Movement Strategies for Learning in Visual Recognition

Edgar J. Bermudez*, Andrew Philippides and Anil K. Seth

University of Sussex. UK
*E.J.Bermudez-Contreras@sussex.ac.uk

Abstract

In this paper we study the role of movement strategies during learning in object recognition models. We show that a simple model, the RBF, can outperform a more complex hierarchical model, the HMAX, when rotation and scale invariance are provided by the training phase. Moreover, we assess the exploitation of temporal information by the RBF using optic flow. The results show that the RBF model can only exploit the temporal information using optic flow when the training and testing trajectories are the same. This work exemplifies the idea that the complexity of the neural mechanisms in object recognition can be understood not only in the brain but also in the interaction between brain, body and environment.

Introduction

Object recognition is a very complex computational task that has been widely studied. Whereas visual systems in nature solve this task with exceptional reliability and speed, the performance of artificial visual systems is still far from their counterpart in nature. We are interested in the exploration of ways active vision can help biologically inspired models of object recognition in autonomous agents. In order to understand the visual processes in the brain and design artificial visual systems, various models have been proposed for object recognition based on different perception theories. These models can be classified as object-based or view-based. The former category describes models that "extract" structural features or parts of the object that are view-invariant in a 3D coordinate system centred on the object. In contrast, view-based models represent objects as a combination or set of features extracted directly from the image.For a review of models and theories of perception see (Riesenhuber and Poggio, 2000; Peters, 2000). Most state-of-the-art models are view-based, which in turn, are divided by the way they extract the view-based features. Some computer-vision based models use statistical regularities extracted from the images, mainly using, template or histogram systems (bag-of-features, nearest-neighbour, etc.) (Wang et al., 2006; Zhang et al., 2006; Lazebnik et al., 2006). Others are biologically inspired, resembling the hierarchical nature of the visual cortex (Riesenhuber and Poggio, 1999; Poggio and Edelman, 1990; Serre et al., 2005; Mutch and Lowe, 2006).

Template-based models perform very well on object recognition of single object category (e.g. faces, cars, etc.). However, these methods show limitations when the object is subject to appearance modifications, suffering from high specificity and therefore, lacking invariance to object transformations. Histogram-based models show a large amount of invariance to transformations but their performance drops for general object recognition tasks (i.e. with multiple object categories) (Serre et al., 2005). Biologically inspired models for object recognition have been gaining interest because they perform very well for general purpose object recognition tasks (Pinto et al., 2008).(Serre et al., 2005), presented a modified hierarchical model based on (Riesenhuber and Poggio, 1999) and reported it to be at least comparable to the best computer vision-based systems.

A common baseline of these systems is that they do not acquire the incoming visual information by themselves, the way the visual information is presented to them is restricted by the experimenter. In some cases, these imposed restrictions can play an important role in the recognition process and hence, in the performance and evaluation of models. For example, in (Bermudez-Contreras et al., 2007), it was shown that a simple model of the primary visual cortex [(RBF), (Howell and Buxton, 1995; Poggio and Edelman, 1990)]can perform just as well as a complex hierarchical model [(HMAX), (Riesenhuber and Poggio, 1999)] when natural conditions are present and the former is augmented by a simple 'attentional mechanism'. In addition, in (Pinto et al., 2008), a comparison between state-of-the-art object recognition systems and a simple V1-like model is carried out. They show that by imposing conditions on the way the visual information is presented to the systems (taken from databases of natural images), the simple biological model outperforms all of the state-of-the-art systems presented.

Given this importance, an additional fact to be considered when studying or modelling visual systems is that, in nature, visual systems are active. In active vision, the control of acquisition of visual information is part of the system.

It is well known that the restrictions imposed by the interaction between body and environment can facilitate visual processing (Aloimonos, 1993). Active vision strategies are important both in recognition and in visual learning. For instance, insects utilise specific movement strategies in order to learn how to perform various visually guided tasks including homing, navigation, and finding conspecifics (Lehrer and Bianco, 2000; Collett and Rees, 1997; Carwright and Collett, 1983). Therefore, in the study and modelling of visual systems, it is important to consider the way incoming visual information is acquired.

There have been relatively only a few studies that analyse the role of motion and object recognition. Arbel and Ferrie (2002, 2001) propose a paradigm to facilitate object recognition of a system. Most of the research work on visual systems in autonomous robots is oriented to navigation, with some exceptions. For example, in (Gvozdjak and Li, 1998), the importance of active vision in an agent for recognition tasks is highlighted using a hierarchical template-based model. In (Andreasson and Duckett, 2003), an exploratory study of object recognition using a mobile robot with an omni-directional camera is presented. The robot tracks extracted low-level features and constructs higher level features for object identification. While these works are exploratory, they show the potential of active vision in object recognition tasks. Furthermore, given the success of object recognition models that reflect the hierarchical nature of the visual cortex, we evaluate the importance of the role of visual information acquisition processes in these models.

In this work, we analyse how different movement strategies during training affect the performance of a version of the HMAX model and the RBF model. We employ a mobile agent in a simulated world with a simple object recognition task. We find that movement strategies are exploited differently for both models. When a movement strategy does not provide the opportunity to develop rotation and translation invariance, the HMAX model performs better than the RBF. However, when such opportunities are provided, the RBF model outperforms the HMAX model. These results suggest that exploiting the dynamics of agent-environment interaction can, in certain circumstances, obviate the need for complex models of visual object recognition. We also consider whether the RBF model performance can be further enhanced by training and testing using dynamic visual signals generated during each movement strategy. We find that such time dependent information is only exploited by the RBF model when the training and testing movement strategies are the same.

Methods

The following experiments involve a simulated agent performing a simple object recognition task. The agent-environment system comprises a simple wheeled agent in a flat planar environment containing two objects (a 'kettle' and a 'bolt'), simulated using the OpenGL library. It is important to mention that the goal of this exploratory study is to investigate how the way of acquiring visual information can affect the recognition process in two models of object recognition rather than comparing their performance. The visual object recognition system of the agent comprises three parts: a 'blob detection mechanism' (BDM), an 'analysis module' consisting of either the HMAX or the RBF model, and a 'classifier module' which classifies the output of the analysis module into one of two categories (kettle' or 'bolt'). Each experiment consisted of two phases. First, a learning phase in which the agent followed one of four different movement strategies (see figure 3) while collecting training views which are used to train either the HMAX or the RBF model. Second, a testing phase, during which the agent follows a separate movement strategy (see figure 2A) while collecting views used to test object recognition performance.

Blob detection mechanism. The BDM selects the area of the visual field containing the object. It is the 'attentional mechanism' referred to in the introduction. Cropped regions returned by the BDM are normalised to 60×80 pixels (a blob) before being processed by the analysis module. The BDM therefore provides some robustness to changes in the size of objects. The order in which blobs are processed by the visual system is determined by the area of the blob detected. The larger the area of the blob in the visual field, the higher the priority of being processed by the visual system [see (Bermudez and Seth, 2007) for a more detailed explanation of the visual system of the agent].

Analysis module. The analysis module processes visual information coming from the BDM (current views). These views are processed by either the HMAX or the RBF model. The RBF model emulates simple cells in the primary visual cortex, V1, based on the function of receptive fields implemented by using Derivative of Gaussian filters with different orientations and sizes. The RBF model uses four different sizes of square filters with sides of 7, 11, 15 and 21 units and 0, 45, 90, and 135 degrees of orientation. There are therefore 16 different filters in total with outputs responding to oriented 'edges' at different spatial scales. Therefore, this model responds only to a collection of simple primary features. In contrast, the HMAX model proposed in (Riesenhuber and Poggio, 1999) is a hierarchical model resembling the ventral pathway in the visual cortex. The HMAX model consists of four layers (S1, C1, S2 and C2) resembling simple and complex cells in the ventral pathway. Units in S1 would correspond to simple features detected by the different filters of the RBF model. The next layer C1, responds to the most salient features in S1 at each orientation and spatial scale. It achieves this by applying max pooling operations (extracting the most salient features across the different orientations and spatial scales) over the selected features in S1. The next layer, S2 combines the output of C1 into a higher order features sets which are passed into C2 where the out-

puts are again max pooled to produce a vector of the dominant features detected along the hierarchy (see details of the original model in (Riesenhuber and Poggio, 1999) and details of this implementation in (Bermudez-Contreras et al., 2007)). By virtue of its hierarchical structure, this model shows a degree of translation and scale invariance.

Classifier module.

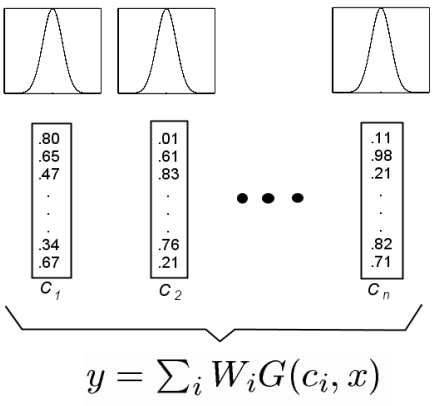

Figure 1: View Tuned Unit (VTU): each view vector c_i is the centre of a Gaussian function. The more similar a vector x is to a centre, the stronger the response of the unit.

The classifier module is based on the work of (Edelman and Duvdevani-Bar, 1997; Poggio and Edelman, 1990). It uses view tuned units (VTU) to recognise objects. There is one VTU for each object. Each VTU is trained to respond so that it responds strongly to test views that are similar to the training views of the object. Each VTU (see figure 1) corresponds to a set of radial basis functions (RBF unit). A RBF unit is a Gaussian function G centered on each view c_i collected during the training phase. The response of each RBF is given by $G(c_i, v) = e^{-\|c_i - v\|/\sigma_i^2}$ where v is the vector that is being classified.

The response y of each VTU for a test vector x is given by $y = \Sigma_i W_i G(v_i, x)$, that is, y is a linear combination of weights W_i and $G(v_i, x)$. The weights W_i are computed using an inversion matrix procedure (the details are described in the appendix section in (Bermudez-Contreras et al., 2007)).

Movement strategies. The training views were collected when the agent was navigating around or approaching the object, following one of four different trajectories (these trajectories are called movement strategies throughout the rest of this paper). The training views were processed by the analysis module (using the RBF model or the HMAX model) and learned and classified by the classifier module.

The properties of the set of training views changed depending upon the movement strategy used during their collection. These strategies were designed in order to provide different properties in the training views (see figure 3). movement strategy 1 allows the agent to exploit the different training distances while using the same point of view. Therefore, the training views using this strategy only provide variance in scale. Strategy 2 provides a small degree of variance in perceived rotation (points of view) and a small degree of variance in scale as well, since the agent is passing in front of the target object. The point of view changes slightly as the distance between the agent and the object changes. Strategy 3 provides only variance in points of view since the distance between the agent and the object is always the same, while the point of view changes for each training view. Strategy 4 provides a combination of variance in scale and point of view since the distance and the perspective of the agent to the object are changing continuously. For each strategy, 16 training views are taken for each object at regular time intervals. Therefore, training phases varied in length from 160 to 200 time steps depending on the movement strategy used.

In the testing phase, the agents followed a trajectory (testing trajectory) that differs from the movement strategies used in the learning phase. The testing trajectory was designed so it would resemble a plausible situation in the real world where the objects are approached in a natural way that provides views of the objects from multiple angles and scales (see figure 2). The testing phase lasted for 200 time steps. During the first 55 steps (period 1) object 1 was present in the visual field and during 125-180 (period 2) object 2 was present in the visual field (see testing trajectory in figure 2).

Optic flow. An important consequence of actively exploring the world is the visual motion that this evokes. Optic flow is defined as this type of motion. In our study, we calculated a simple approximation of optic flow by taking the absolute difference between consecutive views i and j, $F = 1/2 \cdot \|RBF(i) - RBF(j)\|$ after being processed by the RBF model. For the rest of the paper, F is referred to as RBF optic flow.

Experiment 1: Movement strategies

To assess the role of active vision in the object recognition models, we tested the RBF and HMAX using the different movement strategies shown in figure 3 during the learning phase. The models are then tested while the agent traverses the testing trajectory shown in figure 2A. The results are shown in figure 3.

For strategy 1, HMAX outperforms the RBF model. Since this movement strategy presents the objects from a single point of view, the models can only acquire scale invariance. For a simple model like the RBF, this strategy would only work if the objects were viewed from a similar perspective to training during the test phase. Since this is not the case (the point of view is changing and is different from training), the RBF model cannot closely match test views to training. However, the HMAX model is able to generalise when a limited point of view is provided during training. This is

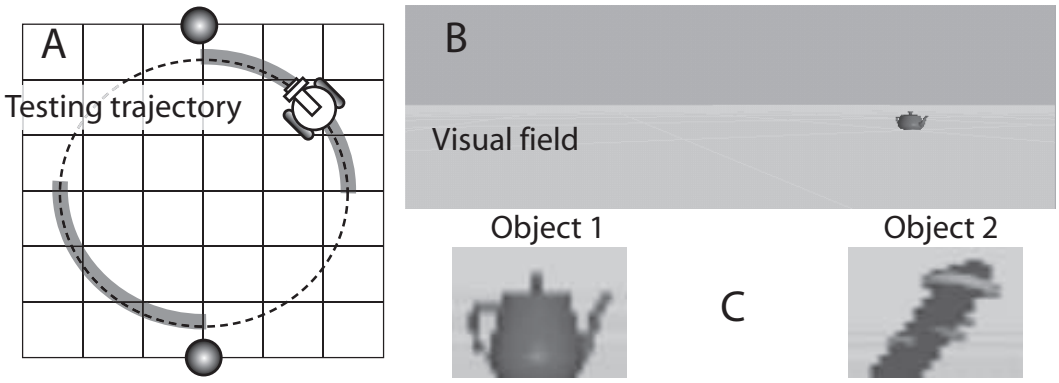

Figure 2: A. Testing trajectory: the grey segments represent the periods where objects were present in the visual field. B. Visual field of the agent: shows object 1 in the field of view. C. Sample views of object 1 and object 2: object 1 is a rounded object so it does not have a significant variability to rotation, in contrast, object 2 has a significantly higher variability to rotation due to its vertical inclination.

because the features extracted by HMAX from a single perspective capture higher order properties of the objects which are in some sense independent of the angle it is viewed at. For strategy 2, the results are very similar to the previous case as the training views are again taken from a limited set of angular positions. However, when the point of view is varied significantly during the training phase in strategy 3, the RBF model's performance increases greatly. Since the number of points of view is significantly increased, the RBF can achieve a close match between the training and the test views. In contrast, HMAX's performance decreases, demonstrating that its discriminability can be reduced when the variability of the training views is increased. Similar results are obtained for strategy 4 where both point of view and scale are changed during training.

The reason the models' performance changes with different movement strategies has to do with the way the objects change with the movement of the agent and also with the features detected by each model. In particular, the variability of the objects to rotation is significantly different. As object 1 is quite round (see figure 2), its image does not change significantly when the agent rotates around it (especially at large distances). In contrast, object 2 has a vertical orientation which makes it very variable when the point of view is changed.

Since the RBF model responds mainly to oriented edges (as it simply comprises a set of differently oriented filters at different spatial scales), its response depends on a close match between the test and the training views and we would expect it to fail when a close match is not possible. When the points of view are limited (strategy 1 and strategy 2), since the features detected for object 1 do not change significantly with the change of perspective (along the testing trajectory), the RBF model has a relatively close match between training and test views. Object 2 is difficult to discriminate, however, as it changes significantly along the testing trajectory. Thus the overall performance on these strategies is around 50%. Because the HMAX model acts on a combination of the dominant features detected by the RBF (since its first layer is the RBF), it responds to a more generalised pattern of features, rather than a close match. Since object 1 does not change significantly, the dominant features will be the ones responding to the main orientation of the object (horizontal). For object 2, if the object is seen from a single point of view, the dominant features will be the ones corresponding to the main orientation of the object, roughly 30 degrees from the vertical in the case of strategy 1. These features (which form the HMAX template for object 2), will be different to the dominant features detected for object 1 (which form the HMAX template for object 1), so the discriminability of the HMAX model is high in this case.

In contrast, when the point of view is varied significantly during the training phase (strategies 3 and 4), the RBF achieves a close match. Since there are more points of views in the training set, the model can cope with object rotation. In the case of the HMAX model, since object 2 changes its orientation during training, the model extracts dominant features in many orientations, which form a very general template and thus decrease object discriminability. This scenario is depicted in figure 4 which shows the models' output after training with strategy 3. The objects are within the field of view in different periods (grey segments in figure 2) during the 200 time step trial. In period 1 (1-55 time steps) object 1 is within the field of view, and in period 2 (125-180 time steps) object 2 is within the field of view. For the RBF model, the agent can correctly discriminate both objects. Note the peak in output that corresponds to a close match between test and training view (around time step 37). In the case of the HMAX model, while there is no problem with period 1, in period 2 discriminability is reduced significantly.

Similarity maps further explain the discrimination ability of the models (figure 5). A similarity map is a diagram representing the similarity between the current view (the one

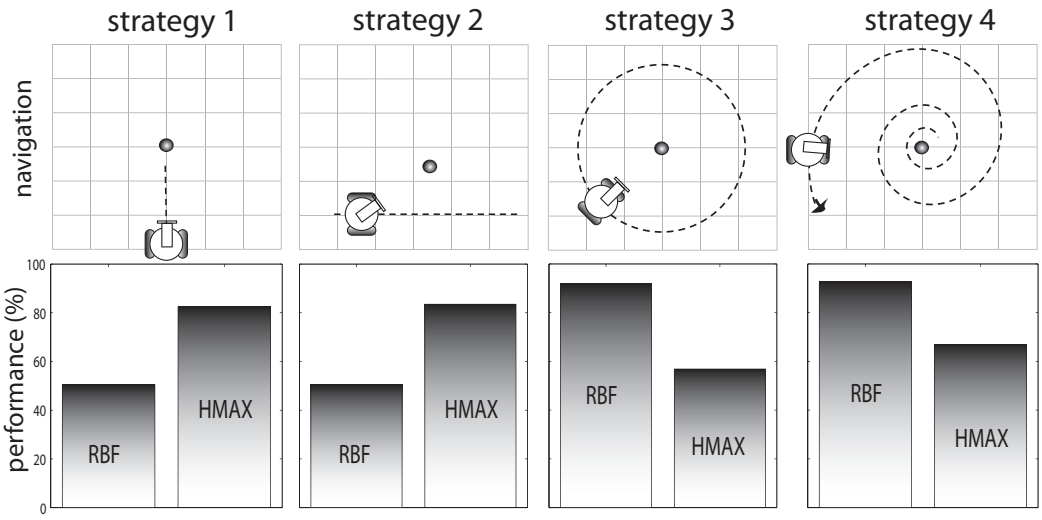

Figure 3: Movement strategies and models' performance. The performance of the models refers to the number of times the model has a correct guess over the test phase. During the following trajectories, the agent takes snapshots at uniform intervals. Strategy 1: the agent approaches the object in a straight line. Strategy 2: the agent passes the object following a straight line. Strategy 3: the agent circles the object with a fixed radius. Strategy 4: the agent spirals the object. The performance of the RBF model increases when the movement strategies allow it to exploit the rotational information during training. In contrast, the HMAX model performance decreases when the model is exposed to multiple rotational views during training in strategies 3 and 4.

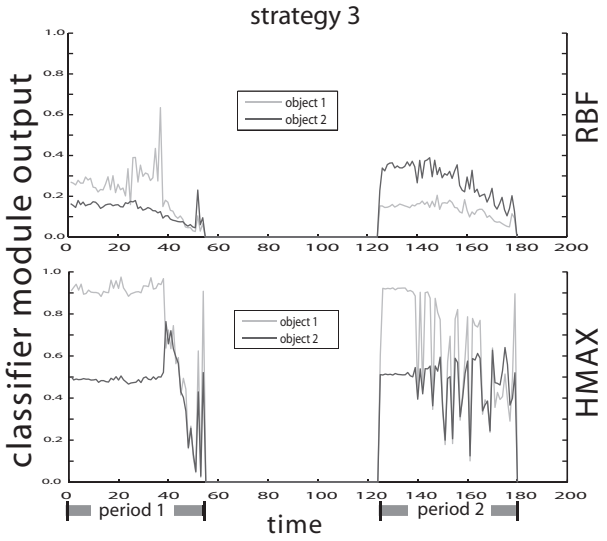

Figure 4: RBF and HMAX models activity during the test phase using strategy 3. When the movement strategy provides multiple points of view during the learning phase, the RBF can have a close match between the training and the test views. In contrast, the HMAX model decreases its discriminability when more points of view are considered. Period 1 represents the time when object 1 is within the visual field. Period 2 is the time when object 2 is within the visual field.

extracted from the visual field) and the training views of the objects (Y axis) at every time step (X axis). Every point in the map has a grey-scale value dependent on the distance between the current view and the training view after processing by the analysis module. The darker a point, the smaller the distance between the views, where distance is the sum of the absolute difference between the views. Each map is divided in two periods which correspond to points where the objects are in the agents' visual field (see figure 2). In the first 55 time steps (period 1), object 1 is present in the visual field and during period 2 (from 125-180), object 2 is in the visual field.

The upper part of figure 5 shows the similarity between views for the RBF, while the lower shows the similarity map for the HMAX model (HMAX views). If a model was responding correctly, we would expect darker areas in the lower region of period 1 and in the upper region of period 2. The similarity map for the RBF has these general features as it has acquired a degree of both rotation and scale invariance from the training trajectory. The responses of the HMAX model however, show that the higher level features extracted for each object are too similar for the two objects to be discriminated reliably.

Thus we see that the performance advantage of the complex HMAX model over the RBF can be achieved by an active vision strategy which uses its motion to provide generalised rotational information. Moreover, we note that HMAX can fail if the views provided to it are too dissimilar.

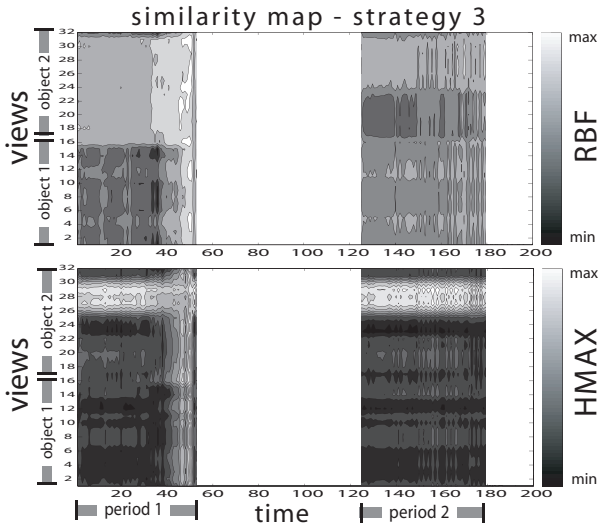

Figure 5: Similarity maps of the models using strategy 3. The darker the regions in each map, the more similar the corresponding views. For the RBF map there is an obvious darker region in the left lower area (corresponding to the views of object 1) for the first period, and a smaller darker region in the right upper area (corresponding to the views of the object 2). In contrast, for the HMAX similarity map dark areas appear during both periods for views associated with both objects.

Experiment 2: RBF Optic flow

Above, we have seen how the RBF exploits multiple viewpoints in training to achieve reliable object discrimination. However, embodied visual systems gather information by moving not only in space (defining the perceived properties of the world) but also in time. In these experiments we explore the role of time dependency in the presentation of the training views during learning. To do this we have assessed the performance of the RBF model when we provided it with optic flow type information (see Methods). The performance of the RBF model with and without optic flow are shown in table 1. The results are broadly similar showing that optic flow information can be exploited by the RBF and provides the same invariances to rotation and scale as when the model was trained on a series of static images (Figure 4).

strategy	non-optic	optic
1	50	51
2	51	56
3	92	73
4	94	95

Table 1: Comparison of the performance (%) of the RBF model with and without optic flow when using the 4 movement strategies. The performance refers to the number of times the models guess correctly over the number of time steps in the test phase.

Time dependency in the recognition process. One of the important properties of optic flow is the time dependency imposed in the recognition process. While the experiment above shows that the model is able to take advantage of differences between successive images, it does not tell us whether it is using this temporal structure. That is, it does not tell us whether the order in which the views are presented during the learning phase is important. To test this, we trained the RBF model using optic flow with views taken using strategy 3 as usual or with the order of training views randomised. To further emphasise the effects of temporal structure the test trajectory used was the same as the training trajectory.

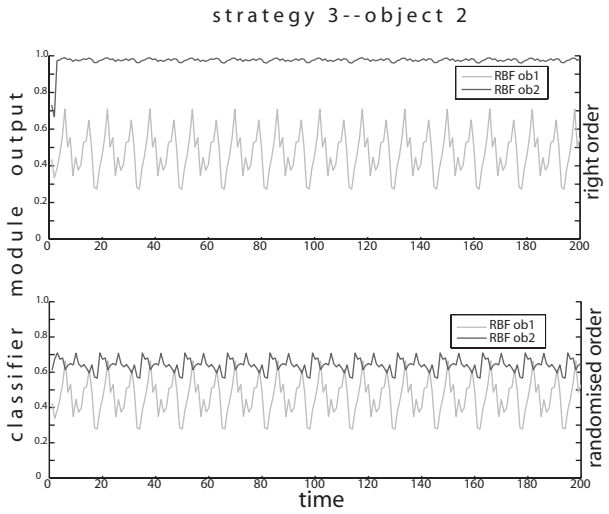

Figure 6: RBF model activity trained using strategy 3 and tested in the same trajectory with randomised ordered training views. (A) model activity for normal conditions (B) model activity of random ordered training views.

Given that object 1 is a rounded object and so appears similar from any perspective while object 2 is more rotationally variable, object 2 was used for this experiment. Figure 6 shows the model activity in using both non-randomized (top) and randomized (bottom) view sequences when circling object 2. While the object can be discriminated in both cases, in the randomised training case outputs from both VTUs are very similar. Results (not shown) confirm that for object 1, variations in the order of the presentation during the training phase are not as relevant as for object 2. These results show that if the same movement strategy is used during training and testing phase, the RBF model with optic flow can exploit the time dependency imposed in the strategy. We next consider what happens when the trajectory is not the same: Is the optic flow-based model robust to changes in the trajectory?

Using a different test trajectory. Robustness in the recognition signals is an important issue when using movement strategies. It is desirable to have some degree of robustness in the movement strategies when testing an object recognition model. In this section, we test the optic flow

strategy in the visual system to certain perturbations in the testing trajectory or in the order of the training views. Initially, to test whether the RBF optic flow changes the activity of the RBF model when having different training and testing trajectories, we used strategies 3 and 4 during the learning phase (we used these strategies because they provide higher rotational and scale variation thus maximizing optic flow), and the testing trajectory during the testing phase (as in Experiment 1). If the order of the presentation of the

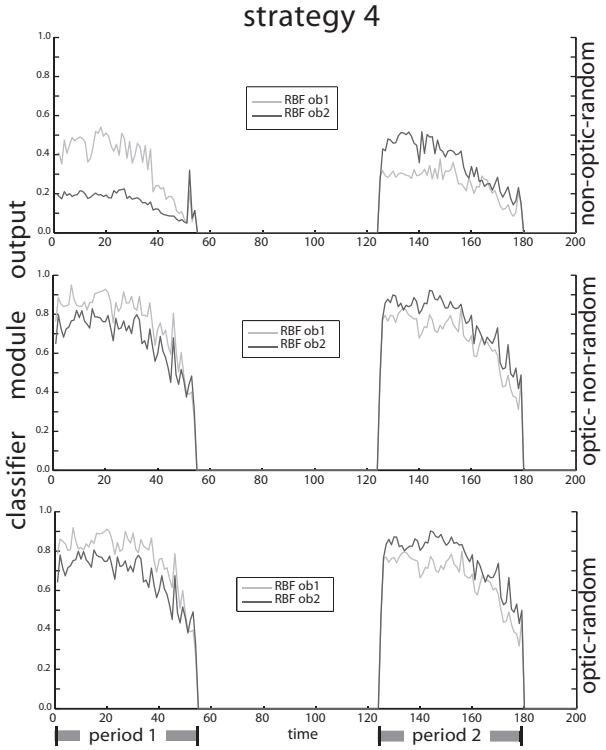

Figure 7: RBF model activity for strategy 4 under various conditions. Top: no optic flow and randomized training view, model activity is the same as in Experiment 1. Middle: optic flow and non-randomized training views. Bottom: optic flow and randomised order of training views.

views during the training phase is important, we expect the model activity to be affected when the order of the training views is randomised. Figure 7 shows model activity in three cases: randomized training views without optic flow (top), non-randomized views with optic flow (middle), and randomized views with optic flow (bottom). In the case of non-optic flow scenario (top), the activation is the same as in Experiment 1 (figure 4) since there is no temporal information present. However, when using optic flow the model activity is not significantly affected by randomizing the training views (compare middle and bottom panels of 7). These results therefore show that the order of the training views does not significantly affect the RBF model activity when using optic flow, and when the training and testing movement strategies are not the same. Thus, in contrast to the previous result (in figure 6), under these conditions the temporal information provided by optic flow is not exploited by the RBF model.

Conclusion

In this paper we have compared the performance of the RBF and HMAX models, on their performance when utilizing embodied movement strategies for training and testing. In the first experiment, four different movement strategies were used to collect the training views and a single, distinct testing strategy was used to assess object recognition performance. Each training strategy offered different degrees of variation in point of view and distance, potentially supporting the development of rotation invariance and scale invariance respectively. When no rotation variance is present in the training views, the HMAX model shows a good performance. However, when more points of view are provided, not only does the RBF model outperform the HMAX model, but the HMAX model performs worse than before. Thus, in what arguably reflects natural viewing conditions, when incorporating variance in both point-of-view and distance, the simple RBF outperforms the more complex HMAX model. In the second experiment, the role of time dependent visual information in the learning process was tested using the RBF model. We found that an RBF model trained on an approximation of optic flow could exploit the temporal information in the difference of consecutive views but only in the restrictive condition in which training and testing trajectories were identical. However, when there are significant differences between training and testing strategies, the RBF model is unable to take advantage of this temporal information. This result suggests that optic-flow style information cannot be assumed to improve visual processing in these conditions and invites further modelling to investigate how such information can best be leveraged by simple models of object recognition.

Our results exemplify the idea that the natural computations underlying adaptive behavior are best understood as being implemented not only in the brain of an organism, but as well in the interactions that cut across brain, body and environment. Improved insights into these natural computations are likely to support the development of enhanced artificial object recognition technologies. This work can be extended in different directions. One is considering more objects in the simulated world. In addition, a study of conditions where temporal information can be exploited to improve object recognition in mobile agents.

Acknowledgements

Edgar Bermudez was funded by the National Council of Science and Technology (CONACyT, Mexico). Andrew Philippides was funded by EPSRC grant GR-T08753-01

References

Aloimonos, Y., editor (1993). *Active Perception*. Erlbaum, Hillsdale, NJ.

Andreasson, H. and Duckett, T. (2003). Object recognition by a mobile robot using omni-directional vision. In *Proc. Eighth Scandinavian Conference on Artificial Intelligence (SCAI 2003)*.

Arbel, T. and Ferrie, F. P. (2001). Entropy-based gaze planning. *Image and Vision Computing*, 19(11):779–786.

Arbel, T. and Ferrie, F. P. (2002). Interactive visual dialog. *Image and Vision Computing*, pages 639–646.

Bermudez, E. and Seth, A. (2007). Simulations of simulations in evolutionary robotics. In Almeida e Costa, F., Mateus Rocha, L., Costa, E., Harvey, I., and Coutinho, A., editors, *Proc. European Conference of Artificial Life (ECAL)*, pages 796–806. Springer-Verlag.

Bermudez-Contreras, E., Buxton, H., and Spier, E. (2007). Attention can improve a simple model for visual object recognition. (In Press) Image and Vision Computing.

Carwright, B. and Collett, T. (1983). Landmark learning in bees: Experiments and models. *Journal of Comparative Physiology*, 151:521–543.

Collett, T. S. and Rees, J. A. (1997). View-based navigation in hymenoptera: multiple strategies of landmark guidance in the approach to a feeder. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 181(1):47–58.

Edelman, S. and Duvdevani-Bar, S. (1997). A model of visual recognition and categorization. *Phil. Trans. R. Soc. Lond. B*, 352(1358):1191–2002.

Gvozdjak, P. and Li, Z. N. (1998). From nomad to explorer: active object recognition on mobile robots.

Howell, J. and Buxton, H. (1995). Receptive fields functions for face recognition. In *Proc. 2nd International Workshop on Parallel Modelling of Neural Operators for Pattern Recognition*, pages 221–226.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. volume 2, pages 2169–2178.

Lehrer, M. and Bianco, G. (2000). The turn-back-and-look behaviour: bee versus robot. *Biological cybernetics*, 83(3):211–229.

Mutch, J. and Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. volume 1, pages 11–18.

Peters, G. (2000). Theories of three-dimensional object perception: A survey. *Recent research developments in pattern recognition*, 1:179–197.

Pinto, N., Cox, D. D., and Dicarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4(1):151–156.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize 3-d objects. *Nature*, 343:263–266.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nature Neuroscience*, 3:1199–1204.

Serre, T., Wolf, L., and Poggio, T. (June, 2005). Object recognition with features inspired by visual cortex. In *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Patter Recognition (CVPR)*.

Wang, G., Zhang, Y., and Fei-Fei, L. (2006). Using dependent regions for object categorization in a generative framework. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1597–1604.

Zhang, H., Berg, A. C., Maire, M., and Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2126–2136.