Automatic Visual Integration: Defragmenting the Face

Luke Barrington (lbarrington@ucsd.edu)
Electrical & Computer Engineering Department, University of California, San Diego
La Jolla, CA 92093 USA

Garrison W. Cottrell (gary@cs.ucsd.edu)
Computer Science & Engineering Department, University of California, San Diego
La Jolla, CA 92093 USA

Abstract

We describe a computational model of saccadic visual memory applied to the task of facial recall. Each saccade provides a mixed-resolution, quasi-stationary snapshot of a visual scene to the striate cortical areas yet the brain recreates spatially and temporally smooth perceptions and memories by combining these visual fragments. We build on the work of (Lacroix et al., 2006) to model this integration of saccades in a facial recall task by selecting fixation points based on low-level saliency of a face image. At each fixation point, this model stores discrete visual samples with multiple resolutions as activation patterns without knowledge of their temporal or spatial origin to create a kernel density estimate of the studied faces. These visual fragments are then integrated and compared to new fixations during recall. We replicate Lacroix's results by demonstrating that the model achieves human levels of performance on the standard psychological memory test of facial recall. We then extend the model to facial identity recognition and examine the task-dependent effects of visual resolution.

Visual Integration

Visual perception begins with retinal sampling information of localized areas of a scene before our eyes. These saccades are discrete in space and time. High-resolution visual information is only available from the fovea, covering only a small portion of the observed scene. In order to perceive all parts of a visual scene with great detail as well as to maintain neural activation in the visual cortex, we repeatedly foveate different areas of the scene, spending the most time fixating on the parts that are most salient or most task-relevant (Yarbus 1967).

It seems reasonable to suppose that the sequences of saccades or *scan paths* that collect the visual data do not, in general, follow exactly the same path and thus can not reconstruct the exact sequence of stored visual memories when examining a previously-viewed object (although the distribution of saccades tends to be similar, see Henderson, Williams, & Falk, 2005). Yet we can mentally comprehend and recreate spatially and temporally complex constructs using only a combination of these static, non-uniform retinal samples stored in memory. How is this discrete information integrated into a continuous, dynamic construct?

The solution to this visual jigsaw puzzle must come from the way that retinal frames are stored in and retrieved from visual memory. Since snapshots of the environment are unlikely to be repeated, a straightforward template matching procedure is unlikely to work. For a simple task like face recognition, it is therefore important that there be a matching procedure that probabilistically assesses how well samples drawn from the current environment could have been generated by previous observations. Here we describe how the model developed by Lacroix et al. (2006) can be viewed as a kernel density estimate of the likelihood that new visual patches of faces were generated by our memory. We then show how we can use the same model for face identification. Finally, we explore the issue of the scale-space representation used and show how different spatial frequencies affect the matching process.

We begin by presenting an improvement to the model of (Lacroix et al., 2006) that uses a saccade selection routine that uses the same filters that are used for the memory representation. In Experiment 1, we demonstrate that this model can solve the challenge of visual integration by modeling the facial recall task. Experiment 2 extends the model to perform a new task; recognition of facial identity. Finally, Experiment 3 examines the effect of allowing the model to process Gabor filter bands separately and indicates a potential method for task-based control of visual attention.

Visual Memory Model

Saccade 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-dependant directives such as attention and expectation. Though many methods (Mozer, Shettel & Vecera 2005; Wolfe, 1994; Zelinsky et. al 2005) have been proposed for how to integrate these cues, in this work we concentrate only on bottom-up salience of static images. We model the saccade selection process using an interest operator for determining the scan paths introduced in (Yamada & Cottrell, 1995). This simplified model uses the rotational variance of low-resolution Gabor filter responses to construct a distribution of the contour complexity (read: *salience*) over all pixels in a given image:

$$Salience\ (i,j) = \frac{\sum_{n=1}^{8}\left(G\left(i,j,\frac{\pi n}{8}\right) - \mu_G\right)^2}{8}$$

where $G(i, j, \theta)$ is the response of a Gabor filter with orientation θ, centered at pixel (i, j) and μ_G is the mean response across all orientations. A similar technique developed by (Renninger et al, 2004) uses entropy rather than variance of local image contours to define salience.

We convert this salience map into a probability distribution by normalizing with the softmax function (Bishop, 1995). A fixation point is then chosen randomly, according to this distribution. We relax the salience around the fixated point by subtracting a univariate Gaussian, centered at the point from the salience distribution and renormalizing. This inhibits repeated fixations at the same location. Figure 1 shows a salience map generated in this manner as well as a sample distribution of fixation points.

This purely bottom-up model is simple but the resulting scan paths for face recognition task qualitatively approximate those observed in humans (Yamada & Cottrell, 1995). The model satisfies 3 of the 5 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. We ignore the other 2 criteria that concern the top-down influence of attention and object recognition on fixation point selection. Future work intends to augment this model by extending the results of (Nelson & Cottrell, 2005) to use top-down feedback to direct the selection of eye-movements by examining which queries (i.e. eye movements) would be most useful in enhancing performance of the current visual task.

Retinal / Cortical Image Transform

Once it has been fixated, an input image undergoes many stages of neural processing before being stored as a pattern of activation in high-level visual cortex. In all experiments below, we use as input 128x192 pixel grayscale image from the FERET database of (Phillips et. al 1998). 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.

Gabor filter responses at 8 orientations and 4 frequencies form our biologically-motivated, V1 processing model (Jones & Palmer, 1987). We transform an image into the Gabor-filtered domain by calculating the response of each filter at every image pixel. In these experiments, we use Gabor filter frequencies of 1/4, 1/8, 1/12 and 1/16 cycles/radian.

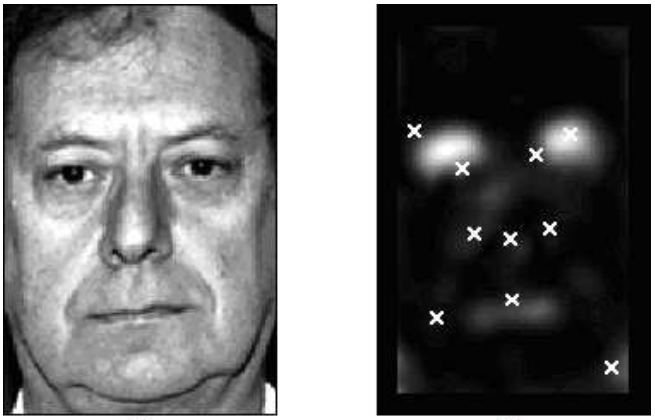

Figure 1(a): An image from the FERET database. Figure 1(b): The corresponding salience map generated using the technique of (Yamada & Cottrell, 1997) with a sample distribution of ten fixation points. Fixations tend to cluster around highly salient areas but relaxation of sampled points enforces an even distribution across the image.

The highest spatial-frequency filter responses correspond to the high-resolution foveated area around the fixation point. The responses of the low-frequency filters are each computed from an area surrounding the fixated pixel that has spatial context greater than that of the foveated patch and thus provide extra-foveal information, corresponding to the low-resolution data from the retinal periphery. By extracting just a square patch from these Gabor response images, we are in fact producing a foveated representation of the fixated point.

The size of the stored 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 fixation patch size of 35x35 pixels (corresponding to a visual angle of 1.5° for a subject about 75cm from a 96 dpi computer monitor: an approximation of the conditions for human studies discussed below), the input feature vector to our model has;

35 x 35 x 8 orientations x 4 frequencies = 39200 dimensions

In order that the memory be able to generalize to recognize familiar faces under new conditions where different fixation points may be chosen and also so that it has the capacity to remember a large number of faces, the dimensionality of the input features is reduced while maintaining the majority of their representational ability by using principal component analysis. This is analogous to the concise encoding of the over-complete retinal and V1 data in higher levels of visual cortex. This feature extraction procedure of wavelet-based image decomposition followed by PCA is a standard

approximation for biologically motivated vision models (Dailey et al., 2002; Palmeri & Gauthier, 2004; Lacroix et al., 2006)

We chose to retain just the first 80 principal components in order to make the model tractable as well as biologically feasible while still accounting for 87% of the variance in the feature data when processing all frequency bands together. Experiment 3 will present an alternative treatment where we do individual PCA decompositions on each of the 4 frequency bands.

Memory storage and retrieval

Given the natural input patterns of activation or image fragments described above, the role of the brain is to analyze them and retrieve similar or related patterns. The nature of this analysis and the methods of storage and recall are the focus of our modeling work.

The memory model used is based on the Natural Input Memory from (Lacroix et. al, 2006). The "memory" in this case is a high-dimensional vector space and "memories" are vectors in this space. Given an input vector derived from an image as described above, the memory storage process is simply to assign this vector to the memory space. This approach of conceiving of memories as patterns of neural activation in a sparse vector space has been successfully applied in many domains of cognitive modeling such as (Nosofsky & Palmeri, 1997 - response-time modeling; Sagi et al., 2002 - speech processing and Dollar et al., 2004 - video behavior analysis), among others.

The retrieval process is instigated when a new sample is input to the system. The patterns of activation of this novel input are compared to all the stored instances in the memory vector space. As with the models in (Lacroix et. al, 2006; Nosofsky & Palmeri, 1997), proximity in this space is designed to relate to similarity in the perceived world. Multi-dimensional patterns can now be compared for similarity using simple vector-based methods such as Euclidean distance, suitable for comparing integral-dimension stimuli.

There is no guarantee that the fixation points chosen in the testing phase by our stochastic interest operator will exactly match those used in training; scan paths are not repeatable (Henderson, Williams, & Falk, 2005). Therefore (as with human vision), fixated samples will rarely be a perfect match for anything stored in memory and we must instead use a more tolerant metric. In order to judge the familiarity of an input pattern, the model uses a form of kernel density estimation (Bishop, 1995) a technique that has been applied by (Lacroix et al, 2006; Dailey, Cottrell & Busey, 1999) for facial memory modeling. For each M-dimensional fragment input to the system, we count the number of stored memories, n_f, that lie within an M-dimensional volume of radius r, centered on that input. r is a free parameter of the model which controls required distinctiveness for an input fragments, that is the strictness required for it to be judged a match. The *familiarity*, F_i, of an image i, is defined as the average number of memories matching each of the S fixated samples taken from the image;

$$F_i = \frac{1}{S}\sum_{f=1}^{S} n_{if} \qquad \forall \quad fragments, f \in image, i$$

As outlined in Bishop (1995), this is an (un-normalized) kernel density estimate of the probability that the new face was generated by the memory. For a subject to decide whether the image is familiar or not, they must threshold this probability. To visually assess the hypothesis that proximity in the memory vector space corresponds to similarity in the world, figure 2(c) shows the distance from a fragment taken from the image in figure 2(a) to all other possible fragments from that image while figure 2(d) shows the distance from the same fragment to all possible fragments from a different image, shown in figure 2(b). Given the strong peak in similarity around the fixation point, it can be seen that fragments that are close to the memorized fragment match very well. Moving away from this point, similarity drops off quickly so that even fragments from similar locations on the unfamiliar image do not respond strongly.

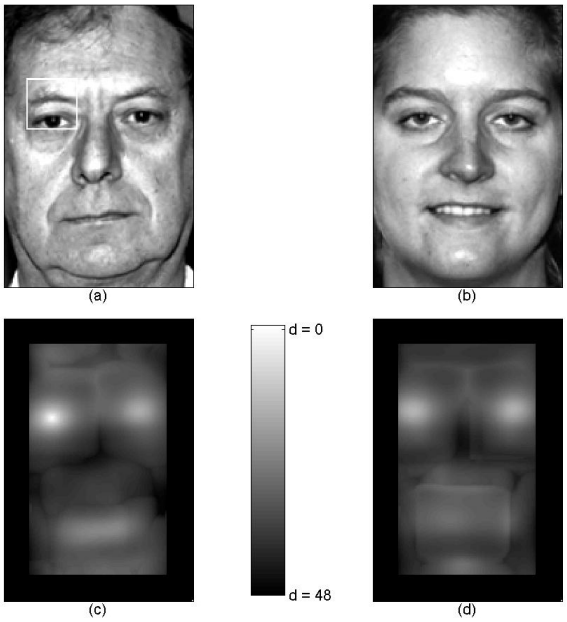

Figure 2: A target and lure image and their similarity maps. Figure 2(c) compares distances from the outlined patch of the face in figure 2(a) to all possible patches from the same image. Figure 2(d) compares the same patch to all possible patches from the face in figure 2(b).

Experiment 1 – Facial Recall

To confirm that our modifications of Lacroix's model still perform tasks similar to those she performed with it, we applied it to a face recognition experiment. In this paradigm the subject studies a sequence of N briefly presented faces. In the test phase, a second list of faces is shown where (typically) half are from the original studied list (targets) and half are unfamiliar distracters (lures). The goal is to test the subject's ability to recognize the studied faces (hits) without classifying the lures as familiar (false alarms).

In order to quantify the relationship between hits and false alarms, we look to signal detection theory and the *detectability index*, *d'* which compares the normalized familiarity scores for the target and lure images (Kay, 1998);

$$d' = \frac{\mu_{F_T} - \mu_{F_L}}{\sqrt{\frac{\sigma^2_{F_T} + \sigma^2_{F_L}}{2}}}$$

where μ_{F_T} and $\sigma^2_{F_T}$ (μ_{F_L} and $\sigma^2_{F_L}$) are the mean and variance of the familiarity scores for images from the target (lure) list. All results are the average of 10 trials.

To compare our model's performance with the results from human subjects, we examine the psychological data from (Lewis & Johnston, 1997). Note that in this experiment, subjects viewed ¾ profile faces whereas in our experiments, we use frontal views. Other human studies testing frontal views such as (Hancock, Burton & Bruce, 1996; O'Toole et al., 1994) used far larger test lists (174 faces) albeit with unconstrained study times and had correspondingly lower d' scores (average 1.37). The results of (Lewis and Johnston, 1997) do give us a baseline comparison level for human performance in a facial recall task. In their work, the study list had N = 20 images, each displayed for 5–10 seconds. Allowing for approximately 4 saccades per second, we allow our model to take 10–40 samples from the studied images. For each saccade, we sample 35x35 patches from the Gabor-filtered responses of the studied 128x192 pixel grayscale image from the FERET database. We then test on 40 images, including the 20 from the study set. Results are shown in Figure 3. These data replicate the human-like performance of the model found by (Lacroix et. al, 2006). These results make the intuitive prediction that as more samples are taken from the study and test lists (corresponding to longer viewing time by the human subjects), recall performance improves.

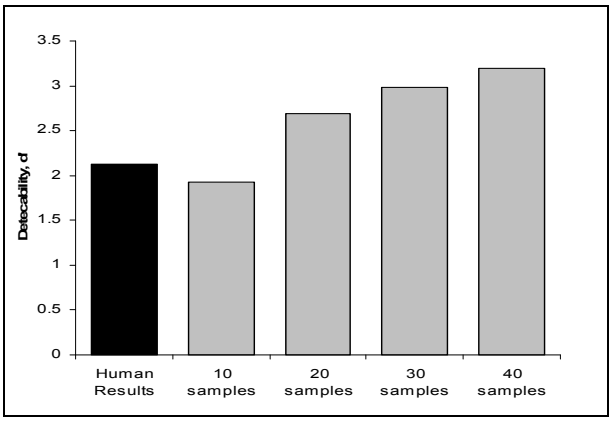

Figure 3: Recall performance for humans as reported by (Lewis & Johnston, 1997) and for our model

Experiment 2 - Identity recognition

The second memory task we tested our model on was identity recognition. In order to frame the task in the same terms as Experiment 1, we again present the model with a set of images for study and then examine its performance on a set of test images. In this paradigm, we can evaluate performance with the same familiarity and d' metrics as before. Here the study list is comprised of 6 different images of the same person (with changes in lighting and expression). Given a test list of 39 images containing 3 novel images of the studied face (the targets) as well as 9 different images of 4 other identities (the lures), the task now is to discriminate this original identity from the unfamiliar faces. For each image, we allow the model to make 20 fixations.

This task introduces an extra challenge for the kernel density estimate model in that it must now generalize to recognize images that it has never seen before. While the stochastic nature of the saccade selection model made it unlikely that same fragment would be examined in study and test in the recall task, this is impossible now. This further tests the capabilities of the kernel density method of memory modeling by requiring the model to integrate fragments from multiple images to form an estimate of the identity and to compare this density estimate to a completely unfamiliar set of fragments.

Figure 4 demonstrates that the model performs successfully on the identification task. The d' scores of around 2 indicate human-like performance. As expected, overall recognition performance was somewhat worse than for the simple recall task (compare the results for this experiment in Figure 4 to the middle bar of Figure 3). Again, reported results are the average of 10 repeated trials where, in each trial, variability arises from the fact that a different sample of saccades is taken from salience distribution and thus a different set of fragments are stored and analyzed for the study and test images.

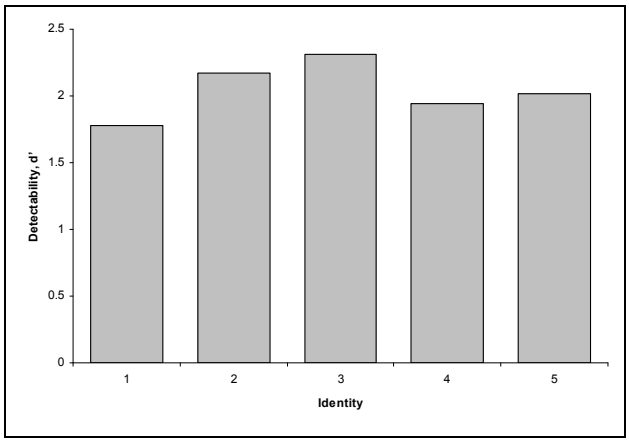

Figure 4: ID Performance (20 fragments per image, radius = 1.5)

Experiment 3 - Multi-resolution processing

Combining resolutions

The results presented above and in (Lacroix et. al, 2006) as well as many of previous models from our lab. (Dailey et. al, 2002) consider each level of resolution together by concatenating them into a single feature vector. Though it is the most straightforward method for processing image data, this approach disregards the fact that the scale most appropriate for a given visual object recognition task can vary greatly, depending on both the object and the task. As an example, low-resolution, peripheral vision might be enough to tell you that there is a page of text in front of you but would be useless for discerning the individual letters printed there. On the other hand, the fine-detailed discrepancies between a friend's face from one year to the next would not prevent you from recognizing her. Thus, it seems necessary to allow a memory model to have control over how it uses information of varying resolution, rather than forcing it to consider all scales as equal. Implementation of this idea can provide insight into how multi-resolution data is combined in visual recognition tasks and gives clues about how to apply task-based (top-down) controls on attention.

We examined methods for improving our model by allowing it to process each frequency band separately. The first observation is that the variance in the feature data is proportional to the frequency of the filter as shown in Table 1. While this may not be surprising of itself (higher frequency filters capture more rapidly varying data), it makes a strong argument for treating each frequency band individually. PCA allocates components in directions where the data has most variance, in this case, the high frequency bands. However, as shown in Figure 5, these are not the most useful bands for recall tasks. Therefore, treating all bands as equal and using a single PCA transform to represent them all will allocate more representational capacity to features that are less useful for the memory task[1].

Table 1: Variance of Gabor filter bands.

Gabor Filter Spatial Frequency	Kernel Radius (r)	Variance Accounted for by PCA to 20 Dimensions
0.0625	1.5	90.3%
0.0883	1.05	86.5%
0.125	0.6	76.1%
0.25	0.15	61.3%

Rather than coalesce all frequency bands into one feature vector, we have extended our model to process each band individually and calculate detectability based on average familiarity values. In this way, fragments which are ambiguously familiar at one scale ($F_{i,scale\ a} \approx 0$) but distinct at another ($F_{i,scale\ b} > 0$) can still be recalled.

The interesting result is that the effect of the 4 frequency bands in the identity task is different from the recall task. Figure 5 shows the detectability score for both tasks using just one frequency band (with 20 principal component features stored for each example). We see that, for recall, the lowest frequency was the most significant in detecting familiar faces, with a gradual fall off as the resolution is increased. Identity recognition relies more on the intermediate-low frequency (though the low is also important) but receives very little useful information from the highest frequency band. This makes the intuitive prediction that, for identifying familiar objects under novel conditions, excessive detail is in fact distracting.

A second result that springs from our multi-resolution analysis is that the distance between fragments (and therefore radius parameter of the kernel, inside of which they will be classified as familiar) changes with scales. This is illustrated in the center column of Table 1. These radii are roughly inversely proportional to the frequencies they correspond to (we found setting the high-frequency radii even lower improved results). This demonstrates that the criterion for successful fragment matches, distinctiveness, is coarser at low spatial frequencies due to the imprecise nature and low variability of these features than those required for features from higher frequency bands and adds weight to the argument for treating bands individually.

[1] We should note here that in previous models from our lab, the filter responses are whitened on a per-filter basis, so all bands have the same energy, and PCA only collects covariances.

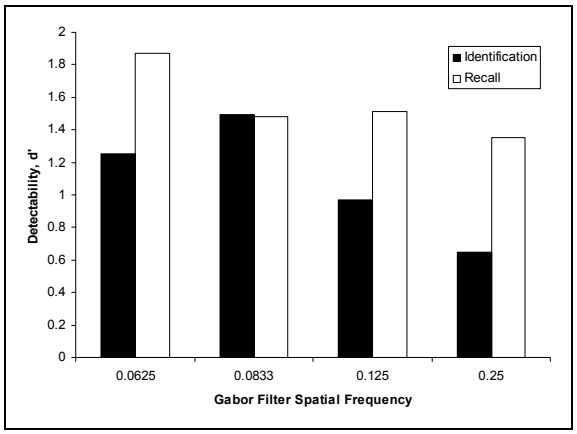

Figure 5: Per-band detectability for recall and identification tasks, from low to high frequency (left to right).

Discussion

We have introduced an extension of the facial recall memory model of (Lacroix et. al, 2006) and shown how it can also be applied successfully to the task of identity recognition. Using inspiration from neurobiology, this model is able to integrate a non-uniform sampling of a visual scene, potentially containing much novelty and without explicit knowledge of the spatial or temporal ordering of the samples it achieves human-levels of memory performance. This discrete sampling, concentrating on the salient parts of images could be the genesis for bottom-up, parts-based object representations where extracted fragments are stored, grouped and recalled according to their locations in our memory vector space.

By examining visual fragments at multiple scales, we have also demonstrated a possible method for implementing top-down, task-specific controls on familiarity. We have shown that the constraints imposed by fragment matches at one scale could be used to set expectations for matches at others dynamically. Our future work plans to incorporate these insights by developing a model that can learn task- and scale-specific match thresholds, corresponding to the versatile development of task-dependant perceptual expertise in humans.

Acknowledgments

This research project was supported by NIMH grant MH57075 to GWC.

References

Bishop, C. (1995) *Neural networks for Pattern Recognition.* Oxford University Press.

Dailey, M.N., Cottrell, G.W. & Busey, T.A. (1998) Facial Memory is Kernel Density Estimation (almost). *Proceedings Neural Information Processing Systems.*

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

Dollar, P., Rabaud, V., Cottrell, G. & Belongie, S. (2005) Behavior recognition via sparse spatio-temporal features. *Proceedings Joint IEEE Workshop on Visual Surveillance & Performance Evaluation of Tracking & Surveillance.*

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

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

Itti, L. & Koch, C. (2001) Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*, 2, No. 3, 194-203.

Jones, J.P. & Palmer, L.A. (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.

Kay, S.M. (1998) *Fundamentals of Statistical Signal Processing, Vol 2: Detection Theory.*

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

Lewis, M.B. & Johnston, R.A. (1997). Familiarity, target set and false positives in face recognition. *European Journal of Cognitive Psychology*, 9, 437-459.

Mozer, M.C., Shettel M. & Vecera S. (2005) Top-Down Control of Visual Attention- a Rational Account. *Neural Information Processing Systems 2005.*

Nelson, J.D. & Cottrell, G.W. (2005) A probabilistic model of eye movements in concept formation. *Neurocomputing*

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

O'Toole, A.J., Defffenbacher, K.a., Valentin, D. & Abdi, H. (1994) Structural aspects of face recognition and the other race effect. *Memory & Cognition,* 22, 208-224

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

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

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

Sagi, B., Nemat-Nasser, S.C., Kerr, R., Hayek, R., Downing, C. & Hecht-Nielsen, R. (2001). A biologically motivated solution to the cocktail party problem. *Neural Computation*, 13, 7, 1575-1602

Seiple, W., Holopigian, K., Szlyk, J.P., & Wu, C. (2004) Multidimensional visual field maps: Relationships among local psychophysical and local electrophysiological measures. *Journal of Rehabilitation Research and Development*, 41 (3a), 359–372

Wolfe, J.M. (1994) Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review*, 1, 2, 202-23

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

Yarbus, A.L. (1967). *Eye Movements and Vision*. Plenum Press, New York.

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