Modeling Visual Classification using Bottom-up and Top-down Fixation Selection

Joyca P. W. Lacroix (jlacroix@fsw.leidenuniv.nl)
Department of Cognitive Psychology, Leiden University,
Wassenaarseweg 52, 2300 RB, Leiden, The Netherlands

Eric O. Postma (postma@micc.unimaas.nl) and **H. Jaap van den Herik (herik@micc.unimaas.nl)**
Department of Computer Science, MICC-IKAT, Maastricht University,
Tongersestraat 6, 6211 LN, Maastricht, The Netherlands

Abstract

This paper describes two initial steps towards the realization of a plausible model of natural visual classification. As a first step, we extend the recently developed Natural Input Memory (NIM) model (Lacroix, Murre, Postma, & van den Herik, 2006) to a classification model of natural visual input called NIM-CLASS and evaluate the model in a face-classification experiment. Our experimental results show that NIM-CLASS is able to recognize and classify faces after a single encounter. In addition, NIM-CLASS is insensitive to variations in facial expressions, illumination conditions, and occlusions. As a second step, we extend NIM-CLASS to NIM-CLASS A by adding an active top-down fixation-selection mechanism. We then assess to what extent NIM-CLASS A improves the performance on the face-classification task. The results show that the incorporation of a selection mechanism improves classification performance, particularly when a limited number of fixations are taken during the classification process. Our results lead us to the conclusion that NIM-CLASS A may provide a suitable basis for a model of natural visual classification.

Keywords: Perception, memory, classification, gaze control.

Introduction

Traditional computational models of cognition (e.g., Shiffrin & Steyvers, 1997) generally operate on an abstract representation space, because they lack a mechanism to derive representations from the physical features of stimuli, i.e., they are not grounded in the real world. In sharp contrast, natural systems ground representations in physical interaction with the world. Acknowledging the importance of the environment for natural cognition, a recent trend in psychologically motivated cognitive models is to focus on grounding representations in terms of their real-world referents (e.g., Pecher & Zwaan, 2005). Following this trend, the recently proposed Natural Input Memory model (NIM; Lacroix et al., 2006) realizes a memory model that operates directly on real-world visual input (i.e., natural digitized images). It builds feature-vector representations on the basis of local samples (i.e., eye fixations) from natural images and uses these to make recognition-memory decisions (e.g., Lacroix et al., 2006).

We aim at extending NIM (Lacroix et al., 2006) to a model of natural visual classification. This paper provides two initial steps towards achieving this objective. The outline of the remainder of this paper is as follows. As a first step, we extend NIM into a classifier of natural images called NIM-CLASS and assess NIM-CLASS's performance in a face-classification experiment. As a second step, we aim to approach the interactive nature of natural vision by extending NIM-CLASS to NIM-CLASS A by adding an active top-down mechanism to select fixation locations. We then evaluate to what extent the use of the top-down mechanism improves classification by testing NIM-CLASS A on the classification task. Subsequently, we discuss top-down gaze control models, and the scalability and sensitivity to changes in viewpoint of the NIM-CLASS models. Finally, we present our conclusion.

Extending NIM to NIM-CLASS

NIM is a model for recognition of natural images (Lacroix et al., 2006). NIM encompasses the following two stages.

1. A perceptual preprocessing stage during which a natural image is translated into feature vectors.
2. A memory stage comprising two processes:
   (a) a storage process that stores feature vectors in a straightforward manner;
   (b) a recognition process that compares feature vectors of a newly presented image with previously stored feature vectors.

Inspired by eye fixations in human vision, the perceptual preprocessing stage selects image samples (i.e., fixations) randomly along the contours in the image. For each fixation, visual input is translated into a feature vector that resides in a similarity space. The translation is realized using a biologically informed method that involves a multi-scale wavelet decomposition (see, e.g., Rao & Ballard, 1995) followed by a principal component analysis. This method from the domain of visual object recognition models the first stages of processing of information in the human visual system (i.e., retina, LGN, V1/V2, V4/LOC; (Palmeri & Gauthier, 2004)). NIM applies the method in a saccadic based manner to build representations of fixated image parts that together constitute the feature-vector representation of an image. The memory stage stores the feature-vector representation (the storage process) and makes recognition decisions by matching an incoming feature-vector representation with previously stored representations. For a more detailed description and a schematic overview of NIM we refer to Lacroix et al., 2006. While NIM is a model for recognition of natural images, here we show that it can readily be adapted into a model for classification of natural images which we call NIM-CLASS. The feasibility

of adapting NIM for classification has been shown recently by Barrington, Marks, and Cottrell (2007) who presented a Bayesian version of NIM called NIMBLE and successfully applied it to face classification. NIM-CLASS uses a slightly different approach that adopts NIM's perceptual preprocessing stage (i.e., the perceptual front-end), but introduces a new memory stage that is expected to be suitable for classification. Below, we discuss the two processes of the NIM-CLASS memory stage: the storage process and the classification process.

The Storage Process

The NIM-CLASS storage process retains (i.e., stores) preprocessed samples of natural images (i.e., fixations) that belong to a certain class. Each natural image is represented by a number of low-dimensional feature vectors (one for each fixation) in a similarity space. In contrast to the original NIM that stores unlabeled feature vectors, NIM-CLASS stores class labels with each feature vector corresponding to the class associated with the image (i.e., '1' for class 1, '2' for for class 2, and so forth).

The Classification Process

The NIM-CLASS classification process employs a naive Bayesian method that is based on an incremental estimate of the class-dependent probabilities (Duda, Hart, & Stork, 2001). During the classification process, each fixation of the test image (i.e., each test feature vector) contributes to an n-bin histogram, the bins of which represent the 'beliefs' in the n different classes. For each test feature vector, the bin that corresponds to the label of its nearest neighbouring stored labeled feature vector (acquired in the storage process) is incremented (e.g., if the stored labeled feature vector that is closest to the test feature vector has label '1', bin 1 is incremented). Finally, upon the last fixation, the class with the largest bin (i.e., belief) determines the classification decision. This heuristic classification process could readily be extended into a Bayesian approach in which each fixation updates class-conditional probabilities according to the Bayes update rule.

The Classification Experiment

The experiment evaluates NIM-CLASS's ability to classify natural images of faces. Below, we discuss the classification task, the data set, and the experimental procedure.

The Classification Task

The classification task entails the identification of a natural image of a frontal face with variations in facial expressions, illumination conditions (location of the light source), and occlusions (sun glasses and scarf). Humans are generally able to identify a face after a single encounter only, despite variations in appearance (e.g., Burton, Jenkins, Hancock, & White, 2005). Inspired by this fact, NIM-CLASS is evaluated on a task in which the training set (i.e., the study list) consists of a single image for each class and the test set (i.e., the test list) of the twelve remaining images. In this respect, our evaluation differs from most evaluations in machine learning, where the training set consists of a much larger fraction of the data set.

Figure 1: Example of the 13 views of one individual from the AR data set.

The Data Set

For the face-classification task, a data set with different images of the same individual was needed. We chose to use the AR data set created by Martinez and Benavente (1998) that contains over 4,000 images corresponding to the faces of 126 individuals. For each individual, the AR data set includes a sequence of 13 images featuring frontal view faces with different facial expressions, illumination conditions, and occlusions. For the experiment, we selected the sequence of 13 images (i.e., views) of the first 10 male individuals of the AR data set as our data set. Fig. 1 shows an example of the sequence of 13 views of one individual. The first (standard) view of each individual was selected for the study list, the remaining 12 views were assigned to the test list.

The Experimental Procedure

The face-classification experiment entailed a study and a test phase. During the study phase, NIM-CLASS was presented with the images from the study list containing the first view of each of the $n = 10$ individuals (i.e., the study faces). For each study face, NIM-CLASS extracted and stored s labeled feature vectors. Then during the test phase, the model was presented with the 120 images from the test list (i.e., the 12 test faces of each of the $n = 10$ individuals). For each of the test faces, the model extracted t test feature vectors to classify the face as one of the $n = 10$ individuals that it had previously encountered. To assess how the NIM-CLASS classification performance varied as a function of the number of storage fixations s and the number of test fixations t, the experiment was repeated for values of s and t in the range 10 to 100, i.e., $s,t \in \{10,20,...100\}$.

Classification with NIM-CLASS

Below, we present the NIM-CLASS results for the face-classification task[1] and compare these with human face-identification results.

Classification Results

Table 1 presents the percentages of test faces classified correctly by NIM-CLASS for a range of values of the number

[1]These results were partly presented at the workshop *Towards Cognitive Humanoid Robots* of the IEEE-RAS International Conference on Humanoid Robots 2006

of storage fixations s and the number of test fixations t. Fig. 2(a) presents the same results as a surface plot. The NIM-CLASS classification performances range from just above chance level (16%) for $s = t = 10$ and reach a good performance of 89.0% for $s = t = 100$. Evidently, NIM-CLASS is capable of exhibiting a good performance provided that a sufficient number of fixations is made.

The results show, not surprisingly, that the performance increases both with the number of storage fixations and the number of test fixations. Increasing the number of storage fixations s, improves the performances more than increasing the number of test fixations t. For small s values, performance hardly increases with t. Evidently, taking more test fixations is only useful when a sufficient number of feature vectors were stored previously. From a statistical perspective this makes sense. A proper approximation of the true distribution of feature vectors in a similarity space associated with a single face requires a sufficient number of samples (fixations) of that face.

Table 1: Percentages of faces classified correctly by NIM-CLASS for a range of values for the number of storage fixations s and the number of test fixations t.

	t									
	10	20	30	40	50	60	70	80	90	100
s										
10	16.0	18.2	20.6	22.1	23.6	23.7	24.4	25.3	25.5	26.2
20	21.3	26.3	29.5	32.1	35.5	38.3	39.3	41.1	42.7	43.5
30	26.5	32.8	38.1	42.5	46.3	49.0	52.0	53.3	55.5	57.3
40	30.0	39.5	45.7	51.1	55.1	58.6	60.8	63.1	64.5	66.8
50	34.0	45.2	51.7	57.0	61.8	64.9	68.0	70.0	71.5	73.7
60	36.7	49.2	57.0	62.7	66.9	70.7	73.7	75.3	77.3	78.5
70	39.8	52.9	61.8	67.7	71.2	75.3	77.8	79.6	80.9	82.5
80	42.7	57.0	65.9	70.9	75.4	77.9	80.7	82.9	84.3	85.4
90	45.7	60.1	68.3	73.8	78.3	81.1	83.3	84.8	85.9	87.4
100	47.6	63.1	71.3	77.0	80.6	83.2	84.7	87.1	87.8	89.0

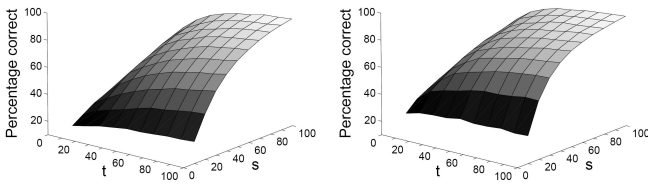

Figure 2: The classification performance as a function of the number of storage fixations s and the number of test fixations t of NIM-CLASS (left), and NIM-CLASS A (right).

To provide some insight into the distribution of beliefs in the different classes for each of the 120 test faces (i.e., 12 test views for each of the 10 individuals in the data set), Fig. 3 presents an overview of the histograms for each of the 120 test faces for $s = t = 100$. Each histogram represents the belief in class 1 (leftmost bin in each histogram) to 10 (rightmost bin in each histogram). In other words, the histograms represent the frequency counts of the labels of the nearest neighbours of the test feature vectors. Each row of histograms corresponds to the view depicted to the left of that row and each column of histograms corresponds to the individual depicted at the top of that column. A face is classified correctly when the index of the largest bin corresponds to the class of the particular face. From Fig. 3 it can be seen that, in most cases, the largest bin corresponds to the class of the test face. Where this is not the case, the largest bin is not considerably larger than the other bins. Therefore, the faces classified falsely can be said to be classified with less certainty than the faces classified correctly.

Comparison with Human Face Identification

Since this paper addresses the suitability of NIM-CLASS as a cognitive controller of a humanoid robot, we compare the NIM-CLASS performance with that of human face identification in a natural setting.

The number of storage and test fixations extracted by NIM-CLASS can be interpreted as the amount of viewing time of the image during study and test, respectively. Dividing the number of fixations by five provides a rough estimate of the number of seconds the image is inspected, since humans make about five fixations per second (see, e.g., Henderson, 2003). As the results show, the NIM-CLASS performance relies heavily on the amount of viewing time during the study phase. This accords with results from several psychological studies indicating that memory for visual information increases with viewing time during study (e.g., Mäntylä & Holm, 2006; Melcher, 2006). Moreover, it is interesting that a considerable percentage of faces (say $\geq 75\%$) is classified correctly after a short viewing time of about 8 seconds (40 fixations) during the test phase, provided that there was a sufficiently long viewing time of about 20 seconds (100 fixations) during the study phase. In additional simulations, we assessed in more detail to what extent NIM-CLASS is able to

Figure 3: Overview of the histograms across the 120 test faces (i.e., 12 views of each of the 10 individuals) for $s = t = 100$.

classify the test faces correctly on the basis of a brief viewing time during the test phase of only 1 second ($t = 5$). The simulation results show that NIM-CLASS is able to reach a considerable classification performance on the basis of a viewing time of 1 second during the test phase, provided that the viewing time during the study phase, s, is sufficiently long (mean percentages of faces classified correctly across all the views ranged from 36.9% to 74.0% when the number of storage fixations s were varied in the range 100 to 1000 , i.e., $s \in \{100, 110, ...1000\}$, corresponding to about 20 to 200 seconds of viewing time). The same holds for human vision, for which it is known that a brief viewing time will allow for correct identification, provided the face is sufficiently familiar to the observer (e.g., Burton et al., 2005).

Overall, the NIM-CLASS classification results demonstrate that natural images of frontal faces under a variety of potentially disturbing conditions can be classified correctly using a classification process that compares (a sufficient number of) stored local image samples (i.e., fixations) acquired during one encounter (i.e., one stored view) to incoming local samples. NIM-CLASS uses a bottom-up fixation-selection mechanism that selects fixations on the basis of their visual saliency (contours). While bottom-up processes are important in human vision too, they are integrated with top-down processes that direct the gaze to relevant locations on the basis of cognitive systems (see, e.g., Henderson, 2003). Below, we explore the use of top-down fixation selection and investigate to what extent top-down fixation selection aids performance on the classification task.

Top-down Fixation Selection: Extending NIM-CLASS to NIM-CLASS A

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., Henderson, Brockmole, Castelhano, & Mack, to appear). The top-down processes are driven by several cognitive systems, including: (1) short-term episodic memory for previously attended visual input, (2) stored long-term knowledge about visual, spatial, and semantic characteristics of classes of items or scenes acquired through experience, and (3) the goals and plans of the viewer (e.g., Henderson, 2003; Mäntylä & Holm, 2006). Inspired by fixation selection in human vision, this section extends NIM-CLASS to NIM-CLASS A by introducing a top-down fixation-selection during the classification process. In order to do so, we rely on the short-term episodic knowledge about previously attended visual input (see, e.g., Henderson, 2003; Mäntylä & Holm, 2006) that is known to operate in human gaze control. Below, we discuss the two processes of the memory stage of NIM-CLASS A: (1) the storage process and (2) the classification process.

The Storage Process

NIM-CLASS A extends NIM-CLASS with top-down fixation selection during the classification process, while featuring the bottom-up (i.e., contour-based) fixation selection of NIM-CLASS during the storage process. Therefore, the storage process is similar to that in NIM-CLASS, except that NIM-CLASS A stores the coordinates of each fixation along with the class label. The coordinate labels are used for the top-down fixation selection during the classification process.

The Classification Process

For the implementation of the top-down fixation selection in NIM-CLASS A, we rely on the notion of Shannon's (1948) entropy. Shannon introduced entropy as a measure of uncertainty. In order to decide in the most efficient way to which class a new item belongs, a system should select new input that minimizes the entropy, i.e., the uncertainty about the class membership. In NIM-CLASS, uncertainty is represented by the histogram in which the heights of the bins represent the beliefs in the different classes. Considering the uncertainty, the top-down fixation-selection mechanism selects those locations that contain the most relevant information to decide upon the class of the face under consideration (i.e., that minimize the entropy or uncertainty about the class). In order to do so, the mechanism relies on short-term episodic knowledge about attended parts of recently encountered faces which is represented by the labeled feature vectors that were acquired during the storage process directly preceding the current classification process.

For each fixation, the top-down fixation selection mechanism first chooses the two most likely classes, A and B, by selecting the two highest bins in the histogram. Subsequently, it selects the fixation location that best discriminates between the two classes A and B (i.e., contains the most relevant visual input with respect to A and B). The idea behind the selection is that spatially adjacent fixations within one class give rise to similar feature vectors. Hence, the fixation mechanism searches for a pair of feature vectors a and b coming from classes A and B, respectively, that originate from relatively close spatial locations and at the same time are relatively distant from each other in the representation space. A detailed specification of the implementation of this idea can be found in Lacroix, Postma, Murre, and van den Herik (in preparation).

Classification with NIM-CLASS A

Below, we present the results for the face-classification task performed with NIM-CLASS A and compare the classification performances of NIM-CLASS and NIM-CLASS A.

Classification Results

Table 2 presents the percentages of test faces classified correctly by NIM-CLASS A for a range of values of the number of storage fixations s and the number of test fixations t. In addition, Fig. 2(b) displays the classification performances as a function of s and t for NIM-CLASS A. The NIM-CLASS A classification performance ranges from 25.2% for $s = t = 10$ to a performance of 91.0% for $s = t = 100$. Overall, the NIM-CLASS A performance is improved compared to the origi-

Table 2: Percentages of faces classified correctly by NIM-CLASS A for a range of values for the number of storage fixations s and the number of test fixations t.

	t									
	10	20	30	40	50	60	70	80	90	100
s										
10	25.2	27.8	28.8	29.1	30.9	29.9	30.1	31.1	30.0	30.6
20	32.5	40.6	43.8	45.9	48.7	50.6	50.7	51.8	53.6	53.4
30	36.9	46.3	53.3	56.6	59.3	62.0	64.0	65.4	66.7	67.1
40	41.0	51.5	58.3	63.9	66.5	68.7	71.2	72.6	73.8	75.4
50	44.1	56.2	63.7	67.8	71.3	73.5	75.7	78.1	79.3	80.9
60	46.7	59.6	66.5	72.3	75.2	77.6	79.9	81.2	82.6	84.2
70	49.1	62.7	69.8	75.0	78.5	80.8	82.2	84.2	85.8	86.7
80	51.3	64.8	72.1	77.4	80.0	83.3	85.2	85.8	86.9	88.2
90	53.4	67.0	75.3	79.3	82.7	84.7	86.6	87.6	88.7	89.8
100	54.9	69.5	77.1	81.5	84.4	85.8	87.8	88.9	90.0	91.0

nal NIM-CLASS performance. As for NIM-CLASS, the results of NIM-CLASS A show that performance increases with the number of storage fixations s and the number of test fixations t and the performance increases more with s than with t. As was demonstrated for NIM-CLASS, the results of NIM-CLASS A show that taking more test fixations t becomes useful when a sufficient amount of feature vectors were stored previously.

Comparison of Classification Results

The results show that extending NIM-CLASS with top-down fixation selection to direct the gaze towards relevant locations, improves the performance on the classification task. In NIM-CLASS A, the top-down fixation selection actively constructs a fixation sequence based on: (1) the task to be solved (i.e., classification), and (2) the stored episodic knowledge about previous encounters with particular faces (i.e., the stored labeled feature vectors). By combining top-down and bottom-up processes for the selection of fixations, NIM-CLASS A acknowledges the influence of the episodic short term knowledge and the goals (i.e., classification) that are known to play a role in human gaze control (see, e.g. Henderson, 2003). The active strategy employed by NIM-CLASS A ensures that the locations are fixated which are known to discriminate well among the two classes considered to be the most likely at that time by the model. Therefore, the model more often makes the correct classification decision. This is particularly so, when a limited number of fixations are taken during the classification process. With a large number of fixations during the classification process, a sufficient amount of relevant visual information is gathered for correct classification even when fixations are taken randomly along the contours. With fewer fixations during the classification process, the probability that a sufficient amount of relevant visual information is gathered for correct classification decreases. Therefore, performance differences between the original NIM-CLASS and the NIM-CLASS A models are most pronounced for small t values.

Discussion

Below, we compare top-down gaze control in NIM-CLASS A with other top-down gaze-control models, and discuss the scalability of the NIM-CLASS models and their sensitivity to changes in viewpoint.

Top-down Gaze-Control Models

NIM-CLASS A that employs a top-down fixation-selection mechanism based on episodic short-term knowledge about previously attended image parts, may be related to probabilistic active vision models for classification (for an overview see, de Croon, Sprinkhuizen-Kuyper, & Postma, 2006). Probabilistic active models either consider all possible fixation selections at each time step (e.g., Denzler & Brown, 2002), consider all possible fixation selections in advance (e.g., Arbel & Ferrie, 2006), or use a fixation selection policy that is acquired on the basis of an extensive training (e.g., reinforcement learning, see, Paletta, Prantl, & Pinz, 1998) or on the basis of an evolutionary algorithm (e.g., de Croon, Postma, & van den Herik, 2006). In contrast, top-down fixation selection in NIM-CLASS A relies solely on the feature vectors that were stored during one encounter with the class instance (during the storage process).

Scalability

In our classification task, NIM-CLASS and NIM-CLASS A deal with 130 objects (i.e., faces) coming from 10 different classes. Obviously, this limited number of objects can hardly be considered to be representative for the enormous number of objects that natural systems encounter in the real world. Ideally, a plausible classification or recognition model should be able to distinguish among large numbers of objects. However, since the different NIM-CLASS models store the complete encountered visual input, classification time grows linearly with the number of encountered objects (see also Bajramovic, Mattern, Butko, & Denzler, 2006). In order to deal with this problem, mechanisms can be incorporated that ensure an efficient use and maintenance of the representation space, e.g., neurally inspired representation techniques including self-organizing maps, radial basis function networks, and spiking neural networks.

Viewpoint Invariance

We have not tested the model's sensitivity to changes in viewpoint. For many object recognition techniques, changes in viewpoint cause major degradations in performance. It has been suggested that the brain brings about invariance across viewpoint through interpolation across the responses of a set of stored global shape templates corresponding to prototypical object views (e.g., Edelman & Duvdevani-Bar, 1997). In contrast, the classical recognition-by-components theory attempted to deal with invariance by representing objects in terms of their invariant parts (Biederman, 1987). However, the extraction of invariant parts from natural images has proved to be computationally challenging, if not infeasible. NIM-CLASS combines both approaches by extracting both local and more global shape information (Lacroix et al., 2006). Further studies should address to what extent this

combined approach copes with the weaknesses of the separate approaches in dealing with invariance. Also, we may consider extending NIM-CLASS with existing statistical techniques that operate on the representation space in order to enhance viewpoint invariance (e.g., Prince & Elder, 2006).

Conclusion

This paper presented two initial steps towards the realization of a plausible model of natural visual classification. As a first step, we extended the recently developed NIM model to a model for classification of natural images called NIM-CLASS. The results obtained by testing NIM-CLASS in a face-classification experiment, demonstrate that NIM-CLASS is able to recognize and classify faces after a single encounter despite variations in facial expressions, illumination conditions, and occlusions. As a second step, we extended NIM-CLASS to NIM-CLASS A by adding an active top-down fixation selection mechanism. The results obtained with NIM-CLASS A demonstrate that using a top-down fixation-selection mechanism can enhance performance on the face-classification task by selecting actively the most relevant fixations. From our results, we may conclude that NIM-CLASS A provides a suitable basis for a model of natural visual classification.

Acknowledgments

The work described in this paper was partially conducted within the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission and partially within the Cognition Program project Events in Memory and Environment (051.02.2002) funded by the Netherlands Organization for Scientific Research (NWO).

References

Arbel, T., & Ferrie, F. P. (2006). Entropy-based gaze planning. *Image and Vision Computing*, *19*, 779–786.

Bajramovic, F., Mattern, F., Butko, N., & Denzler, J. (2006). A comparison of nearest neighbor search algorithms for generic object recognition. In *Proceedings of the advanced concepts for intelligent vision systems (ACIVS 2006)* (pp. 1186–1197).

Barrington, L., Marks, T. K., & Cottrell, G. W. (2007). NIMBLE: A kernel density model of saccade-based visual memory. In *Proceedings of the 29th annual meeting of the cognitive science society (CogSci 2007).*

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*, 29–73.

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representation for face recognition: The power of averages. *Cognitive Psychology*, *51*, 256–284.

de Croon, G., Postma, E. O., & van den Herik, H. J. (2006). A situated model for sensory-motor coordination in gaze control. *Pattern Recognition Letters: Special Issue on Evolutionary Computer Vision and Image Understanding*, *27*, 287–314. (Guest Editor G. Olague)

de Croon, G., Sprinkhuizen-Kuyper, I. G., & Postma, E. O. (2006). *Comparing active vision models* (Tech. Rep. No. 06-02). MICC-IKAT, Universiteit Maastricht.

Denzler, J., & Brown, C. M. (2002). Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 145–157.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification.* New York, NY: Wiley & Sons Inc.

Edelman, S., & Duvdevani-Bar, S. (1997). Similarity-based viewspace interpolation and the categorization of 3d objects. In *Proceedings of the edinburgh workshop on similarity and categorization.*

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 *Eye movements: A window on mind and brain.* Oxford, UK: Elsevier.

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.

Lacroix, J. P. W., Postma, E. O., Murre, J. M. J., & van den Herik, H. J. (in preparation). Active classification with NIM-CLASS.

Mäntylä, T., & Holm, L. (2006). Gaze control and recollective experience in face recognition. *Visual Cognition*, *13*, 365–386.

Martinez, A., & Benavente, R. (1998). The ar face database. *CVC Technical Report #24*.

Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision*, *6*, 8–17.

Paletta, L., Prantl, M., & Pinz, A. (1998). Reinforcement learning for autonomous three-dimensional object recognition. In *Proceedings of the 6th symposium on intelligent robotics systems* (pp. 63–72). Edinburgh, UK.

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

Pecher, D., & Zwaan, R. A. (2005). *Grounding cognition.* Cambridge, UK: Cambridge University Press.

Prince, S. J. D., & Elder, J. H. (2006). Tied factor analysis for face recognition across large pose changes. In *Proceedings of the british machine vision conference* (Vol. 3, pp. 889-898).

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem: Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.