Age of Acquisition in Facial Identification: A Connectionist Approach

Brenden M. Lake (brenden@stanford.edu)
Torrey Pines High School, 710 Encinitas Blvd.
Encinitas, CA 92024 USA

Garrison W. Cottrell (gary@ucsd.edu)
UCSD Computer Science and Engineering, 9500 Gilman Dr.
La Jolla, CA 92093-0114 USA

Abstract

Age of Acquisition (AoA) is the phenomenon that acquiring a certain piece of information earlier than another results in a faster response time in adulthood. AoA has been shown to have a significant role in a variety of human studies. Recently, it has been demonstrated that connectionist networks that abstractly model reading and arbitrary mappings can also show AoA effects, and we extend this to facial identification. We present a connectionist model of facial identification that demonstrates strong AoA effects by allowing faces to be acquired in their natural order and by staging face presentation. This extends previous work by showing that a network that simply classifies its inputs also shows AoA effects. We manipulate the staged model in two ways, by either assuming outputs for the late set are trained to be off early in learning (or not) and by assuming the representation developed for the early set is used for the late set (or not). In three of these cases, we find strong AoA effects, and in the fourth, we find AoA effects with a recency control.

Introduction

The latency between a visual stimulus (an object, a face, or a printed word) and the correct response has been the subject of numerous studies. In an object-naming task, Oldfield & Wingfield (1964) reported naming latency to be dependent upon the frequency of the object's name in the word corpora. Carroll & White (1973) reanalyzed Oldfield & Wingfield's data, establishing that word frequency was not significant when AoA was accounted for. Since then, both AoA and frequency have drawn considerable attention. Despite many conflicting reports, it is generally concluded that both play significant roles.

In a celebrity face-naming study, Moore & Valentine (1998) showed independent effects of both AoA and frequency on naming latency. Facial distinctiveness, surname frequency, and number of phonemes in the celebrity's full name were not significant predictors of naming latency. Additionally, Moore & Valentine (1999) found AoA effects when frequency was not significant in a variety of tasks. They proposed, along with Morrison & Ellis (1995), that connectionist networks would be incapable of showing similar AoA effects due to "catastrophic interference" from more recently presented material.

Ellis & Lambon Ralph (2000) demonstrated AoA effects in a connectionist network that abstractly models reading. This task mapped patterns of random binary bits to similar patterns with some bits flipped. Since orthography and phonology are highly correlated but not identical (Jared, McRae, & Seidenberg, 1990), the model was designed to represent the inconsistencies in skilled reading. Ellis & Lambon Ralph were able to show AoA effects by training the network on some patterns initially and adding an additional set further into training. They found that the network is more "plastic" at the beginning of training, allowing the earlier patterns to have a larger impact on the weights. They also provide evidence that "catastrophic interference" does not occur unless the earlier patterns are completely removed from training.

In a similar task, Smith, Cottrell, & Anderson (2001) and Anderson & Cottrell (2001) demonstrated AoA effects by presenting all patterns at the beginning of training and measuring the naturally occurring AoA for each pattern. They found that earlier acquired patterns had a stronger correlation with the rest of the training set and had lower final errors than later acquired patterns. Smith et al. also found independent effects of frequency and AoA when the simulation was subjected to a frequency manipulation.

Zevin & Seidenberg (2002) did not find AoA effects in a connectionist network modeling reading with normal spelling-sound regularities. However, when they used early words that were completely unrelated to later words (like the arbitrary binary mappings of Ellis & Lambon Ralph and Smith et al.), AoA effects were shown. Zevin & Seidenberg claimed that AoA effects only exist if little information is carried over from the early patterns to the late patterns. Additionally, they suggested that learning the names associated with faces would produce genuine AoA effects, because no information would be shared between the stages of learning. Anderson and Cottrell (2004) have found in replicating their work that if one measures AoA in their simulations one still finds AoA effects. This is because words that they considered to be "late" by their manipulation were actually acquired early through generalization. Hence, we continue to use measured AoA in our first simulation below.

In this study, we provide a connectionist model of a facial identification task. Our model demonstrates strong AoA effects. We show that the model produces "natural" AoA effects (patterns that happen to be learned early show AoA effects), and we investigate the effects of staging the presentation of the faces. Unlike the reading simulations discussed above, in this case there is no opportunity for generalization to the late set early in training, as the network

is simply performing an identification task. Hence, we actually can "control" AoA.

Experiment 1: Natural AoA

Our first experiment models a facial identification task using a simple feedforward neural network. Neither frequency manipulation nor pattern staging is used in this experiment, allowing the faces to be acquired in their natural order. All of the faces in the training set are introduced at the beginning of training and are presented once per epoch. We carry out multiple replications on different networks to simulate replicating a human study over multiple subjects. The training process is identical for each replication, differing only on the initial random weights and the random presentation of the faces. We see if faces acquired earlier by the network have a lower Sum Squared Error, equivalent to adult naming latency (Ellis & Lambon Ralph, 2000; Zevin & Seidenberg, 2002), at the completion of training.

Model

Our model is a multilayer image classification system (Figure 1) that has been used in previous work (Dailey et al., 2002; Zhang & Cottrell, 2004). The raw images are first aligned. They are then filtered by 2-D Gabor filters, and Principal Component Analysis (PCA) is used to reduce the dimensionality of the filter responses. In the final stage, a backpropagation network learns to identify the faces.

Training Set The training set consists of 26 Caucasian individuals, 13 male and 13 female (Lee, 2004). There are 4 pictures of each person, all of frontal orientation. The 4 pictures consist of a happy expression, a sad expression, a neutral expression, and a neutral expression with his/her eyes looking 15 degrees to the right. Examples are shown in Figure 2. The raw images are first converted to grayscale from the original color. They are then rotated, scaled, translated, and cropped so that the eyes in every image are in the same location. The mouth is also aligned by the y-coordinate (height) value. The aligned images are 240 pixels wide and 292 pixels high, and the mean is set to zero with standard deviation one across the image spatially.

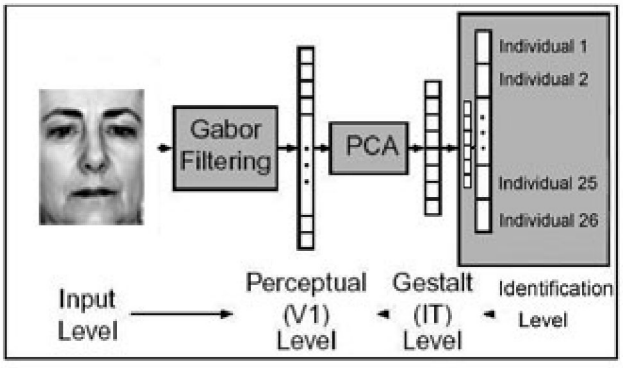

Figure 1: Our facial identification model, modified from Dailey et al. (2002).

Figure 2: Female and male subject with the following facial expressions: angry, neutral, happy, and eyes averted.

Perceptual Level As in previous models by our group, we model the early striate cortex processing via Gabor filters, which are 2-D sine and cosine waves restricted by a gaussian, resulting in "wavelets." One of these filters (either a sine or a cosine) corresponds well with simple cell processing (Daugman, 1985; Jones & Palmer, 1987). We use the magnitude of the sine/cosine pair, which provides a small amount of translation invariance, and hence is a simple model of complex cells in V1. We use a rigid 35 by 29 grid of overlapping 2-D Gabor filters (Daugman, 1985) in quadrature pairs at five scales and eight orientations (Dailey et al., 2002). We thus obtained 35 X 29 X 5 X 8 = 40,600 filter responses in this layer, which we call the perceptual layer (Dailey et al., 2002). We then z-score each filter response across the images on a per-filter basis (a local operation).

Gestalt Level We then perform Principal Component Analysis (PCA) on the Gabor filter responses from the last level. PCA reduces the dimensionality of the filter responses. Daily et al. (2002) suggest this is biologically plausible since it can be learned by neural networks using Hebbian learning. We project the Gabor filter responses down to 50 dimensions. Finally, each principal component is z-scored and multiplied by 0.8. Therefore, each component has a mean of 0 and a standard deviation of 0.8. This step normalizes the inputs for the tanh hidden units for more efficient learning (LeCun et al., 1998).

Identification Level Each person is assigned a number from 1 to 26. A value of 1 in the corresponding output unit with 0's for the rest indicate a correct response. We use a simple 50-20-26 feedforward network implementing the backpropagation learning algorithm. The tanh activation is used for the hidden layer and logistic activation for the output layer. We optimize the Sum Squared Error (SSE) criterion. For each epoch, each face pattern is presented once and the weights are updated accordingly. At the end of the epoch, the patterns are presented a second time, at which the errors are recorded, but no weight changes are made. These errors are then averaged across the 4 pictures for each person. If this average SSE drops below the threshold of 0.1, the person is said to be "acquired" by the network, and

the current epoch of training is recorded as the person's AoA. Therefore, AoA is determined for each person, not for each image. This is realistic because once someone is able to identify an individual, one can generally identify that person regardless of facial expression. At this point in training, the network will be very close to the correct response when identifying the acquired individual from a variety of images. We use a learning rate of 0.005 and momentum of 0.9. The weights are initialized randomly between -0.1 and 0.1 using a uniform distribution. The task is trained on 10 networks with different initial weights. Training lasts for 150 epochs, and patterns are presented in a different random order for each epoch.

Results

We find strong evidence for AoA effects in our model. All the individuals are acquired by the 150th epoch in the 10 runs. We correlate AoA with final SSE of each person for each individual run. We then average these correlation coefficients over the 10 runs. We find the epoch of acquisition to be strongly correlated with final SSE (average $r = 0.81$, $p < .001$ for all 10 runs). At the end of training, faces acquired earlier have lower errors than faces acquired later (Figure 3). This shows that faces that make a larger impact on the weights early in training maintain their advantage through extended training.

We then check to see if people are learned in a similar order in the 10 runs despite the random initial weights. We correlate the AoA of each person in one run with the AoA of each person in the other nine runs. We find that people are not learned in a similar order. On average, any two given runs yield a correlation coefficient of -0.0030 (average $p = 0.54$). This suggests that properties of the training set do not determine which faces are acquired first in our simulation.

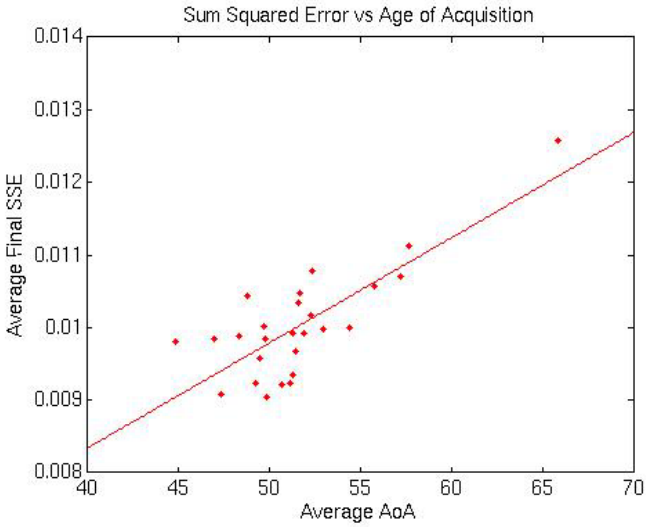

Figure 3: Relationship between final SSE and AoA for the 26 individuals averaged across the 10 runs in Experiment 1.

Experiment 2: Staged AoA

After demonstrating naturally occurring AoA effects, we introduce staging of face presentation into our model. Humans generally encounter new faces at different times throughout their lives. We investigate if staging can produce AoA effects in our connectionist model.

Model

In our simulation, we randomly divide our 26 individuals into two groups of 13 people each. One group is labeled the "early" group, and the other the group is labeled the "late" group. Different random groups are used for each run. Therefore, the 10 replications of the simulation differ in the initial random weights and in the distribution of the early and late groups. We train the network for 200 epochs using the same frequency distribution as Ellis and Lambon Ralph (2000) for our early and late set. The early set is presented once per epoch for the entire length of training. The late set is not presented to the network at all for the first 100 epochs, but it is presented twice per epoch for the remaining 100 epochs. Patterns are presented in a random order for each epoch and multiple presentations of a pattern are distributed randomly throughout the epoch as well. At the completion of training, the cumulative frequency of presentation is identical for all patterns. Otherwise, the model is identical to the one used in Experiment 1.

Results

All patterns are acquired in each run. For each run, we measure the average AoA for the early and late sets. We then perform a paired t-test comparing the difference in means between these averages across the 10 runs. We find a highly reliable difference, $t(9) = -106$, $p < .001$. When averaged across the 10 runs, the mean epoch for an early face to be acquired is 37 (sd = 2.6), and the mean epoch for a late face to be acquired is 133 (sd = 1.4). Therefore, the early set is acquired significantly before the late set, and our manipulation of AoA "worked."

Our network shows significant AoA effects. For each run, we measure the average final SSE for the early and late sets. We then perform a paired t-test comparing the difference in means between these averages across the 10 runs. We find a reliable difference, $t(9) = -2.7$, $p < .05$. The mean final SSE for the early set is 0.0071 (sd = 0.00035), and the mean final SSE for the late set is 0.0075 (sd = 0.00024). The early set has a significant advantage over the late set. This is plotted in Figure 4.

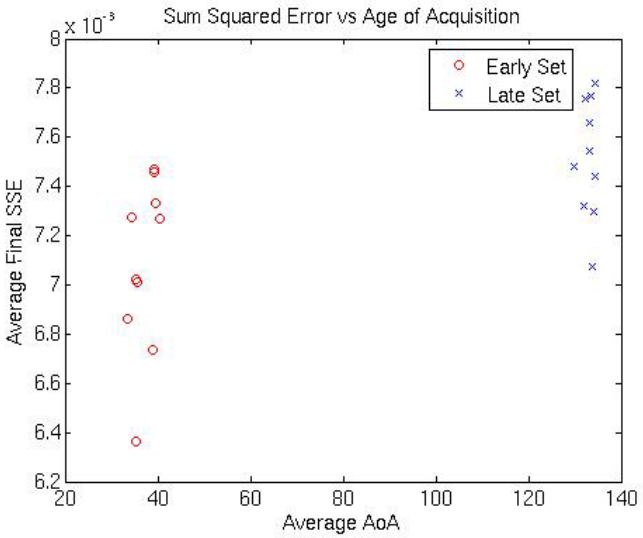

Figure 4: Relationship between average final SSE and average AoA for the 10 runs in Experiment 2.

Experiment 3: Staged AoA Revisited

During the presentation of the early set exclusively in the first half of training, the network received feedback from the late set output units. The network received constant feedback from these units saying "not this person," since the late set was not presented yet. This is unrealistic. Humans are not constantly reinforced with "not this person" if the subject is unaware of the late set person's existence at the time. Therefore, in this simulation, the network treats the 13 late set output units as though they do not exist while training exclusively on the early set. This is a more accurate model of adding additional faces to a person's repertoire of identifiable faces.

Model

Late set output units are not trained during presentation of the early set. Therefore, the weights between the hidden layer and these output units remain unchanged. Additionally, these units have no effect in the calculation of the early set's SSE. During the second half of training when the late set is introduced, the network receives feedback from all the output units as usual. The same staging of face presentation is used from Experiment 2, except training lasts for a total of 250 epochs. As it turns out, these networks actually require longer to learn. Therefore, there are 125 epochs with the early set presented once per epoch. Another 125 epochs follow with the early set presented once per epoch and the late set presented twice per epoch. As we describe below, there are no AoA effects at the end of 250 epochs. To check the strength of this result, we add a recency control by training for another 50 epochs in which both early and late sets are presented once per epoch. Otherwise, the model is identical to the one used in Experiment 2.

Results

All patterns are acquired in each run before the period of recency control begins. For each run, we measure the average AoA for the early and late sets. We then perform a paired t-test comparing the difference in means between these averages across the 10 runs. We find a highly reliable difference, $t(9) = -329$, $p < .001$. The mean epoch for an early face to be acquired is 27 (sd = 1.2), and the mean epoch for a late face to be acquired is 146 (sd = 0.73). Therefore, the early set is acquired significantly before the late set.

Without the recency control period, our model does not show statistically significant AoA effects. In fact, the late set has a statistically significant lower mean error. For each run, we measure the average SSE at the 250th epoch for the early and late sets. We then perform a paired t-test comparing the difference in means between these averages across the 10 runs. We find a highly reliable difference, $t(9) = 10.5$, $p < .001$. The mean SSE for the early set is 0.0055 (sd = 0.00028), and the mean SSE for the late set is 0.0047 (sd = 0.00012). The late set has a clear advantage over the early set without the period of recency control. This is plotted in Figure 5.

During the recency control period, the early set error drops faster than the late set error, allowing the AoA effects to emerge. Our network shows very strong AoA effects at the 300th epoch. For each run, we measure the average final SSE for the early and late sets. We then perform a paired t-test comparing the difference in means between these averages across the 10 runs. We find a highly reliable difference, $t(9) = -9.5$, $p < .001$. The mean final SSE for the early set is 0.0036 (sd = 0.00016), and the mean final SSE for the late set is 0.0040 (sd = 0.00013). The early set has a significant advantage over the late set with the period of recency control. This is plotted in Figure 6.

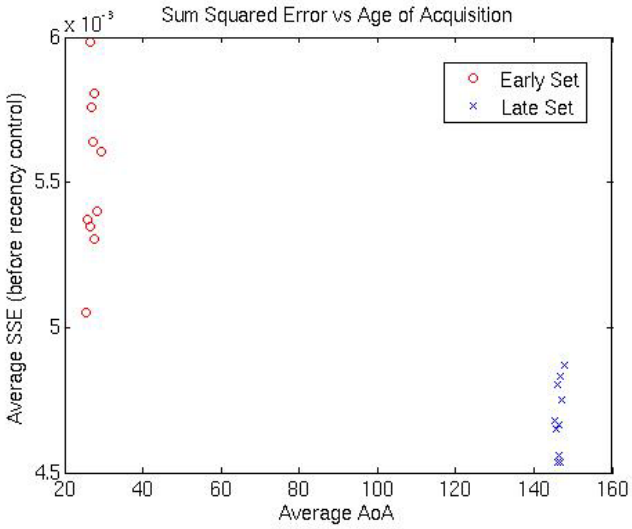

Figure 5: Relationship between average SSE (at the 250th epoch) and average AoA for the 10 runs in Experiment 3.

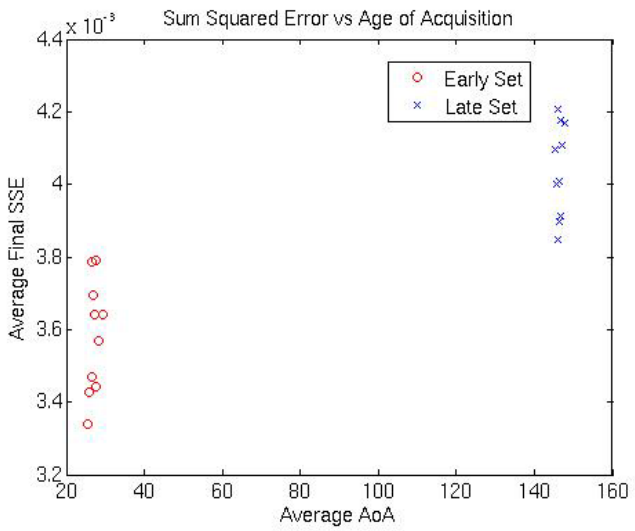

Figure 6: Relationship between average final SSE (at the 300th epoch) and average AoA for the 10 runs in Experiment 3.

Experiments 4&5: Effects of an Early Perceptual Representation

We assume that participants' perceptual representation depends on their experience with stimuli. In particular, models of the Other Race Effect (O'Toole et al. 1994) have used principal components analysis of faces to model the effect. The main idea of these models is that fewer faces of the other race than the own race go into the PCA, resulting in a poorer representation of the other race. This gives rise to the false alarms commonly found in recognition experiments. Could AoA for faces be explained in this way? In this simulation, we assume that the perceptual representation given to the model (Figure 1) in the PCA is formed from the early set. That is, the principal component eigenvectors are formed from the early set, and the late set is simply projected onto these and given as input to the network. In these simulations, we also vary whether the output units for the late set are trained to be off or not during acquisition of the early set. We find that the late set takes longer to learn this way, so we train the early and late sets for 175 epochs each.

Results

When replicating Experiment 2 in which all the outputs are trained throughout, one face is not learned in the 7th run, so that run is not used in the statistics. Using the same analysis procedure as before, we find the early set is acquired before the late set, $t(8) = -106$, $p < .001$. The mean epoch for an early face to be acquired is 35 (sd = 0.96), and the mean epoch for a late face to be acquired is 294 (sd = 7.9). Again, using the same analysis as before, we obtain significant AoA effects, $t(8) = -7.6$, $p < .001$. The mean final SSE for the early set is 0.016 (sd = 0.0026), and the mean final SSE for the late set is 0.026 (sd = 0.0055). The early set has a significant advantage over the late set.

The replication of Experiment 4 with the late set outputs untrained during the early phase shows similar effects as before. The early set is acquired early (26 epochs, sd = 0.78, $t(9) = -103$, $p < .001$), compared to the late set (262 epochs, sd = 7.7). Again, significant AoA effects in the final SSE are found, $t(9) = -7.0$, $p < .001$. The mean final SSE for the early set is 0.011 (sd = 0.0011), and the mean final SSE for the late set is 0.016 (sd = 0.0026). When assuming the representation developed by the early set is also used for the late set, our networks show strong AoA effects, regardless of whether or not the network receives feedback from the late set outputs during early set training.

Discussion

Our results suggest that connectionist networks trained as classifiers can show significant AoA effects in the same way as other network mappings. Previous work with autoencoders mapped high dimensional patterns to identical or very similar high dimensional patterns (Anderson & Cottrell, 2001, 2004; Ellis & Lambon Ralph, 2000; Smith et al., 2001). Anderson and Cottrell (2001) showed that even completely random high dimensional mappings show AoA effects, and these survive very strong frequency manipulations. Anderson and Cottrell (2004) showed that mapping in quasi-regular domains such as from spelling to sound also show robust AoA effects. In the current work, we show that for the facial identification task, where a high dimensional pattern is mapped to an output layer that merely selects a single output unit, AoA effects are also found. This suggests that AoA effects in connectionist networks have a wider scope than previously explored.

In Experiment 1, we show AoA effects by allowing the faces to be acquired in their natural order. Despite further training after all the patterns were acquired, the earlier acquired patterns maintained lower errors. We also find that the faces were not acquired in a similar order across replications with different initial random weights. We would like to explore if the natural order of acquisition can be predicted by properties of the training set in a facial identification model. Initial experiments have revealed that faces may be acquired in similar order across multiple runs if fewer principal components are used in the Gestalt Level. Additionally, if the training set includes more variance, it would be expected that some individuals would be easier to acquire than others. For example, a training set containing multiple races or some individuals with exaggerated facial expressions (imagine the variance in Jim Carrey's face!) and others with no facial expression (e.g., Natalie Portman in Star Wars Episode I) might allow the network to acquire individuals in a similar order across multiple simulations.

In contrast to our first simulation, however, humans may be exposed to different people at different times in their lives. For example, Moore and Valentine (1998) tested human subjects on a celebrity face-naming task similar to the one solved by our computational model. They divided celebrities into early and late groups based on rated AoA, while familiarity was controlled. They found participants named early celebrities significantly faster than late celebrities. In our Experiments 2-5, where we manipulated the age of first exposure, the faces in the early set had significantly lower errors than the faces in the late set, except for Experiment 3 where a recency control was

needed. This suggests that early training, as in humans, is critical in determining the mature network performance.

Of the manipulations we performed, it seems to us that the most realistic situation is where the early set determines the representation of face space, such as Experiments 4 and 5. In both experiments, we found strong AoA effects. In Experiments 2 and 3, when the late set was included in the face space representation, a significant effect was found in one case and a recency control restored it in the second. In the case of the recency control, it appears that the underlying AoA effects are briefly covered up by an advantage for recently presented material. In Moore and Valentine's (1998) study, it is impossible to know when each participant last viewed each celebrity in the early group, but it seems unlikely that such an inadvertent "recency control" occurred for the subjects. The need for a recency control period is a potential discrepancy between our computational results and human studies. However, we suggest here that the true situation may be closer to that in Experiments 4 and 5, where the perceptual representation is biased towards the early set.

Acknowledgements

We wish to thank Lingyun Zhang and Gary's Unbelievable Research Unit for their contributions. Garrison W. Cottrell is supported by NIH grant MH57075.

References

Anderson, K. L., & Cottrell, G. W. (2001). Age of Acquisition in Connectionist Networks. *Proceedings of the 23rd Annual Cognitive Science Conference* (pp. 27-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, K. L. and Cottrell, G. W. (2004). Measured Age of Acquisition effects in quasi-regular domains. Poster presented at the Annual meeting of the Cognitive Neuroscience Society, San Francisco, CA.

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture naming latency. *Quarterly Journal of Experimental Psychology*, *25*, 85-95.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience,* 14(8), 1158-1173.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spacial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of American A*, 2, 1160–1169.

Ellis, A. W., & Lambon Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 1103-1123.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, 29, 687-715.

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.

LeCun, Y., Bottou, L., Orr, G. B., & Muller, K. R. (1998). Efficient backprop. In G. B. Orr & K. R. Muller (Eds.), *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer-Verlag.

Lee, K. (2004). Multi-race face database. Department of Psychology, University of California, San Diego.

Moore, V., & Valentine, T. (1998). The effect of age of acquisition on speed and accuracy of naming famous faces. *The Quarterly Journal of Experimental Psychology*, 51A, 485-513.

Moore, V., & Valentine, T. (1999). The effects of age of acquisition in processing famous faces: Exploring the locus and proposing a mechanism. *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 416-421). Mahwah, NJ: Lawrence Erlbaum Associates.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 116-153.

Oldfield, R.C. & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, 17, 273-281.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., Abdi, H. (1994). Structural aspects of face recognition and the other-race effect, *Memory & Cognition*, 22(2), 208-224.

Smith, M. A., Cottrell, G.W., & Anderson, K. L. (2001). The early word catches the weights. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems.* Cambridge, MA: MIT Press.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of Acquisition Effects in Word Reading and Other Tasks. *Journal of Memory and Language*, 47, 1-29.

Zhang, L., & Cottrell, G. W. (2004). When holistic processing is not enough: Local features save the day. *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1506-1511). Mahwah, NJ: Lawrence Erlbaum Associates.