Selective Attention in the Learning of Invariant Representation of Objects

Muhua Li
Centre for Intelligent Machines
McGill University
Montreal, Quebec Canada H3A 2A7
limh@cim.mcgill.ca

James J. Clark
Centre for Intelligent Machines
McGill University
Montreal, Quebec Canada H3A 2A7
clark@cim.mcgill.ca

Abstract

Selective attention plays an important role in visual processing in reducing the problem scale and in actively gathering useful information. We propose a modified saliency map mechanism that uses a simple top-down task-dependent cue to allow attention to stay mainly on one object in the scene each time for the first few shifts. Such a modification allows the learning of invariant object representations across attention shifts in a multiple-object scene. In this paper, we will first introduce this saliency map mechanism and then propose a neural network model to learn invariant representations for objects across attention shifts in a temporal sequence.

1. Introduction

Processing massive visual information seems rather frustrating. Selective attention could simplify such processing by permitting the focusing on a small fraction of the total input visual information ([7], [10]), thus breaking down the problem into several sequential smaller-scale visual analysis sub-problems. Shifting of attention enables the visual system to actively, and efficiently, acquire useful information from the external environment for further processing. Hafed's ([4]) work shows evidence that saccade target features are attended as a result of the preparation to move the eyes and such shifting of attention is important to aid the visual system in processing the recently foveated saccade target after a saccade ends. His work also reveals a possible temporal association mechanism across attention shifts.

Temporal association is influential in the development of transformation invariance when we consider the importance of the continuous properties of an object in both space and time domain in the world. An object at one place on the retina might activate feature analyzers at the next stage of cortical processing. Psychophysical studies by Wallis and Bülthoff ([14]) also revealed the importance of temporal information in object recognition and representation, which suggests that humans are continuously associating views of objects to support later recognition, and the recognition is not only based on the physical similarity but also the correlated appearance in time of the objects.

There are some models where the visual input is filtered into a focus of attention (therefore an object of interest is pop out in the center of the attention window) and then fed into a recognition system for position or scale invariant recognition ([11], [5]). The dynamic routing circuits employed in these models efficiently select out the regions of Focus of Attention (FOA) to perform position and scale invariant recognition in an associative (or knowledge) network. However, these models focus on the recognition of features such as a whole object in the FOA, which ignore the facts that attention goes to not only between objects but also within object. We will study the more general cases of attention shifts over objects and the learning of invariant representations of objects across attention shifts.

In this paper, we will first propose a saliency map that use both the bottom-up saliency cues and a simple task-dependent cue that enables attention to stay mainly on an object of interest for the first few shifts. Then we will apply this saliency map to generate a sequence of attention shifts, to guide the process of the temporal learning of invariance.

2. System composition

The overall system is composed of two sub-modules, as illustrated in Figure 1. One is the attention control module, which generates attention-shift signals according to a saliency map. The module obtains as input local feature images from the raw retinal images via a dynamically position-changing attention window. The second sub-module is the learning module, which performs the learning of invariant neuronal representations across attention shifts in temporal sequences.

3. Attention shift control

The traditional saliency map mechanism follows the idea that human attention is mostly likely to focus on the most salient features in the scene. It is mainly based on bottom-up image-based saliency cues ([6]). There is another im-

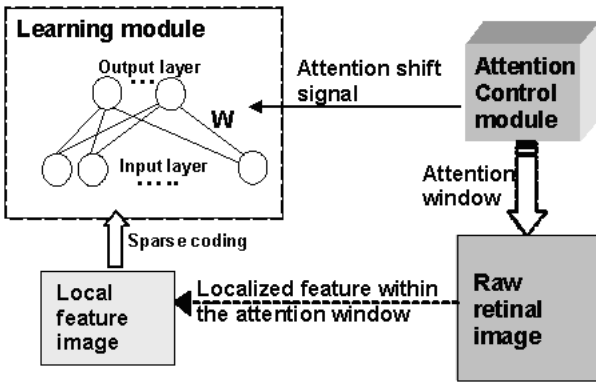

Figure 1: The system is composed of two modules: an attention control module and a learning module. The attention control module is an attention shift mechanism that generates attention shift signals to trigger the learning processes in the learning module. It also selects local features, which are part of the raw retinal image falling within the attention window, as input to the learning module.

portant factor to be kept in mind, however, which is that humans also tend to keep the attention on the attended object or its proximity within a very short time period ([1], [7]), even when the points to be attended following the first attention shift have no more saliency than other points in the scene. This consideration is very helpful when in a short time interval we need a sequence of attention shifts remaining mostly fixed on a targeted object when multiple objects are present in the scene. Such a requirement in the attention shift control can be implemented by introducing a top-down task-dependent cue. The following few paragraphs describe the implementation of the saliency map mechanism with an extension to force the first few attention shifts to stay on the same object.

The saliency map is a weighted sum of the intensity features and the orientation features. The algorithm to calculate these features is that proposed by Itti *et al.* ([6]), which we will describe briefly in the next paragraph.

Intensity features, $I(\sigma)$, are obtained from an 8-level Gaussian pyramid computed from the raw input intensity, where the scale factor σ ranges from [0..8]. Local orientation information is obtained by convolution with oriented Gabor pyramids $O(\sigma, \theta)$, where $\sigma \in [0..8]$ is the scale and $\theta \in [0^\circ, 45^\circ, 90^\circ, 135^\circ]$ is the preferred orientation. Feature maps are calculated by a set of "centre-surround" operations, which are implemented as the difference between fine (at scale $c \in [2, 3, 4]$) and coarse scales (at scale $s = c + \delta$, with $\delta \in [3, 4]$). In total, 30 feature maps, 6 for intensity and 24 for orientation, are calculated and combined into two "conspicuity maps", $\bar{I}$ and $\bar{O}$, at the scale ($\theta = 4$) of the saliency map, through a cross-scale addition where all feature maps are down-sampled into scale four and made an element-by-element addition.

In addition to the intensity and orientation features, we introduce a center-region priority R which has high values in the center of the image. This is used because, in practice, objects in the center of the view are much more likely to attract attention for humans. Such eccentricity effect is interpreted by Wolfe and his colleges ([16]) as an attentional bias that allocates attention preferentially to central items. R is expressed in the form of a two-dimensional Gaussian function:

$$R = e^{-[\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}]} \quad (1)$$

where x_0 and y_0 are the center coordinates of the retinal image, and σ_x and σ_y is the standard deviation in horizontal and vertical directions respectively.

The initial saliency map is formed by:

$$S = \frac{\bar{I} + \bar{O} + R}{3} \quad (2)$$

Once the saliency map is calculated, a competitive Winner-Take-All (*WTA*) algorithm ([13]) is used to determine the location of the currently most salient feature in the saliency map. In the *WTA* algorithm, a unit with the highest value wins the competition and the rest are suppressed. This winner then becomes the target of the next attention shift. An Inhibition-Of-Return (*IOR*) mechanism is added to prevent immediate attention shifts back to the current feature of interest, to allow other parts of the object to be explored. In our implementation, instead of inhibiting the region near the current fixation point, the *IOR* function inhibits all these small regions around the fixation points in a recent history trace of the fixation points. Therefore in the algorithm, we will keep a trace of these fixation points in a vector called *tp*. When an overt attention shift occurs, the image point with fixed world reference coordinates will have a coordinate translation accordingly in the retinal image coordinate reference system. The information of the coordinate offset resulting from each attention shift is used to update the whole trace, reflecting the newest positional change on the fixation points in the history appearing in the new retinal image.

In order to solve the problem how the attention stays on the same object during the learning process, we introduce into the calculation of the saliency map a spatial constraint which forces the next attention target to stay close to the current fixation point. The spatial constraint (*SC*) is implemented by adding a trace of neighbours of the fixation points in the history of the observation duration:

$$SC(t) = \alpha \times SC(t-1) + \sum_{p \in tp} NB(p, t) \quad (3)$$

where *SC(t)* is a spatial constraint function of time *t*, and *NB(p,t)* is a function that puts a neighboring region at high values around the fixation point *p* at time *t* from the trace list *tp*, which is likely to receive high saliency of attention. In our method, for simplicity, we choose *NB(p,t)* to have high values uniformly distributed in a small rectangular region centered at the current fixation point and with low values elsewhere.

Each time after an attention shift, the saliency map is updated by:

$$S'(t) = S(t) \bigotimes SC(t) \tag{4}$$

where $\bigotimes$ is an element-by-element multiplication between two matrices.

We include the time index here because we want to emphasize that the saliency map is dynamically changed each time an attention shift occurs to foveate the target. The attention shifting consequently causes changes in the input retinal image, and in its corresponding saliency map as well. This is the reason why we need to keep a trace of the positions of the previous fixation points in the history and transform their relative positions in the retinal coordinates to maintain consistency with each shifting. Similarly, we need to re-calculate the *SC* function each time, as well as the *IOR* function, because they all depend on their positions on the retinal images.

The spatial constraint helps to focus on the same object during the first few attention shifts (here we use five shifts) over an object. This assumption is consistent with the result of neurophysiological studies of attention shift. In the real world, objects are typically be viewed for 0.5 - 1 sec or more, with a saccade occuring every 200 - 300 msec ([15]). Therefore, statistically there would be around 2 - 5 shifts of overt attention over the object during the observation.

4. Temporal learning of attention shift invariant

A naïve approach to learn invariant representation across a sequence of attention shifts would be that responses to local features of the same object be correlated temporally, such as a Hebbian rule with a trace mechanism proposed by Földiák ([3]), or a learning rule that applies a temporal stability constraint to require the output layer neuronal responses remain constant over time in the form as follows:

$$\Delta W(t) = \gamma \times [\tilde{C}(t) - C(t)] \times S(t) \tag{5}$$

where *S* is the input neural responses, *C* is the output layer neural responses, $\tilde{C}$ is the short-term memory trace keeping a history record of *C*, and *W* is the updating rule of the weight matrix.

However, motivated by the temporal learning of position invariance as proposed by the authors in previous papers ([8], [9]), we propose that we are able to use a similar learning rule if we could find a proper candidate for the canonical representation as the reinforcement reward during the observation duration of an object across attention shift. The time interval between attention shifts is rather short when compared with time taken during self-motions of the object or even of the observer. An assumption could be made that within the duration of the first few attention shifts on a targeted object, there are no changes in the viewing condition of the object, either due to its self-movements or the observer's slow head or body motions. The specific view of the targeted object during the first few attention shifts therefore remains almost same, except for the slight positional displacement due to the attention shifting over the object. The representation of an object from one view in the scene at the coarse resolution therefore becomes a good candidate for the canonical representation, especially when position invariance is already achieved ([8]).

Based on the consideration above, we are able to give the weight updating rule to learn invariance across attention shifts. The learning rule is composed of two terms, one is a Temporal-Difference (TD) reinforcement learning term as in ([2]) , and the other is a temporal perceptual stability constraint.

$$\begin{aligned}\Delta W(t) = \eta \times [(R(t) + \gamma \times C(t) - \tilde{C}(t-1)) + \\ \kappa \times (\tilde{C}(t-1) - C(t)] \times \tilde{S}(t)\end{aligned} \tag{6}$$

which can be simplified into:

$$\Delta W(t) = \eta \times [R(t) + (1-\kappa) \times (\chi \times C(t) - \tilde{C}(t-1)] \times \tilde{S}(t) \tag{7}$$

with

$$\chi = \begin{cases} \frac{\gamma - \kappa}{1-\kappa} & \text{if } \kappa \neq 1 \\ 1 & \text{otherwise} \end{cases}$$

and

$$\Delta \tilde{C}(t) = \alpha_1 \times (C(t) - \tilde{C}(t-1))$$

$$\Delta \tilde{S}(t) = \alpha_2 \times (S(t) - \tilde{(}S)(t-1))$$

Here *R(t)* is the canonical representation as the reinforcement reward, and the parameters η, α_1 and α_2 are learning rates with predefined constant values. The weight update rule correlates this reinforcement reward *R(t)* and (an estimate of) the temporal difference of the output layer neuronal responses with the memory trace of the input layer neuronal responses. The constraint of temporal perceptual stability also requires that updating is necessary only when there is a difference between current neuronal response and previous neuronal responses kept in the short-term memory trace. The parameter κ is an importance factor and lies in the range [0, 1]. It is used to emphasize the importance of the perceptual stability constraint in driving the learning towards a better performance. When the value of κ is

near zero, the constraint term has no effect on the learning rule. The updating of the weight matrix relies totally on the TD reinforcement-learning term, in which case it is similar to the approach in ([2]), except for the longer time scale of the temporal difference used in this rule. Conversely, a value near one will give the constraint term the same importance as the TD reinforcement-learning term. We also use a sparse coding approach ([12]) to ensure a sparsely distributed neuronal responses to the input image patches.

To bound the growth of the weight matrix, the matrix can be either explicitly normalized, as in many competitive networks, or by using a local weight bounding operation ([3], [15]), the implementation of which is more biologically relevant.

5. Simulation and Results

5.1. Effects of the modified saliency map mechanism

The spatial constraint (Equation 3) is aimed at forcing the attention to stay close to the region previously visited, to some extent guaranteeing that the attention will shift within the same object for a certain duration. In this section we will examine the effect of the spatial constraint on the saliency map mechanism during the attention shift, which confines the shifts to stay near the same object in a multi-object scene. The scene is relatively simple in the sense that the scene is static, as all objects within the scene have a low probability of overlapping, and a black background is used to eliminate any distraction from the background. The images we used in this experiment are 320x240 pixels in size, and the attention window is 60x60 pixels in size. The *IOR* region is 72x72 pixels in size, and the spatial constraint is applied to a region of 90x90 pixels in size centered at the fixation point.

We use a scenario where three toys are displayed before a black background. We compare the result of attention shifts based on the saliency map without and with the spatial constraint respectively. In Figure 2, A and C show the post-attention-shift saliency maps with the *IOR* regions without and with the spatial constraint respectively. The small black rectangles in the figure are the regions influenced by the *IOR*. The saliency map is shifted accordingly when an attention shift is executed to put the target point in the center of the view window. B and D show the local features falling within a rectangle attention window accordingly. In the scene, the right-most toy has the most salient feature; therefore the first attention shift is focused on it. Without the spatial constraint, attention is likely to be shifted from the focused object to other objects that have high salient values during the observation (Figure 2 B). However, the problem can be fixed when we introduce the spatial constraint into the saliency map. As shown in Figure 2 D, the first several attention shifts stay on the same object.

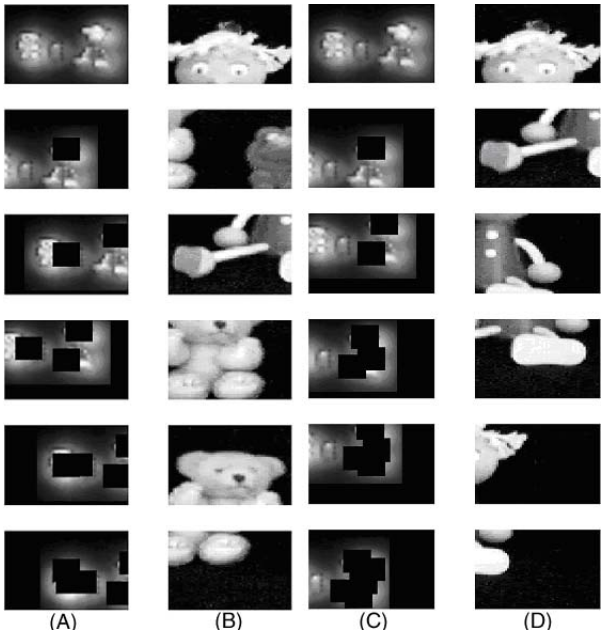

Figure 2: A sequence of attention shifts on a scene with three objects. Attention shifting is guided by a saliency map without (A shows the saliency map and B shows the local features) and with (C shows the saliency map and D shows the local features) the spatial constraint.The small black rectangles in the figure are the regions influenced by the *IOR*.

From the above demonstration, we are able to declare that with the spatial constraint employed in recalculation of the saliency map during the sequence of attention shifts, it is possible for attention to stay mostly on the same object in a relatively simple multiple-object scene. Therefore, an adequate attention shift sequence can be performed to guide the learning of position and attention-shift invariance for the following experiments.

The modified saliency map mechanism is very useful in gathering valid training data sets as input to our proposed neural network. A limitation of this method would be that it requires the distribution of the objects in a scene to be sparse, i.e., having no overlap between objects. If any two objects are placed very close, they are likely to be deemed as one object due to their spatial closeness.

One useful property is revealed from the study on this saliency map mechanism. That is that the position differences of an object on the images can be screened out when we focus on only the local features obtained across attention shifts, using the modified saliency map mechanism to perform the task of position-invariant object representation and recognition. Each time with attention selecting out the local

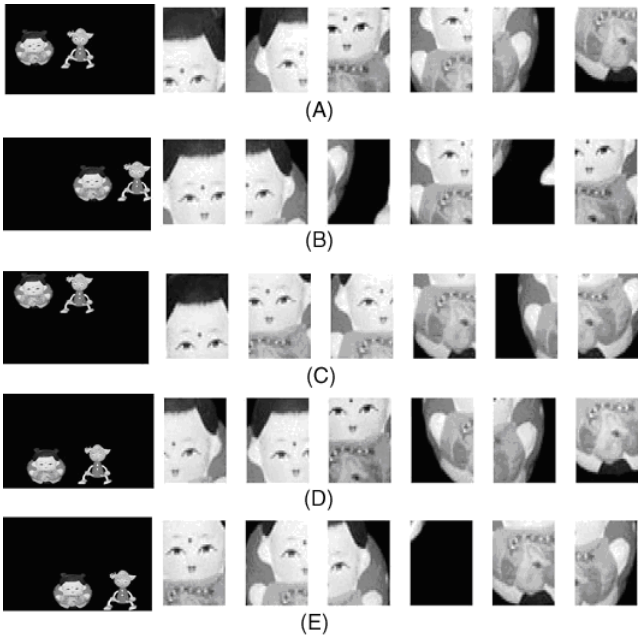

Figure 3: Local features obtained after the first six attention shifts for the same two objects appearing at five different positions.

feature associated with current fixation point, the global position information of the object is of no importance. What really matters is the content of the local feature and its relative position to the object. Two objects were placed at five different positions, and the first six attention shifts were observed following the saliency map mechanism. We notice in Figure 3 that the first few attention shifts usually select similar local features of an object appearing at different positions due to its saliency map distribution. This observation leads us to think that at a fine detail level of object representation, via temporally correlating local features of an object across attention shifts, the global position difference can be canceled out by focusing on only the attended parts of an object.

5.2. Invariance over attention shifts

Attention usually goes more easily to some unwanted features from the distractions of the background in a real world environment. To eliminate such distractions and focus solely on the objects themselves, in this experiment we will use a simple multiple-object scene where three objects are sparsely arranged in front of a black background.

The first five attention shifts were performed following the guide of the saliency map with the spatial constraint. Local features of an object with the highest saliency in the saliency map were recorded. In this implementation, after five attention shifts on an object, the region covering the

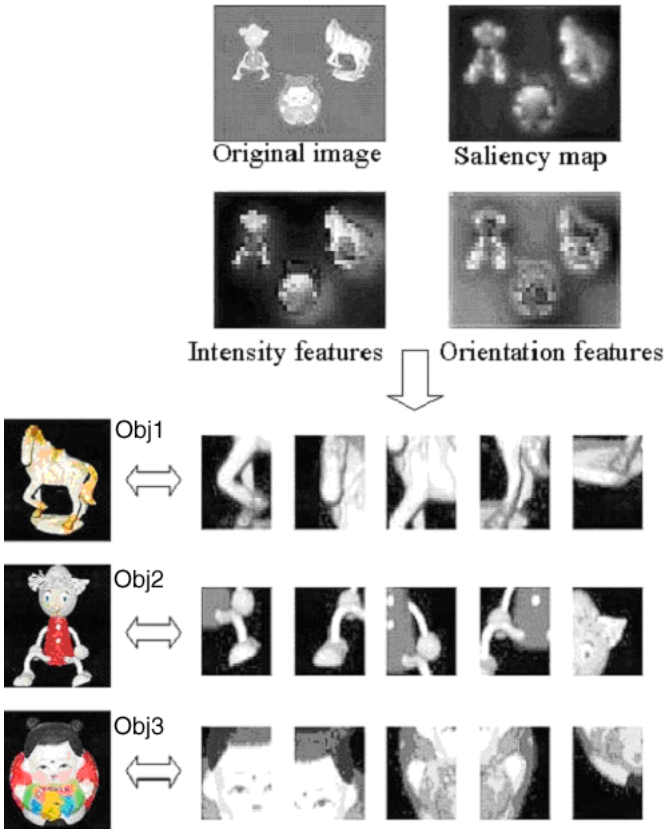

Figure 4: Sequences of attention shifts over three objects in the scene. Following the saliency map calculated as shown at the top, the attention first stays on obj1, then moves to obj2, and so on.

object will be inhibited so that the attention goes to another object in the scene. Figure 4 shows iterations of attention shifts over the three objects in the scene and their corresponding local features following the shifts. These local features are fed into the network as the training data.

In this experiment we use the sparse coding strategy for the output layer neuronal representation, so the neuronal responses to the local features across attention shifts are sparsely distributed. To understand the activity of the neurons, their responses to local features are plotted with respect to the first five attention shifts from one object at a time. The activities of the eight most active neurons are shown in Figure 5. The activity curves show each neuron favors one specific object during the attention shifts.

Finally we compare the performance of the proposed learning rule as in Equation 7 with the Hebbian trace rule ([3]) and the temporal stability constraint rule (Equation 5). The three learning rules run on the same training data set and the weight matrices are initialized to the same values as well. The learning results are illustrated in Figure

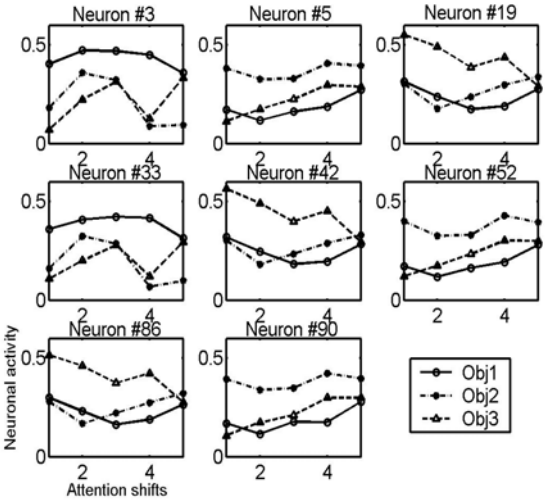

Figure 5: Neuronal activities of the eight most active neurons responding to local features belonging to three objects in the scene across five attention shifts.

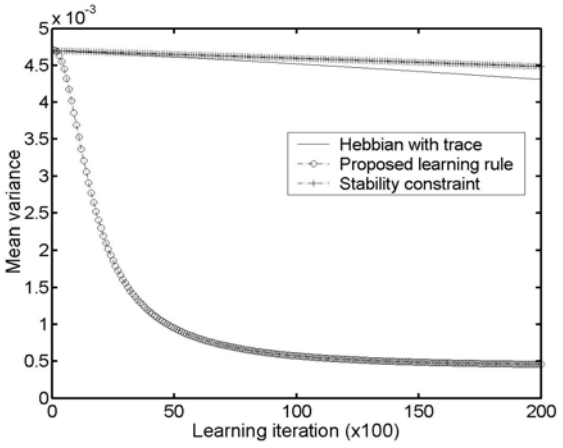

Figure 6: Comparison of three attention-shift invariance learning rules in the measurement of the mean variance over the learning iteration.

6 in the measurement of mean variance over the learning iteration. The mean variance of the output neuronal responses is sampled every 100 iterations, over the network output with respect to a set of input stimuli resulting from a sequence of attention shifts. The value of the mean variance stays low when the neuron tends to maintain a constant response to the temporal sequence of local features across attention shifts; while a higher value means less stability for the neuronal responses across attention shifts. In other words,if the model is to exhibit attention-shift-invariance, the output neuron responses should remain nearly constant and therefore have a low variance. The proposed learning rule converges faster than the other two learning rules, and its mean variance is lower than the other two as well. This simulation demonstrates that the proposed learning rule is able to learn the invariance with respect to changes resulting from attention shifts in a more efficient way.

5.3. The influence of the temporal factor κ on learning

We examine the learning performance with (when $\kappa = 1$) and without (when $\kappa = 0$) the temporal perceptual stability constraint term in the learning rule. Again the performance is evaluated by the measurement of the mean variance of the output neuronal responses in both cases over the learning iteration. The value is sampled every 25 simulation iterations. As seen from Figure 7, the learning both with and without the perceptual stability constraint term converges to a certain point with a low standard deviation, demonstrating the correctness of the learning direction. But from the figure we can also observe that, although the two curves descend over time, the one with $\kappa = 1$ descends faster than the other and reaches a lower value of standard deviation. This result reinforces the importance of the perceptual stability constraint in achieving a better and faster performance in the learning of invariance in our approach, and it also demonstrates that this proposed approach surmounts the performance of the approach in ([2]).

6. Conclusions

In this paper, we have presented a modified saliency map mechanism that uses a simple top-down task-dependent cue (a neighborhood of the current fixation point is likely to attract most attention within a short observation period), which enables attention to stay mainly on an object of interest for the first several shifts in a multiple-object scene. Then the saliency map mechanism is applied to a neural network model that learns invariant representations of objects temporally across attention shifts.

Experimental simulations have demonstrated that the modified saliency map mechanism is able to generate a sequence of attention shifts that stay mostly on a single object during the short period of observation. And the proposed neural network model performs well in the learning of invariant representations for objects in a scene with respect to position variance and attention shifts. However, invariance to scale is not considered here, and future work needs to be done on this aspect.

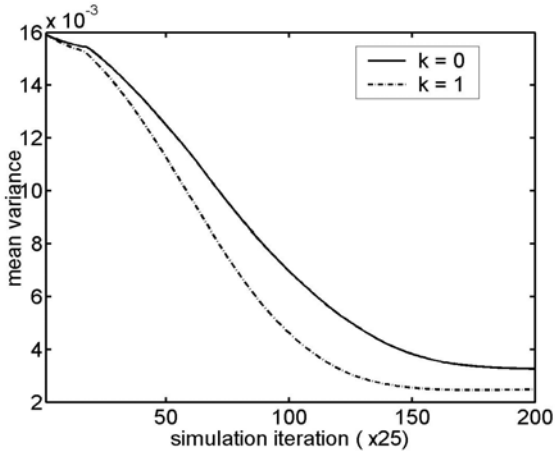

Figure 7: Comparison of learning performance using the perceptual stability constraint $(\kappa = 1)$ and not using it $(\kappa = 0)$ by the measurement of mean variance over the simulation iteration.

References

[1] Carrasco, M., and Chang, I., The interaction of objective and subjective organizations in a localization search task. *Perception and Psychophysics* **57(8)** (1995) 1134 – 1150

[2] Clark, J.J. and O'Regan, J.K., A Temporal-difference learning model for perceptual stability in color vision. *Proceedings of 15th International Conference on Pattern Recognition* **2** (2000) 503–506

[3] Földiák, P., Learning invariance from transformation sequences. *Neural Computation* **3** (1991) 194–200

[4] Hafed, Z. M., Motor theories of attention: How action serves perception in the visual system. *Ph.D Thesis, McGill University, Canada.* (2003)

[5] Heinke, D., and Humphreys, G. W., Attention, spatial representation and visual neglect: Simulating emergent attention and spatial memory in the Selective Attention for Identification Model (SAIM). *Psychological Review* **110(1)**, (2003) 29–87

[6] Itti, L., Koch, C. and Niebur, E., A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20(11)**, (1998) 1254–1259

[7] Koch, C., and Ullman, S., Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* **4**, (1985) 219–227

[8] Li, M. and Clark, J.J., A temporal stability approach to position and attention shift invariant recognition. *Neural Computation* **16(11)** (2004) 2293–2321

[9] Li, M. and Clark, J.J., Learning of position-invariant object representation across attention shifts. *In: Lucas Paletta, John K. Tsotsos, Erich Rome, et al. (Eds.): Attention and Performance in Computational Vision: Second International Workshop, WAPCV 2004, Revised Selected Papers: Springer-Verlag Berlin Heidelberg, LNCS* **3368** (2005) 57–70

[10] Maunsell, J.H.R., and Cook, E.P., The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* **357(1424)**, (2002) 1063–1072

[11] Olshausen, B.A., Anderson C.H., and Van Essen D.C., A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience* **13(11)** (1993) 4700–4719

[12] Olshausen, B. A. and Field, D. J., Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* **37**, (1997) 3311–3325

[13] Rumelhart, D.I., and Zipser, D., A complex-cell receptive-filed model. *Journal of Neurophysiology* **53** (1985) 1266–1286

[14] Wallis, G., and Blthoff, H.H., Effect of temporal association on recognition memory. *Proceedings of the National Academy of Science, USA* **98**, (2001) 4800–4804

[15] Wallis, G., and Rolls, E.T., Invariant face and object recognition in the visual system. *Progress in Neurobiology* **51**, (1997) 167–194

[16] Wolfe, J.M., and O'Neill, P., Why are there Eccentricity Effects in Visual Search? *Perception and Psychophysics* **60(1)**, 1998: 140-156