Knowledge-driven Gaze Control in the NIM Model

Joyca P. W. Lacroix (j.lacroix@cs.unimaas.nl)
Eric O. Postma (postma@cs.unimaas.nl)
Department of Computer Science, IKAT, Universiteit Maastricht
Minderbroedersberg 6a, Maastricht, The Netherlands

Jaap M. J. Murre (jaap@murre.com)
Department of Psychology, University of Amsterdam
Roeterstraat 15, 1018 WB Amsterdam, The Netherlands

Abstract

In earlier work, we proposed a recognition memory model, the Natural Input Memory (NIM) model, that operates directly on digitized natural images. When presented with a natural image, the NIM model employs a biologically-informed perceptual preprocessing method that takes local samples (i.e., eye fixations) from the image and translates these into a similarity-space representation. Recognition is based on a matching of incoming and previously stored representations. In this paper, we investigate whether it is possible to extend the NIM model with a gaze control mechanism to select relevant eye-fixation locations based on scene-schema knowledge and episodic knowledge. We perform two experiments. In the first experiment, we test whether the similarity-space representations can be used to infer scene-schema knowledge of a specific category of natural stimuli, i.e., natural face images. In the second experiment, we examine how the model can use the scene-schema knowledge in combination with stored episodic knowledge to direct the gaze toward relevant spatial locations when performing a categorization task. Our results show that the spatial structure of face images can be inferred from the NIM model's similarity-space representations, i.e., scene-schema knowledge can be acquired for the category of face images. Moreover, our results show that extending the NIM model with a gaze control mechanism that combines scene-schema knowledge with stored episodic knowledge, enhances performance on a categorization task.

Introduction

We proposed a recognition memory model called the Natural Input Memory (NIM) model (Lacroix, Murre, Postma, & Van den Herik, 2004, 2006). The NIM model differs from existing memory models in that it encompasses a biologically-informed perceptual front-end. As a consequence, the NIM model can take realistic image-like stimuli as input. The perceptual front-end takes local samples (i.e., eye fixations) at selected locations in a natural image and translates these into feature vectors. The extracted feature vectors contain information on oriented edges at multiple scales and form an efficient basis for object recognition (see, e.g., Rao & Ballard, 1995). Perceptually similar image regions result in similar feature vectors. A representation space in which similarity corresponds to proximity is often called a 'similarity space' (e.g., Shepard, 1957; Edelman, 1995; Edelman & Duvdevani-Bar, 1997; Goldstone & Son, 2005). Recognition decisions in the model are based on a matching of stored and incoming similarity-space representations. By fixating different image regions, the NIM model has a natural way of dealing with spatial (overt) attention. Since humans tend to place fixations at or near contours in a visual scene (e.g., Norman, Phillips, & Ross, 2001), the initial versions of the NIM model 'fixated' randomly along the contours in an image. However, random sampling along contours can hardly be considered to agree with the active context-dependent scanning of a visual scene that is performed by humans (e.g., Rajashekar, Cormack, & Bovik, 2002). In the dynamic process of actively scanning the visual scene, eye fixations are guided by bottom-up and top-down processes (Karn & HayHoe, 2000; Henderson, 2003; Oliva, Torralba, Castelhano, & Henderson, 2003). Several studies showed that bottom-up processes draw the eyes toward salient visual features such as high edge density (see, e.g., Mannan, Ruddock, & Wooding, 1996) and local contrast (see, e.g., Parkhurst & Niebur, 2003). Based on these findings, many models of gaze control employed a bottom-up approach (e.g., Braun, Koch, Lee, & Itti, 2001; Rao, Zelinsky, Hayhoe, & Ballard, 2002). Often, a so-called 'saliency map' is constructed that marks those image regions that are visually distinct from their surround in one or more visual features (Koch & Ullman, 1985). Gaze is directed to locations that are marked as highly salient on the saliency map. While visual saliency has often been used to control gaze in artificial systems, the use of top-down processes has not received as much attention. Top-down processes rely on stored knowledge to select the most informative location to orient the eyes (see, e.g., Henderson, 2003). Several studies showed that human gaze control relies more on top-down processes than on bottom-up processes when performing an active visual task with meaningful stimuli (see, e.g., Oliva et al., 2003). The top-down processes are driven by two types of knowledge. One type of knowledge is episodic knowledge, which is obtained from one or more encounters with the stimulus or visual scene (Henderson, 2003). A second type is knowledge about the general spatial arrangement of a certain category of stimuli or scenes (i.e., scene-schema knowledge; Henderson, 2003). Since natural objects and scenes within a particular category contain spatial regularities, the similarity-space representations contain information about both the object identity and its spatial structure.

In this paper, we investigate whether it is possible to combine episodic and scene-schema knowledge in the NIM model to guide the selection of relevant fixation locations. Spatial knowledge is not represented explicitly in our model. However, since natural objects and scenes within a particular category contain spatial regularities, the similarity-space representations contain information about both the object identity and its spatial structure (i.e., scene-schema knowledge).

For our investigations, we performed two experiments. In the first experiment, we test whether the similarity-space representations can be used to infer scene-schema knowledge of a specific category of natural stimuli, i.e., natural face images. In the second experiment, we examine how the model can use the scene-schema knowledge in combination with stored episodic knowledge to direct the gaze toward relevant spatial locations in the image.

The outline of the remainder of this paper is as follows. In the following section, we briefly discuss how similarity space representations are built in the NIM model. Then, in the next section, we attempt to train a neural network to map the similarity-space representation to the spatial representation (the coordinates of the image area) for face images. Subsequently, we examine how the model can use the mapping in combination with episodic knowledge of previously encountered stimuli to direct gaze toward relevant spatial locations in the image. Finally we compare gaze control in the extended NIM model with knowledge-driven gaze control in existing systems and draw two conclusions. The first conclusion is that spatial structure of faces can be extracted from their similarity-space representations. The second conclusion is that a gaze control mechanism that combines scene-schema knowledge with episodic knowledge of a specific set of encountered stimuli, can enhance performance on a face-categorization task.

The NIM model

The NIM model encompasses: (1) a preprocessing stage that translates a natural image into a feature-vector representation, and (2) a memory stage that either stores the representation or makes a recognition decision based on a matching of an incoming representation and previously stored representations. Fig. 1 shows a schematic diagram of the preprocessing and the memory stages of the NIM model. The face image is an example of a natural image. The left and right side of the diagram correspond to the two stages of the NIM model: the perceptual preprocessing stage (left) and the memory stage (right).

Since our first experiment focuses on the NIM model's similarity-space representations, we first discuss the extraction of representations (i.e., the perceptual preprocessing stage). The memory stage will be discussed later when the model is tested on a categorization task.

The preprocessing stage in the NIM model is based on the processing of information in the human visual system. Motivated by eye fixations in human vision, the preprocessing stage takes local samples (i.e., fixations) at selected locations in a natural image. The features extracted at fixation locations in the face image consist of the responses of different derivatives of Gaussian filters. The filter responses model the responses of neurons in the visual processing area V1. By applying a set of filters at multiple scales and orientations, a representation of the image area centered at the fixation location is obtained (Freeman & Adelson, 1991). After feature-vector extraction, we apply a principal component analysis (PCA) to the extracted feature vectors in order to reduce their dimensionality. We selected the first 50 principal components, since it has been shown that approximately 50 components are sufficient to accurately represent faces (e.g., Hancock, Burton, & Bruce, 1996).

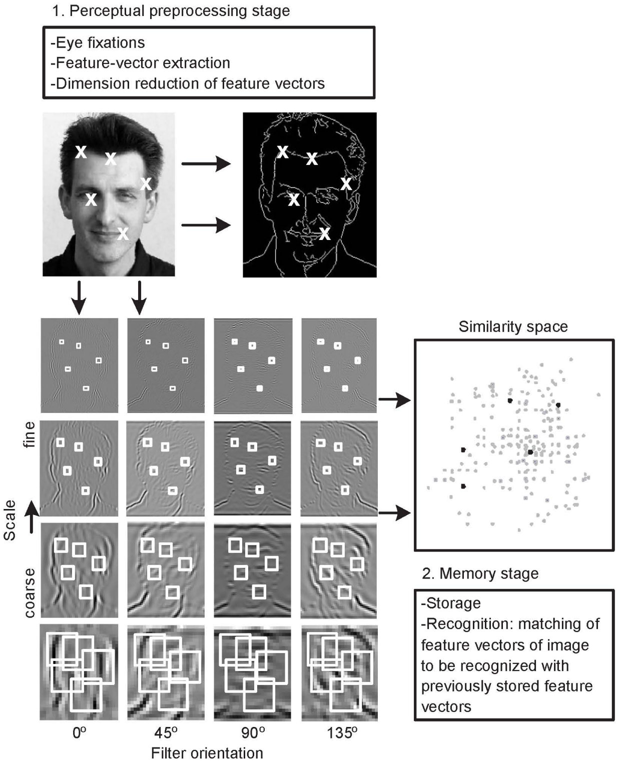

1. Perceptual preprocessing stage
-Eye fixations
-Feature-vector extraction
-Dimension reduction of feature vectors
X
fine
Scale
coarse
0°
45°
90°
135°
Filter orientation
Similarity space
2. Memory stage
-Storage
-Recognition: matching of feature vectors of image to be recognized with previously stored feature vectors

Figure 1: The Natural Input Memory (NIM) model.

In Fig 1, the sets of white squares in each of the filtered images correspond to the fixated image regions (centered at the fixation locations shown as crosses in the face image) and the five bold dots in the similarity space represent the corresponding feature vectors (after PCA) that contain the contents of the filtered images within the fixated regions. A preprocessed image is thus represented by a number of low-dimensional feature vectors, each containing partial information about the image.

The multi-scale wavelet decomposition followed by PCA is a biologically plausible model of visual processing in the brain (cf., Palmeri & Gauthier, 2004). Moreover, several studies showed that the distances between representations in the similarity space that results from preprocessing the input with the multi-scale wavelet decomposition and PCA, agree well with dissimilarities as perceived by humans (see, e.g., Dailey, Cottrell, Padgett, & Adolphs, 2002; Lacroix et al., 2006).

Extending the NIM model

In our first experiment, we investigated whether it is possible to learn the spatial regularities of the visual features of one category of natural stimuli, i.e., face images. In particular, we assessed to what extent a mapping can be learned from similarity-space representation of a face image region to its spatial location, i.e., the similarity-to-spatial mapping. Such a mapping can incorporate the scene-schema knowledge of an extended NIM model on the basis of which relevant eye fixation locations can be selected. In our experiment, we used a set of 60 digitized gray-scale images of male human faces without glasses that were used in several other studies (e.g., Busey & Tunnicliff, 1999; Lacroix et al., 2004, 2006). All face images were downscaled to 156 × 198 pixels. Three examples of these faces are shown in Fig. 2. Below, we discuss the experimental set-up for learning the similarity-to-spatial mapping and present the results.

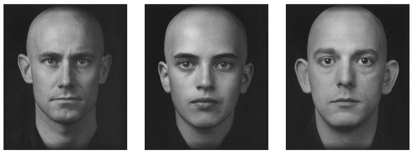


Figure 2: Three examples of faces from the set of faces used in the experiments.

Learning the similarity-to-spatial mapping

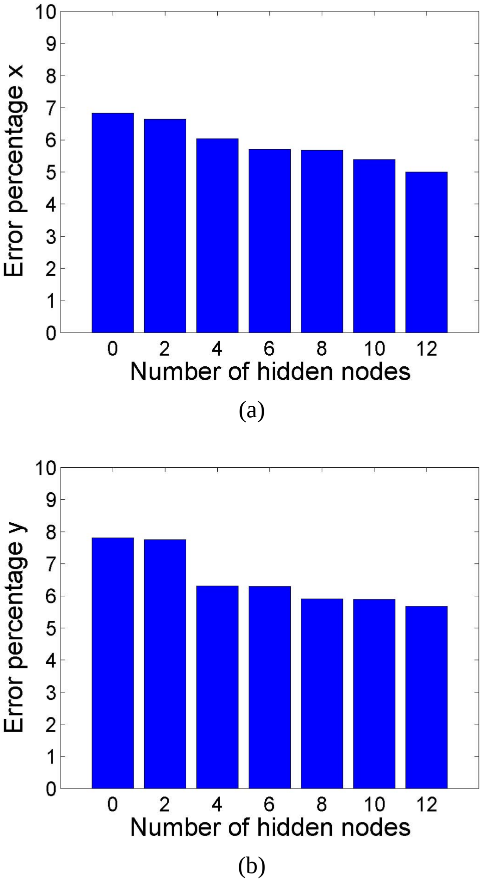
We trained different feedforward neural networks to map similarity-space representations (i.e., fixation vectors) onto spatial representations (i.e., image coordinates). This entails learning a mapping between the 50-dimensional feature vectors of an image area and their associated 2-dimensional spatial locations. Therefore, each network was dimensioned accordingly: an input layer of 50 nodes (encoding the feature vector), a hidden layer of h nodes, and 2 output nodes (encoding the x and y coordinates). Since the different faces covered about the same area within each image, the network was trained on absolute spatial coordinates x and y. Both the x and y coordinates where mapped onto the unit interval. Within each network, the transfer functions for the hidden and output nodes were defined as hyperbolic tangent sigmoid functions (range $[-1,+1]$) and standard sigmoid functions (range $[0,1]$), respectively.

The data set of face images was subdivided into two parts: a training set and a test set. From these sets, fixation vectors were extracted yielding training feature vectors and test feature vectors, respectively. Feature vectors were extracted randomly along the contours in the images. During training, the network was fed with 30,000 training feature vector coordinate pairs. After training the network was tested on 30,000 test feature vectors. Each test vector was submitted to the input layer of the trained network yielding an estimate of the spatial coordinates as output. We varied the number of nodes in the hidden layer, h, from 0 to 12 to assess how this affected the accuracy of the mapping.

Results

Figs. 3(a) and 3(b) show the mean error percentages for the x and y coordinates for networks with $h = 0$, 2, 4, 6, 8, 10, and 12 hidden nodes after 400 training epochs. The error percentages are the distances between the mapped spatial x and y coordinates and the actual spatial x and y coordinates of the fixation vector.

Overall, all networks were able to learn the similarity-to-spatial mapping quite accurately. Adding hidden nodes resulted in marginally better accuracy. The networks with $h = 0$ hidden nodes (i.e., single-layer perceptrons) showed an average spatial error of 10.65 pixels (6.8%) in the x direction and an average error of 15.44 (7.8%) pixels in the y direction. Adding hidden nodes slightly decreased the average spatial error to a minimum of 8.07 pixels (5.2%) in the x direction and 11.21 pixels (5.7%) in the y direction for the networks with $h = 12$ hidden nodes. The correlation between actual spatial locations and those predicted by the networks with 12 hidden units is $R = 0.96$ in both the x and y directions.



(a)

(b)

Figure 3: Mean error percentages for networks with $h = 0$, 2, 4, 6, 8, 10, and 12 hidden nodes after 400 training epochs, (a) for the x coordinates, and (b) for the y coordinates.

The results indicate that the spatial structure of face images can be inferred quite reliably from the NIM model's similarity-space representations. Knowledge about the spatial origin of visual input is implicitly present in the similarity-space representations. The explanation for this is twofold. First, the spatial regularity is considerable across the faces; similar spatial locations in different images, contain similar information. Second, the overlap structure in the fixation vectors ensures that the feedforward networks perform well on fixations that the model has not been trained on. Spatially adjacent fixations give rise to similar feature vectors. Hence, they provide a cue for the scene schema.

Using a simple feedforward neural network, it is possible to learn a mapping between the similarity-space representation of specific visual input (from objects within one category) and its spatial location. After many encounters with visual input from objects within a particular category, a network can predict the spatial location of specific visual input quite accurately. Therefore, it can be said that the network has acquired knowledge about the spatial arrangement of objects within the category, i.e., scene-schema knowledge (Henderson, 2003).

In the second experiment, we examine the NIM model extended with a gaze control mechanism based on the acquired scene-schema knowledge (i.e., the similarity-to-spatial mapping) to guide the selection of relevant eye-fixation locations.

Categorization with the extended NIM model

We extended the NIM model with the similarity-to-spatial mapping acquired in the first experiment. Using the mapping, the extended NIM model can select novel fixation locations by means of a control mechanism operating on the similarity space only. In the second experiment, the extended NIM model was tested on a face-categorization task. In the categorization task, the model categorized a face as one of the n faces that it had previously encountered. In order to assess to what extent categorization can profit from knowledge-driven gaze control, we compared the categorization performance of the extended NIM model with that of the NIM model without the gaze control extension.

The NIM model was originally developed to model behaviorally obtained recognition performance for individual items and general findings from recognition-memory studies (Lacroix et al., 2004, 2006). Therefore, in order to model categorization rather than recognition the model needs to be adapted slightly. Below, we discuss the categorization version of the (extended) NIM model and discuss the performance of the extended NIM model with the gaze control mechanism on the categorization task.

Adapting the NIM model for categorization

The NIM model encompasses two stages: the perceptual preprocessing stage and the memory stage. In addition, the extended NIM model employs a gaze control mechanism to direct gaze to relevant image regions. The preprocessing stage was discussed previously. Below, we discuss the two processes of the memory stage of the categorization version of the NIM model: the storage process and the categorization process. Subsequently, we discuss the gaze control mechanism.

The storage process In the NIM model, samples of natural images are retained (i.e., 'stored') in a similarity space. The storage process stores preprocessed natural images. As stated before, a preprocessed natural image is represented by a number of fixations, i.e., low-dimensional feature vectors in a similarity space, each corresponding to the pre-processed image contents at the fixation location. For each of the n faces that the model encountered in the categorization task, s feature vectors were selected randomly along the contours in the image and stored with the appropriate label (i.e., '1' for face 1, '2' for face 2, and so forth).

The categorization process The categorization process of the NIM model relies on the distance between newly encountered (unlabeled) and stored (labeled) feature vectors. Given an image to be categorized, the categorization process consists of the following four steps:

1. Select a central fixation location F.
2. Define the $W \times H$ pixels region centered at F as the 'target region'.
3. Determine the feature vectors for all fixation locations within the target region that lie on a contour. These feature vectors are called 'target feature vectors'.
4. For each category c, find the target feature vector with the smallest Euclidean distance d_c to a labeled feature vector of category c. The 'belief' in a particular category c is defined as $1/d_c$.

The four steps are repeated f times, where f represents the number of fixations. The category with the largest belief value after f fixations is defined as the category associated with the image to be categorized.

The gaze control mechanism The main goal of the gaze control mechanism is to direct gaze to spatial locations that are relevant to solve the categorization task. To direct gaze, the mechanism uses two types of knowledge: (1) scene-schema knowledge (i.e., the similarity-to-spatial mapping, and (2) episodic knowledge (i.e., previously stored representations). The scene-schema knowledge is used to infer the spatial locations of the fixations that were stored during previous encounters with the faces (i.e., the episodic knowledge). From the previously stored fixations, the gaze control mechanism selects the most heterogeneous (spatial) cluster of fixations (i.e., the cluster that contains the largest variety of labels). For each subsequent location, the gaze control mechanism selects the most heterogeneous cluster of stored fixations that were not in a previously selected cluster, and so forth.

The categorization experiment

In the categorization experiment, the NIM model and the extended NIM model were presented with $n = 10$ faces. In both models, for each face $s = 10$ feature vectors were extracted along the contours in the image and stored. Then, during categorization, both models produced a categorization response to a test face on the basis of f fixations. In step 1 of the categorization process (described above), the fixation location is selected. While the original NIM model selects fixation locations randomly along the contours in the image, the extended NIM model selects fixation locations with the gaze control mechanism. For each experimental run, the n faces were randomly selected from the 30 faces in the dataset that were not used to learn the similarity-to-spatial mapping. The W and H values of the target region were set to 21 and 27 pixels, respectively. Experiments employed $f = 2$ and $f = 3$ fixations.

Results

Fig. 4 shows the mean performances of the original and the extended NIM models on the categorization task. Error bars indicate the standard error of the mean. Categorization performance is expressed as the percentage of correctly categorized faces. The extended NIM model performed significantly better than the original one on both the experiments employing $f = 2$ and $f = 3$ selected fixation locations. Based on the scene-schema knowledge incorporated in the similarity-

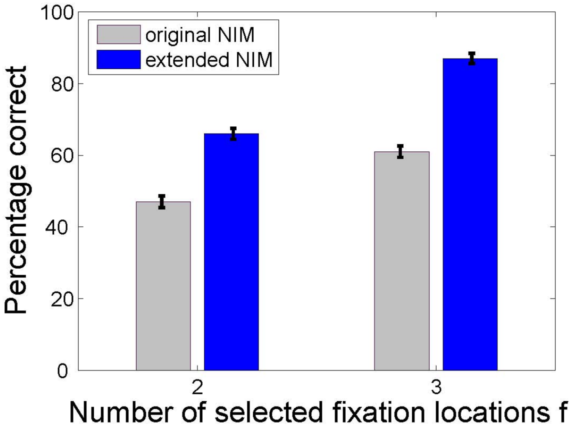


Figure 4: Categorization performance as a function of the number of selected fixation locations of the original and extended NIM models. Error bars indicate the standard error of the mean.

to-spatial mapping, the NIM model is able to infer the spatial origin of the episodic knowledge that it has stored. Using the spatial information present in the similarity-space representations, it can select fixation locations in regions that are most likely to contain the visual information needed to solve the categorization task. Our results show, that extending the NIM model with a gaze control mechanism that uses scene-schema knowledge in combination with episodic knowledge enhances performance on a straightforward categorization task.

Discussion

The results from our first experiment show that the spatial structure of face images (i.e., scene-schema knowledge) can be inferred from the NIM model's similarity-space representations. The results of our second experiment illustrate how the gaze control mechanism can enhance performance on a categorization task. As mentioned previously, two types of gaze control models can be distinguished: (1) image-driven models, and (2) knowledge-driven models. Below, we first discuss image-driven gaze control models. Then, we compare knowledge-driven gaze control in the NIM model with other gaze control models that use a knowledge-driven approach.

Image-driven gaze control

Until now, the image-driven approach has been the dominant approach to model gaze control. Image-driven gaze-control models generally assume that fixation locations are selected in a bottom-up manner based on the image properties (e.g., Itti & Koch, 2000; Rao et al., 2002). These models create a saliency map that marks the saliency of each image location. Saliency is defined by the distinctiveness of a region from its surround on certain visual dimensions. Since locations with a high visual saliency are assumed to be informative, gaze is directed toward highly salient locations. Often, the visual dimensions that are used to generate a saliency map are similar to the visual dimensions that are known to be processed by the human visual system such as color, intensity, contrast, orientation, edge junctions, and motion (see, e.g., Koch & Ullman, 1985; Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002). Also, in order to discover certain important visual dimensions for generating a saliency map, a few studies analyzed which visual dimensions best distinguish fixated image regions from non-fixated regions (see, e.g., Mannan et al., 1996; Parkhurst & Niebur, 2003; Henderson, Brockmole, Castelhano, & Mack, to appear).

Several studies showed that, under some conditions, fixation patterns predicted by image-driven gaze-control models correlate well with those observed in human subjects (see, e.g., Parkhurst et al., 2002). In their study, Parkhurst et al. (2002) recorded human scan paths when viewing a series of complex natural and artificial scenes. They found that human scan paths could be predicted quite accurately by stimulus saliency which was based on color, intensity, and orientation. While the image-driven approach was successful in predicting human fixation patterns in some tasks, it is inaccurate predicting fixation patterns in an active task that uses meaningful stimuli (see, e.g., Oliva et al., 2003; Turano, Geruschat, & Baker, 2003; Henderson et al., to appear). For example, Turano et al. (2003) showed that a saliency model performed as accurate as a random model in predicting the scan paths of human subjects during a real-world activity. In contrast, they found that a model that used only knowledge-driven (i.e., top-down) gaze control outperformed the random model. Obviously, visual saliency alone cannot account for the human fixation patterns when performing certain tasks. Similar results were found by Henderson et al. (to appear) who analyzed eye movements of subjects that viewed images of real-world scenes during an active search task. They found that a visual saliency model did not predict fixation patterns any better than a random model did. They concluded that visual saliency does not account for eye movements during active search and that top-down (i.e., knowledge-driven) processes play the dominant role.

Knowledge-driven gaze control

Knowledge-driven gaze-control models rely on stored knowledge to select the most informative location to direct gaze (see, e.g., Henderson, 2003). The extended NIM model uses a knowledge-driven gaze-control mechanism that combines acquired scene-schema knowledge with episodic knowledge. A few other models have used a knowledge-driven approach (e.g., Rybak, Gusakova, Golovan, Podladchikova, & Shevtsova, 1998). The main difference between knowledge-driven gaze control in the extended NIM model and existing knowledge-driven gaze control models concerns the representation of spatial knowledge. Whereas existing models of knowledge-driven gaze control include separate 'what' (episodic memory) and 'where' (spatial and motor memory) representation spaces, the NIM model relies solely on the structure of the similarity space representations. Since the faces contain spatial regularities, the similarity-space representations contain information about both the face identity and the spatial origin. While there might be good neurobiological evidence to support separate what and where systems, the results presented in this chapter show that the NIM model can acquire knowledge about the spatial arrangement of objects within a particular category based on the structure of the similarity-space representations.

Conclusion

This paper investigated whether the NIM model could be extended with a knowledge-driven gaze control mechanism that selects relevant eye-fixation locations based on scene-schema knowledge and episodic knowledge. From our results we conclude that the spatial structure of natural images can be inferred from the NIM model's similarity-space representations, i.e., scene-schema knowledge can be acquired on the basis of the similarity-space representations. Moreover, we showed that extending the NIM model with a gaze control mechanism that combines the acquired scene-schema knowledge with stored episodic knowledge, enhances performance on a categorization task. In our future work, we will use the extended NIM model for modeling human scan paths in a variety of visual tasks.

Acknowledgments

The research project is supported in the framework of the NWO Cognition Program with financial aid from the Netherlands Organization for Scientific Research (NWO). It is part of the larger project: 'Events in memory and environment: modeling and experimentation in humans and robots' (project number: 051.02.2002).

References

Braun, J., Koch, C., Lee, D. K., & Itti, L. (2001). Perceptual consequences of multilevel selection. In J. Braun, C. Koch, & J. L. Davis (Eds.), *Visual attention and cortical circuits* (pp. 215-241). Cambridge, MA: MIT Press.

Busey, T. A., & Tunnicliff, J. (1999). Accounts of blending, typicality and distinctiveness in face recognition. *Journal of Experimental Psychology: Learning Memory and Cognition*, *25*, 1210–1235.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*, 1158–1173.

Edelman, S. (1995). Representation, similarity, and the chorus of prototypes. *Minds and Machines*, *5*, 45–68.

Edelman, S., & Duvdevani-Bar, S. (1997). Similarity, connectionism, and the problem of representation in vision. *Neural computation*, *9*, 701–720.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. IEEE *Trans. Pattern Analysis and Machine Intelligence*, *13*, 891–906.

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 1–29). Cambridge, MA: Cambridge University Press.

Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: human perception and principal components analysis. *Memory and Cognition*, *24*, 26–40.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Science*, *7*, 498–504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (to appear). Visual saliency does not account for eye movements during search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain.* Elsevier.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

Karn, K. S., & HayHoe, M. M. (2000). Memory representations guide targeting eye movements in a natural task. *Visual Cognition*, *7*, 673–703.

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

Lacroix, J. P. W., Murre, J. M. J., Postma, E. O., & Van den Herik, H. J. (2004). The natural input memory model. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the cognitive science society (*CogSci *2004)* (pp. 773–778). Mahwah, NJ: Lawrence Erlbaum Associates.

Lacroix, J. P. W., Murre, J. M. J., Postma, E. O., & Van den Herik, H. J. (2006). Modeling recognition memory using the similarity structure of natural input. *Cognitive Science*, *30*, 121-145.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*, 165–188.

Norman, J. F., Phillips, F., & Ross, H. E. (2001). Information concentration along the boundary contours of naturally shaped solid objects. *Perception*, *30*, 1285–1294.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In IEEE *proceedings of the international conference on image processing* (Vol. 1, pp. 253–256).

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*, 291–303.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 107–123.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*, 125–154.

Rajashekar, U., Cormack, L. K., & Bovik, A. C. (2002). Visual search: Structure from noise. In *Proceedings of the eye tracking research & applications symposium 2002* (pp. 119–123). New Orleans, LA.

Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, *42*, 1447-1463.

Rao, R. P. N., & Ballard, D. H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, *78*, 461–505.

Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., & Shevtsova, N. A. (1998). A model of attention-guided visual perception and recognition. *Vision Research*, *38*, 2387–2400.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, *43*, 333–346.