

# NIMBLE: A kernel density model of saccade-based visual memory

**Luke Barrington**

Department of Electrical and Computer Engineering,  
University of California, San Diego, La Jolla, CA, USA



AQ1

**Tim K. Marks**

Department of Computer Science and Engineering,  
University of California, San Diego, La Jolla, CA, USA



AQ1

**Janet Hui-wen Hsiao**

Department of Computer Science and Engineering,  
University of California, San Diego, La Jolla, CA, USA



AQ1

**Garrison W. Cottrell**

Department of Computer Science and Engineering,  
University of California, San Diego, La Jolla, CA, USA



AQ1

We present a Bayesian version of J. Lacroix, J. Murre, and E. Postma's (2006) Natural Input Memory (NIM) model of saccadic visual memory. Our model, which we call NIMBLE (NIM with Bayesian Likelihood Estimation), uses a cognitively plausible image sampling technique that provides a foveated representation of image patches. We conceive 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 combine 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. We also use NIMBLE to examine the change in beliefs as more fixations are taken from an image. Using fixation data collected from human subjects, we directly compare the performance of NIMBLE's memory component to human performance, demonstrating that using human fixation locations allows NIMBLE to recognize familiar faces with only a single fixation.

Keywords: computational modeling, eye movements, face recognition, object recognition, memory, kernel density estimation

Citation: Barrington, L., Marks, T. K., Hui-wen Hsiao, J., & Cottrell, G. W. (2008). NIMBLE: A kernel density model of saccade-based visual memory. *Journal of Vision*, 0(0):1, 1–14, <http://journalofvision.org/0/0/1/>, doi:10.1167/0.0.1.

## Introduction

In human visual perception, we repeatedly foveate different areas of a visual scene, concentrating fixations on the parts that are most salient or task-relevant (Yarbus, 1967). It is still mysterious how we nevertheless are able to recognize objects from these samples. It would be slightly less mysterious if we fixated exactly the same locations each time we viewed an image and thus extracted identical fragments. However, while the scan paths we use may be similar between two observations of the same image (Foulsham & Underwood, 2008), it is by no means a requirement for high recognition rates that we fixate the same locations (Henderson, Williams, & Falk, 2005). For example, Henderson et al. (2005) showed that subjects do not make significantly different fixations during face recognition compared to control subjects after having their fixations artificially restricted during face learning. In other words, the scan paths generated during

recognition may not be just replicating those followed during learning, as proposed by the scan path theory of Noton and Stark (1971). Thus, simple exemplar matching of information from new saccades to stored memories cannot be relied upon to account for human capacities for object recognition, as the exemplars may differ between study and test.

Lacroix, Murre, Postma, and Van den Herik (2004) and Lacroix, Murre, and Postma (2006) have proposed the Natural Input Memory (NIM) model to account for humans' ability to recognize faces from fixations. (We use the term face *recognition* in the sense used in the experimental psychology literature. It refers to the ability to discriminate previously seen faces from novel faces, based on a study list. In contrast, we use face *identification* to refer to the ability to identify face images as particular individuals). NIM is an exemplar model of memory (Raaijmakers & Shiffrin, 2002), in that it stores memories as points in a vector space and compares memories based on distances in this space. However,

NIM differs from standard mathematical psychology models in that it (a) it uses actual facial images as input and (b) it is based on the idea of storing fixation-based face fragments, rather than whole face exemplars. The NIM model's memory is reminiscent of a kernel density estimator but differs in important details from a true probabilistic model in the way that the estimates from individual fragments are combined. In this paper, we present a Bayesian version of the NIM model that uses naive Bayes to combine the likelihood estimates from individual fragments. We further extend the model to perform 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 also performs multi-class face and object identification tasks with high accuracy. Bayesian combination of individual fragment likelihoods outperforms the combination method from the original NIM model in most cases, and the new kernels far outperform those used in NIM.

Though there are few cognitive models of saccade-based visual memory, fragment-based models are common in computer vision. Supporting the idea that the whole image can be recovered from sampling only at interesting fixation points, the work of Raj, Geisler, Frazor, and Bovik (2005) on entropy minimization of natural scenes demonstrates that images can be reconstructed from fragments. Ullman, Vidal-Naquet, and Sali (2002) used the mutual information between an image fragment and the class label of the object from which it is sampled to show that fragments of intermediate complexity (fragments that are smaller than the total object but much larger than a pixel) are most useful for image classification. The SIFT features proposed by Lowe (2004) are based on finding key points in images that are invariant to changes of scale, orientation, and illumination, then describing each point using histograms of image gradients in the region surrounding the point. Belongie, Malik, and Puzicha (2002) find interest points in images and model the correspondence between the shapes described by these points to compare and classify images. Applying fragment-based representation to video, Dollar, Rabaud, Cottrell, and Belongie (2005) find interest points in three-dimensional video signals and extract spatio-temporal fragments (called cuboids) for use in behavior classification. Thus, not only are image fragments a biologically plausible representation for image classification, they have also been used quite successfully in computer vision applications.

In the [Methods](#) section, 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 compare human and model performances on visual memory tasks in the [Results](#) section, and conclude the paper with a discussion.

## Methods

### 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 (such as sound), as well as top-down task-dependent directives such as attention and expectation. Though many computer models (e.g., Mozer, Shettel, & Vecera, 2006; Wolfe, 1994; Zelinsky, Zhang, Yu, Chen, & Samaras, 2006) have been proposed for how to integrate top-down and bottom-up cues, in this work we select fixations based only on the bottom-up salience of static images. We model the fixation selection process using a local interest operator for determining the scan paths (Yamada & Cottrell, 1995). This 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:

$$\text{Salience}(i, j) = \frac{1}{8} \sum_{n=1}^8 (G(i, j, \theta) - \mu_G(i, j))^2, \quad (1)$$

where  $G(i, j, \theta)$  is the magnitude response of a Gabor filter with orientation  $\theta$  centered at pixel  $(i, j)$ , and  $\mu_G(i, j)$  is the mean response across all eight orientations. A similar technique developed by Renninger, Coughlan, Verghese, and Malik (2005) 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 point by subtracting from the salience distribution a rotationally symmetric Gaussian, centered at the fixation point, and then renormalizing. This inhibits repeated fixations of the same location and prevents fixations from clustering in areas that are initially highly salient. A given image will always have the same initial salience map, but by randomly sampling from this distribution and inhibiting return to fixated locations, repeated viewings of the same image will result in different scan paths.

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 and Koch (2001) for a computational model of visual attention: it derives

132  
133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

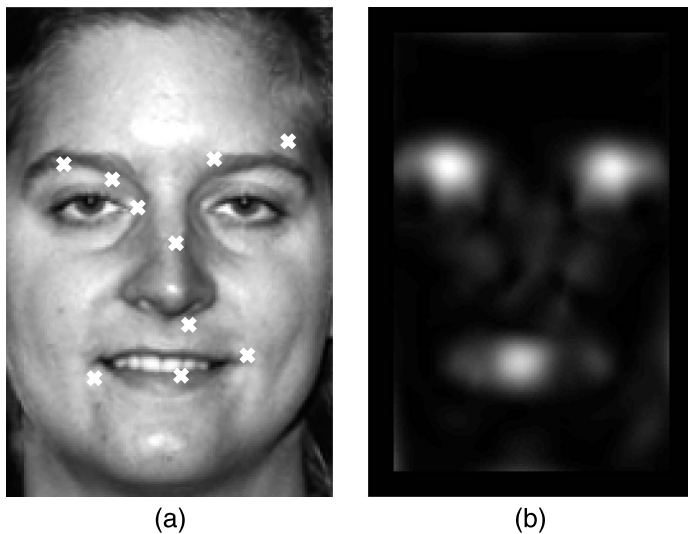


Figure 1. (a) An image from the FERET database with 10 sample fixation points. (b) The corresponding saliency map generated using the technique of Yamada and Cottrell (1995). The fixations shown in (a) were chosen randomly according to this saliency map, and while they tend to cluster around highly salient areas, inhibition of return enforces a more even distribution of fixations across the image.

181 perceptual salience of a fixation point from the surround-  
 182 ing context, it creates a saliency distribution over the  
 183 visual scene, and it inhibits return to previously attended  
 184 locations. In this paper, we ignore the remaining two  
 185 criteria, which concern top-down influences on fixation  
 186 point selection. Since this saliency model is fully  
 187 probabilistic, it could be combined with top-down feed-  
 188 back to direct eye movements, for example, by extending  
 189 the results of Nelson and Cottrell (2007) to determine the  
 190 fixations that would be most useful in enhancing perfor-  
 191 mance on the current visual task.

192 We have tested NIMBLE using various alternative  
 193 mechanisms for computing visual salience (for selecting  
 194 fixations). NIMBLE using the saliency operator of Itti and  
 195 Koch (2001) results in roughly the same face recognition  
 196 performance as NIMBLE using the saliency operator  
 197 (Equation 1) of Yamada and Cottrell (1995), but the latter  
 198 approach has the advantage that the same mechanism  
 199 (Gabor filters) is used for computing salience as for  
 200 processing images (see the [Retinal/cortical image trans-  
 201 form](#) section). Purely random selection of fixations  
 202 (corresponding to a uniform saliency map) reduces  
 203 performance by 30%. Sampling fixations from the Canny  
 204 edge map of the image, as was done in the NIM model,  
 205 reduces performance by 20%. We also tested NIMBLE  
 206 using the locations of actual human fixations, which were  
 207 recorded from the same face images using an eye tracker  
 208 (Hsiao & Cottrell, [in press](#)). A detailed description of this  
 209 experiment is given in the [Results](#) section.

### Retinal/cortical image transform

211 A fixated patch of an input image undergoes many  
 212 stages of neural processing before being represented as a  
 213 pattern of activation in high-level visual cortex. Our  
 214 biologically inspired model of the processing in primary  
 215 visual cortex (V1) uses the magnitude responses of Gabor  
 216 filters at 8 orientations and 4 frequencies (Jones & Palmer,  
 217 1987). At each orientation, we use Gabor filter frequencies  
 218 of  $\frac{1}{16}$ ,  $\frac{1}{12}$ ,  $\frac{1}{8}$ , and  $\frac{1}{4}$  cycles/pixel (corresponding to 8, 10  $\frac{2}{3}$ , 16,  
 219 and 32 cycles/face) to approximate the varying resolution  
 220 available to the retina. We transform an image into the  
 221 Gabor-filtered domain by calculating the response of each  
 222 of these 32 filters (8 orientations  $\times$  4 frequencies) at every  
 223 image pixel.

224 The model approximates a foveated representation of  
 225 the fixated location by extracting square patches from  
 226 these Gabor response images. The highest spatial fre-  
 227 quency filter responses correspond to the high-resolution  
 228 foveated area centered at the fixated location. The low-  
 229 frequency filter responses at a given pixel within the  
 230 square patch are computed from an image area with  
 231 spatial context that extends beyond the borders of the  
 232 patch—a low-resolution filter placed at the edge of a patch  
 233 responds with one quarter of its peak magnitude at 52  
 234 pixels from the fixation point or 40% of the width of the  
 235 test images, corresponding to a visual angle of  $3.2^\circ$   
 236 (assuming a real face viewed from a distance of 1 m  
 237 occupies  $8^\circ$  of the visual field; Henderson et al., 2005).  
 238 Thus, this patch-based representation includes extra-  
 239 foveal information corresponding to the low-resolution  
 240 data from the retinal periphery.

241 The size of the extracted patch of filter responses and  
 242 the number of fragments that the model may examine for  
 243 each image are experimental parameters that correspond,  
 244 in human vision, to the distance of the eye from the image  
 245 (and thus the size of the foveated area) and the time spent  
 246 studying the image (determining the number of fixations  
 247 made). For a fragment size of  $35 \times 35$  pixels (corre-  
 248 sponding to a visual angle of  $2^\circ$  for a subject about 1 m  
 249 from a real face, an approximation of the human studies  
 250 discussed below), the model's input feature vector has  
 251  $35 \times 35 \text{ pixels} \times 8 \text{ orientations} \times 4 \text{ frequencies} = 39,200$   
 252 dimensions. This fragment size was chosen to approx-  
 253 imate the experimental conditions experienced by the  
 254 human subjects whose data we model in the [NIMBLE  
 255 using human fixations](#) section, but the NIMBLE model  
 256 could use image fragments of any size. For efficiency and  
 257 good generalization, we use principal component analysis  
 258 (PCA) to reduce the size of this vector to 80 components,  
 259 retaining about 90% of the data variance (depending on  
 260 the data set—see the [Results](#) section). This feature  
 261 extraction procedure of wavelet-based image decomposi-  
 262 tion followed by PCA is a standard approximation for  
 263 biologically motivated vision models (Dailey, Cottrell,  
 264 Padgett, & Adolphs, 2002; Lacroix et al., 2006; Palmeri &  
 265 Gauthier, 2004).

## 269 The Natural Input Memory (NIM) model

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

281 Following the lead of many cognitive memory models  
 282 (Dailey, Cottrell, & Busey, 1998; Hintzman, 1984;  
 283 Nosofsky & Palmeri, 1997), the NIM model’s memory  
 284 process stores the feature-transformed representation of  
 285 fixated image fragments as vectors in a high-dimensional  
 286 memory space. Memories are compared to each other as  
 287 well as to new image fragments by comparing the  
 288 Euclidean distance between their vector representations  
 289 in this memory space. The NIM model computes the  
 290 familiarity of a new fragment by calculating the propor-  
 291 tion of previously stored memories that lie within a radius  
 292  $r$  (a model parameter) of the new fragment in the memory  
 293 space. Averaging these familiarities over all samples from  
 294 a new image produces an estimate of the probability that  
 295 the image is from the class known to the memory. The  
 296 memory space introduced by the NIM model has been  
 297 shown to achieve the best known correlation with human  
 298 judgments of perceptual similarity (Lacroix et al., 2006),  
 299 and the retrieval methods exhibit human performance  
 300 effects (such as list length and list strength) on face  
 301 recognition memory tasks (Lacroix et al., 2004).

302 The NIM memory retrieval method (Lacroix et al.,  
 303 2006) determines the familiarity of a newly examined  
 304 fragment by counting how many of the stored memories,  
 305  $\{m_1, \dots, m_M\}$ , lie within a radius  $r$  of the new image  
 306 fragment. Thus, the familiarity of the new fragment,  $f$ , is  
 307 defined by

$$fam(f) = \sum_{j=1}^M I_r(\|m_j - f\|_2), \quad (2)$$

308 where

$$I_r(x) = \begin{cases} 1, & x \leq r \\ 0, & \text{otherwise} \end{cases}. \quad (3)$$

310

### 312 NIM combination of fragment familiarities

313 An image is represented as a set of  $N$  fragments  $F = \{f_1,$   
 314  $\dots, f_N\}$ . In the NIM model, Lacroix et al. (2006) define the

familiarity of a test image as the mean of the familiarities 315  
 of all  $N$  fragments taken from that image: 316

$$fam(F) = \frac{1}{N} \sum_{i=1}^N fam(f_i). \quad (4)$$

They use a logistic function to transform this mean 318  
 familiarity value into a probability value between 0 and 1: 319

$$P(\text{familiar image}) = \frac{1}{1 + \beta e^{-\theta fam(F)}}, \quad (5)$$

where  $\beta$  and  $\theta$  are parameters of the model used to fit the 320  
 performance to human data. 322

The NIM model formulation (Lacroix et al., 2006) only 323  
 attempts to make judgments about the familiarity of a 324  
 studied image by comparing a set of fragments extracted 325  
 from the image to all previously stored memories. Since 326  
 these memories are stored without labels, the resulting 327  
 familiarity value must be compared to a threshold to 328  
 decide whether the image is familiar or unfamiliar. Our 329  
 extension of NIM, described in [A more NIMBLE](#)  
[approach](#) section, stores class labels with each exemplar 330  
 and can return explicit posterior probabilities for each 331  
 class given the image fragments, permitting multi-class 332  
 and hierarchical memory tasks in addition to the familiar/  
 unfamiliar recognition memory task of Lacroix et al. 333  
 (2006). More recent versions of NIM (e.g., NIMCLASS 334  
 (Lacroix, Postma, & Van den Herik, 2007) have also been 335  
 extended to classification. These more recent versions are 336  
 compared to our model in the [Discussion](#) section. 337  
 338  
 339  
 340  
 341

## A more NIMBLE approach 342

Having sampled and processed a new image as 343  
 described above in the [Retinal/cortical image transform](#)  
 section, we want to evaluate the probability of the 344  
 resulting set of  $N$  fragments,  $F = \{f_1, \dots, f_N\}$ , under the 345  
 models for each of a number of image classes. For 346  
 instance, we handle the previously described familiar/  
 unfamiliar faces task as a two-class problem and can 347  
 additionally handle other classification tasks such as 348  
 Alice/Bob/Carol/Dan/unknown or dogs/not dogs. For each 349  
 class,  $c$ , we use Bayes rule to compute the posterior 350  
 distribution: 351  
 352  
 353

$$p(c|F) = \frac{p(F|c)p(c)}{p(F)}. \quad (6)$$

In this case,  $p(F|c)$  is the likelihood of the set of image 354  
 fragments under the density model for class  $c$ , and  $p(c)$  is 355

357 the class prior which may be learned from experience with  
358 training data.

359 We compute the likelihood of the set of image  
360 fragments,  $p(F|c)$ , by combining the likelihoods of each  
361 individual fragment,  $p(F_i|c)$ , as explained in the [Naive](#)  
362 [Bayes fragment combination](#) section. Each of these class-  
363 conditional fragment likelihoods is computed using kernel  
364 density estimation (see the [Kernel density estimation](#)  
365 section).

### 366 [Naive Bayes fragment combination](#)

367 In the NIMBLE model, we make the naive Bayes  
368 assumption of conditional independence between each  
369 fragment  $f_i \in F$ , given the class, and take the product of  
370 the individual fragment likelihoods to obtain an estimate  
371 of the overall likelihood function:

$$p(F|c) = \prod_{i=1}^N p(f_i|c). \quad (7)$$

373 By integrating fragment likelihoods using the naive Bayes  
374 combination ([Equation 7](#)), we can obtain a parameter-free  
375 estimate of the posterior probability of each class given the  
376 image.

377 In contrast, if we conceive of the familiarity of image  
378 patches in the NIM model as fragment likelihoods, then  
379 we can think of NIM's fragment integration method  
380 ([Equation 4](#)) as defining the likelihood of an image to be  
381 the mean of its fragment likelihoods:

$$p(F|c) = \frac{1}{N} \sum_{i=1}^N p(f_i|c). \quad (8)$$

383  
384 Combining fragment likelihoods by taking their prod-  
385 uct using the naive Bayes method ([Equation 7](#)) assumes  
386 that the probabilities of observing each patch are condi-  
387 tionally independent, given the class. While this assump-  
388 tion is clearly not true (e.g., if our first glimpse of a  
389 picture of a randomly chosen friend reveals Alice's left  
390 eye, it is very likely that a further saccade will reveal  
391 Alice's right eye), explicitly modeling the class-conditional  
392 dependencies of all possible sets of observable fragments  
393 is computationally intractable. While it is unwarranted, at  
394 least the conditional independence assumption is explicit  
395 in the model, and this probabilistic framework allows for  
396 the inclusion of higher level dependencies. In contrast, the  
397 probabilistic interpretation of the implicit assumptions  
398 about dependence between fragments in the mean famil-  
399 iarity method of combination ([Equation 4](#)) is unclear,  
400 though Kittler, Hatef, Duin, and Matas (1998) have  
401 indicated that such ad hoc methods can perform well in  
403 practice.

### [Bayesian classification](#)

404  
405 The classification decision is made by comparing the  
406 log ratio of the class and non-class posteriors:

$$\begin{aligned} \log \frac{p(c|F)}{p(\bar{c}|F)} &= \log \frac{p(F|c)p(c)}{p(F|\bar{c})p(\bar{c})} \\ &= \log \frac{p(F|c)}{p(F|\bar{c})} + \log \frac{p(c)}{p(\bar{c})}. \end{aligned} \quad (9)$$

408 The first term on the right-hand side of [Equation 9](#)  
409 compares the relative likelihoods of the observed frag-  
410 ments under the class and non-class models. The second  
411 term controls the bias or prior weight that the model or  
412 subject puts on seeing images from class  $c$  versus all other  
413 images. The Bayes decision rule classifies the image as  
414 coming from class  $c$  when [Equation 9](#) is positive and from  
415 class  $\bar{c}$  otherwise. In the multi-class framework, the  
416 Bayes-optimal rule is to choose the class with the largest  
417 posterior probability:

$$c^* = \operatorname{argmax}_c p(c|F). \quad (10)$$

### [Kernel density estimation](#)

422 Kernel density estimation centers a kernel function at  
423 the point in memory space corresponding to every  
424 memorized fragment and computes the probability density  
425 of a new point (new fragment)  $f$  under each of these  
426 kernels. The sum of these probabilities forms the overall  
427 estimate of the likelihood of the new fragment,  $p(f|c)$ .  
428 The choice of kernel function and the parameters that  
429 control its shape are design features of the model, which  
430 we will consider below.

431 We may interpret the original NIM (Lacroix et al., 2006)  
432 measure of a new fragment's familiarity ([Equation 2](#)) as a  
433 kernel density estimate that centers a hypersphere of  
434 radius  $r$ , with uniform density, at the location of each  
435 stored exemplar in memory space. The familiarity of a  
436 new fragment,  $f$ , can be viewed as summing its density  
437 under all of these uniform kernels. By casting the problem  
438 of memory retrieval as a kernel density estimation task,  
439 we can explore the model's performance under a variety  
440 of kernel functions beyond the hypersphere in [Equation 2](#).  
441 Indeed, this NIM kernel prohibits using the naive Bayes  
442 combination of fragment likelihoods ([Equation 7](#)), since if  
443 a test fragment  $f$  was to find no stored points within radius  
444  $r$ , it would be assigned zero likelihood. In that case, even  
445 if all other fragments were strongly predictive of the class,  
446 the resulting product of fragment likelihoods would be  
447  $p(F|c) = 0$ .

Kernel	Face ID accuracy (%)		Object ID accuracy (%)	
	Naïve Bayes	Mean familiarity	Naïve Bayes	Mean familiarity
Gaussian ( $\sigma = 1, 10$ )	85.6 $\pm$ 2	72.2 $\pm$ 2	87 $\pm$ 1	73.7 $\pm$ 2
kNN ( $k = 1$ )	89.2 $\pm$ 0.6	85.5 $\pm$ 2	92.7 $\pm$ 0.7	87 $\pm$ 0.4

Table 1. Model accuracy for NIMBLE identification 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.

448 We have implemented the NIMBLE model using two  
449 alternative kernel functions. The first is a Gaussian kernel:

$$p(f|c) = \frac{1}{|M_c|} \sum_{j=1}^{|M_c|} N(f, m_j, \sigma), \quad (11)$$

450 where  $N(f, m_j, \sigma)$  represents the normal probability density  
452 function of  $x$  with mean  $\mu$  and  $\sigma$  variance, and  $M_c = \{m_1,$   
453  $m_2, \dots, m_{|M_c|}\}$  is the set of previously memorized  
454 fragments from class  $c$ . The second is a  $k$ -nearest-neighbor  
455 (kNN) kernel:

$$p(f|c) \propto \frac{k_c}{|M_c|V}, \quad (12)$$

456 where  $V$  is the minimum volume centered at  $f$  that  
458 contains  $k$  stored memories, of which are from class  $c$   
459 (Bishop, 1995).

460 NIMBLE's Bayesian framework can accommodate both  
461 naive Bayes combination of fragment likelihoods (Equa-  
462 tion 7) and NIM's averaging method of combining  
463 fragment likelihoods (Equation 8). In Tables 1 and 2, we  
464 refer to these two methods for obtaining an overall image  
465 likelihood from fragment likelihoods as *Naïve Bayes* and  
466 *Mean familiarity*, respectively. In each table, we also  
467 indicate the best parameter setting for each kernel, where  
468 optimization over the parameters was performed using 10  
469 random trials.

## Results

473 In this section, we present the results of applying  
474 NIMBLE to a face identification task, an object identi-  
475 fication task, and a face recognition (memory) task.

### Experimental data

478 In the experiments described below, we consider both  
479 face and object data sets. For facial identity and memory  
480 tasks, we use as input  $128 \times 192$  pixel grayscale images  
481 from the FERET database (Phillips, Wechsler, Huang, &  
482 Rauss, 1998). Images of male and female Caucasian faces  
483 without facial hair or glasses were chosen, and the images  
484 were centered and normalized to have common eye  
485 positions and equal contrast. An example may be seen in  
486 Figure 1a. For object identification tasks, we use  $128 \times$   
487  $128$  pixel grayscale images of 20 objects from the COIL-  
488 100 data set (Nene, Nayar, & Murase, 1996).

### Multi-class face and object identification

490 Since the NIMBLE model allows memories to be stored  
492 with class labels, unlike the original NIM model (but like  
493 other NIM extensions such as NIM-CLASS; Lacroix  
494 et al., 2007), we can apply NIMBLE to multi-class  
495

Kernel	Fragment combination	ROC area	
		10-D BG	80-D BG
Gaussian (=1)	Naïve Bayes	0.94 $\pm$ 0.03	0.58 $\pm$ 0.02
	Mean familiarity	0.97 $\pm$ 0.02	0.62 $\pm$ 0.13
kNN ( $k = 1$ )	Naïve Bayes	0.93 $\pm$ 0.05	0.97 $\pm$ 0.02
	Mean familiarity	0.96 $\pm$ 0.02	0.96 $\pm$ 0.01

Table 2. Model ROC area for face recognition memory. Image likelihoods are determined by combining the familiarities of image fragments using either naive Bayes (Equation 7) or the mean of the fragment familiarities (Equation 8). The likelihood of an image given the distracter class is found using a background model with either 10 or 80 dimensions. Standard errors of the mean are computed over 5 random trials.

identification tasks. In this paradigm, the model is trained using  $N = 10$  fragments to represent 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. In this *identification* task, the model is presented with a novel test image that it has never seen before, and it must identify which category this novel image belongs to, based on previously studied images from the same face or object category. This is unlike the face *recognition* task described below in which the model must recognize the exact face images that it has previously studied.

For each test image, the output of the model is the posterior probability for each class, and the classification decision is made using Equation 10. For this multi-class problem, we assign equal prior probability to each of the classes and evaluate overall performance as the average accuracy across all classes. Note that the optimal parameter,  $s$ , 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 across patches sampled from aligned faces (FERET). We fit this parameter by 10-fold cross validation on randomly sampled image sets. Identification task results are shown in Table 1. Our model demonstrates high performance on these multi-class tasks. For example, our best object recognition model (kNN with Naive Bayes) achieves a respectable performance of almost 93%. A state-of-the-art computer vision system for object recognition, Belongie’s shape context system (Belongie et al., 2002), achieves 97.6% accuracy on the same task. However, that system uses far more complex and less biologically plausible—methods for selecting and matching correspondence points.

We use this multi-class task to demonstrate the advantage of using the naive Bayes method for combination of fragment likelihoods (Equation 7) over the mean familiarity method (Equation 8) used by 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, simply adding this fragment’s likelihood to a running average over fragments (Equation 8) provides less useful modification to the ultimate posterior than does the naive Bayes updating method of multiplication (Equation 7). This is illustrated in Figure 2, which plots the mean posterior probability of the correct class in the facial identification task, averaged over all 29 facial identities. For this figure, we use an online version of NIMBLE to update the posterior,  $p(c|F)$ , as each fragment is added to  $F$ . With more information, the posterior for the correct class using naive Bayes likelihood combination (Equation 7) rises toward 1, while the posterior calculated using mean familiarity (Equation 8) 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 therefore, the

Bayes decision rule (Equation 9) almost always results in correct classification. For comparison, random guessing would set  $p(c|F) = \frac{1}{29}$ . Note that the results shown in Figure 2 reveal that, on average, a single fixation is enough to correctly identify a face.

## Face recognition

Having validated NIMBLE’s capabilities as a saccade-based face and object identification model, we now test it on a standard human memory task. We begin by testing NIMBLE’s memory performance on a simple face recognition task. We follow the standard formulation of many human experiments, including Hsiao and Cottrell (in press), which we will examine further below. In the study phase (training phase) of our simulations, NIMBLE extracts  $N = 10$  fragments (approximating the number of saccades a human makes in 3 s) from each of 32 target images of faces. NIMBLE samples each of the 32 target faces and stores the resulting 320 fragments in the model’s memory space. During the testing phase, NIMBLE extracts a new set of  $N$  fragments from 64 test face images, of which 32 are the original targets and 32 are novel distracters, known as lures. The model’s task is to classify each image in the test phase as target (familiar) or lure (unfamiliar).

When viewing an image, NIMBLE selects a set of fixation points that are independent samples from the

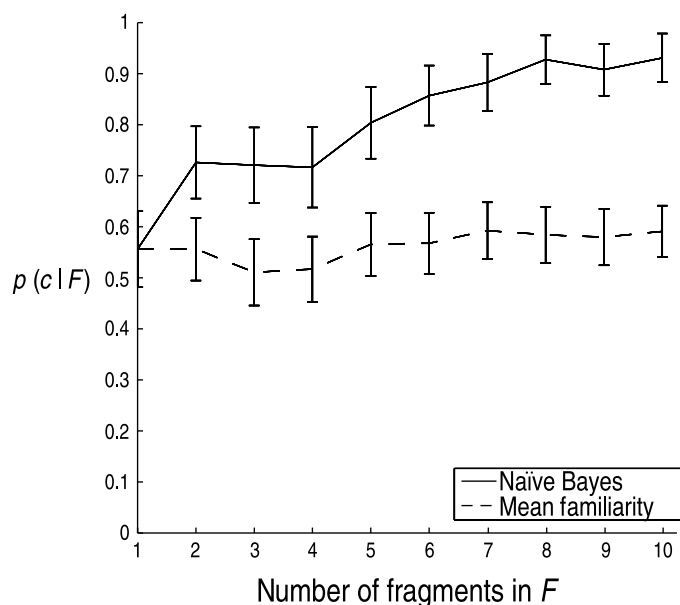


Figure 2. NIMBLE’s posterior probability of the *correct* face class vs. number of fixations in the 29-class face identification task. Posteriors are computed using both naive Bayes combination of fragment likelihoods (Equation 7) and mean familiarity combination of fragment likelihoods (Equation 8). The very low probabilities of the 28 incorrect classes are not shown.

580 salience distribution for that image. Although the fixation  
 581 points chosen during training and testing are sampled  
 582 from the same underlying salience distribution, the  
 583 stochasticity involved in this process, as well as the  
 584 renormalization that results from the suppression of  
 585 previously fixated locations, means that the actual fixation  
 586 locations are almost always different. As Henderson et al.  
 587 (2005) demonstrated, human scan paths in facial memory  
 588 retrieval are not just replicating those generated during  
 589 memory encoding, and so simple exemplar matching may  
 590 not perform well. In our experiments, the mean distance  
 591 from a point sampled from a face during encoding to the  
 592 nearest point from the same face studied during retrieval  
 593 was 8.8 pixels or 0.5 degrees of visual angle.

594 Since we do not restrict our model to discrete kernel  
 595 functions such as Equation 2, in which only a subset of the  
 596 stored memories contribute to the old/new decision, all  
 597 stored memories from a given class contribute to the  
 598 estimate of the posterior probability of the class. In order  
 599 to apply the Bayes decision rule (Equation 9) to this one-  
 600 class recognition task, we recast it as a two-class  
 601 classification task. Fragments from training images are  
 602 stored with a class label that indicates they have been seen  
 603 in the study phase.

604 We need to estimate the likelihood that an image  
 605 fragment was generated by the lure (distracter) class,  
 606  $p(f|\bar{c})$ . To estimate this probability, we use a multivariate  
 607 Gaussian whose variance in each feature dimension is set  
 608 equal to the principal component (eigenvalue) obtained by  
 609 performing PCA on fragments extracted from 55 face  
 610 images not used in the study or test phases. We used this  
 611 method because it approximates storing a large number of  
 612 face patches that a subject might see over her lifetime but  
 613 is computationally faster than explicitly sampling from an  
 614 extra set of non-task images. We compared the effect of  
 615 using two different background models to estimate  
 616  $p(f|\bar{c})$ : a low-dimensional background model using the  
 617 first 10 principal component dimensions and a high-  
 618 dimensional background model using the first 80 principal  
 619 component dimensions. In Table 2, we refer to these as  
 620 10-D BG and 80-D BG, respectively. The Gaussian kernel  
 621 suffers a drop in performance when using the high-D  
 622 background model because the extra dimensions of the  
 623 80-dimensional background model (which account for the  
 624 least variance in the data) are quite susceptible to noise.  
 625 When categorizing a new input, the kNN model ( $k = 1$ )  
 626 uses only one data point, unlike the Gaussian model which  
 627 takes input from every point in memory. As a result, the  
 628 kNN model is less affected by noise.

629 For each set of test fragments, we compute the posterior  
 630 probability that these image fragments were generated by  
 631 the target  $p(c|F)$  and lure distributions  $p(\bar{c}|F)$ . Comput-  
 632 ing the (log) ratio of these probabilities (as in Equation 9)  
 633 for each image provides a ranking of the images in order  
 634 of how likely they are to be a familiar target image; larger  
 635 values of  $\frac{p(c|F)}{p(\bar{c}|F)}$  are more likely to be targets rather than  
 636 lures. We quantify the model's results using the area

under the receiver operating characteristic (ROC) curve. 637  
 The ROC curve compares the rate of correct detections to 638  
 false alarms at each point in the ranking. By varying the 639  
 prior probabilities for each class,  $p(c)$  and  $p(\bar{c})$  (which 640  
 comprise the second term on the right-hand side of 641  
 Equation 9), we can trace out every point on the ROC 642  
 curve and compute the area under the curve. This ROC 643  
 area provides a single number that describes the accuracy 644  
 of the ranking (often estimated non-parametrically in 645  
 psychophysics experiments as  $A'$ ). For example, when 646  
 $p(c) = 0$ , all images are deemed to be lures, whereas when 647  
 $p(\bar{c}) = 0$ , all images are recognized as targets. A perfect 648  
 ranking (i.e., all the target images at the top of the ranking 649  
 and all the lures at the bottom) would result in an ROC 650  
 area equal to 1. Ranking images randomly, we would 651  
 expect the ROC area to be 0.5. 652

The results for the recognition memory task are shown 653  
 in Table 2. We can see that in this task the mean 654  
 familiarity model performs quite well. When tested on 655  
 face images, normal human subjects achieved  $A'$  (a bias- 656  
 free, nonparametric estimate of ROC area) in the range of 657  
 0.9 to 1.0 for this task (e.g., Duchaine & Nakayama, 2005; 658  
 Hsiao & Cottrell, in press), and NIMBLE performs 659  
 similarly. A more detailed analysis of NIMBLE's per- 660  
 formance in comparison to humans is given in the 661  
 NIMBLE using human fixations section. 662

## 663 NIMBLE using human fixations 664

The memory experiments described above demonstrate 665  
 that NIMBLE performs well on standard memory tasks: 666  
 NIMBLE approaches a more sophisticated computer 667  
 vision model's object identification abilities, and NIM- 668  
 BLE recognizes familiar faces as well as humans. This 669  
 performance arises in spite of NIMBLE's simple, bottom- 670  
 up model of visual salience. In order to test the saccade- 671  
 based memory component of the model in isolation 672  
 (independent of the salience model), we examine NIM- 673  
 BLE's performance on a memory task for which we know 674  
 the exact fixation locations used by humans. 675

## 676 Human face recognition with varying numbers 677 of fixations

We conducted a human experiment to examine how 678  
 many fixations are required to recognize a face (Hsiao & 679  
 Cottrell, in press). During an otherwise standard face 680  
 recognition task, participants were allowed a variable 681  
 number of fixations (one, two, three, or no restriction/free 682  
 viewing) before the stimulus was masked. The stimuli 683  
 consisted of 16 male and 16 female Caucasian face 684  
 images, taken from the FERET database (Phillips et al., 685  
 1998). During the experiment, the image size on the 686  
 screen spanned about 8 degrees of visual angle, equivalent 687  
 to the size of a real face at a viewing distance of 1 m; an 688

689 area approximately the size of one eye on the face may be  
 690 foveated at a time. The eye movements of eight male and  
 691 eight female right-handed participants were recorded with  
 692 an EyeLink II eye tracker. The tracking mode was pupil  
 693 only, with a sample rate of 500 Hz.

694 As in the NIMBLE face recognition experiments  
 695 described above, the human experiments had a study  
 696 phase and a test phase and used the same stimuli as the  
 697 NIMBLE experiments. During the study phase, partic-  
 698 ipants saw the 32 faces, one at a time, for 3 s in a random  
 699 order. During the test phase, they saw the same 32 faces as  
 700 well as 32 lures, one at a time. For each test stimulus, they  
 701 were asked to judge, as quickly and accurately as possible,  
 702 whether or not they had seen the face during the study  
 703 phase by pressing “YES” or “NO” buttons within 3 s. The  
 704 faces were divided into the four fixation restriction  
 705 conditions evenly, counterbalanced through a Latin square  
 706 design: one fixation, two fixations, three fixations, and no  
 707 restriction (free viewing). In each trial, the image  
 708 remained on the screen until either the participant’s eyes  
 709 moved away from the last permissible fixation (if a  
 710 restriction was imposed), or the response occurred, or  
 711 the end of 3 s, whichever came first. The image was  
 712 masked by the average face (an image created using the  
 713 pixel-wise average of all of the stimuli from the study)  
 714 after the last permissible fixation; the mask remained until  
 715 the response or the end of 3 s. The participants were not  
 716 informed of the association between the mask and the  
 717 number of fixations they made during the experiment. One

718 major difference from the NIMBLE experiments is that  
 719 the human subjects had to saccade to the face after it  
 720 appeared above or below a fixation point. Initially, the  
 721 average face appeared, then during the subjects’ first  
 722 saccade it was replaced by the test face. This method was  
 723 used to ensure that the subjects were unable to extract any  
 724 identity information from peripheral vision. NIMBLE, on  
 725 the other hand, is given the entire face to start with and  
 726 computes salient points directly on the facial image.

727 **Figure 3** shows the distribution of the first three  
 728 fixations and overall fixations from all subjects during  
 729 the study and the test phases. The participants’ discrim-  
 730 ination performance as measured by  $A'$ , a bias-free  
 731 nonparametric measure of sensitivity that estimates ROC  
 732 area, showed that the optimal human recognition perform-  
 733 ance was achieved with two fixations—performance did  
 734 not improve with additional fixations. This is illustrated in  
 735 **Figure 4**: The  $A'$  in the two-fixation condition was  
 736 significantly larger than that in the one-fixation condition  
 737 ( $F(1, 15) = 44.435, p < 0.001$ ); in contrast,  $A'$  in the two,  
 738 three, and no restriction (4+ fixations) conditions were not  
 739 significantly different from each other (there were no  
 740 statistically significant differences between any two of the  
 741 three).  
 742

#### 743 **NIMBLE face recognition using human fixations**

744 In the human memory experiment just described (Hsiao & Cottrell, [in press](#)), we recorded the eye movements of  
 745

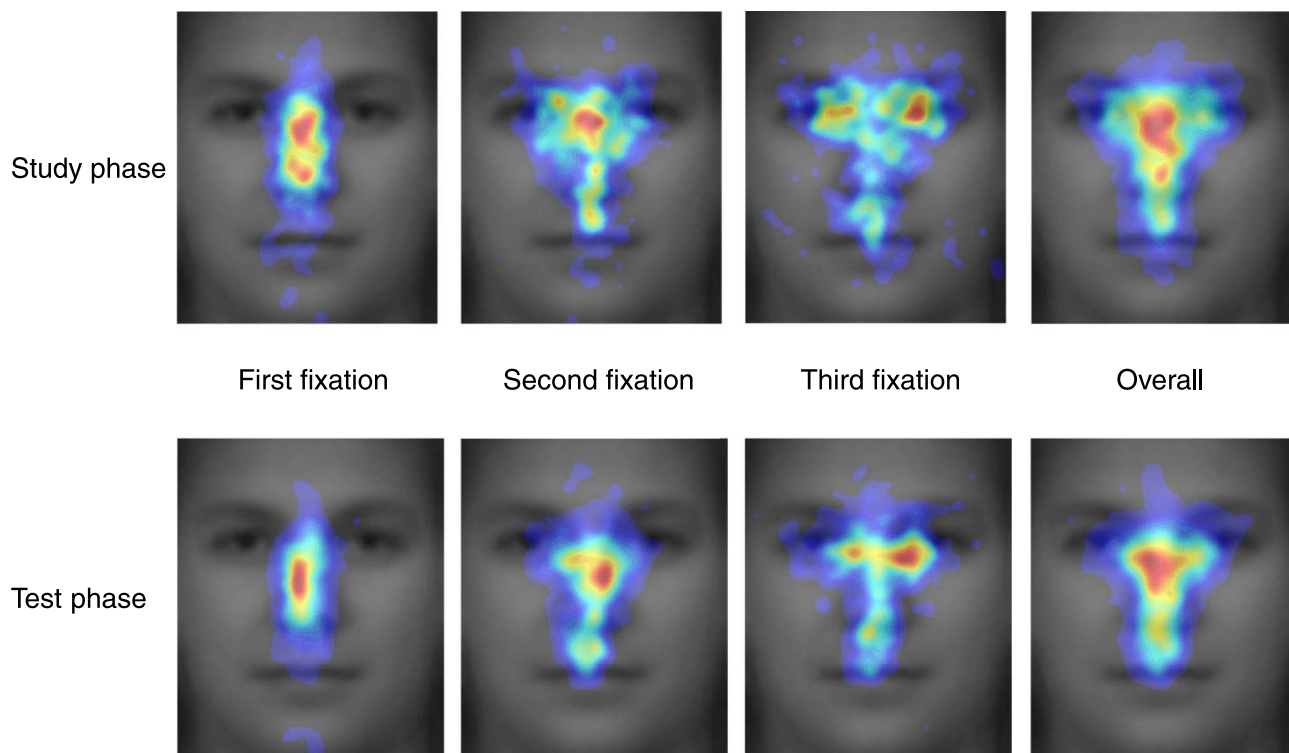


Figure 3. Distribution of the first three fixations in all trials and overall fixations from all subjects during the study and the test phases (from Hsiao & Cottrell, [in press](#)).

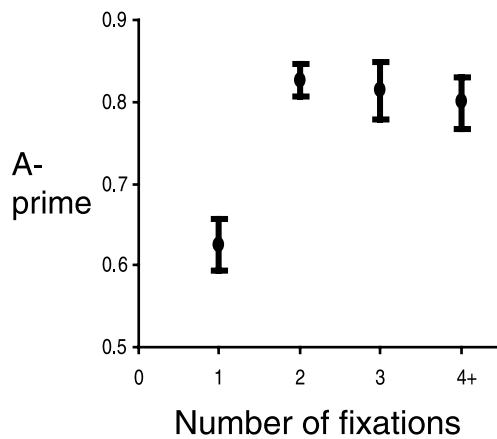


Figure 4. Human participants' discrimination performance measured by  $A'$  in different fixation restriction conditions: one fixation, two fixations, three fixations, and the free viewing condition (4+). Error bars show standard errors.

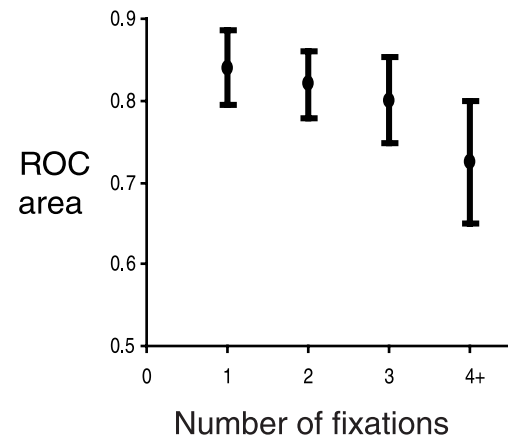


Figure 5. NIMBLE's discrimination performance using human fixations, measured by ROC area in different fixation restriction conditions: one fixation, two fixations, three fixations, and the free viewing condition (4+ fixations). Error bars show standard errors.

16 subjects as they performed both the study and test phases of a face recognition memory task. It is clear that the distribution of human fixations (Figure 3) differs significantly from the fixation locations chosen using NIMBLE's current model of visual salience (demonstrated in Figure 1). These differences most likely arise from two sources: 1) The human subjects start off the face and saccade to it, making the center of the face a more likely target; and 2) the simplified salience model (Equation 1) of Yamada and Cottrell (1995) takes no account of the top-down, task-specific cues used in directing human saccades. By using the human fixation locations recorded from the eye tracking experiment described above, rather than fixations drawn from a computed salience map, we can test the fragment memory component of NIMBLE in isolation from the salience model.

Using the same images and the same test paradigm as Hsiao and Cottrell (in press), we use the recorded human fixation locations as the basis for NIMBLE's fragment selection. During the study phase, we extract an image fragment from each of the locations where a human subject fixated a study image and store these fragments in NIMBLE's memory. During the test phase, we extract fragments from the locations in the target and lure images that subjects fixated during the test phase under each of the 4 fixation conditions and compute the likelihood of these fragments under NIMBLE's kernel density estimate of the target and lure classes. We calculate NIMBLE's ROC curves as before, this time using the fixation locations from each human subject, and compare the performance in each of the four fixation conditions. The results are shown in Figure 5, where we use a Gaussian kernel with  $\sigma = 0.1$  to best fit the human data. (Note that the results for the original NIMBLE face recognition experiments in Table 2 are very insensitive to the value of this parameter, and setting  $\sigma = 0.1$  with the computed salience map provides similar results to those shown in Table 2.)

For comparison with Figure 4 (which shows human subjects' performance) and Figure 5 (which shows the performance of the NIMBLE model using actual human fixations), we have included Figure 6, which shows the NIMBLE model's performance using one, two, three, and the mean of four to ten fixations that were selected using the visual salience model (Equation 1) of Yamada and Cottrell (1995).

Figures 4 and 5 demonstrate that with limited fixations, NIMBLE using human fixations recognizes faces with similar accuracy to humans (with ROC area between 0.8 and 0.9). However, the two graphs also show that NIMBLE's trend in performance differs from humans. In particular, NIMBLE achieves maximum performance using just the first human fixation. Note that this is not the case when NIMBLE chooses fixations using the model

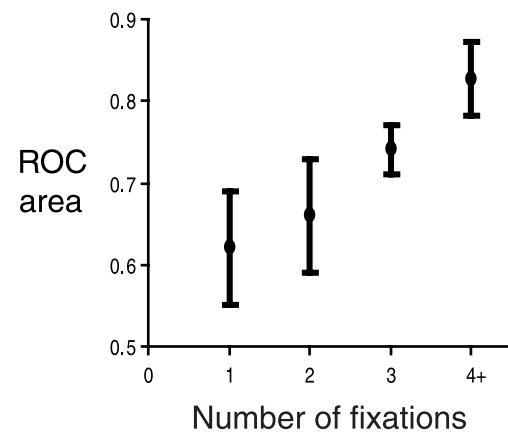


Figure 6. NIMBLE's discrimination performance using the model of visual salience, measured by ROC area in different fixation restriction conditions: one fixation, two fixations, three fixations, and the average of 4 to 10 fixations (4+). Error bars show standard errors.

of visual salience (Equation 1), as shown in Figure 6. This is consistent with the claim that the location chosen by humans for the first fixation (the bridge of the nose) may be the optimal viewing position for face recognition (Hsiao & Cottrell, *in press*).

A possible explanation for the discrepancy between the human results in Figure 4 and NIMBLE's performance using the same fixations in Figure 5 is that, as shown in Figure 3, the first and second fixations made by the human subjects tended to land in very similar locations (around the bridge of the nose), while it was not until the third fixation that attention was directed to the eyes. This suggests that all of the information required for face recognition in this task was obtainable by fixating the bridge of the nose, but that perhaps the subjects were not able to obtain all of the information required for face recognition during the duration of a typical fixation. Since we move our eyes about three times per second (Henderson, 2003; in the human experiment described above, the first fixation lasted 295 ms and the second fixation lasted 315 ms on average), it may be that a second fixation in a nearby location is required to accumulate more information and thus achieves the best face recognition performance. The need for a second fixation nearby the first may also be due to task switching from localizing to exploring for recognition. This task-switching effect could result from subjects planning a localizing saccade from the center of the screen to the target face stimulus before the first fixation; this localizing saccade has been shown to have a central bias (e.g., Renninger, Verghese, & Coughlan, 2007; Tatler, 2007). Obviously, the NIMBLE model does not experience this limitation.

While the performance on the second and third fixations are almost identical for human subjects and NIMBLE with human fixations, the performance of NIMBLE using human fixations degrades in the unrestricted viewing condition (4+ fixations, in which the model used up to 3 s of recorded human fixations on each image). This drop in performance may be because most human subjects had recorded enough information to match the study face using just the first two fixations, as evidenced by the plateau in their results in Figure 4. After this, subjects could continue to actively fixate the image or simply let their gaze wander while they considered and entered their response. Thus, perhaps many of the fixations recorded by the eye tracker in this condition were not actively being used by the subjects to make their decision (were not task-relevant), which could explain why including them hampered NIMBLE's performance.

## Discussion

Our Bayesian version of NIM provides several extensions to the original cognitive model (Lacroix et al.,

2006). First, we recast NIM into a probabilistic framework, placing it on a firmer theoretical foundation. In this setting, NIM's spherical kernel becomes the likelihood function (a density estimator). We showed how this could be replaced by a more robust Gaussian kernel or by a nearest-neighbor density estimator. The resulting NIMBLE model demonstrates good performance on identification tasks for both faces and objects. We also showed that in face identification, the model's belief in the correct answer increases in a reasonable way as more evidence is gathered through further fixations, a property not enjoyed by the original NIM model's method of evidence combination.

In addition, we cast the problem of memory for faces as a classification problem, where we assume that faces from the study set are encoded with the context of the experiment (here, simply labeled as being from the study set). Our NIMBLE model demonstrates human levels of performance on a facial memory (familiar/unfamiliar) task. However, a more detailed examination of the relationship between NIMBLE and human processing revealed differences in both the apparent salience map, which is the front end of our model, and the memory component of the model when human fixations are used. These differences will require further refinements to the model in order to better match the human data.

One obvious difference is the methodology used in the two cases. Whereas the human subjects begin the experiment looking at a screen location off the face and saccade to the face, NIMBLE starts by computing a complete salience map from the entire full-resolution face image. A more reasonable paradigm might be for NIMBLE to also start off the face and compute the salience of the image using only the low-resolution information about the face that would be available from an object in the periphery. Clearly, the model would find the face salient and saccade to it. Before the first saccade, the representation of the face would be low resolution because it is out of the fovea, and so we hypothesize that the salience would be maximal near the center of the face (corresponding to human subjects first saccading to near the bridge of the nose). This idea also suggests potential future research to develop a more realistic salience model that computes salience based only on the currently available image information, rather than one computed on the entire image in parallel in advance (as is currently done).

The first-fixation performance discrepancy between the human subjects and NIMBLE may be due to the humans' imperfect perception of the studied face during a single fixation (the duration of the first fixation is 295 ms on average in the human experiment). Perhaps the current model moves its beliefs too far based on the current input, resulting in the high accuracy after the first fixation. Humans appear to be more "skeptical" in their updating until they have a second fixation. We could account for this using a generative model that is a mixture of noise

912 and actual beliefs, as was done by Nelson and Cottrell  
913 (2007) to model the gradual acquisition of concepts in a  
914 learning task.

915 Alternatively, the discrepancy may also be due to task  
916 switching from localizing to exploring for recognition that  
917 occurs during the first fixation. Subjects begin by planning  
918 a localizing saccade from the center of the screen to the  
919 center of the stimulus (Renninger et al., 2007) and may  
920 not take in all of the information from this first fixation  
921 location that is required for recognition. Future work  
922 could test this hypothesis by instructing and/or restricting  
923 participants to make only one fixation during the face  
924 recognition task and increasing the preview time of the  
925 average face in the periphery to reduce the task-switching  
926 cost. Based on NIMBLE's results, our hypothesis is that  
927 this will improve the human subjects' performance in the  
928 one-fixation condition.

929 Similar to what we have proposed here, Lacroix et al.  
930 (2007) have developed an extension of their original  
931 NIM model, called NIM-CLASS, which can be applied  
932 to multi-class identification tasks by storing a class label  
933 with each memorized fragment. They assign a label to a  
934 new fragment by finding the peak in the histogram of its  
935 nearest neighbors' labels. However, the model is still  
936 heuristic rather than probabilistic. A further extension in  
937 NIM-CLASS that is beyond what we have done here is  
938 to include the influence of top-down attention by also  
939 storing the image coordinates of each memorized frag-  
940 ment as well as a low-resolution, global representation  
941 of the image that captures the gist of the image. Using  
942 this global representation, they direct future fixations to  
943 the area in the image that is most likely to reduce  
944 classification uncertainty. The computational complex-  
945 ity of this step makes it somewhat implausible for  
946 fixation planning. Bayesian methods exist for augment-  
947 ing the salience calculation by including scene con-  
948 text (e.g., priors on target locations; Torralba, Oliva,  
949 Castelhana, & Henderson, 2006) and top-down influences  
950 (e.g., the likelihood of the features given the class;  
951 Zhang, Tong, Marks, Shan, & Cottrell, *in press*). Such  
952 models could easily be integrated into NIMBLE's  
953 probabilistic framework.

954 Future improvements to NIMBLE will include more  
955 detailed modeling of the retinal and lateral geniculate  
956 nucleus (LGN) transformation used to convert fixated  
957 image locations into features for the memory model. This  
958 will move from the oversimplified square patch samples  
959 of the Gabor filter responses to a properly foveated retina  
960 model with lower resolution samples from the periphery.  
961 In addition, we plan to improve the selection of fixation  
962 points by integrating learned, task-specific feedback to  
963 direct NIMBLE's "eye" movements to sample from image  
964 locations with top-down interest as well as bottom-up  
965 salience. Since NIMBLE is a fully probabilistic model, it  
966 will be straightforward to integrate these more complex  
967 systems into the existing model.  
968

## Conclusions

969  
970

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's  
performance is comparable to human performance on  
standard identification and recognition memory tasks and  
that this biologically inspired model approaches the best  
machine vision results. In addition, the sequential sam-  
pling version of NIMBLE demonstrates that, like humans,  
our system can achieve correct identification and recog-  
nition of faces and objects after a very small number of  
fixations. Implementing NIMBLE using actual human  
fixations improves performance considerably, suggesting  
that in a face recognition memory task, humans fixate  
locations that are optimally suited to solving the problem.

986  
987

## Acknowledgments

This work is supported by NIMH Grant MH57075 to  
GWC, a James S. McDonnell Foundation Grant to the  
Perceptual Expertise Network (Isabel Gauthier, PI), and  
NSF Grant #SBE-0542013 to the Temporal Dynamics of  
Learning Center (GWC, PI). LB was supported by an  
IGERT fellowship (NSF DGE-0333451 to GWC).

A002  
995

Commercial relationships: none.  
Corresponding author:  
Email:  
Address:

996  
997

## References

998  
999

- Belongie, S., Malik, J., & Puzicha, J. (2002). Shape  
matching and object recognition using shape contexts.  
*Pattern Analysis and Machine Intelligence*, 24-4,  
509–522. 1000  
1001  
1002  
1003
- Bishop, C. (1995). *Neural networks for pattern recogni-  
tion*. Oxford University Press. 1005  
A003
- Dailey, M., Cottrell, G., & Busey, T. (1998). Facial  
memory is kernel density estimation (almost). In  
M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.),  
*Advances in neural information processing systems*  
(vol. 11, pp. 24–30). Cambridge, MA: MIT Press. 1008  
1009  
1010  
1011  
1012
- Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R.  
(2002). EMPATH: A neural network that categorizes  
facial expressions. *Journal of Cognitive Neuro-  
science*, 14, 1158–1173. [PubMed] 1014  
1015  
1016  
1018
- Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005).  
Behavior recognition via sparse spatio-temporal 1019  
1020

- 1021 features. *Proceedings of the Joint IEEE Workshop on*  
 1022 *Visual Surveillance & Performance Evaluation of*  
 1023 *Tracking & Surveillance.*
- 1025 Duchaine, B., & Nakayama, K. (2005). Dissociations of  
 1026 face and object recognition in developmental prosopagnosia. *Journal of Cognitive Neuroscience*, *17*,  
 1027 249–261. [PubMed]
- 1030 Foulsham, T., & Underwood, G. (2008). What can saliency  
 1031 models predict about eye movements? Spatial and  
 1032 sequential aspects of fixations during encoding and  
 1033 recognition. *Journal of Vision*, *8*(2):6, 1–17, [http://](http://journalofvision.org/8/2/6/)  
 1034 [journalofvision.org/8/2/6/](http://journalofvision.org/8/2/6/), doi:10.1167/8.2.6.  
 1035 [PubMed] [Article]
- 1037 Henderson, J. M. (2003). Human gaze control during real-  
 1038 world scene perception. *Trends in Cognitive Science*,  
 1039 *7*, 498–504. [PubMed]
- 1041 Henderson, J. M., Williams, C. C., & Falk, R. J. (2005).  
 1042 Eye movements are functional during face learning.  
 1043 *Memory & Cognition*, *33*, 98–106. [PubMed]
- 1045 Hintzman, D. (1984). MINERVA 2: A simulation model  
 1046 of human memory. *Behavior Research Methods,*  
 1047 *Instruments and Computers*, *16*, 96–101.
- 1049 Hsiao, J., & Cottrell, G. (in press). Two fixations suffice in  
 1050 face recognition. *Psychological Science.*
- 1052 Itti, L., & Koch, C. (2001). Computational modelling of  
 1053 visual attention. *Nature Reviews, Neuroscience*, *2*,  
 1054 194–203. [PubMed]
- 1055 Jones, J. P., & Palmer, L. A. (1987). An evaluation of the  
 1056 two-dimensional Gabor filter model of simple recep-  
 1057 tive fields in cat striate cortex. *Journal of Neuro-*  
 1058 *physiology*, *58*, 1233–1258. [PubMed]
- 1059 Kittler, J., Hatef, M., Duin, R., & Matas, J. (1998). On  
 1060 combining classifiers. *IEEE Transactions on Pattern*  
 1061 *Analysis and Machine Intelligence*, *20*, 226–239.
- 1062 Lacroix, J., Murre, J., & Postma, E. (2006). Modeling  
 1063 recognition memory using the similarity structure of  
 1064 natural input. *Cognitive Science*, *30*, 121–145.
- 1065 Lacroix, J., Murre, J., Postma, E., & Van den Herik, H.  
 1066 (2004). The natural input memory model. *Proceed-*  
 1067 *ings of the 26th Annual Meeting of the Cognitive*  
 1068 *Science Society.*
- 1069 Lacroix, J., Postma, E., & Van den Herik, H. (2007).  
 1070 Modeling visual classification using bottom-up and  
 1071 top-down fixation selection. *Proceedings of the 29th*  
 1072 *Annual Meeting of the Cognitive Science Society.*
- 1073 Lowe, D. (2004). Distinctive image features from scale-  
 1074 invariant keypoints. *International Journal of Com-*  
 1075 *puter Vision*, *20*, 91–110.
- 1076 Mozer, M., Shettel, M., & Vecera, S. (2006). Top-down  
 1077 control of visual attention: A rational account. In  
 1078 Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in*  
 1079 *neural information processing systems*, (vol. 18,  
 1080 pp. 923–930). Cambridge, MA: MIT Press.
- Nelson, J., & Cottrell, G. (2007). A probabilistic model of  
 eye movements in concept formation. *Neurocomput-*  
*ing*, *70*, 2256–2272.
- Nene, S., Nayar, S. K., & Murase, H. (1996). Columbia  
 object image library (COIL-100) (Tech. Rep.).
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-  
 based random walk model of speeded classification.  
*Psychological Review*, *104*, 266–300. [PubMed]
- Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye  
 movements while viewing and recognizing patterns.  
*Vision Research*, *11*, 929–942. [PubMed]
- Palmeri, T. J., & Gauthier, I. (2004). Visual object  
 understanding. *Nature Reviews, Neuroscience*, *5*,  
 291–303. [PubMed]
- Phillips, J., Wechsler, H., Huang, J., & Rauss, P. (1998).  
 The FERET database and evaluation procedure for  
 face-recognition algorithms. *Image & Vision Comput-*  
*ing*, *16*, 295–306.
- Raaijmakers, J., & Shiffrin, R. (2002). Models of memory.  
 In H. Pashler & D. Medin (Eds.), *Stevens' handbook*  
*of experimental psychology* (3rd ed., vol. 2), *Memory*  
*and cognitive processes* (pp. 43–76). Wiley.
- Raj, R., Geisler, W. S., Frazor, R. A., & Bovik, A. C.  
 (2005). Contrast statistics for foveated visual systems:  
 Fixation selection by minimizing contrast entropy.  
*Journal of the Optical Society of America A, Optics,*  
*Image Science, and Vision*, *22*, 2039–2049. [PubMed]
- Renninger, L., Coughlan, J., Verghese, P., & Malik, J.  
 (2005). An information maximization model of eye  
 movements. In L. K. Saul, Y. Weiss, & L. Bottou  
 (Eds.), *Advances in neural information processing*  
*systems*, (vol. 17, pp. 1121–1128). Cambridge, MA:  
 MIT Press.
- Renninger, L. W., Verghese, P., & Coughlan, J. (2007).  
 Where to look next? Eye movements reduce local  
 uncertainty. *Journal of Vision*, *7*(3):6, 1–17, [http://](http://journalofvision.org/7/3/6/)  
[journalofvision.org/7/3/6/](http://journalofvision.org/7/3/6/), doi:10.1167/7.3.6.  
 [PubMed] [Article]
- Tatler, B. W. (2007). The central fixation bias in scene  
 viewing: Selecting an optimal viewing position  
 independently of motor biases and image feature  
 distributions. *Journal of Vision*, *7*(14):4, 1–17, [http://](http://journalofvision.org/7/14/4/)  
[journalofvision.org/7/14/4/](http://journalofvision.org/7/14/4/), doi:10.1167/7.14.4.  
 [PubMed] [Article]
- Torralba, A., Oliva, A., Castelano, M. S., & Henderson,  
 J. M. (2006). Contextual guidance of eye movements  
 and attention in real-world scenes: The role of global  
 features in object search. *Psychological Review*, *113*,  
 766–786. [PubMed]
- Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual  
 features of intermediate complexity and their use in  
 classification. *Nature Neuroscience*, *5*, 682–687.  
 [PubMed]

- 1137 Wolfe, J. (1994). Guided search 2.0: A revised model of  
1138 visual search. *Psychonomic Bulletin & Review*, *1*,  
1139 202–238.
- 1140 Yamada, K., & Cottrell, G. (1995). A model of scan paths  
1141 applied to face recognition. *Proceedings of the 17th*  
1142 *Annual Cognitive Science Conference*, 55–60.
- 1143 Yarbus, A. (1967). *Eye movements and vision*. New York:  
1144 Plenum Press.
- 1145 Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D.  
1146 (2006). The role of top-down and bottom-up  
1154 processes in guiding eye movements during visual  
search. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.),  
*Advances in neural information processing systems*  
(vol. 18, pp. 1569–1576). Cambridge, MA: MIT Press.
- Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G.  
(in press). SUN: A Bayesian framework for saliency  
using natural statistics. *Journal of Vision*. **AQ5**

## **AUTHOR QUERIES**

**AUTHOR PLEASE ANSWER ALL QUERIES**

AQ1: Please provide Web address, if any.

AQ2: Please provide data for author's correspondence.

AQ3: Please provide publication location here.

AQ4: Please update the publication status of Hsiao and Cottrell, in press.

AQ5: Please update the publication status of Zhang et al., in press.

**END OF AUTHOR QUERIES**