NIM as a Brain for a Humanoid Robot

Joyca Lacroix
Department of Cognitive Psychology
Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands
Email: jlacroix@fsw.leidenuniv.nl

Eric Postma
Department of Computer Science/MICC
Maastricht University
P.O. Box 616, 6200 MD Maastricht
The Netherlands
Email: postma@micc.unimaas.nl

Bernhard Hommel
and Pascal Haazebroek
Department of Cognitive Psychology
Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands
Email: hommel@fsw.leidenuniv.nl;
phaazebroek@fsw.leidenuniv.nl

***Abstract*— In the context of the PACO+ project (http://www.paco-plus.org/), we aim at extending the recently developed Natural Input Memory (NIM) model [11] to a controller for a humanoid robot that translates real-world visual input into actions. The NIM controller can be conceived of as the 'brain' of the robot. This paper describes the initial step towards realizing a controller by extending NIM to a classifier that learns to map visual instances onto classes. The extended model, called NIM-CLASS is evaluated in an experiment that involves the classification of face images. The results of the experiment 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. These results lead us to the conclusion that NIM-CLASS provides a suitable basis for controlling the actions of a humanoid robot. In future work, we will extend NIM-CLASS to a controller that maps the classified visual inputs to actions.**

I. INTRODUCTION

Traditional computational models of cognition (e.g., [31], [19], [23]) focus on the isolated processes underlying cognition without taking the environmental context into consideration. These models 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. Therefore, the traditional models fall short as models of cognitive natural systems, which are known to relie on the interaction with the environment for learning and survival (e.g., [4], [29]). 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., [28]). Following this trend, the recently proposed Natural Input Memory model (NIM; [11]) realizes a memory model that operates directly on real-world visual input (i.e., natural digitized images). NIM encompasses a perceptual front-end that takes local samples (i.e., eye fixations) from natural images and translates these into feature-vector representations. These representations are used to make memory-based decisions such as recognition decisions (e.g., [11]) and classification decisions (e.g., [12], [13]). Extending a memory model with a perceptual front-end that derives representations directly from natural input is an important step toward a new type of cognitive system that is able to operate directly on natural human environments. The PACO+ project (http://www.paco-plus.org/) is a currently ongoing project that focusses on constructing such a cognitive system within a humanoid robot. In the context of the PACO+ project, we aim to extend NIM [11] to the cognitive controller (i.e., the brain) of the humanoid robot. This paper describes the initial step towards realizing a controller by extending NIM to a classifier that learns to map visual instances onto classes. Classifying perceptual input into semantic, cognitive, and perceptual classes is central to cognition (see, e.g., [22]) and constitutes one of the key objectives within the PACO+ project. While NIM was originally introduced as a recognition-memory model, it can readily be extended to obtain a model of classification. This paper extends NIM into a classifier of natural images called NIM-CLASS and assesses NIM-CLASS's performance in a classification experiment that involves the classification of face images.

The outline of the remainder of this chapter is as follows. In section II, we extend NIM into a classification model of natural visual input, called NIM-CLASS. This is followed in section III by a description of the classification experiment that was used for our classification studies. Subsequently, in section IV, the NIM-CLASS classification performance is evaluated in the classification experiment. Then, section V discusses the selection of visual input in NIM-CLASS and the scalability of NIM-CLASS. Finally, section VI presents our conclusion.

II. EXTENDING NIM TO NIM-CLASS

NIM is a model for recognition of natural images [11]. NIM encompasses the following two stages.

1) A perceptual preprocessing 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.

Fig. 1 presents a schematic overview of NIM. The face image is an example of a natural image. The left and right

side of the figure correspond to the perceptual preprocessing stage (left) and the memory stage (right), respectively. During the perceptual preprocessing stage, eye fixation locations are selected randomly along the contours in the image. At each eye-fixation location, visual input is translated into a feature-vector representation that resides in a similarity space. The translation is realized using a biologically informed method that involves a multi-scale wavelet decomposition followed by a principal component analysis. This is an often applied method in the domain of visual object recognition to model the first stages of processing of information in the human visual system (i.e., retina, LGN, V1/V2, V4/LOC; [25]). The feature-vector representation forms the input for the memory stage. The memory stage comprises two processes: storage and recognition. During storage, the memory stage stores the feature-vector representation. During recognition, the memory stage compares the feature-vector representation with previously stored representations in order to make a recognition decision. For a more detailed description of NIM we refer to [11]. While NIM is a model for recognition of natural images, it can readily be adapted into a model for classification of natural images. Classifciation and recognition are closely related cognitive processes, because both processes operate by assessing the similarity between an item and previously encountered items. Here we extend NIM to a model for classification called NIM-CLASS. NIM-CLASS combines NIM's perceptual preprocessing stage (i.e., the perceptual front-end) with a new memory stage that is suitable for classification.

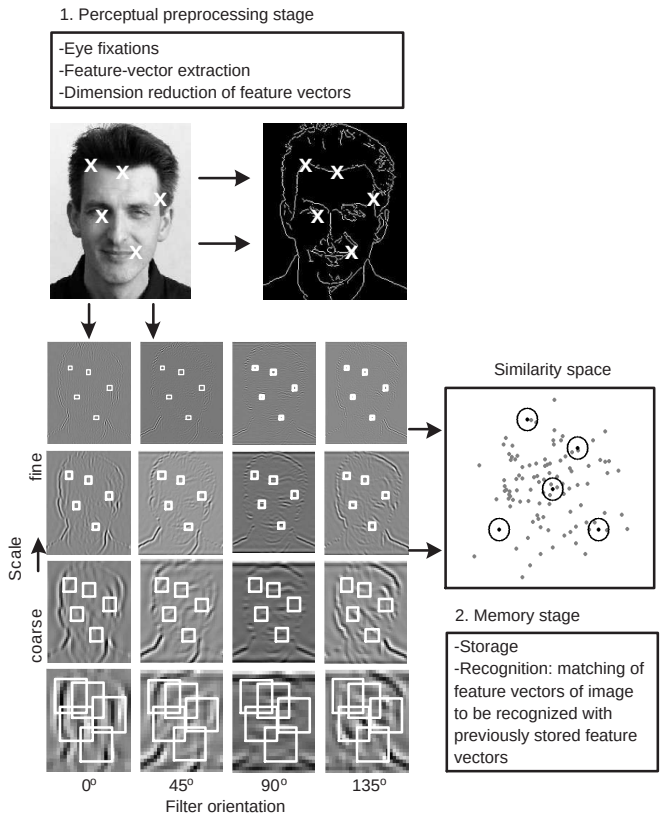

Fig. 1. The Natural Input Memory model (NIM). Reproduced from [11]

NIM-CLASS features NIM's perceptual preprocessing stage to transform fixated image parts into feature-vector representations. For NIM-CLASS, each image represents an instance of a class. Therefore, the images and the feature vectors obtained by fixating the images are labelled with a class label. NIM-CLASS differs from NIM in the design of the memory stage only. Below, we discuss the two processes of the NIM-CLASS memory stage: the storage process (II-A) and the classification process (II-B).

A. 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 unlabelled 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 storage and classification processes correspond to the training and the testing stage, respectively, which are commonly distinguished in supervised learning (see, e.g., [5]).

B. 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 [5]. 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 labelled 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.

III. THE CLASSIFICATION EXPERIMENT

In our experiments, the ability of NIM-CLASS to classify natural images of faces is evaluated. Below, we discuss the classification task (III-A), the data set (III-B) and the experimental procedure (III-C).

A. 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., [2]). 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.

B. 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 [18] 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. All face images were downscaled to 165×165 pixels. Fig. 2 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.

C. 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 labelled feature vectors. Then during the test phase, the model was presented with the images from the test list (i.e, the 12 test faces) of each of the $n = 10$ studied 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 asses 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., $s, t \in \{10, 20, ...100\}$.

Fig. 2. Example of the 13 views of one individual from the AR data set. We selected 10 individuals (with 13 views each) from the AR data set as the data set for the classification experiment.

IV. CLASSIFICATION WITH NIM-CLASS

Below, we present the NIM-CLASS results for the face-classification task (IV-A). Subsequently, we discuss how the number of fixations and the fixation selection in NIM-CLASS relate to that in human vision (IV-B).

A. Classification results

Table IV-A presents the percentages correctly classified test faces for a range of values of the number of storage fixations s and the number of test fixations t. Fig. 3 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 similarity space associated with a single face requires a sufficient number of samples (fixations) of that face.

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. 4 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 neigbours of the test feature vectors. Each row of histograms corresponds to the view depicted to the left of that row and each column of

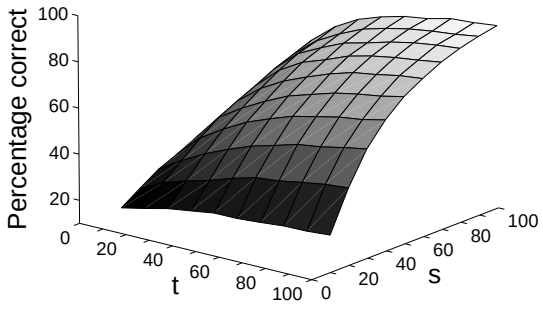

Fig. 3. Percentages correctly classified faces as a function of the number of storage fixations s and the number of test fixations t.

TABLE I

PERCENTAGES CORRECTLY CLASSIFIED FACES FOR A RANGE OF VALUES OF THE NUMBER OF STORAGE FIXATIONS s AND THE NUMBERS OF TEST FIXATIONS t.

s \ t	10	20	30	40	50	60	70	80	90	100
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

histograms corresponds to the individual depicted at the top of that column. A face is correctly classified when the index of the largest bin corresponds to the class of the particular face. From Fig. 4 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, it can be said that the falsely classified faces were classified with less certainty than the correctly classified faces.

B. Discussion and analysis of classification results

The results show that NIM-CLASS is able to classify faces quite accurately despite variations in facial expressions, illumination conditions, and occlusions. The model reaches a performance of 89% for $s = t = 100$ storage and test fixations. Since this paper addresses the suitability of NIM-CLASS as a brain of a cognitive humanoid robot, it is interesting how the NIM-CLASS performance compares with that of human face identification in a natural setting. Below, we briefly discuss the NIM-CLASS performance in relation to human face identification.

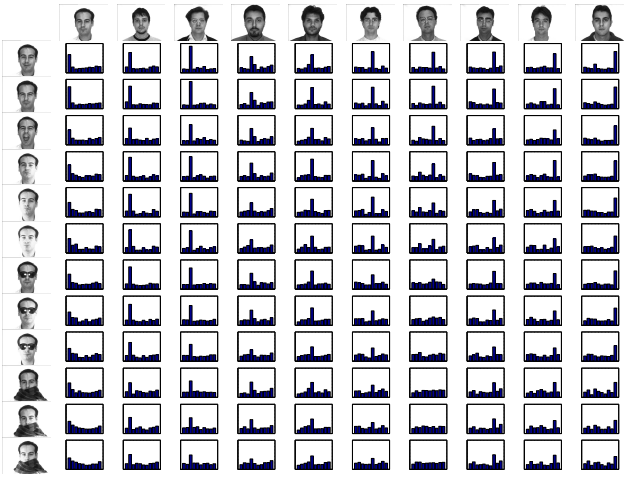

Fig. 4. Overview of the histograms obtained in the classification experiment across the 120 test faces (i.e., 12 views of each of the 10 individuals) for $s = t = 100$.

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., [7], [20]). As the results show, the NIM-CLASS performance relies heavily on the amount of viewing time during study. This accords with results from several psychological studies indicating that memory for visual information increases with viewing time during study (e.g., [15], [17], [21]). Moreover, it is interesting that a considerable percentage of faces (say $\leq 75\%$) is classified correctly after a short viewing time of about 8 seconds (40 fixations) during testing, provided that there was a sufficiently long viewing time of about 20 seconds (100 fixations) during study.

To assess in more detail to what extent NIM-CLASS is able to correctly classify the test faces on the basis of a brief viewing time during testing, we performed additional simulations In these simulations, the experiment was repeated for values of s in the range 10 to 1000, i.e., $s \in \{10, 20, ...1000\}$ which corresponds to about 2 to 200 seconds of viewing time, and the number of test fixations were set to $t = 5$, which corresponds to approximately one second of viewing time during testing. Fig. 5 presents the NIM-CLASS performance for a fixed number of test fixations $t = 5$ as a function of the number of storage fixations s. The results show that NIM-CLASS is able to reach a considerable classification performance on the basis of a brief viewing time during testing, provided NIM-CLASS has studied the face for a sufficiently long 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., [6], [2]).

Overall, the NIM-CLASS classification results show that NIM-CLASS is able to correctly classify faces under a variety of unfavorable conditions on the basis of one encounter (i.e., one stored view).

V. DISCUSSION

The NIM-CLASS classification results demonstrate that natural images of frontal faces under a variety of potentially

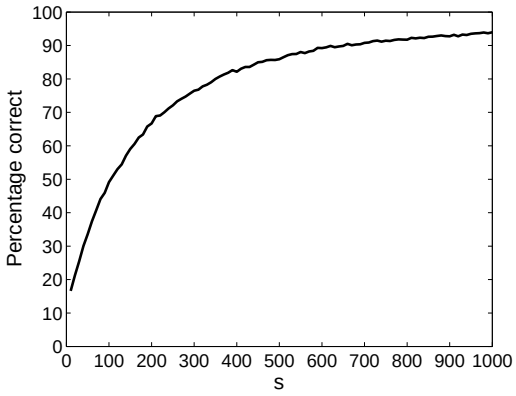

Fig. 5. Percentages correctly classified faces for a fixed number of test fixations $t = 5$ as a function of the number of storage fixations s.

disturbing conditions can be classified correctly using a classification process that compares (a sufficient number of) stored local image samples to incoming local image samples. NIM-CLASS employs a straightforward contour-based selection of image samples (eye fixations). Below we first discuss bottom-up and top-down fixation selection in humans and models of gaze control (V-A). Subsequently, we address the scalability of NIM-CLASS in terms of the number of classes (V-B).

A. Bottom-up and Top-down Fixation Selection

In NIM-CLASS, the samples (i.e., the eye fixations) are selected randomly along the contours in the image. The contour-based selection of fixations can be regarded as a realization of a bottom-up approach in which contours are the salient features. Until now, the saliency-based or bottom-up approach has been the dominant approach to model gaze control. Bottom-up gaze-control models generally assume that fixation locations are selected on the basis of particular image properties (e.g., [9], [30]). 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 in terms of certain visual dimensions. Since locations with a high visual saliency are assumed to be informative, gaze is directed towards 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 colour, intensity, contrast, orientation, edge junctions, and motion (see, e.g., [10], [9], [26]). Also, in order to discover certain important visual dimensions for generating a saliency map, a few studies analysed which visual dimensions best distinguish fixated image regions from non-fixated regions (see, e.g., [16], [27], [8]). Several studies showed that, under some conditions, fixation patterns predicted by bottom-up gaze-control models correlate well with those observed in human subjects (see, e.g., [26]). In their study, [26] 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 colour, intensity, and orientation. While the bottom-up 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., [24], [32], [8]). For example, [32] 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. Similar results were found by [8] who analysed 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.

Evidently, human fixation patterns do not rely solely on bottom-up processes when performing certain tasks. Rather, they are integrated with top-down processes that direct gaze to relevant locations (e.g., [7]). The top-down processes are driven by several cognitive systems, including: (1) short-term episodic memory for previously attended visual input (e.g., [3], [7]), (2) stored long-term knowledge about visual, spatial, and semantic characteristics of classes of items or scenes acquired through experience (e.g., [7]), and (3) the goals and plans of the viewer [33], [14], [7]). A psychologically plausible brain of a humanoid robot should incorporate a fixation selection mechanism that uses bottom-up as well as top-down processes to select informative visual input. In a recent study [13] extended NIM-CLASS with top-down fixation selection that relies on two types of knowledge known to operate in human gaze control: (I) the short-term episodic knowledge about previously attended visual input (e.g., [3], [7], [17]), and (II) the long-term knowledge about a class of items acquired through experience with instances from the class (e.g., [7]). Their results showed that extending NIM-CLASS with top-down fixation selection to direct gaze towards informative locations, improves performance on the face-classification task.

B. Scalability of NIM-CLASS

In our studies we have not examined how the NIM-CLASS performances scale up with the number of classes. Below we offer some perspective on the aspects that relate to the scalability of the model.

In our classification task, NIM-CLASS deals 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 humanoid robot brain should be able to distinguish among large numbers of objects. However, since NIM-CLASS stores the complete encountered visual input, classification time is linear in the amount of encountered objects (see also [1]). In order to deal with this problem, NIM-CLASS should be extended with mechanisms that use the representation space in an efficient way and that ensure the maintenance of an

efficient representation space. The recent NIM-CLASS extensions proposed by [13] involved a top-down fixation selection mechanism that operates on the representation space in an efficient way by actively searching for the most relevant information in the representation space. Moreover, they introduced a mechanism that maintains an efficient representation space by selecting and storing visual input on the basis of its relevance or informativeness. Using such mechanisms leads to more discriminable class representations. Therefore, we may assume that their incorporation makes the upscaling to a larger number of classes more feasible.

VI. CONCLUSION

In the context of the PACO+ project (http://www.paco-plus.org/), this paper presented an initial step toward the realization of a cognitive controller (i.e., a brain) for a humanoid robot that operates on real-world visual input. The controller extends the recently developed Natural Input Memory model (NIM) 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. On the basis of these results we conclude that NIM-CLASS provides a suitable basis for the cognitive controller of a humanoid robot. Future work will extend NIM-CLASS to a controller that maps the classified visual inputs to actions in order to approach the perception-action cycle characteristic of natural behaviour.

ACKNOWLEDGMENT

This work was partly funded by the European Union (PACO-PLUS integrated grant, IST-FP6-IP-027657; www.paco-plus.org).

REFERENCES

[1] F. Bajramovic, F Mattern, N Butko, and J Denzler. A comparison of nearest neighbor search algorithms for generic object recognition. ..., 2006.

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

[3] M. M. Chun. Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4:170–178, 2000.

[4] W. J. Clancey. *Situated cognition: On human knowledge and computer representations*. Cambridge University Press, New York, NY, 1997.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley and Sons Inc, Huntington, NY, 2001.

[6] J. M. Findlay and I. D. Gilchrist. *Active vision: The psychology of looking and seeing*. Oxford University Press, New Work, NY, 2003.

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

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

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

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

[11] J. P. W. Lacroix, J. M. J. Murre, E. O. Postma, and H. J. Van den Herik. Modeling recognition memory using the similarity structure of natural input. *Cognitive Science*, 30:121–145, 2006.

[12] J. P. W. Lacroix, E. O. Postma, and J. M. J. Murre. Knowledge-driven gaze control in the NIM model. Mahwah, NJ, 2006. Lawrence Erlbaum Associates.

[13] J. P. W. Lacroix, E. O. Postma, J. M. J. Murre, and H. J. Van den Herik. Active classification with NIM-CLASS. in preparation.

[14] M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everday activities? *Vision Research*, 41:3559–3565, 2001.

[15] G. R. Loftus. Eye fixations and recognition memory for pictures. *Cognitive Psychology*, 3, 1972.

[16] S. K. Mannan, K. H. Ruddock, and D. S. Wooding. 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, 1996.

[17] T. Mäntylä and L. Holm. Gaze control and recollective experience in face recognition. *Visual Cognition*, 13:365–386, 2006.

[18] A.M. Martinez and R. Benavente. The ar face database. *CVC Technical Report #24*, 1998.

[19] J. L. McClelland and M. Chappell. Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105:724–760, 1998.

[20] E. McSorley and J. M. Findlay. Saccade target selection in visual search: Accuracy improves when more distractors are present. *Journal of Vision*, 3:877–892, 2003.

[21] D. Melcher. Accumulation and persistence of memory for natural scenes. *Journal of Vision*, 6:8–17, 2006.

[22] J. M. J. Murre. *Learning and categorization in modular neural networks*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1992.

[23] R. M. Nosofsky and S. R. Zaki. A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29:1194–1209, 2003.

[24] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *IEEE Proceedings of the International Conference on Image Processing*, volume 1, pages 253–256, 2003.

[25] T. J. Palmeri and I. Gauthier. Visual object understanding. *Nature Reviews Neuroscience*, 5:291–303, 2004.

[26] D. J. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42:107–123, 2002.

[27] D. J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16:125–154, 2003.

[28] D. Pecher and R. A. Zwaan. *Grounding cognition*. Cambridge University Press, New York, NY, 2005.

[29] R. Pfeifer and C. Scheier. *Understanding intelligence*. The MIT Press, Cambridge, MA, 1999.

[30] R. P. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in iconic visual search. *Vision Research*, 42:1447–1463, 2002.

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

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

[33] A. L. Yarbus. *Eye movements and vision*. Plenum Press, New York, NY, 1967.