Computational Explorations of Split Architecture in Modeling Face and Object Recognition

Janet Hui-wen Hsiao (jhsiao@cs.ucsd.edu)
Garrison W. Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Drive #0404, La Jolla, CA 92093, USA

Danke Shieh (danke@ucsd.edu)
Department of Cognitive Science, University of California San Diego
9500 Gilman Drive #0515, La Jolla, CA 92093, USA

Abstract

Anatomical evidence shows that our visual field is initially split along the vertical axis and contralaterally projected to different hemispheres. It remains unclear at which stage the split information converges. In the current study, we applied the Double Filtering by Frequency (DFF) theory (Ivry & Robertson, 1998) to modeling the visual split; the theory assumes a right hemisphere/low frequency bias. We compared three cognitive architectures with different timing of convergence and examined their cognitive plausibility to account for the left side bias effect in face perception observed in human data. We show that the early convergence model failed to show the left side bias effect. The left side bias effect was also observed in Greeble recognition. The modeling hence suggests that the convergence may take place at an intermediate or late stage, at least after information has been extracted/transformed separately in the two hemispheres; it also provides testable predictions about whether the left side bias effect may also be observed in (expertise-level) object recognition.

Keywords: Connectionist modeling; face recognition; hemispheric differences; split modeling.

Introduction

Because of the partial decussation of optic nerves, our visual system is initially vertically split and the two visual hemifields are initially contralaterally projected to different hemispheres. A fundamental question in cognitive science is whether this initial split has any functional significance; that is, whether the effect of initial splitting can extend far enough to influence our cognition? A second question is at what stage does the information converge?

A functional split

Evidence from visual word recognition supports a functional split. The general finding is that the two hemispheres have contralateral influence on responses driven by the first and last halves of the stimuli, which are initially projected to different visual hemifields (e.g., Lavidor, Ellis, Shillcock, & Bland, 2001; Lavidor & Walsh, 2003; Hsiao & Shillcock, 2005a; Hsiao, Shillcock, & Lavidor, 2006). There is also evidence from face recognition supporting a functional split. For example, a left side bias effect has been frequently reported in face perception. The classical experiment is to ask participants to judge the similarity between a face and chimeric faces made from the two left halves (left chimeric face) or the two right halves (right chimeric face) of the original face (from the viewer's perspective; Figure 1). The results show that the left chimeric face is usually judged more similar to the original face than the right chimeric face, especially for highly familiar faces (Brady, Campbell, & Flaherty, 2005). Consistent with this result, other studies have argued for a right hemisphere (RH) bias in face perception (e.g., Rossion, Joyce, Cottrell, & Tarr, 2003). Nevertheless, it remains unclear how far the split effect extends. Although it has been shown that our visual system is organized as a set of hierarchically connected regions, and the receptive field sizes of the neurons increase by a factor of about 2.5 at each succeeding stage (Rolls, 2000), the initial trajectory of visual activation flow is a fast and widespread sweep and continues through iterations of feedback loops for further processing in the sensory area (Foxe & Simpson, 2002); hence, it is unclear yet whether the split influences high-level cognition. In visual word recognition, Hsiao and Shillcock (2005a) showed that this split effect can reach far enough to interact with sex differences in brain laterality for phonological processing. Thus, the split seems to influence high-level cognition.

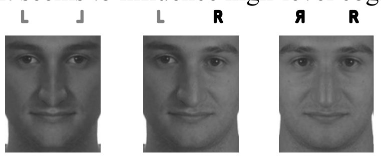


Figure 1: Left chimeric, original, and right chimeric faces.

Split modeling & timing of convergence

In order to address the splitting effects observed in visual word recognition, Shillcock and Monaghan (2001) proposed a split fovea model (Figure 2) and showed that some psychological phenomena in visual word recognition can be better accounted for by the split architecture, such as exterior letter effects in English word recognition and eye fixation behavior in reading English (Shillcock, Monaghan, & Ellison, 2000). Hsiao and Shillcock (2005b) further showed that the split and nonsplit architectures in modeling

Chinese character recognition exhibited qualitatively different processing, and the results were able to account for the sex differences in naming Chinese characters in human data.

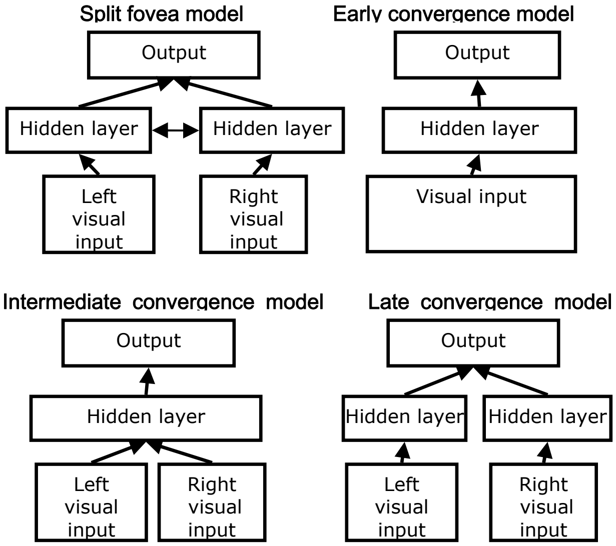


Figure 2: Architectures of different models.

In the current study, we apply the split fovea model to face and object recognition. In contrast to previous models in visual word recognition, which act on a relatively abstract level of representation (i.e., localist representation of letters or stroke patterns), we incorporate some aspects of visual anatomy into the modeling. We use Gabor responses over the input image to simulate neural responses of complex cells in the early visual system (Lades et al., 1993). We then reduce the dimension of this perceptual representation with Principal Component Analysis (PCA), which has been argued to be a biologically plausible linear compression technique (Sanger, 1989; cf. Dailey et al., 2002); this is the visual input shown in Figure 1. With this level of abstraction, convergence of the initial split may happen at three different stages: *early*: after Gabor filters in the early visual system (i.e., at the input layer), *Intermediate*: after information extraction through PCA (i.e., at the hidden layer), and *late*: at the output layer (Figure 2). In the early convergence model, the left and right Gabor filters are processed as a whole through PCA (i.e., nonsplit input representation; Figure 3). In the intermediate convergence model, PCA is applied separately to the left and right Gabor filters and the convergence is at the hidden layer. In the late convergence model, in addition to the split input layer, the hidden layer is also split, and the information converges at the output layer. According to this categorization, the split fovea and nonsplit models proposed first in Shillcock and Monaghan (2001) can be considered as late and intermediate convergence models respectively[1]. Here we conduct a more general comparison between these three architectures and examine their performance and cognitive plausibility.

DFF theory & face recognition

In order to account for various psychological phenomena involving hemispheric differences, Ivry and Robertson (1998) proposed a Double Filtering by Frequency (DFF) theory. The theory argues that information coming into the brain goes through two frequency filtering stages. The first stage involves attentional selection of task-relevant frequency information, and at the second stage the two hemispheres have asymmetric filtering processing: the left hemisphere (LH) amplifies high frequency information (i.e., a high-pass filter), whereas the RH amplifies low frequency information (i.e., a low-pass filter).

There has been an ongoing debate regarding whether the brain processes faces differently from objects. Evidence for this argument comes from studies showing that the fusiform face area (FFA) in the brain selectively responds to face stimuli (e.g., McKone & Kanwisher, 2005), whereas other studies have suggested that several phenomena that were thought to be unique to face recognition may be due to expertise (e.g., Gauthier et al., 1999). Thus, the left side bias effect observed in face perception may be due to a designated face processor located in the RH, or the reliance on low spatial frequency (LSF) processing in the RH (according to the DFF theory) once the expertise is acquired. The split architectures introduced here enable us to apply the DFF theory to modeling face and object recognition. We first examine whether the DFF theory is able to account for the left side bias effect in face perception. A positive result will suggest the RH reliance in face processing is due to the low frequency bias in the RH. We then examine whether the left side bias effect can also be obtained in expert object recognition. The objects under examination are Greebles, a novel class of objects that have been frequently used in studies of object recognition and perceptual expertise (e.g., Gauthier et al., 1999). If the left side bias effect also exists in modeling expert Greeble recognition, the results will provide testable predictions regarding whether faces and objects are processed differently in the brain.

Models and Results

Representations and modeling details

To simulate responses of complex cells in the early visual system, the input image (135 x 100 pixels) was first filtered with a rigid grid (16 x 12) of overlapping 2D Gabor filters (Daugman, 1985) in quadrature pairs at six scales and eight orientations (Figure 3). The six scales corresponded to 2 to 64 cycles per face. Given the width of the image (100 pixels), this frequency range hence can be considered as the

[1] The late convergence model differs from the split fovea model in that it does not have interconnections between the two hidden layers. We removed these interconnections here for comparison reasons; in separate simulations, we found that adding these interconnections did not change the effects we reported here.

task-relevant frequency range (the seventh scale would have 128 cycles, which exceeded the width of the image). Hence the first stage of the DFF is implemented by simply giving this input to all of our models. The paired Gabor responses were combined to obtain Gabor magnitudes. In the nonsplit input representation, this 9,216 (16 x 12 Gabor filters x 6 scales x 8 orientations) element perceptual representation was compressed into a 50-element representation with PCA. In the split input representation, the face was split into left and right halves, and each had 16 x 6 Gabor filters (4,608 elements). The perceptual representation of each half was compressed into a 50-element representation (hence in total there were 100 elements)[2]. After PCA, each principal component was z-scored to equalize the contribution of each component in the models. In the three models, the early convergence model had a nonsplit input representation, whereas both the intermediate and late convergence models had a split input representation. In order to equalize their computational power, the hidden layer of the early and the intermediate convergence models had 20 units, and each of the two hidden layers of the late convergence model had 10 units; in the intermediate convergence model, half of the connections from the input layer to the hidden layer were randomly selected and removed. Hence, the three models had exactly the same number of hidden units and weighted connections. To implement the second stage of the DFF theory, we used a sigmoidal filter (Figure 4) after the Gabor filters to bias the Gabor responses on the left half face (RH) to LSF and those on the right half face (LH) to high spatial frequency (HSF).

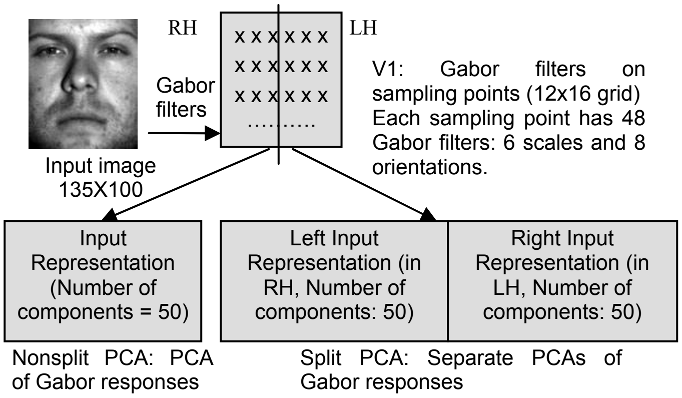

RH
LH
Gabor filters
x x x x x x
x x x x x x
x x x x x x
……….
V1: Gabor filters on sampling points (12x16 grid) Each sampling point has 48 Gabor filters: 6 scales and 8 orientations.
Input image 135X100
Input Representation (Number of components = 50)
Left Input Representation (in RH, Number of components: 50)
Right Input Representation (in LH, Number of components: 50)
Nonsplit PCA: PCA of Gabor responses
Split PCA: Separate PCAs of Gabor responses

Figure 3: Nonsplit and split visual input representations.

In short, we tried to bring the model architecture as close to the visual anatomy as possible. The Gabor filters correspond to V1, the PCA can be thought of as analogous to the Occipital Face Area (i.e., structural representation of faces), the hidden layer of the network has been associated with the Fusiform Face Area. Finally, the output layer has a unit for each individual subject. For the following simulations, we ran each model 80 times and analyzed its behavior with ANOVA after 100-epoch training (their performance on the training set all reached 100% accuracy). The training algorithm was discrete back propagation through time. (Rumelhart, Hinton & Williams, 1986), and the learning rate was 0.1. Performance was analyzed at the end of 7 time steps (cf. Shillcock & Monaghan, 2001; Hsiao & Shillcock, 2005b)[3].The independent variables were architecture (early, intermediate, and late convergence) and frequency bias (unbiased vs. biased). The dependent variables were accuracy and size of left side bias effect. To examine the size of left side bias effect, we took output node activation for a particular individual as a measure of similarity between the chimeric face and the original face. After training, we presented the networks with left and right chimeric faces using test set images. The size of left side bias effect was measured as the difference between the activation of the output node for the original face when the left chimeric face was presented and when the right chimeric face was presented (note that output activation ranged from 0 to 1). For each simulation, the materials consisted images of 30 different individuals (so there are 30 output nodes; see the following sections for simulation details). Two datasets were created for training and testing and the order was counterbalanced across the simulation runs. In order to eliminate any side bias effect due to the baseline difference between the two sides of the images, in half of the simulation runs the mirror images of the original images were used.

f(x,a) = 1 / (1 + exp (- a (x - 3.5))

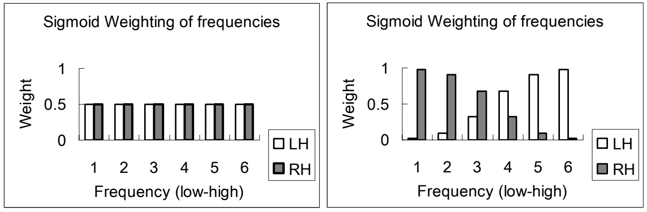


Figure 4: Sigmoidal filters: unbiased (a = 0) and biased conditions (a = 1.5).

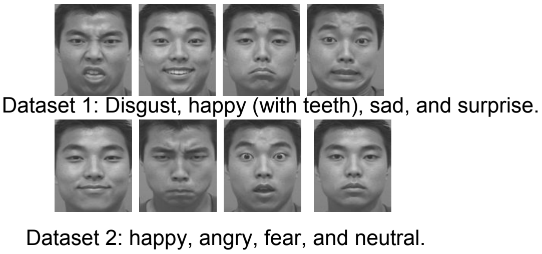
Face recognition with expression changes

We first examined face recognition with different expressions. Each of the two datasets created contained four images with different expressions (Figure 5), for a total of 120 training and 120 test images. These images were taken from CAlifornia Facial Expressions dataset (CAFÉ; Dailey, Cottrell, & Reilly, 2001). The generalization accuracy

[2] Although the split and nonsplit representation had different number of dimensions in the input layer, they both contained information from the first 50 principal components. This equalizes the information contained in each representation better than increasing the number of dimensions in the nonsplit representation to match that of the split representation. In fact, with 100 components, the nonsplit model performs worse.

[3] Although the networks do not have recurrent connections, we used discrete back propagation through time to be consistent with the split fovea model (Shillcock & Monaghan, 2001), which has recurrent connections between the two hidden layers. We found that adding these interconnections did not change the effects we reported here.

results showed a significant interaction between architecture and frequency bias ($F(2, 474) = 8.513$, $p < 0.001$; Figure 6(a)). In general, biased spatial frequency hurt performance, and the later the convergence, the bigger the effect. As for the left side bias effect for chimeric faces, there was a main effect of architecture ($F(2, 474) = 141.457$, $p < 10^{-48}$), a main effect of frequency bias ($F(2, 474) = 709.651$, $p < 10^{-95}$), and a significant interaction between architecture and frequency bias ($F(2, 474) = 144.386$, $p < 10^{-48}$; Figure 6(b)). None of the models had a left side bias effect in the unbiased condition. In the biased condition, the frequency bias significantly induced the left side bias effect in the late and intermediate convergence models. In contrast, the early convergence model did not have the left side bias effect. This suggests that in the biased condition, converging at an early stage (i.e., nonsplit representation) may still extract balanced low and high frequency information for recognition, whereas converging at a later stage (i.e., split representation) allows more low frequency information from the left half face to the hidden layer, and consequently brings about the left side bias effect[4].


Dataset 1: Disgust, happy (with teeth), sad, and surprise.

Dataset 2: happy, angry, fear, and neutral.

Figure 5: Face images for facial expression recognition.

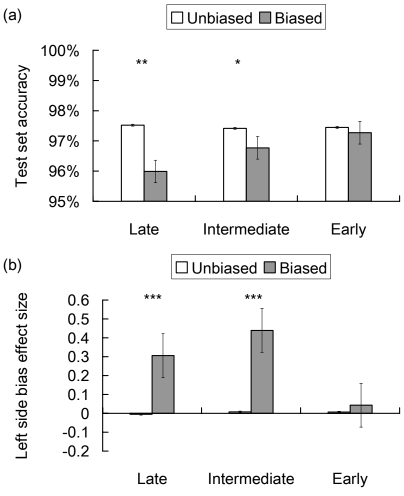


Figure 6: (a) Performance and (b) Left side bias effect in the three models. Error bars show standard errors (* $p < 0.01$; ** $p < 0.001$; *** $p < 0.0001$).

Faces under natural lighting changes

In order to reconfirm the left side bias effects we obtained, we conducted another simulation: face recognition under different lighting changes. We selected face images from Yale face database (Georghiades, Belhumeur, & Kriegman, 2001) with a light source moving from right to left. Each individual had eight different lighting conditions (Figure 7); the lighting conditions in the training and test datasets had the same azimuths but different altitudes.

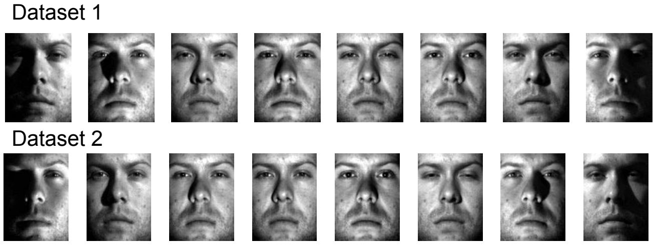


Figure 7: Face images for training. From left to right, the azimuths are: -60, -35, -20, -10, +10, +20, +35, and +60. Altitudes range from -20 to 20.

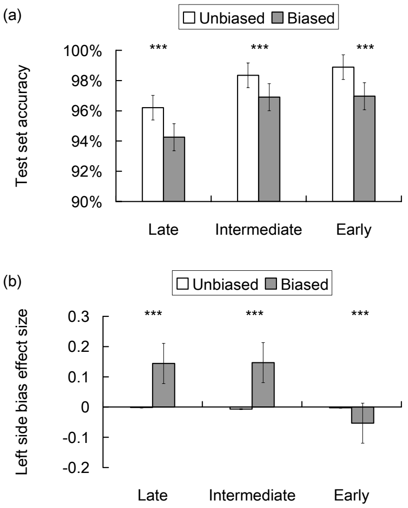


Figure 8: (a) Performance of the three models. (b) Left side bias effect in the three models. Error bars show standard errors (*** $p < 0.0001$).

The results showed that the late convergence model performed the worst ($F(2, 234) = 158.918$, $p < 10^{-52}$), and performance in the unbiased condition was better than the biased condition ($F(2, 234) = 172.075$, $p <10^{-33}$; Figure 8(a)). As for the left side bias effect for chimeric faces, there was a main effect of architecture ($F(2, 234) = 233.286$, $p < 10^{-70}$), a main effect of frequency bias ($F(2, 234) = 369.360$, $p < 10^{-60}$), and a significant interaction between architecture and frequency bias ($F(2, 234) = 242.055$, $p < 10^{-72}$; Figure 8(b)). The frequency bias again significantly induced the left side bias effect in the late and intermediate convergence models; the

[4] The only difference between the early and intermediate convergence models was the input representation (nonsplit vs. split). In a separate simulation, we used a simple perceptron (i.e., the hidden layer was removed) to examine the baseline behavior between the two representations, and the split representation indeed had a left side bias effect whereas the nonsplit representation did not; this effect was consistent across the three simulations we reported here.

early convergence model failed to show the left side bias effect; in fact, it exhibited a slight right side bias effect. The results hence confirmed again that the intermediate convergence model had the strongest left side bias effect, and the early convergence model was not able to exhibit the left side bias effect in human data.

Greebles under natural lighting changes

We turned to see whether the same effects can be obtained in Greeble recognition. Objects such as Greebles do not have expressions; one of the most common object recognition tasks we perform in the real world is probably to recognize the same object under different lighting conditions. Hence, we examined the networks' performance on recognizing Greebles under different lighting changes. We considered the sun as the major source of light in nature, and its azimuth increases during a day and its altitude first increases and then decreases from midday. In each of the two datasets created, each Greeble had eight images under the eight different lighting conditions shown in Figure 9.

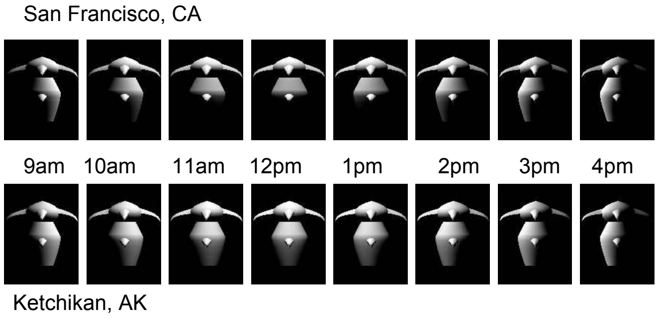


Figure 9: A Greeble, facing south, under the sun in San Francisco, California (latitude, longitude: 27.618, 122.373) and Ketchikan, Alaska (55.342, 131.647), from 9 am to 4 pm.

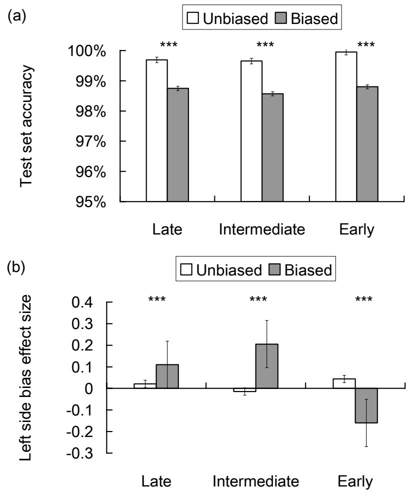


Figure 10: (a) Performance of the three models. (b) Left side bias effect in the three models. Error bars show standard errors (*** p < 0.0001).

The first set was Greebles under the San Francisco sun, and the other was the same Greebles in Ketchikan. Hence, in the two datasets, the sun positions had different azimuths and altitudes. The accuracy results showed that there was a main effect of frequency bias ($F(2, 234) = 94.089$, $p < 10^{-19}$): performance in the unbiased condition was better than the biased condition (Figure 10(a)). As for the left side bias effect for chimeric Greebles, there was a main effect of architecture ($F(2, 234) = 100.768$, $p < 10^{-36}$), a main effect of frequency bias ($F(2, 234) = 13.891$, $p < 0.001$), and an interaction between architecture and frequency bias ($F(2, 234) = 178.707$, $p < 10^{-57}$; Figure 10(b)). Similar to the previous simulation, the frequency bias significantly induced the left side bias effect in the late and intermediate convergence models; the early convergence model did not have the left side bias effect; it exhibited right side bias instead. The results suggest that the left side bias effect may also exist in Greeble recognition.

Conclusion and Discussion

In the current study, we explored split architecture in modeling face and object recognition. We applied the DFF theory to the split modeling of visual processing; for the input representation, we first selected a task relevant frequency range, and then biased the information coming into the RH (i.e., left half of the input) to low frequency and that coming into the LH (i.e., right half of the input) to high frequency through a sigmoidal filter. We then compared performance and cognitive plausibility of three cognitive architectures with different timings of convergence. We showed that, in this computational exploration, the combination of the spatial frequency bias and the splitting of the information between left and right are sufficient to show the left side bias effect, but neither alone can show the effect. This is consistent with the observation that there is a low spatial frequency bias in face identification, both in humans and computational models (Schyns & Oliva, 1999; Dailey et al., 2002). This is reflected in the higher activation of the identity unit when the model's right hemisphere receives the same side of the face it was trained upon, compared with when it does not.

The failure of the early convergence model in exhibiting the left side bias effect suggested that the initially split visual input may converge at an intermediate or late stage, at which at least certain type of information extraction/transformation has been applied separately in each hemisphere, perhaps after the Occipital Face Area. This result is consistent with several behavioral studies showing that each hemisphere seems to have dominant influence on the processing of the visual information presented in the visual hemifield to which it has direct access (e.g., Brady et al, 2005; Hsiao et al., 2006).

The results from modeling Greeble recognition also showed a left side bias effect in both the intermediate and late convergence models, but not in the early convergence model. In human data, the left side bias effect has never been shown in recognition tasks other than faces, and hence has been considered a face-specific effect. However, it may also be due to our expertise in face processing (cf. Gauthier et al., 1999). The modeling result hence provides a testable prediction about whether a left side bias effect can also be observed in object recognition once expertise is acquired.

The models we propose here unavoidably involve abstraction and assumptions about the underlying neural complexity, but they nevertheless address the issue under examination here. The study provides a computational explanation of the cognitive implausibility of the early convergence model, which has been the most typical model for face/object/word recognition in the literature (e.g., Dailey et al, 2002; Harm & Seidenberg, 1999). The fact that the initial split has a functional significance has been overlooked in connectionist modeling of cognitive processes; the current study shows that this fact does have significant impact on how modeling is able to explain and predict human behavior.

As future directions, the proposed models can also be applied to visual word recognition (cf. Shillcock et al., 2000) to examine the current debates about the foveal splitting phenomena (e.g. Lavidor & Walsh, 2004). We will also examine a fundamental question about why such a split and frequency bias exists in the brain. The current simulations seem to suggest that frequency bias deteriorates performance. There may exist an optimal frequency bias setting that is able to boost performance, or the advantage of split and frequency bias may be observed when the system (i.e., the brain) has to deal with tasks with different frequency requirements; for example, word recognition may rely more on high frequency information processing in contrast with face recognition. These issues are currently under examination.

References

Dailey, MN., Cottrell, GW., Padgett, C., & Ralph, A. (2002) EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, 14, 1158-1173.

Dailey, MN., Cottrell, GW., & Reilly, J. (2001). CAlifornia Facial Expressions (CAFE). http://www.cs.ucsd.edu/users/gary/CAFE/.

Foxe, JJ. & Simpson, GV. (2002). Flow of activation from V1 to frontal cortex in humans: a framework for defining 'early' visual processing. *Experimental Brain Research*, *142*, 139–150.

Gauthier, I., Tarr, MJ., Anderson, AW., Skudlarski, P., & Gore, JC. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nat. Neurosci.*, *2*, 568–573.

Georghiades, AS., Belhumeur, PN., & Kriegman, DJ. (2001). From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, *6*, 643-660.

Harm, MW. & Seidenberg, MS. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psych. Rev.*, *106*, 491–528.

Hsiao, JH. & Shillcock, R. (2005a). Foveal splitting causes differential processing of Chinese orthography in the male and female brain. *Cognitive Brain Research*, *25*, 531-536.

Hsiao, JH. & Shillcock, R. (2005b). Differences of split and non-split architectures emerged from modelling Chinese character pronunciation. *Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society* (pp. 989-994). Mahwah, NJ: Lawrence Erlbaum Associates.

Hsiao, JH., Shillcock, R., & Lavidor, M. (2006). A TMS examination of semantic radical combinability effects in Chinese character recognition. *Brain Research*, 1078, 159-167.

Ivry, R. & Robertson, LC. (1998). *The Two Sides of Perception*. Cambridge: MIT Press.

Lades, M., Vorbruggen, JC., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, RP., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, *42*, 300–311.

Lavidor, M., Ellis, AW., Shillcock, R., & Bland, T. (2001). Evaluating a split processing model of visual word recognition: Effects of word length. *Cognitive Brain Research*, *12*, 265-272.

Lavidor, M. & Walsh, V. (2003). A magnetic stimulation examination of orthographic neighbourhood effects in visual word recognition. *Journal of Cognitive Neuroscience, 15,* 354-363.

Lavidor M., Walsh, V. (2004). The nature of foveal representation. *Nat. Rev. Neurosci.*, *5*, 729-735.

McKone, E. & Kanwisher, N. (2005). Does the human brain process objects of expertise like faces? A review of the evidence. In S. Dehaene, J. R. Duhamel, M. Hauser, & Rizzolatti (Eds), *From Monkey Brain to Human Brain*. Cambridge, Massachusetts: the MIT Press.

Rolls, ET. (2000) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron*, *27*, 205–218.

Rossion, B., Joyce, CA., Cottrell GW., & Tarr, MJ. (2003). Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. *Neuroimage*, *20*, 1609-1624.

Rumelhart, DE, Hinton, GE, & Williams, RJ. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533-536.

Sanger, T. (1989). An optimality principle for unsupervised learning. In Touretzky, D. (ed), *Advances in Neural Information Processing Systems,* vol. 1, pp. 11–19, San Mateo: Morgan Kaufmann.

Schyns P.G. & Oliva A. (1999) Dr. Angry and Mr. Smile: when categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, *69*, 243-265.

Shillcock, RC. & Monaghan, P. (2001). The computational exploration of visual word recognition in a split model. *Neural Computation*, *13*, 1171-1198.

Shillcock, R., Ellison, TM., & Monaghan, P. (2000). Eye-fixation behavior, lexical storage, and visual word recognition in a split processing model. *Psychological Review*, *107*, 824-851.