model of attention, the Selective Tuning Model [4], that extends naturally to the binocular domain.

The Selective Tuning model was designed to specifically address the issues described in section 2. Although many models of attention recognize the need for dealing with the problem of complexity, few consider the secondary problem of interference inherent in pyramid processing. This is an issue that is especially important in a binocular context, since there is often mismatch between the two eyes making selection of an appropriate signal for further processing paramount. Further, a means of initiating appropriate movements of the two eyes is a significant issue in active machine vision.

Selective Tuning resolves the issues discussed thus far by implementing spatial selection via inhibition of relevant connections in a gating network, and feature selection through the inclusion of a bias network tied to the interpretive units in the computational architecture. The connectivity between two layers is depicted in figure 2. It is worth noting that this represents only a small portion of the overall pyramid architecture showing only a subset of units from only two layers. Green units correspond to interpretive units that respond maximally to a particular feature such as an edge, or a particular degree

of disparity. Orange units correspond to bias units, with each bias unit connected to an interpretive unit, and bias units from the previous layer. Black units consist of units in the gating network. The overall process of Selective Tuning proceeds as follows:

i. Input gives rise to activation in the interpretive units at the lowest layer of the pyramid. In our implementation, these correspond to simple oriented Gabor filters at orientations of 0, 45, 90, and 135 degrees acting as simple edge detectors.

ii. Activation in layer one gives rise to activation in layer 2, and subsequently higher layers. Interpretive units from these layers compute more complex features. Details of the units involved in our implementation follow the description of the operation of selective tuning.

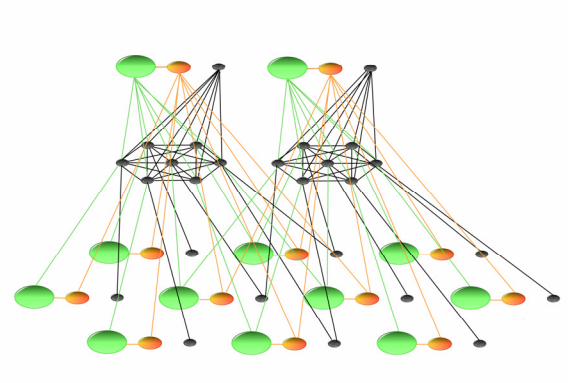

Figure 2. The computational architecture assumed by Selective Tuning. Green corresponds to interpretive units, orange to bias units, and black to gating units.

iii. Units at the highest layer of the pyramid are activated, and a winner-take-all (WTA) competition takes place among all units at this layer. The details of the WTA process are described in [4]. The nature of the winner-take-all competition is such that each unit involved in the competition inhibits all others by a degree proportional to its own response. A condition on this inhibition is that a unit inhibits another only if its value (firing rate) is greater than the to-be inhibited unit by at least θ. This guarantees convergence of the WTA competition and also allows selection of multiple winning units so that arbitrarily sized or shaped regions may be selected at the input layer.

iv. Upon selection of a winning unit(s), gating units that project to the winning unit from the previous layer compete in WTA competition to determine which unit(s) gave rise to the winning activation. This process is repeated at each layer down to layer 1. Any gating unit that competes in WTA competition and loses, no longer passes on the input that it receives from the interpretive unit that projects to it. The result of this process, is that the output of the winning unit at the highest layer is derived entirely from units throughout the pyramid that contribute appreciably to the resulting winner at the top layer. This means that signal interference from other events at the input layer is dramatically reduced. In particular, the problems depicted in figure 1 may be resolved. Further details on this last comment are described in [4].

v. Following WTA competition at layer 1, winning units shut off for one time cycle so that a new region may be attended.

Selective Tuning as described allows the successive selection of attended regions, addressing the issue of complexity, while also handling signal interference. The subsection that follows describes how such an architecture may implement the selection of attended regions in a binocular context, and the particular details of the interpretive units involved in such an implementation.

3.1. Selective Tuning in Stereo Vision

Specification of the stereo architecture we have assumed simply requires an explanation of interpretive units involved and connectivity among such units. The architecture is largely based on a primate model of binocular neural processing proposed by Ohzawa and Freeman [5,6]. It is important to state that the particular model employed for stereo computation is not central to the argument presented in this paper. The model of Ohzawa et al. has been selected on the basis of its popularity, and supporting neurophysiological evidence. The necessary conditions for the attentional framework described are that there exists many to one connections in the interpretive network, and at some point, inputs from the two eyes converge. Figure 3 demonstrates the various layers involved, and connectivity between feature planes. The following describes details of the implementation.

Layer 1 Gabor maps corresponding to the four orientations, and 8 spatial frequencies (3 by 2 to 17 cycles per 100 pixels) are derived from each of the left and right eye input. Such feature maps are produced for 4 different types of Gabor filters including 0 degrees phase positive, 0 degrees phase negative, 90 degrees phase positive, and 90 degrees phase negative.

Layer 2 Binocular simple cells tuned to various disparities (-40 to +40 pixels in increments of 10) are derived by summing the output of a Gabor filter at a particular spatial frequency, and orientation acting on each of the left and right eye input, and shifted by the degree of disparity in question. This gives rise to 1152 feature maps (4*8*9*4). Such binocular simple cells are found

among early visual areas, and often involve a differential contribution from each of the two eyes. In the context of our implementation, including this consideration would give rise to an appreciably larger number of feature maps and no generality is lost in our argument in assuming equal weighted inputs from each eye.

Layer 3 Complex binocular cells are produced by summing the squared output of 4 simple binocular neurons corresponding to the 4 different Gabor filter types for a particular orientation, spatial frequency and disparity. The choice of this operation is biologically motivated and means that output is not sensitive to contrast polarity, and disparity sensitivity is constant for all stimulus positions in the receptive field.

Layer 4 Responses are combined across orientation, and spatial frequency giving rise to 9 feature maps corresponding to the 9 disparities considered. This operation is suggested in the stereo model of Fleet et al. [7] which is also inspired by the energy model of Ohzawa [5]. Combining across orientation and spatial frequency reduces false peaks inherent in narrow band signals. This operation is also appropriate since most models of stereo vision will presumably rely on some form of pooling at a later stage of processing.

Layer 5 The 9 disparity maps are averaged to produce a single representation of disparity related activity.

In the context of Selective Tuning, WTA competition begins at layer 5 down to layer 1 eventually localizing the stimulus that gave rise to winning activation at the input layer. It may not yet be clear how the combination of the aforementioned binocular computational architecture and Selective Tuning achieves selection in three dimensional space and resolved issues of binocular interference. These issues are elucidated in the sections that follow accompanied by implementation results.

4. Shifting Attention in Depth

In this section, we describe how the aforementioned attentional architecture may localize stimuli in 3D space. Following this, implementation results derived from a number of synthetic, and one real image depict selection in the stereo domain.

It is straightforward to describe the manner in which selection of appropriate units in each eye is achieved given the principles of selective tuning, and the computational circuitry involved in the implementation presented here. For simplicity, we have restricted the number of winners at layers 2 through 5 to one (i.e. $\theta = 0$). This does not imply any loss of generality in the argument and simplifies the description considerably. At layer one, all of the units that project to the winning simple binocular cell at layer 2 compete for selection and multiple winners are allowed. This involves units from both the left and right eye. θ is equal to 0.3 times the value (firing rate) of the strongest unit involved in competition for figure 4, and 0.8 times in all other figures. These values were selected for visibility of the resulting figures. Selection involves a feedforward pass in which the response of each layer of the pyramid in figure 3 is computed. Following this, selection occurs for layers 5 down to 1. When a winner is selected for layer 2, the algorithm has localized a particular spatial frequency, orientation and disparity. The WTA process that follows for layer 1 determines if this selection is consistent with the input received by each eye, or arises

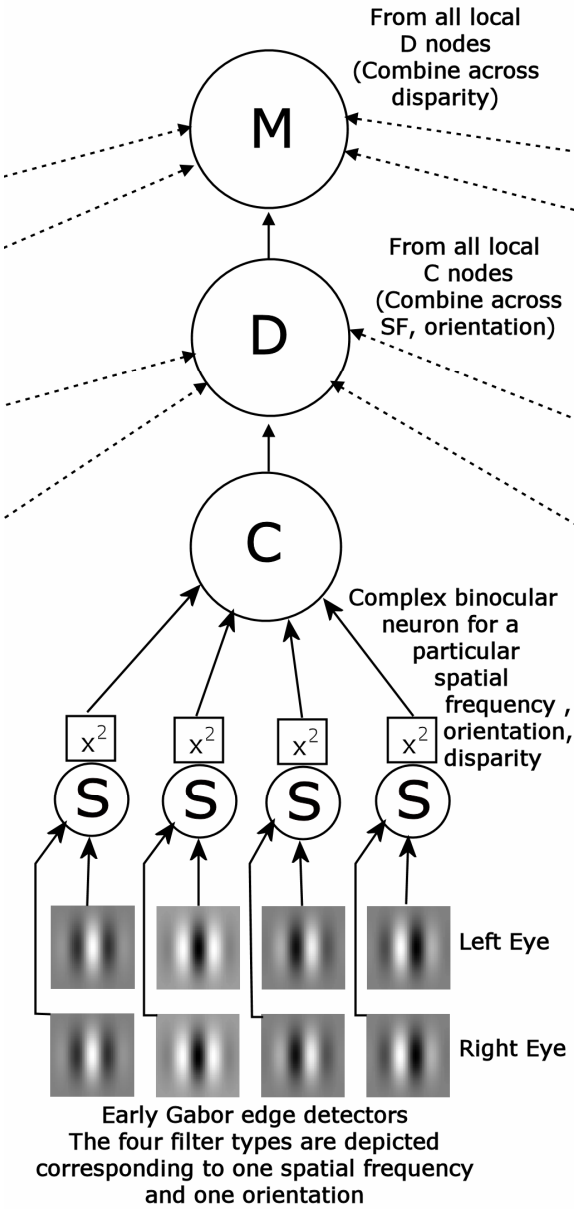

Figure 3. The computational architecture underlying disparity based computation.

from a strong response in one eye only. This might occur for a sufficiently bright edge that is occluded in one eye for example. The reason that appropriate coordinates in each eye are typically selected, is that the greatest activation at the highest layer results from a true correspondence which is traced back to layer 2 via WTA selection, and tied to the non-corresponding retinal coordinates that gave rise to such activation in layer 1, corresponding to a single scene event. Although the chosen architecture is a considerable simplification of the mechanism existent in primates, or even that which may achieve stereo correspondence in machine vision, it provides sufficient conditions to demonstrate how Selective Tuning achieves binocular selection. Any architecture that involves the convergence of inputs from the two eyes onto a single computational unit may achieve appropriate routing of information, and localization through Selective Tuning. Figures 4-8 demonstrate the results of one pass of Selective Tuning applied to the images corresponding to the left and right eyes in each case. Any pixels selected at the input, and their surrounding 8 neighbors (with the exception of Fig. 4) are labeled magenta. Figure 4 consists of a random dot stereogram with a square that appears above the background. Bias in this case is equal for all non-zero disparities and heavily against zero disparity. This results in correct localization of the square in each eye. Figure 5 demonstrates another simple synthetic case in which four passes of the algorithm with different bias parameters (listed in the caption) result in selection of the four squares that each appear at a different disparity. In this case, the selection of the appropriate orientation that gave rise to activation is visible as a single edge of the square is selected in each case. It is worth noting that both edges of the selected square and orientation might be localized if the receptive field size at layer 1 were larger, or if multiple winners were allowed at higher layers. Figure 6 depicts the real stereo pair on which figures 7-10 are based. Figure 7 demonstrates localization performed by the algorithm in the absence of bias. A strong edge that appears in both eyes shifted only slightly is selected. In Figure 8, a vertical bias is introduced, resulting in localization of a nearby vertically oriented edge. Note that the correspondence in the localized features is imperfect. This is attributable to the simplicity of the model, which makes no attempt to solve the correspondence problem, and does not consider rotational disparities. It is clear in each case that selection is bound to the localized stimulus and not a single set of coordinates for each eye.

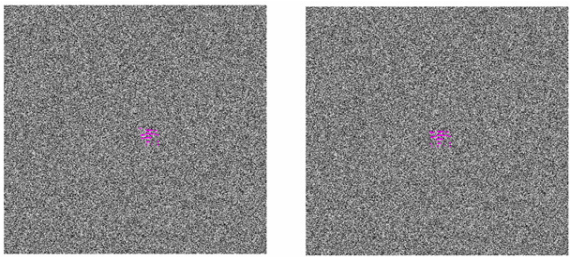

Figure 4. Selection of corresponding regions in each eye based solely on disparity information. Disparity corresponds to a shift of 20 pixels of a foreground patch on the background.

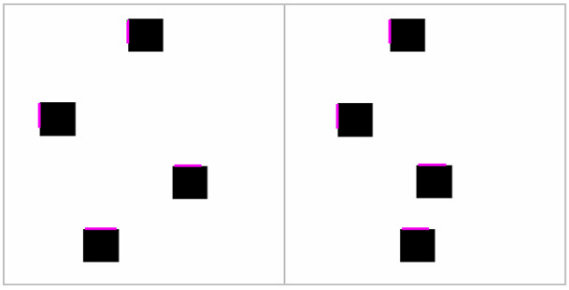

Figure 5. Outcome of four independent runs of selective tuning. In each case, a strong bias is initiated in favor of horizontal or vertical and one of –40, -20, 20, 40 pixels disparity. Correct correspondence is achieved in each case and corresponding event coordinates selected in each eye.

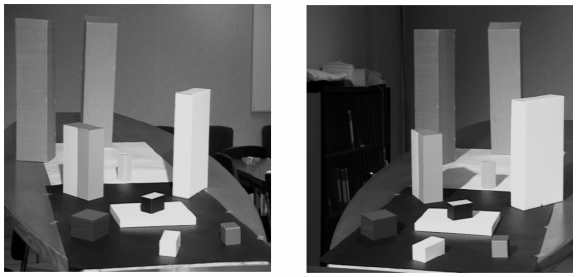

Figure 6. A stereo image containing a number of blocks, on which the output presented in figures 7-10 is based.

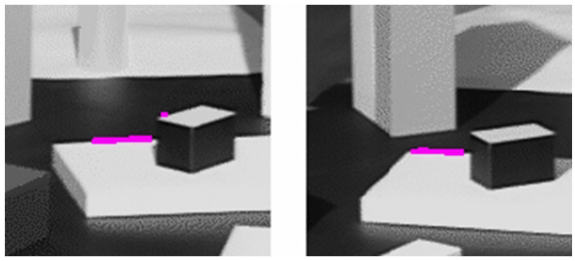

Figure 7. Selective Tuning on the image depicted in figure 6 with no *a priori* bias involved. A strong edge that appears in both the left and right eye is selected and correctly localized at the level of the input. A subsection of the image is depicted for increased visibility.

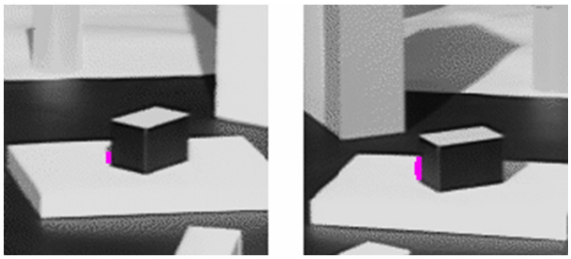

Figure 8. Selective Tuning on the image depicted in figure 6. with a bias for vertical edges. Because of the rotation of the block, and cast shadows, activation in the left eye is marginally weaker than the right and less pixels are selected as contributing to the winning unit at layer two. A subsection of the image is depicted for increased visibility.

5. Resolving Rivalry

Recently, attention has been demonstrated as having an important role in resolving conflict between the inputs received by each eye [8-10]. When the two eyes each view a different image, perception typically alternates between one image and the other. Given a pyramid processing architecture, the issue of interference is also apparent in the context of machine vision. A mobile robot with stereo vision is likely to sporadically encounter monocular occlusion while navigating its surroundings. Implementation results in this section employ the same parameters as those appearing in section 4. Conflict between the two eyes is resolved anywhere that neurons from the two eyes converge on a single unit. If the response of the binocular unit is derived primarily from the left or right eye, selection that ensues at the monocular level will eliminate interference from the eye that does not contribute to the winning signal. Implementation results demonstrating this phenomenon are depicted in figures 9 and 10. In figure 9, bias forces localization of a horizontal edge in the top half of the image. A strong edge appearing in the left eye gives rise to a winning response at the highest layer corresponding to a feature that is occluded in the right eye. The result is the selection of the feature appearing in the left eye only. In figure 10, bias is initiated in favor of edges oriented at 45 degrees. A very strong edge oriented at 45 degrees appearing in the right eye gives rise to the winning activation at the highest layer of the pyramid. Because selection means that a particular orientation and spatial frequency is bound before selection at layer 1, the corresponding edge in the left eye is not selected since it is not oriented close enough to 45 degrees. This might not be the case if multiple winners were allowed at each layer since WTA competition might not converge on a single orientation.

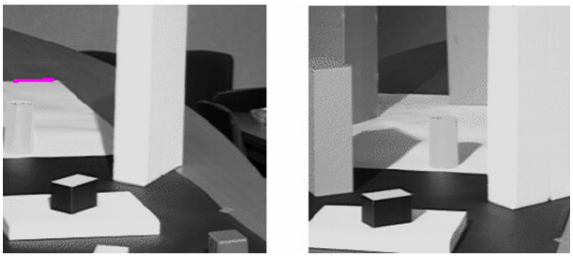

Figure 9. Selective Tuning on the image depicted in figure 6. with a bias for horizontal, and the focus of attention restricted to the top half of the image. A strong edge appearing in one eye is occluded resulting in no appropriate correspondence. As such, a strong edge is localized in one eye with the response of the other eye suppressed.

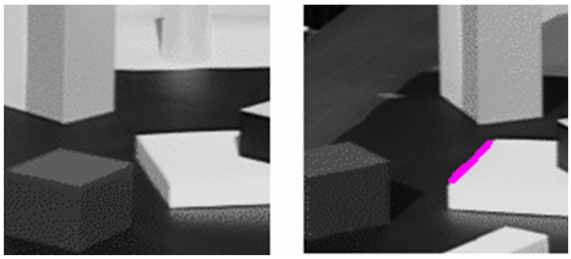

Figure 10. Selective Tuning on the image depicted in figure 6. with a bias for 45 degrees. In this case, a strong edge at 45 degrees appearing in one eye contributes to the winner at the highest layer. Because there is a significant rotational disparity in this edge at the input, the feature is selected in one eye only.

6. Stereo Vision and Computational Models of Attention

There exist numerous models of attention that range from purely computational to purely biological. As mentioned, Selective Tuning is a natural choice for realizing stereo attention since in addition to its biological plausibility, the existing rules asserted by Selective Tuning need only to be applied to the binocular domain with no changes required to the structure or principles of the model to accommodate stereo input. This brings to light an interesting consideration: To what degree might other existing computational models lend themselves to attentional function in the stereo domain, and what issues arise in this context.

Attention models may necessarily be divided into two distinct categories:

i. Models for which attentional selection takes place within the same feedforward interpretive network that realizes feature extraction.
ii. Models that rely on a selective attention procedure external to the extraction of basic features.

In principle, any model that falls into class i. might be modified to incorporate selection based on stereo input since it is possible to directly discern the hierarchy of computation that gave rise to the winning selection. As such, it should be possible to determine the locus of the stimuli that gave rise to the winning activation in each eye. Modification in this regard may be trivial for some models and prohibitively difficult for others, although there is no fundamental reason to rule out the possibility of achieving attentional deployment in three dimensions.

Consideration of the second class of models in the domain of stereo localization brings to the forefront a potential confound with such models in their current form. Models of attention belonging to class ii. consist largely of saliency based attention models [11-13]. Such models figure prominently in the attention literature and typically involve basic feature extraction to produce a *Master* feature map which is employed to execute shifts of attention.

A general formulation of saliency based attention models might be expressed as follows:

For a saliency map S, derived from visual inputs $\boldsymbol{l}$ and $\boldsymbol{r}$ corresponding to the view of the left and right eyes respectively, a general expression for S is as follows:

$$S = f_1(\boldsymbol{l},\boldsymbol{r}) \oplus f_2(\boldsymbol{l},\boldsymbol{r}) \oplus \ldots \oplus f_n(\boldsymbol{l},\boldsymbol{r})$$

Where $f_k(\boldsymbol{l},\boldsymbol{r})$ denotes the k^{th} feature extractor, and $\oplus$ indicates an operation that combines the various feature maps. Attentional selection is achieved by selection based on the two dimensional topographical saliency map S, which quantifies the salience of each position in the image. As salience is based on the projection of various features onto a two dimensional plane, information concerning correspondence is necessarily discarded. As such, there is no means of selecting non-corresponding retinal coordinates in each eye corresponding to a single world event.

Advocates of a saliency based architecture might suggest that stereo attention may be achieved by considering a representation of salience residing in three dimensional space. Such a representation would require close re-examination from the perspectives of biological plausibility and computational complexity. Detailed analysis of implications of stereo input in the context of specific models of attention will be the subject of future work.

7. Discussion

It is apparent that issues that require attention in the context of a single viewpoint extend to the binocular domain. Further, an appropriate architecture must accommodate shifts in the position of an attended event from one eye to the other. We have demonstrated that this may be accomplished via an architecture that includes neurons tuned to a variety of disparities, and that preserves connectivity among interpretive units in this network. In the absence of explicit memory of connectivity, there is no means of exactly determining the source of activation among neurons in higher layers, and no means of selecting an appropriate locus of attention in the two-eyed frame of reference. Within the proposed framework, maximum activation at a higher layer of the pyramid corresponding to a particular disparity, orientation, and spatial frequency may necessarily be traced back to the stimulus that gave rise to such activation in both eyes. The importance of this consideration is especially apparent if one aims to employ attention as a means of directing eye movements. We have also highlighted the possibility that this consideration may pose problems for certain classes of attention models.

A secondary role of attention in the domain of stereo vision, is the resolution of conflict between the two eyes resulting from occlusion or sufficiently large disparities. We have demonstrated that Selective Tuning may accommodate binocular rivalry by virtue of the basic principles put forward by the general model. In a slightly different light, one might say that Selective Tuning predicts that binocular rivalry will be resolved by attention in primates. As discussed, there is mounting evidence in this regard further suggesting the described model affords an appropriate description of attentional function in primates in addition to a scheme for achieving such behavior in machines.

8. References

[1] P. Burt, Attention mechanisms for vision in a dynamic world, in: Proceedings Ninth International Conference on Pattern Recognition, Beijing, China, 1988.

[2] J.K. Tsotsos, Analyzing vision at the complexity level, Behavioral Brain Science, 13(3), 423-469, 1990.

[3] J.K. Tsotsos, A Complexity Level Analysis of Immediate Vision, International Journal of Computer Vision, Marr Prize Special Issue, Vol. 2, No. 1, 303 - 320, Sept. 1988.

[4] J.K. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artifical Intelligence 78(1-2), p 507 - 547, 1995.

[5] I. Ohzawa, G.C. DeAngelis, R. Freeman, Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. Science 249: 1037-1041, 1990.

[6] I. Ohzawa, Mechanisms of stereoscopic vision: the disparity energy model, Current Opinion Neurobiology 8: 509-515, 1998.

[7] D. Fleet, H. Wagner, D. Heeger, Neural encoding of binocular disparity: Energy models, Vision Research, 36(12) 1839-1857, 1996.

[8] E.D. Lumer, K.J. Friston, and G. Rees, Neural correlates of perceptual rivalry in the human brain. Science, 280:1930-1934, 1998.

[9] J.F. Mitchell, G.R. Stoner, and J.H. Reynolds. Object-based attention determines dominance in binocular rivalry. Nature, 429:410-413, 2004.

[10] T.L. Ooi, Z.J. He, Binocular rivalry and visual awareness: the role of attention. Perception, 28:551-574, 1999.

[11] A. Treisman, G. Gelade. A feature-integration theory of attention. Cognitive Psychology, 12(1), 97-136, 1980.

[12] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20:11, 1254-1259, 1998.

[13] J. W. Wolfe, Guided Search 2.0: A revised model of visual search, Psychonomic Bulletin and Review, 1, 202-238, 1994.