NIMBLE: A Kernel Density Model of Saccade-Based Visual Memory

Luke Barrington[⋆], Tim K. Marks[‡] and Garrison W. Cottrell[‡]
[⋆]Electrical and Computer Engineering Department, [‡]Computer Science and Engineering Department
University of California, San Diego
La Jolla, CA 92093
{lbarrington, tkmarks, gary}@ucsd.edu

Abstract

We present a Bayesian version of Lacroix et al.'s natural input memory (NIM) model of saccadic visual memory (Lacroix, Murre, & Postma, 2006). Our model uses a cognitively plausible image sampling technique that provides a foveated representation of image patches. We conceive of these memorized image fragments as samples from image class distributions and model the memory of these fragments using kernel density estimation. Using these models, we derive class-conditional probabilities of new image fragments and integrate individual fragment probabilities to classify images. Our Bayesian formulation of the model extends easily to handle multi-class problems. We validate our model by demonstrating human levels of performance on a face recognition memory task and high accuracy on multi-category face and object identification.

Introduction

Human visual perception begins with saccades, a small number of local, foveated patches sampled from a visual scene. In order to perceive all parts of a visual scene with great acuity as well as to maintain neural activation in the visual cortex, we repeatedly foveate different areas of the scene, concentrating fixations on the parts that are most salient or task-relevant (Yarbus, 1967). Not only is our sampling of visual objects fragmented in space and time, the sequence of saccades (scan path) we follow is unlikely to be repeated in future exposures to the same object or object class (Henderson, Williams, & Falk, 2005). Hence, simple exemplar matching of information from new saccades to stored memories cannot be relied upon to account for human capacities for object recognition.

Lacroix et al. (Lacroix, Murre, Postma, & Herik, 2004; Lacroix et al., 2006) have proposed the Natural Input Memory (NIM) model to account for humans' ability to recognize faces.[1] The model is in the mathematical psychology tradition, but is unusual for this sort of model in that (a) it uses actual facial images as input, and (b) it is based on the idea of storing saccade-based face fragments, rather than whole face exemplars. The model's memory is reminiscent of a kernel density estimator, but differs in important details in the way the estimates from individual fragments are combined. In this paper, we present a Bayesian version of the NIM model that uses naïve Bayes to combine the likelihood estimates from individual fragments. We further extend the model to multi-class visual memory tasks, and to use a variety of kernels for density estimation. Our model, which we call NIMBLE (for NIM with Bayesian Likelihood Estimation), achieves human levels of performance on a standard face recognition task and can also successfully perform multi-class face and object identification tasks. Bayesian combination of individual fragment likelihoods outperforms the combination method from the NIM model in most cases and the new kernels far outperform those used in NIM.

We begin by describing our biologically-motivated image sampling and transformation procedure. We then describe the NIM model. Next, we explain our Bayesian version of the model, NIMBLE, including a variety of extensions. We present human and model performance on visual memory tasks and conclude the paper with a discussion.

Visual Input Simulation

Fixation Point Selection

Given a current fixation point, the choice of where to saccade to next is driven by a number of external cues including motion, peripheral complexity and non-visual stimuli (e.g. sound) as well as top down task-dependent directives such as attention and expectation. Though many computer models (Wolfe, 1994; Mozer, Shettel, & Vecera, 2005; Zelinsky, Zhang, B. Yu, & Samaras, 2005) have been proposed for how to integrate top-down and bottom-up cues, in this work we concentrate only on bottom-up salience of static images. We model the fixation selection process using an interest operator for determining the scan paths (Yamada & Cottrell, 1995). This simplified model uses the rotational variance of eight low-resolution Gabor filter responses to construct a distribution of the contour complexity (salience) over all pixels in a given image:

$$\text{Salience}(i,j) = \frac{1}{8}\sum_{n=1}^{8}\left(\mathcal{G}(i,j,\theta) - \mu_{\mathcal{G}}(i,j)\right)^2$$

where $\mathcal{G}(i,j,\theta)$ is the response of a Gabor filter with orientation θ centered at pixel (i,j) and $\mu_{\mathcal{G}}(i,j)$ is the mean response across all orientations. A similar technique developed by (Renninger, Coughlan, Verghese, & Malik, 2004) defines salience as the entropy, rather than the variance, of local image contours.

We convert this salience map into a probability distribution using the softmax function (Bishop, 1995). A fixation point is then chosen randomly according to this distribution. Figure 1 shows a salience map generated in this manner as well as a sample distribution of fixation points. After each fixation point is chosen, we reduce the salience around the fixated

[1] *Face recognition* refers to the ability to discriminate previously seen faces from novel faces, based on a study list. In contrast, *face identification* or *person identification* refers to the ability to identify face images as particular individuals.

point by subtracting a univariate Gaussian, centered at the fixation point, from the salience distribution and renormalizing. This inhibits repeated fixations of the same location.

Despite the simplicity of this purely bottom-up model, the resulting scan paths for the face recognition task qualitatively approximate those observed in humans (Yamada & Cottrell, 1995). The model satisfies three of the five criteria identified by (Itti & Koch, 2001) for a computational model of visual attention: it derives perceptual saliency of a fixation point from the surrounding context, it creates a salience distribution over the visual scene, and it inhibits return to previously attended locations. In this paper, we ignore the remaining two criteria, which concern top-down influences on fixation point selection. In future work, we intend to incorporate top-down feedback to direct eye movements, by extending the results of (Nelson & Cottrell, 2007) to determine the fixations that would be most useful in enhancing performance on the current visual task.

We have tested NIMBLE using various alternative mechanisms for computing visual salience. The salience operator of (Itti & Koch, 2001) results in roughly the same face recognition performance as that of (Yamada & Cottrell, 1995), but the latter uses the same mechanism for computing salience (Gabor filters) as for processing images (see next section). Purely random selection of fixations reduces performance by 30%. Sampling fixations from the Canny edge map of the image reduces performance by 20%. We also tested NIMBLE using the actual locations of human fixations, which were recorded from the same face images using an eye-tracker, and found the resulting memory performance to be comparable to using the salience operator described.

Retinal / Cortical Image Transform

A fixated patch of an input image undergoes many stages of neural processing before being stored as a pattern of activation in high-level visual cortex. Our biologically-inspired model of the processing in primary visual cortex (V1) uses the magnitude responses of Gabor filters at 8 orientations and 4 frequencies (Jones & Palmer, 1987). We transform an image into the Gabor-filtered domain by calculating the response of each of these 32 filters at every image pixel. We use Gabor filter frequencies of $\frac{1}{16}, \frac{1}{12}, \frac{1}{8}$ and $\frac{1}{4}$ cycles/pixel (corresponding to 8, $10\frac{2}{3}$, 16 and 32 cycles/face).

Square patches extracted from these Gabor response images constitute our foveated representation of the fixated point. The highest-spatial-frequency filter responses correspond to the high-resolution foveated area centered at the fixation point. The responses of the low-frequency filters are each computed from an area centered at the fixated pixel that has spatial context greater than that of the foveated patch. Thus this patch representation includes extra-foveal information, corresponding to the low-resolution data from the retinal periphery.

The size of the extracted patch of filter responses and the number of patches that the model may examine for each image are experimental parameters that correspond, in human vision, to the distance of the eye from the image (and thus the size of the foveated area) and the time spent studying the image (determining the number of saccades made). For a patch size of 35x35 pixels (corresponding to a visual angle of 1.5° for a subject about 75cm from a 256μm-pixel computer monitor, an approximation of the human studies discussed below), the input feature vector to our model has 35 x 35 pixels x 8 orientations x 4 frequencies = 39200 dimensions. For efficiency and good generalization, we use principal components analysis (PCA) to reduce the size of this vector to 80 components, which retain about 90% of the variance depending on the dataset (see NIMBLE Results below). This feature extraction procedure of wavelet-based image decomposition followed by PCA is a standard approximation for biologically motivated vision models (Dailey, Cottrell, Padgett, & Adolphs, 2002; Palmeri & Gauthier, 2004; Lacroix et al., 2006).

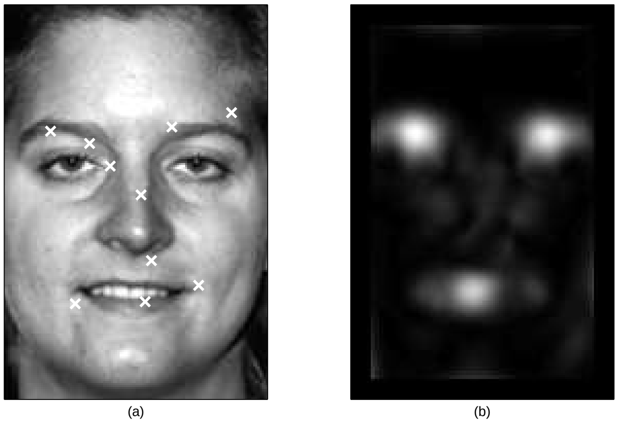

Figure 1: (a) An image from the FERET database with 10 sample fixation points. (b) The corresponding salience map generated using the technique of (Yamada & Cottrell, 1995). Fixations tend to cluster around highly salient areas but relaxation of sampled points enforces an even distribution across the image.

Natural Input Memory (NIM)

The inspiration for our model of saccade-based vision comes from the work of (Lacroix et al., 2004, 2006). Their Natural Input Memory (NIM) model is so-called since it takes saccade-like samples from a studied image as input. Their sampling method differs slightly from ours in that they sample from the contours of an image, determined by Canny edge-detection, and then process the sampled patches with the steerable pyramid transform, a multi-scale wavelet-based transform that is similar to Gabor filtering. They apply PCA to these features before storage in the memory.

Following the lead of many cognitive memory models (Hintzman, 1984; Nosofsky & Palmeri, 1997; Dailey, Cottrell, & Busey, 1998), the NIM model's memory process stores the feature-transformed representation of fixated image fragments as vectors in a high-dimensional memory space. Memories are compared to each other as well as to new image fragments by comparing the Euclidean distance between their vector representations in the memory space. The NIM model computes the familiarity of a new fragment by calculating the proportion of previously stored memories that lie within a radius r (a model parameter) of the new fragment in the memory space. Averaging these familiarities over all samples from a new image produces an estimate of the probability that the image is from the class known to the memory. The memory space introduced by the NIM model has been shown to achieve the best known correlation with human judgements of perceptual similarity (Lacroix et al., 2006), and the retrieval methods exhibit human performance effects (such as list length and list strength) on face recognition memory tasks (Lacroix et al., 2004).

The NIM memory retrieval method (Lacroix et al., 2006) determines the familiarity of a newly examined fragment by counting how many of the stored memories, $\{m_1, \ldots, m_M\}$, lie within a radius r of the new image fragment. Thus the familiarity of the new fragment, f, is defined by:

$$fam(f) = \sum_{j=1}^{M} I_r(||m_j - f||_2), \tag{1}$$

where

$$I_r(x) = \begin{cases} 1, & x \leq r \\ 0, & otherwise. \end{cases}$$

NIM Combination of Fragment Familiarities

An image is represented as a set of fragments $\mathcal{F} = \{f_1, \ldots, f_N\}$. In the NIM model, (Lacroix et al., 2006) define the familiarity of an image as the mean of the familiarities of all N patches taken from that image:

$$fam(\mathcal{F}) = \frac{1}{N} \sum_{i=1}^{N} fam(f_i). \tag{2}$$

They use a logistic function to transform this mean familiarity value into a probability:

$$P(\text{familiar image}) = \frac{1}{1 + \beta e^{-\theta fam(\mathcal{F})}},$$

where β and θ are parameters of the model used to fit the performance to human data.

The NIM model formulation (Lacroix et al., 2006) only attempts to make judgements about the *familiarity* of a studied image by comparing a set of fragments extracted from it to all previously stored memories. Since these memories are stored without labels, the resulting familiarity value must be compared to a threshold to decide whether the image is familiar or unfamiliar. Our extension of NIM, described in the next section, stores class labels with each exemplar, and can return explicit posterior probabilities for each class given the image fragments, permitting multi-class and hierarchical memory tasks in addition to the familiar / unfamiliar recognition memory task in (Lacroix et al., 2006).

A More NIMBLE Approach

Having sampled and processed a new image as described above, we want to evaluate the probability of the resulting set of N fragments, $\mathcal{F} = \{f_1, \ldots, f_N\}$, under the models for each of a number of image classes (e.g., familiar/unfamiliar faces, Alice/Bob/Carol/Dan/unknown, dogs/not dogs). For a class, c, we use Bayes rule to compute the posterior distribution:

$$P(c|\mathcal{F}) = \frac{P(\mathcal{F}|c)P(c)}{P(\mathcal{F})}.$$

In this case, $P(\mathcal{F}|c)$ is the likelihood of the set of image fragments under the density model for class c, and $P(c)$ is the class prior which may be learned from experience with training data.

We compute the likelihood of the set of image fragments, $P(\mathcal{F}|c)$, by combining the likelihoods of each individual fragment, $P(f_i|c)$. We compute these class-conditional fragment likelihoods using kernel density estimation (see below).

Naïve Bayes Fragment Combination

In the NIMBLE model, we can consider the naïve Bayes assumption of conditional independence between each fragment $f_i \in \mathcal{F}$, given the class, and take the product of the individual fragment likelihoods to get an estimate of the overall likelihood function:

$$P(\mathcal{F}|c) = \prod_{i=1}^{N} P(f_i|c). \tag{3}$$

By integrating fragment likelihoods using the naïve Bayes combination (3), we can obtain a parameter-free estimate of the posterior probability of each class given the image.

In contrast, if we consider the familiarity of image patches in the NIM model to be fragment likelihoods, we can think of NIM's fragment integration method as defining the likelihood of an image to be the mean of its fragment likelihoods:

$$P(\mathcal{F}|c) = \frac{1}{N} \sum_{i=1}^{N} P(f_i|c). \tag{4}$$

However, it is difficult to interpret this formulation's implicit assumptions about dependence between fragments.

Bayesian Classification

The classification decision is made by comparing the log ratio of the class and non-class posteriors:

$$\log \frac{P(c|\mathcal{F})}{P(\bar{c}|\mathcal{F})} = \log \frac{P(\mathcal{F}|c)P(c)}{P(\mathcal{F}|\bar{c})P(\bar{c})} = \log \frac{P(\mathcal{F}|c)}{P(\mathcal{F}|\bar{c})} + \log \frac{P(c)}{P(\bar{c})}. \tag{5}$$

The first term on the right-hand side of Equation (5) compares the relative likelihoods of the observed fragments under the class and non-class models. The second term controls the bias or prior weight that the model or subject puts on seeing images from class c versus all other images. The Bayes decision rule classifies the image as coming from class c when

Equation (5) is positive and from class $\bar{c}$ otherwise. In the multi-class framework, the Bayes-optimal rule is to chose the class with the largest posterior probability:

$$c^* = \underset{c}{\operatorname{argmax}} P(\mathcal{F}|c). \tag{6}$$

Kernel Density Estimation

Kernel density estimation places a kernel function at the point in memory space corresponding to every memorized fragment and computes the probability density of the new point f under each of these kernels. The sum of these probabilities forms the overall estimation for the *likelihood* of the new fragment, $P(f|c)$. The choice of kernel function and the parameters that control its shape are design features of the model, which we will consider below.

We may interpret the NIM (Lacroix et al., 2006) measure of a new fragment's familiarity (1) as a kernel density estimate that centers a hypersphere of radius r, with uniform density, at the location of each stored exemplar in memory space. The familiarity of a new fragment, f, can be viewed as summing its density under all of these uniform kernels.

By casting the problem of memory retrieval as a kernel density estimation task we can explore the model's performance under a variety of kernel functions beyond the hypersphere in (1). Indeed, this kernel prohibits using the naïve Bayes combination of fragment likelihoods (3) since, if a test fragment f were to find no stored points within radius r, it would be assigned zero likelihood. In that case, even if all other fragments were strongly predictive of the class, the resulting product of fragment likelihoods would be $P(\mathcal{F}|c) = 0$.

We implement the NIMBLE model using two alternative kernel functions. The first is a Gaussian kernel:

$$P(f|c) = \frac{1}{|\mathcal{M}_c|} \sum_{j=1}^{|\mathcal{M}_c|} \mathcal{N}(f, m_j, \sigma) \tag{7}$$

(here $\mathcal{N}(x,\mu,\sigma)$ represents the normal distribution of x with mean μ and variance σ). The second is a k-nearest-neighbor (kNN) kernel:

$$P(f|c) \propto \frac{k_c}{|\mathcal{M}_c|V}, \tag{8}$$

where V is the minimum volume centered at f that contains k_c of the $|\mathcal{M}_c|$ memories from class c.

NIMBLE's Bayesian framework can accommodate both naïve Bayes combination of fragment likelihoods (3) and NIM's averaging method of combining fragment likelihoods (4). In Tables 1 and 2, we refer to these two methods for obtaining an overall image likelihood from fragment likelihoods as *Naïve Bayes* and *Mean familiarity*, respectively. We also indicate the best parameter setting for each kernel.

Table 1: Model ROC area for face recognition memory. Image likelihoods are determined by combining the familiarities of image fragments using either naïve Bayes (3) or the mean of the fragment familiarities (4). The likelihood of an image, given the distractor class is found using a background model with either 10 or 80 dimensions. Standard errors of the mean are computed over 5 random trials.

Kernel	Fragment Combination	ROC area	
		10-D BG	80-D BG
Gaussian ($\sigma = 1$)	Naïve Bayes Mean familiarity	0.94±.03 0.97±.02	0.58±.02 0.62±.13
kNN ($k = 1$)	Naïve Bayes Mean familiarity	0.93±.05 0.96±.02	0.97±.02 0.96±.01

NIMBLE Results

In our simulations of memory tasks below, we consider both face and object datasets. For facial memory tasks, we use as input 128x192 pixel grayscale images from the FERET database (Phillips, Wechsler, Huang, & Rauss, 1998). Images of 95 male and female Caucasian faces without facial hair or glasses were chosen and the images were centered and normalized to have common eye positions and equal contrast. An example may be seen in Figure 1(a). For object memory tasks, we use 128×128 pixel grayscale images of 20 objects from the COIL-100 data set (Nene, Nayar, & Murase, 1996).

Face Recognition

The first experiment used to test memory performance is a simple face recognition task. We follow the formulation of (Duchaine & Nakayama, 2005) who used this method to evaluate the face and object memory performance of normal and prosopagnosic human subjects. The study phase of their task presented subjects a sequence of 10 target images, each displayed for 3 seconds. This target list was repeated for a total of 20 image viewings. In the test phase, subjects were presented with 50 images where 10 were the original targets (again, shown twice) and 30 were novel distractors, the lures. The subjects' task was to classify each image in the study phase as old or new. When tested on face image categories, normal human subjects achieved receiver operating characteristic (ROC) curve areas in the range 0.9 to 1.0 for this task.

In the study phase of our simulations, we extract $N = 10$ fragments from each of the target faces, approximating the number of saccades a human makes in 3 seconds. We sample each of the 10 target faces twice and store the resulting 200 fragments in the model's memory space. During the testing phase, we extract a *new* set of N fragments from the test faces. Given the stochastic nature of our interest operator, the exact fragments extracted from previously viewed target images are unlikely to be seen again. As (Henderson et al., 2005) demonstrated, human scan paths are not repeated in facial memory encoding and retrieval, and so simple exemplar matching may not perform well. In our experiments, the mean distance from a point sampled from a test face to the nearest study point from the same face was 8.8 pixels.

Table 2: Model accuracy for identity recognition memory tasks. Face ID uses 29 identities from FERET, Object ID uses 20 classes from COIL-100. (Optimal gaussian variance for object ID is 10-times greater than for face ID). Standard errors of the mean are computed over 5 random trials.

Kernel	Face ID Accuracy (%)		Object ID Accuracy (%)	
	Naïve Bayes	Mean familiarity	Naïve Bayes	Mean familiarity
Gaussian ($\sigma = 1, 10$)	85.6±2	72.2±2	87±1	73.7±2
kNN ($k = 1$)	89.2±.6	85.8±2	92.7±.7	87±.4

Since we do not restrict our model to discrete kernel functions such as (1), in which only a subset of the stored memories contribute to the old/new decision, all stored memories from a given class contribute to the estimate of the posterior probability of the class. In order to apply the Bayes decision rule (5) to the one-class recognition task described above, we recast it as a two-class classification task. Training image fragments are stored with a class label that indicates they have been seen in the study phase.

We need to be able to estimate the likelihood that an image fragment was generated by the lure (distractor) class, $P(f|\bar{c})$. To estimate this probability, we use a multivariate Gaussian whose variance in each feature dimension is set equal to the principal component (eigenvalue) obtained by performing PCA on fragments extracted from 55 face images not used in the study or test phases. We used this method because it approximates storing a large number of face patches that a subject might see over her lifetime but is computationally faster than explicitly sampling from an extra set of non-task images. We compared the effect of using two different background models to estimate $P(f|\bar{c})$: a low-dimensional background model using the first 10 principal component dimensions, and a high-dimensional background model using the first 80 principal component dimensions. In Table 1, we refer to these as *10-D BG* and *80-D BG*, respectively. The Gaussian kernel suffers a drop in performance when using the high-D background model since the extra dimensions of the 80-dimensional background model (which account for the least variance in the data) are quite susceptible to noise. When categorizing a new input, the kNN model (k = 1) uses only one data point, unlike the Gaussian model which takes input from every point in memory. Thus, the kNN model is less affected by noise.

For each set of test fragments, we compute the posterior probability that these image fragments were generated by the target and lure distributions. By varying the prior values for each class, $P(c)$ and $P(\bar{c})$, we can generate receiver operating characteristic (ROC) curves for the recognition memory task. The area under the ROC curve is computed and results are shown in Table 1. Normal human performance for face recognition results from (Duchaine & Nakayama, 2005) show ROC areas between 0.9 and 1.0, and NIMBLE performs similarly.

Image Identification

Having extended NIM to allow memories to be stored with class labels, we now apply NIMBLE to multi-class memory tasks. In this paradigm, the model is trained using 3 images (with different lighting, expressions or orientations) from 29 different FERET face identities or 20 COIL-100 object classes, and tested on 3 unseen images from each of these classes. Unlike in the recognition task, the model must now learn to identify images it has never seen before. The output of the model is the posterior probability for each class, and the classification decision is made using (6). For this multi-class problem, we assign equal prior probability to each of the classes and evaluate performance as the average accuracy over all classes. Note that the optimal parameter, σ, for the Gaussian kernel depends on the class of images to be identified since the within-class variance of patches taken from rotating objects (COIL-100) is much higher than the variance between patches sampled from aligned faces (FERET). Identification task results are shown in Table 2. Our model demonstrates high performance on these multi-class tasks. For example, our best object recognition model (kNN with Naïve Bayes) approaches state-of-the-art computer vision models for object recognition; (Belongie, Malik, & Puzicha, 2002) report 97.6% accuracy on the same task.

We use this multi-class task to demonstrate the advantage of using the naïve Bayes method for combination of fragment likelihoods (3) over the mean familiarity method (4) used in (Lacroix et al., 2006). For certain images, a given fragment may be either diagnostic of its true class or useful in excluding another class. In both cases, adding this fragment's likelihood to a running average over fragments (4) provides less useful modification to the ultimate posterior than does the probabilistically valid naïve Bayes updating method of multiplication (3). This scenario is illustrated in Figure 2 for the mean posterior probability of the correct class in the facial identification task, averaged over all 29 facial identities. We use an online version of NIMBLE to update the posterior, $P(c|\mathcal{F})$, as each fragment is added to $\mathcal{F}$. With more information, the posterior for the correct class with naïve Bayes likelihood combination (3) rises towards 1, while the posterior calculated using mean familiarity (4) remains roughly constant. The posterior probabilities of the 28 incorrect classes are not shown but, since the sum over all 29 classes must equal unity, it is clear that each incorrect class has very low probability and that the Bayes decision rule in Equation 5 almost always results in correct classification. Random guessing would set $P(c|\mathcal{F}) = \frac{1}{29}$. Note that these results make the prediction that, on average, a single saccade is enough to correctly identify a face!

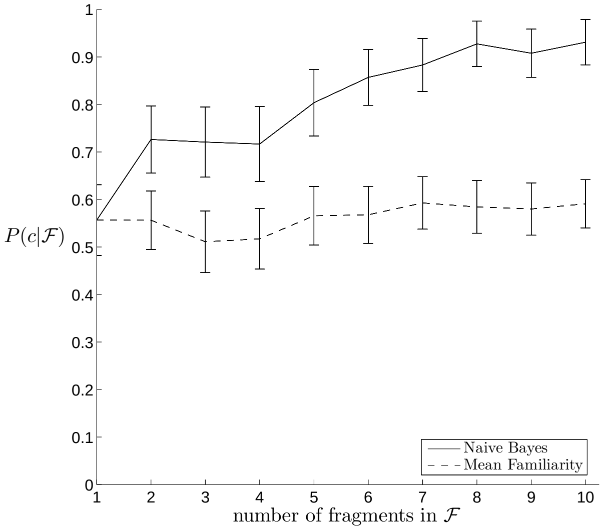

Figure 2: NIMBLE's posterior probability of the correct face class vs. number of fixations in the 29-class face identification task. Posteriors are computed using both naïve Bayes combination of fragment likelihoods (3) and mean familiarity combination of fragment likelihoods (4). The very low probabilities of the 28 incorrect classes are not shown.

Discussion

Using the NIM model (Lacroix et al., 2006) as our starting point, we developed NIMBLE, a biologically-inspired, saccade-based Bayesian model of face and object recognition. We have demonstrated that NIMBLE performance is comparable to human performance on a standard recognition memory task and that this biologically-inspired model approaches the best machine vision results. In addition, the online version of NIMBLE demonstrates that, like humans, our system can achieve correct identification and recognition of faces and objects after a very small number of fixations.

In future work, we plan to integrate top-down feedback into the system to direct fixations to sample from image locations with top-down interest as well as bottom-up salience. Because NIMBLE is a fully probabilistic model, it will be straightforward to integrate the existing model into more complex systems in the future.

References

Belongie, S., Malik, J., & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, 24-4,* 509-522.

Bishop, C. (1995). *Neural networks for pattern recognition.* Oxford University Press.

Dailey, M., Cottrell, G., & Busey, T. (1998). Facial memory is kernel density estimation (almost). *Neural Information Processing Systems.*

Dailey, M., Cottrell, G., Padgett, C., & Adolphs, R. (2002). Empath: a neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience. 14, 1158-1173.*

Duchaine, B., & Nakayama, K. (2005). Dissociations of face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience, 17, 249-261.*

Henderson, J., Williams, C., & Falk, R. (2005). Eye movements are functional during face learning. *Memory & Cognition, 33, 98-106.*

Hintzman, D. (1984). Minerva 2: A simulation model of human memory. *Behavior research methods, instruments and computers, 16,* 96-101.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2, No. 3, 194-203.*

Jones, J., & Palmer, L. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58(6) 1233-1258.*

Lacroix, J., Murre, J., & Postma, E. (2006). Modeling recognition memory using the similarity structure of natural input. *Cognitive Science, 30,* 121-145.

Lacroix, J., Murre, J., Postma, E., & Herik, H. J. V. den. (2004). The natural input memory model. *Proc. of the 26th annual meeting of the Cognitive Science Society.*

Mozer, M., Shettel, M., & Vecera, S. (2005). Top-down control of visual attention- a rational account. *Neural Information Processing Systems.*

Nelson, J., & Cottrell, G. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing.*

Nene, S., Nayar, S. K., & Murase, H. (1996). *Columbia object image library (coil-100)* (Tech. Rep.).

Nosofsky, R., & Palmeri, T. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104, 2, 266-300.*

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

Phillips, J., Wechsler, H., Huang, J., & Rauss, P. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image & Vision Computing, 16, 5, 295-306.*

Renninger, L., Coughlan, J., Verghese, P., & Malik, J. (2004). An information maximization model of eye movements. *Neural Information Processing Systems.*

Wolfe, J. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1, 2, 202-238.*

Yamada, K., & Cottrell, G. (1995). A model of scan paths applied to face recognition. *Proc. of the 17th Annual Cognitive Science Conference 55-60.*

Yarbus, A. (1967). *Eye movements and vision.* Plenum Press, New York.

Zelinsky, G., Zhang, W., B. Yu, X. C., & Samaras, D. (2005). The role of top-down and bottom-up processes in guiding eye movements during visual search. *Neural Information Processing Systems.*