Evaluation of Selective Attention under Similarity Transforms

Bruce A. Draper and Albert Lionelle
Department of Computer Science
Colorado State University
Fort Collins, CO, 80523, U.S.A.
draper,lionelle@cs.colostate.edu

Abstract

Computational selective attention systems have mostly been developed as models of human attention, and they have been evaluated on that basis. Now, however, they are being used as front ends to object recognition systems, and in particular to appearance-based recognition systems. As such, they need to be evaluated by other criteria. A common goal for object recognition systms in invariance to 2D similarity transformations (i.e. in-plane translations, rotations, reflections and scales). This implies that attention systems used as front ends should also be invariant to similarity transforms. This paper evaluates the Neuromorphic Vision Toolkit (NVT), a well known and publicly available selective attention system, and finds it to be highly sensitive to 2D similarity transforms. Further investigation, however, suggests that this sensitivity is an artifact of the publicly available implementation, and not of the neuromorphic principles it is based on. Therefore we have developed a new system, called SAFE (Selective Attention as a Front End), that is conceptually similar to NVT. However, SAFE is largely invariant to 2D similarity transformations of the source image and selects scales as well as spatial locations for fixations, implementing a combined "zoom-spotlight" model of attention.

1. Introduction

In the primate visual system, it is believed that only select locations of a scene are processed in detail, [1, 3, 18] and these locations are commonly referred to as the salient fixations of a scene. Cognitive scientists believe that fixations are selected both bottom-up, in response to the contents of the current scene, and top-down, reflecting current goals [18, 24]. There are a number of computational systems that attempt to model aspects of selective attention, including [13, 14, 24, 23, 19, 8, 22, 16].

Most of these systems were intended as cognitive models of human attention, and have therefore been evaluated by cognitive science measures, such as the "pop-out" effect [6], their speed in relation to human performance [6], and fidelity to human eye tracking experiments [20]. More and more, however, computational models of selective attention are utilized as the front end to automatic object recognition systems, and in particular to appearance-based systems (e.g. [26]). In this case, they need to be evaluated by other criteria.

Appearance-based recognition systems match features extracted from image patches to features stored in memory. In general, appearance-based matching is not invariant to 3D geometric transformations. 3D object recognition is implemented by matching to features extracted from many different views of an object, often interpolating between viewpoints (e.g. [15]).

To keep the number of appearance-based templates manageable, however, it is helpful if appearance matching is invariant to 2D similarity transformations (i.e. 2D translation, rotation, reflection and scale[1]) of the image. This is the advantage of using a selective attention system as a front end; it focuses on objects to match, regardless of their position in the (2D) image. It implies, however, that selective attention systems should be invariant to 2D similarity transforms, at least in this context.

This paper presents two simple methods for empirically measuring the invariance of selective attention systems in response to 2D transformations. The first looks for gross errors, and measures how often fixations extracted from a transformed version of a test image corresponds to fixations extracted from the original image. The second measures the median drift in the position of fixations extracted from transformed and non-transformed versions of the same image. It should be noted that neither of these methods determine whether a fixation location is good or bad; they simply

[1] This is sometimes called a 4DOF affine transformation, but since the term "affine" usually refers to transformations with 6 DOF (the extra two degrees of freedom permit shearing), we stick with the older terminology of similarity transformations here.

measure whether the selective attention system is sensitive to 2D similarity transforms.

Using these measures, we evaluate the Neuromorphic Vision Toolkit (NVT), a well-known and publicly available selective attention system developed at CalTech and USC (see ilab.usc.edu/bu). We find, unfortunately, that NVT is highly sensitive to 2D similarity transformations, and is therefore not a good candidate to serve as the front end to an object recognition system, even though it has been used for this purpose [26]. The extreme sensitivity to 2D transformations evident in this study is not an inherent feature of neuromorphic principles, however, but rather an artifact of the publicly available implementation. We have therefore implemented a new selective attention system, called SAFE (Selective Attention as a Front End) based on similar neuromorphic principles. However, SAFE is more robust to similarity transformations and selects in scale in addition to spatial coordinates following a combined "zoom-spotlight" model of attention.

Figure 1. A visual display illustrates the difficulty of attending to two different regions and scales at the same moment. Attention shifts are needed to see the entire image or to focus on individual words and letters.

2. Background: Models of Attention

It has long been theorized that the primate visual system does not passively process all the information in the visual field, but rather selectively attends to specific locations (for a review, see [18], Chapter 11). Within the selective attention paradigm, cognitive scientists have proposed two competing models of visual attention, the so-called "spotlight" and "zoom lens" models. The spotlight model can be metaphorically described as a pen light moving across a dark scene. When the spotlight stops at a location, that location is illuminated or attended to. It is a strictly spatial view of attention that moves with the inner-eye. The zoom lens model, on the other hand, suggests a metaphor of looking at a scene through the zoom lens of a camera. The viewer can attend to tiny objects by close inspection, or "zoom-out" and attend to larger or coarser objects [5, 18, 17]. In essence, the attention window selects a scale. However, both these models could be considered limiting, so Palmer refers to a model that selects both location and scale as a combined model [18]. We refer to this combined model as a "zoom-spotlight" model of attention.

There is growing biological evidence for the zoom-spotlight model. Oliva and Schyns, for example, demonstrate bottom-up queueing of selective attention at specific spatial frequencies [17]. Figure 1 helps illustrate such differences. When we "zoom-out" we can see concentric rings, but it is difficult to discern the wording unless one focuses in on each letter and follows the path. Notice that when one is focused on "here" it is difficult to focus on "focus" without switching our attention, yet a moment ago one was able to focus on all the rings.

There are advantages for appearance-based object recognition systems to the zoom-spotlight model. Appearance-based systems match patches of the current image to patches of previously seen images, and focus of attention mechanisms are required for selecting the patches. In the pen light model, the selected image patch will always be at a fixed resolution, since scale is not selected as part of the attention system. This can be a problem if the system needs to recognize objects of varying sizes. In the zoom-spotlight model, on the other hand, the attention system can vary the resolution of the selected image patch, allowing the system to match a wider range of objects. We therefore believe the zoom-spotlight model may be better suited when implimenting visual attention as a front end to an appearance-based matching.

The zoom-spotlight model is a unitary model of selective attention; it assumes that humans can focus attention on only one fixation point at a time. There is evidence, however, that people may be able to simultaneously focus attention on multiple fixation points [21, 25]. If this is true, there is a critical question about how multiple attention windows are interpreted. If they are interpreted independently, then the difference between unitary and multiple-focus models is a matter of scheduling and capacity (whether the windows are processed concurrently or sequentially). If there are interactions between the interpretation processes, however, then the difference between unitary and multiple-focus models is more significant. It is not clear, however, how to take significant advantage of a variable number of concur-

rent attention windows within an appearance-based recognition paradigm, so we restrict ourselves to unitary (or multiple-focus but independent) models here.

3 Background: Neuromorphic Vision

A currently popular model of visual attention is embedded in the Neuromorphic Vision Toolkit (NVT), developed by Koch, Itti et al. at the California Technical Institute and the University of Southern California [20, 11, 10, 8, 7, 6], and based on a line of research dating back to work by Koch and Ullman in 1985 [12]. The goal of NVT is to design vision systems organized similarly to biological vision systems. Primate vision systems are constructed of highly specified cell organizations. In this organization, key characteristics have been found; most notably the excitatory and inhibitory opponent-processing interaction between the early processing cells. This exhibits an on-center and off-surround effect in association with different types of stimuli. For example, green is excitatory if the surround is red within the retinal ganglion cells [1, 18]. Also, within the superior colliculus and intraparietal sulcus there appear to be several neuronal maps which specifically encode the salience of visual scenes [2]. Koch and colleagues have therefore organized their neuromorphic vision system around two principles. First, feature maps are constructed of on-center and off-surround differences within different stimuli. Second, feature maps are combined into "saliency" maps, where the maximum value in the salience map is attended to by the attention spotlight, and saliency maps evolve to allow for multiple feature selection [6].

The design of NVT is shown in Figure 2. The image is split into three independent channels, one for intensity, one for oponent colors, and one for edge information. The opponent color channel is then divided into two subchannels, one for red vs. green opponency, and the other for blue vs. yellow. The edge channel is also divided, this time into four subchannels according to edge orientation (0, 45, 90 and 135 degrees). Image pyramids are used to simulate multi-scale processing in each subchannel.

Every scale of every subchannel is conceptually convolved with an on-center, off-surround mask. (The implementation is slightly different; see below.) The responses to this mask are normalized based on the responses of other pixels within a neighborhood, and summed across scales to produce a single saliency map for every subchannel. The subchannel saliency maps are then normalized and summed to produce one saliency map per channel. These channel saliency maps are renormalized again and summed to produce a single saliency map for the image. A neural network selects the first fixation from the saliency map. The values around this location are then suppressed, and the neural network is called again to choose the next fixation, in a process

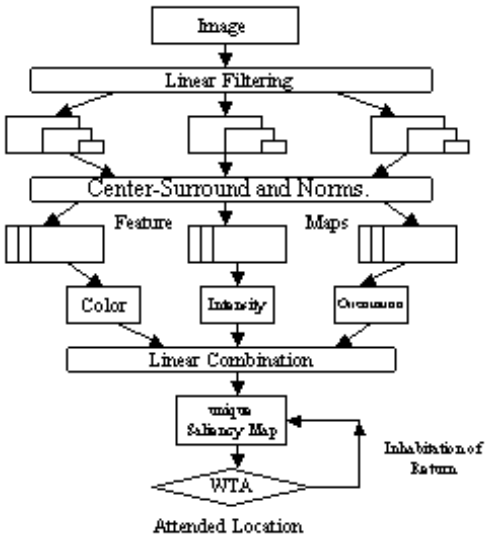

Figure 2. The model presented.

that repeats until the desired number of fixations have been selected. For more information, see [11, 10, 8, 6, 20].

It is important to note that there are other computational models of selective attention. Maki et al present a system that exploits motion and stereo depth perception for selection attention [13]. Tsotsos et al. present a more general neural network model that uses selective tuning [23]. Park et al. recently introduced a system similar to NVT, except that they use independent component analysis to impliment a feature competition scheme [19]. The mentioned here are just a few in a quickly growing feild of visual attention models each with inheriant strengths and limitations.

4. Evaluating Visual Attention Models

Computational models of selective attention such as NVT serve two purposes, as cognitive models of primate vision and as components of computer vision systems. They are difficult to evaluate in either context. Evaluating attention systems as cognitive models is hard because of the lack of ground truth data. Eye tracking systems record eye movements, but cannot measure the "inner eye's" attentional fixations (or selected scales). Nonetheless, there is presumably some relation between ocular and attentional fixations, and Parkhurst et. al. are able to demonstrate that the fixations selected by NVT correspond better to human eye tracking data than do random signals [20]. Attention selection systems have also been evaluated cognitively in terms of response time and the visual "pop out" effect [6].

Evaluating attention systems as components of computer vision (or object recognition) systems is no easier. Once

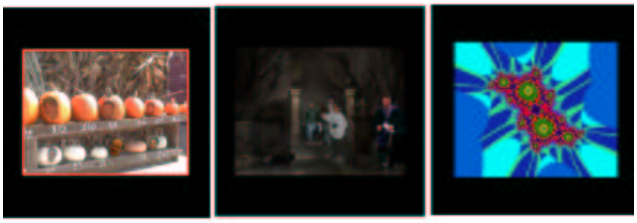

Figure 3. Image 1: Pumpkins, Image 2: to-Class, Image 3: Fractal 7

again, the optimal sequence of attention windows is unknown, so there is no ground truth data. We can, however, test for invariance. We would like a selective attention system to be invariant to 2D similarity transformations. In other words, we would like the system to attend to the same features in a scene, whether or not the scene has been translated, rotated, reflected or scaled. This is particularly important for appearance based recognition systems, since an invariant focus of attention system has the property that it solves the 2D registration problem.

To test the invariance of NVT to similarity transformations, we apply it to test images and to transformed versions of the test images. Ideally, NVT should return the same 2D fixations for both images, once we compensate for the geometric transformation. Since NVT uses image size as a processing parameter, both the source and transformed images are the same size[2]. To make sure that rotations and translations do not alter the content of the scene, every test image is surrounded by a large black border.

The image transformations tested are translation (down and to the right) by 1, 7 and 32 pixels, rotation from 45 to 315 degrees in increments of 45 degrees, and reflections about the horizontal and vertical axes. (Scale will be discussed in Section 7.) Thus we tested 12 transformations for every test image. The test images are shown in Figure 3. The first image is bright, with many potential interest points. The second image is fairly dark, and therefore provides low overall stimulation. The third image is a fractal image, with many bright colors and strong edges. Since all of the (untransformed) test images have black borders, none of the fixations are mapped to coordinates outside the image by any of the transformations.

We quantify the performance of an attention system through two measures. The first looks for gross errors, and records the percentage of fixations in the test image that are not within a threshold radius of any fixation in the transformed image, once the geometric transformation is compensated for. The reported number is therefore a gross error rate. (In these experiments, the radius threshold was 1/48th of the image size, roughly between 17-18 pixels.)

The second measure is a form of the Hausdorf distance metric. It measures positional noise, assuming that at least half the fixations identified in the test image are also found in the transformed image. In particular, it is the median positional drift, as measured by:

$$\begin{aligned} & max(h(A,B), h(B,A)) \\ h(A,B) &= median(min||a-b||_2) \end{aligned} \quad (1)$$

where A is the set of fixations from the original test image, and B is the set of compensated fixations from the transformed test image. As long as fewer than half the locations are outliers, this statistic reflects the positional noise between the two sets of fixations. If more than half the points are outliers, as measured by our first statistic, then this measure is not meaningful and is not reported.

5. Evaluating NVT

Using these tests, we evaluated the Neuromorphic Vision Toolkit (NVT) discussed above. NVT was parameterized to extract 25 fixations for every image. Since NVT has non-deterministic components in the neural network, we repeated each test 10 times. Table 1 shows the average gross error rate and average median positional drift for every image and transformation. The gross error rates are particularly startling. Simple image rotations lead to error rates between 12% and 84%; if we average across images and rotations, we get an average gross error rate of over 44%. This suggests that if we apply NVT to an image and its rotation, on average 11 of the 25 features identified as fixations in the first image will not be identified as fixations in the rotated image. Poor performance is also demostrated in translations and reflections. Although a small translation of one pixel down and to the right had no effect in terms of gross errors on two of the test images, on the fractal image even this simple transformation caused 4 of the 25 fixations to be lost. Translations of 7 and 32 pixels caused an average 12.7% and 17.3% of fixations to disappear, respectively. Reflection behaved like a large rotation, with almost half the fixations disappearing.

When the same image features are selected in both the original and transformed images, they drift. For example, when the image is translated seven pixels down and to the right, the new position of a fixation should be seven pixels down and to the right of the original. Instead, we find that the new position of the fixation is an average of 11 pixels away from this predicted position.

An interesting number not reported in Table 1 is the variance between runs on a single pair of images (test and transformed test). For 26 of 36 image pairs, there is no variance

[2]Except when testing for scale invariance; see Section 7.

	Gross Error %			
Transform	Img 1	Img 2	Img 3	Avg
Translate 1	0.0	0.0	16.0	5.3
Translate 7	2.0	24.0	12.0	12.7
Translate 32	8.0	20.0	24.0	17.3
Reflect H	48.0	36.0	64.0	49.3
Reflect V	48.0	40.0	60.0	49.3
Rotate 45	28.0	16.0	52.0	32.0
Rotate 90	48.0	24.0	60.0	41.3
Rotate 135	48.0	48.0	52.0	49.3
Rotate 180	56.0	44.0	52.0	50.6
Rotate 225	60.0	44.0	84.0	62.6
Rotate 270	48.0	32.0	64.0	46.6
Rotate 315	12.0	24.0	52.0	29.3
Averages	33.8	30.6	49.3	37.9
	Median Drift			
Transform	Img 1	Img 2	Img 3	Avg
Translate 1	1.4	1.4	1.4	1.4
Translate 7	11.4	11.4	11.4	11.4
Translate 32	0.0	16.0	16.0	10.6
Reflect H	-	14.0	-	14.0
Reflect V	-	16.1	-	16.1
Rotate 45	9.8	11.7	-	10.7
Rotate 90	-	17.0	-	17.0
Rotate 135	-	-	-	-
Rotate 180	-	17.1	-	17.1
Rotate 225	-	17.8	-	17.8
Rotate 270	-	15.0	-	15.0
Rotate 315	17.0	14.0	-	15.5
Averages	7.9	13.7	9.6	13.3

Table 1. Average Error across all modifications for three Images

between runs. For these images, if you run NVT twice on the same image, you get the same set of fixations. For the remaining images, however, NVT is non-deterministic, producing different fixations on some runs than others. We do not know why this happens; we suspect the neural network that selects the final fixations. We do know that ten of the eleven cases where this happens involve transformations of the fractal test image. The last row of Table 1 shows the gross error rates and median drifts for each test image, averaged over the eleven transformations. Clearly the fractal image is the most challenging image for NVT.

These results are both disappointing and surprising, since the description of NVT's algorithm in Section 2 suggests that it should be invariant to rotation, translation and reflection. Obviously, edge detectors are not perfectly insensitive to rotations, but NVT uses four edge detectors at 45 degree intervals, so it should be invariant to 45 degree rotations, which is what we tested. Instead, we believe that the extreme sensitivity to these transformations evident here is a result of trade-offs made in the implementation between accuracy and processing speed. For example, NVT approximates convolution with an on-center, off-surround (OCOS) mask by subtracting pixels at one level of the pyramid from pixels at a higher level. This is a computationally efficient approximatation to convolution with an OCOS mask, and is also used in other systems (e.g. [19]). Unfortunately, the OCOS mask it best approximates is square, not circular. As a result, the system is sensitive to rotations.

We also have suspicions about the consistency of the neural network used to select fixations. It is the only component we can identify that might cause it to select different fixations on two runs on the same image, as happened with some versions of the fractal test image. Logically, it may be introducing part of the inconsistency between transformed images as well. Certainly the combined system is very, very sensitive to 2D transformations. Given that processors keep getting faster, it may be that some of the compromises made in the implementation of NVT for speed were not wise.

There are also reported strengths of NVT that we did not test for. These include is robustness to Gaussian noise, and its speed in relation to human performance. For further details of these properties, see [9, 11, 6].

6. Reimplementing Selective Attention : SAFE

Our original intention was to use NVT as the front end to an appearance based object recognition system (see [4]). Unfortunately, our experience with NVT summarized above convinced us that it would be an unreliable method for selecting focus of attention regions. It also fails to select scales, a property that would be helpful to the object recognition system.

The sensitivity to 2D transformations evident in this study appear to be more the result of design decisions made in the implementation than the underlying neuromorphic theory, however. We therefore decided to implement a new selective attention system, based on the same underlying ideas as NVT. Since the motivation of our system is to serve as the front end for an object recognition system, we call it SAFE (Selective Attention as a Front End). The design of SAFE is shown in Figure 4. Like NVT, it divides processing into three channels: intensity, opponent color, and edge. As with NVT, the opponent color channel is subdivided into red vs. green and blue vs. yellow channels. Also as with NVT, an image pyramid is created for each channel (and subchannel).

At every level of the pyramid, the salience of a channel is determined by convolving it with a circularly symmetric on-center off-surround mask, formed by a difference of gaussians. The absolute values of the salience images are

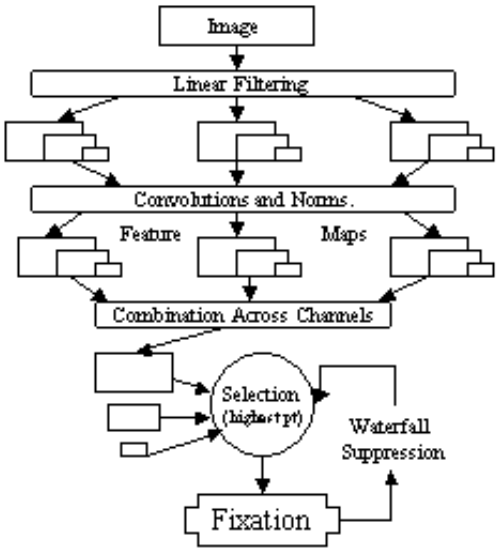

Figure 4. Model of Selective Attention as a Front End

normalized by their standard deviation (other normalization techniques are being considered), and then smoothed by convolution with a gaussian. In the case of the two color subchannels, the same process is followed and then the salience images are summed. Unlike NVT, salience images are not combined across scales within a channel. Instead, salience images are combined across channels within each scale, producing a pyramid of salience maps. The first fixation is the position (in x, y and scale) of the maximum value in the salience pyramid. When a fixation is selected, a watershed algorithm is used to suppress all other salience values at the same scale that are part of the same peak. Then the remaining maximum value is the next fixation.

We evaluated SAFE using the same tests and protocols that we used for NVT; the results are shown in Tables 2. Unlike NVT, SAFE is largely invariant to 2D transformations. Figure 5 shows the attention windows extracted by NVT and SAFE for an image and its rotation. Clearly, the fixations are more consistent between the rotated and non-rotated image for SAFE than for NVT. There are other differences as well, however. For SAFE, the attention windows vary is size, depending on the scale of the fixation; all attention windows selected by NVT are the same size. This reflects the difference between a pen light and zoom-spotlight model. Whether the attention windows selected by SAFE are "better" than the ones selected by NVT is matter of subjective judgement.

	Gross Error %			
Transform	Img 1	Img 2	Img 3	Avg
Translate 1	4.0	0.0	4.0	2.6
Translate 7	0.0	0.0	4.0	1.3
Translate 32	4.0	0.0	4.0	2.6
Reflect H	0.0	0.0	0.0	0.0
Reflect V	0.0	0.0	4.0	1.3
Rotate 45	4.0	0.0	4.0	2.6
Rotate 90	0.0	0.0	4.0	1.3
Rotate 135	4.0	0.0	4.0	2.6
Rotate 180	0.0	0.0	4.0	1.3
Rotate 225	4.0	0.0	4.0	2.6
Rotate 270	0.0	0.0	0.0	0.0
Rotate 315	4.0	0.0	4.0	2.6
Averages	1.6	0.0	3.3	1.6
	Median Drift			
Transform	Img 1	Img 2	Img 3	Avg
Translate 1	0.0	0.0	0.0	0.0
Translate 7	0.0	0.0	0.0	0.0
Translate 32	0.0	0.0	0.0	0.0
Reflect H	0.0	1.0	1.0	0.6
Reflect V	1.0	1.0	1.0	1.0
Rotate 45	1.0	1.0	1.0	1.0
Rotate 90	1.0	1.4	1.0	3.1
Rotate 135	1.0	1.0	1.0	1.0
Rotate 180	1.0	2.2	1.0	1.4
Rotate 225	1.4	1.0	4.0	2.1
Rotate 270	1.0	1.0	1.0	1.0
Rotate 315	1.0	1.0	1.0	1.0
Averages	0.7	0.8	2.0	1.2

Table 2. Average Error across all modifications for three Images

7. Scale Invariance

So far, we have evaluated NVT and SAFE for invariance to translation, rotation and reflection, but not scale. This is because scale invariance is the one area where the design goals of NVT and SAFE differ. NVT embodies a pen light model of selective attention; it selects the positions of fixations, but not the scales. In fact, it explicitly integrates information across scales. SAFE, on the other hand, implements a zoom-spotlight model and selects both positions and scales.

NVT is also parameterized by the image size. As a result, if one reduces the resolution of an image feature without changing the overall image size (for example, by moving the object farther away from the camera), it may no longer select the same image feature. NVT is not designed to be scale invariant in this sense, even though it does multi-scale

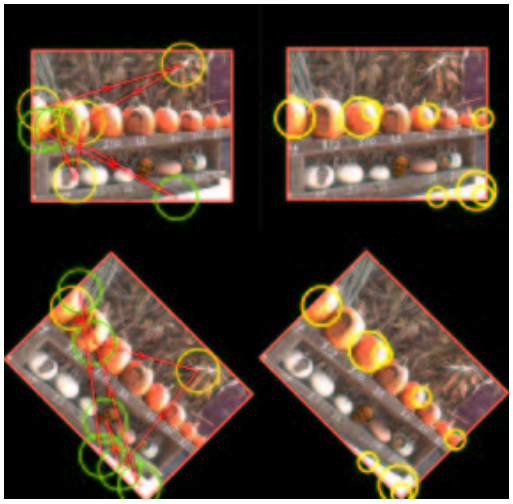

Figure 5. Visual output of NVT (shown on the left) compared to the output of SAFE (shown on the right).

	Gross Error %			
Transform	Img 1	Img 2	Img 3	Avg
Scaled 1/2	24.0	36.0	56.0	38.6
Scaled 1/4	68.0	88.0	96.0	84
Averages	46.0	62.0	76.0	61.3
	Median Drift			
Transform	Img 1	Img 2	Img 3	Avg
Scaled 1/2	16.0	16.0	-	16.0
Scaled 1/4	-	-	-	-
Averages	16.0	16.0	-	16.0

Table 3. NVT: Average Error across all scale modifications for three Images

processing. SAFE, on the other hand, is not parameterized by image size, and selects scales as well as positions. If the resolution of an image feature is reduced, one would hope to find the same feature in the new image, but at a reduced resolution. We therefore have to be careful when comparing the relative scale invariance of NVT and SAFE. In principle, if the scale of an entire image is reduced, NVT should select the same feature locations as in the original, since the image size will be reduced by the same factor as the image features. Unfortunately, Table 3 shows that NVT is no more invariant to scale in this sense than it was to translation, rotation or reflection. In SAFE, the size of the masks is a user parameter. If these masks are reduced by the same factor as the image, it should behave like NVT, i.e. it should select the same feature locations in the both

	Gross Error %			
Transform	Img 1	Img 2	Img 3	Avg
Scaled 1/2	4.0	4.0	16.0	8.0
Scaled 1/4	12.0	8.0	16.0	12.0
Averages	8.0	6.0	16.0	10.0
	Median Drift			
Transform	Img 1	Img 2	Img 3	Avg
Scaled 1/2	1.0	1.0	1.4	3.1
Scaled 1/4	3.6	3.1	3.1	3.2
Averages	2.3	2.0	2.2	3.1

Table 4. SAFE: Average Error across scale modifications for three Images

the original and reduced image. As shown in Table 4, this is essentially what happens for SAFE.

8. Conclusion

The Neuromorphic Vision Toolkit (NVT) is a well-known and publicly available computational model of selective attention. In at least one paper it has been used as the front end to an object recognition system [26]. Our studies suggest, however, that the publicly available implementation is highly sensitive to 2D transformations, and is therefore not a good candidate for the front end of an object recognition system.

To address these problems, we have created a new selective attention system called SAFE based on roughly the same neuromorphic principles as NVT. SAFE has the advantage that it is largely invariant to 2D transformations. Also, it selects scales as well as locations for fixations, making it a zoom-spotlight (rather than pen light) model of selective attention.

SAFE is publicly available and can be downloaded from our web site along with basic datasets[3]. For speed, it is implemented using Intel's IPPI library, which is commercially available for Intel processors running either Windows or Linux.

References

[1] N. R. Carlson. *Foundations Physiological Psychology*. Allyn and Bacon, Needham Heights, Massachusetts, 1995.

[2] C. L. Colby and M. E. Goldberg. Space and attention in parietal cortex. *Annual Review of Neuroscience*, 1999.

[3] R. De Valois. Spatial and form vision: Early processing. In W. Backhaus, editor, *Neuronal Coding of Perceptual Systems*. World Scientific, 1998.

[3] www.cs.colostate.edu/∼vision/safe

[4] B. A. Draper, K. Baek, and J. Boody. Implementing the expert object recognition pathway. In *International Conference on Vision Systems*, Graz, Austria, April 2003. Springer-Verlag. To Appear.

[5] C. W. Eriksen and J. D. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, 1986.

[6] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, Pasadena, California, Jan 2000.

[7] L. Itti. Modeling primate visual attention. In J. Feng, editor, *Computational Neuroscience: A Comprehensive Approach*. CRC Press, Boca Raton, in press.

[8] L. Itti. Visual attention. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks, 2nd Ed.* MIT Press, in press.

[9] L. Itti, C. Gold, and C. Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9):1784–1793, Sep 2001.

[10] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, May 2000.

[11] L. Itti, C. Koch, and 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, Nov 1998.

[12] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–27, 1985.

[13] A. Maki, P. Nordlund, and J.-O. Eklundh. Attentional scene segmentation: Integrating depth and motion from phase. *Computer Vision and Image Understanding*, 2000.

[14] R. Milanese, S. Gil, and T. Pun. Attentive mechanisms for dynamic and static scene analysis. *Optical Engineering*, 1995.

[15] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 1995.

[16] V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tuebingen, Germany*, in-press.

[17] A. Oliva and P. G. Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 1997.

[18] S. E. Palmer. *Vision Science – Photons to Phenomenology*. MIT Press, Cambridge, Massachusetts, 1999.

[19] S.-J. Park, J.-K. Shin, and M. Lee. Biologically inspired saliency map model for bottom-up visual attention. In *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tubingen, Germany*, 2002.

[20] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 2002.

[21] Z. Pylyshyn. Some primitive mechanisms of spatial attention. *Cognition*, 1994.

[22] O. Stasse, Y. Kuniyoshi, and G. Cheng. Development of a biologically inspired real-time visual attention system. In *Proc. 1st Workshop on Biologically Motivated Computer Vision (BMCV'00), Soul, Korea*, 2000.

[23] J. K. Tsotsos, M. S. Culhane, W. Yan Kei Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artifical Intelligence*, 1995.

[24] D. C. Van Essen, B. Olshausen, C. Anderson, and J. L. Gallant. Pattern recognition, attention and information bottlenecks in the primate visual system. *Visual Information Processing: From Neurons to Chips*, 1991.

[25] L. Viswanathan and E. Mingolla. Attention in depth: Disparity and occlusion cues facilitate multi-element visual tracking. Technical report, Department of Cognitive and Neural Systems, Boston University, 1998.

[26] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition - a gentle way. In *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tubingen, Germany*, 2002.