A New Angle on the EMPATH Model: Spatial Frequency Orientation in Recognition of Facial Expressions

Rentao Li (reli@ucsd.edu)
Department of Computer Science, 9500 Gilman Drive
La Jolla, CA 92093 USA

Garrison Cottrell (gary@eng.ucsd.edu)
Department of Computer Science, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

Many have investigated the sensitivity of face processing to both spatial frequencies and face orientation, but few have researched the sensitivity of face processing to the orientation of spatial frequencies. One recent exception has been Yu, Chai, & Chung (2011), which investigated facial expression recognition in regards to the orientation of spatial filters and showed that most information is contained in the horizontal orientation. Here, we model the Yu, Chai, & Chung (2011) study using the EMPATH model, a feed-forward neural network that has been used to model facial expression recognition (Dailey, Cottrell, Padgett, & Adolphs 2002). We used the NimStim set of facial expressions, which were the basis for the Yu, Chai, & Chung (2011) experiment, and followed their method of filtering images through different spatial orientations. Our results show that this simple, biologically plausible model produces very similar results to that of human subjects in their study.

Keywords: emotions; facial expressions; spatial frequency; neural network; face recognition.

Introduction

Many studies have been conducted regarding the role of spatial frequencies in human face recognition (Näsänen, 1999; Costen, Parker, & Craw, 1996; Gold, Bennett, & Sekuler, 1999), although the uniqueness of sensitivity of faces to spatial frequency has been debated (Williams, Willenbockel, & Gauthier (2009). However, few have explored how different *orientations* of spatial frequencies impact recognition of facial images. One such experiment was done by Yu, Chai, & Chung (2011), who passed facial images through orientation filters from -60 to 90 degrees in increments of 30 degrees. They found that the spatial information near horizontal (between -30 and 30 degrees) were the most important for normally-sighted human respondents to recognize facial expressions.

Background: Yu, Chai, & Chung's Experiment (2011)

The aim of the Yu, Chai, & Chung (2011) experiment was to determine which spatial orientations on the face contained the most information for identifying emotions. The four emotions they tested were the closed-mouth forms of anger, fear, happiness, and sadness. Images were obtained from the NimStim set of facial expressions (Tottenham, Tanaka, Leon, McCarry, Nurse, Hare, Marcus, Westerlund, Casey, & Nelson 2009), and were distorted with an orientation filter of bandwidth 23^o in the Fourier domain, where the center of the filter ranged from -60^o to 90^o in increments of 30^o. Unfiltered images were used for comparison.

Their experiment consisted of having 15 normally-sighted human subjects try to recognize the expression displayed by each image under a four-way forced choice. The results indicate that the human observers had the most success with images filtered at orientations near the horizontal (-30^o, 0^o, and 30^o), suggesting that horizontal spatial information is most important for recognizing facial expressions. One modest exception to this trend is the fearful face; the human subjects tended to be significantly biased towards labeling a face as fearful as the orientation filter approached 90^o, which seems to indicate that much of the information for fear is represented vertically.

The purpose of this current experiment is to determine if a neural network model can produce similar results as the human subjects, especially in regards to the increased recognition performance for horizontal orientations and the preference towards fear for vertical orientations. Such evidence would provide greater support for Yu, Chai, & Chung's (2011) findings and further validate EMPATH's flexibility and accuracy in modeling human face recognition.

Methods

The Model

The neural network used for this experiment closely followed the EMPATH model developed by Dailey et al. (2002), consisting of a biologically plausible, three-layer, feed-forward perceptron. EMPATH has been shown to have remarkable face recognition performance on aligned, grayscale images from Ekman and Friesen's POFA (1976). Without being tuned specifically to those images, the network classified the emotions Anger, Disgust, Fear, Happiness, Sadness, and Surprise with 90% accuracy on average, compared to 91.6% for human subjects (Dailey et al., 2002). For this experiment, we kept much of the settings (outlined below) identical to those of the original EMPATH model, so the network was not tailored for the spatially filtered images or the NimStim dataset.

The first layer consisted of a set of model neurons based on the magnitude of Gabor filters, which have become a standard way to model complex cells in the early visual cortex (Daugman, 1985). In all, 40 different Gabor filters were used, in combinations of 5 scales and 8 orientations; filtering was done by passing the face images through a 29 by 35 "grid" of filters, resulting in 40,600 responses per image. Note that the orientations of the Gabor filters were the same as in the original EMPATH model, and were not changed to fit with the spatial filtering used in this study.

In order to reduce the dimensionality of the data set, we performed principal component analysis (PCA) on the Gabor filter outputs, producing 50 principal components (again based on the original EMPATH model). In this second layer, the principal components capture the distinguishing features of each facial expression but abstract away from details unique to each face; hence they allow the network to generalize to novel faces that are not part of the training set. As Dailey et al. stated, these components are similar to face cells in the inferior temporal cortex (2002).

Lastly, the principal components were fed into the third layer, consisting of a simple linear perceptron with six softmax outputs representing anger, fear, happiness, surprise, disgust, and sadness. This perceptron was trained using stochastic gradient descent with the cross-entropy error criterion. We used an "all-or-none" teaching signal that had "1" for each correct expression and "0" for the incorrect expressions. In order to replicate the four-way forced-choice employed by Yu, Chai, and Chang (2011), we only took the results of the four relevant emotions via a process described in "Training, Validating, and Testing."

As stated in Dailey et al., we acknowledge that this perceptron is very simplistic (2002). However, since it was powerful enough to map the principal components to emotion categories, we did not feel that a non-linear classifier was needed.

The Images

The images used in the testing set by Yu, Chai, & Chung (2011) consisted of morphs of images taken from the NimStim set (2009), which reduced variations among faces (such as race, gender, etc.). Without access to these exact morphs, however, we simply used some of the original NimStim images to create our training and testing sets. Given this difference, our results still closely matched those of Yu, Chai, & Chung (2011).

Our testing and training sets consisted of grayscale images of 30 different people (17 male, 13 female). Two of the images are shown in Figure 1. These images were judged to be the most frontally aligned, making them the most suitable for EMPATH. Each image was 240 x 292 pixels in size and was cropped closely about the face.

Both the testing and training sets contained images of six different emotions for each of the 30 people. These expressions were comprised of both open and closed-mouth forms of anger, fear, happiness, and sadness; the open-mouth form of surprise; and the closed-mouth form of disgust.

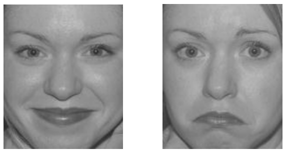


Figure 1: Two of the cropped images, corresponding to Happiness and Sadness

The additional expressions were chosen for the training set so that the network would have a more comprehensive exposure to the range of different emotions, making it more similar to the experience of the human subjects in the experiment. Although the testing set also contained six expressions, our method for producing the network's output was able to emulate a four-way forced choice among the four expressions used by Yu, Chai, & Chung (2011), which effectively limited the output to only those four choices (detailed in "Training, Validating, and Testing").

Processing the training set: In order to better replicate the images used by Yu, Chai, & Chung (2011), the 30 sets of images were closely cropped about the face using an oval mask so that only an oval-shaped portion of the face was visible. The parts that were cropped out were filled in with a uniform gray color of RGB value 127, and the entire image was adjusted to have a root-mean-square contrast value of 0.096, as per specifications given in Yu & Chung (2011). Examples of images used in the training set are shown in Figure 2.

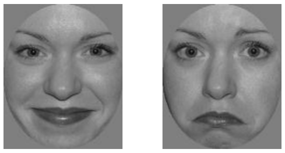


Figure 2: Two images from the training set

Processing the testing set: The process for creating the testing set was very similar to that of the training set. The 30 original sets of images were first cropped, aligned, and then processed using Yu, Chai, & Chung's (2011) filters at six orientations from -60^{o} to 90^{o} in increments of 30^{o}, which selectively pass information at the specified orientations. Afterwards, the oval "mask" was applied to the images in the same way as that of the training set, and all of the images were again normalized to have the same root-mean-square contrast of 0.096. Examples of the test images are shown in Figure 3. These were then processed by the same Gabor filters as used in the training set, and the resulting filter responses were projected onto the 50 eigenvectors using the PCA that was computed on the training set.

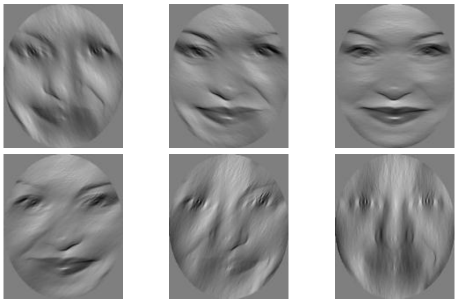


Figure 3: All 6 filtering conditions shown horizontally from -60^o (top left) to 90^o (bottom right).

Training, validating, and testing

The last layer of the model was a 50-input, 6-output single layer perceptron with softmax outputs trained using cross entropy. This procedure leads to outputs that compute the conditional probability of the category given the inputs (Bishop, 1995). Hence the output of the network is a probability distribution over the facial expression categories.

Cross-validation and early stopping were used to prevent overfitting the network to the training set. Since there are thirty individuals in the training set, we performed thirty instances of cross validation, each time holding out a different individual in the training set to use for early stopping. This comprehensive cross-validation made the testing less prone to unevenness among the images. Overall, there were 30 independent test cycles; each time 1 set was chosen for testing, 1 chosen for validation, and the remaining 28 were used as the training set. The aggregate performance from these 30 sets of tests constituted the results for each filtering condition. The validation set was taken from the processed test set as a guide to know when to stop training (of course the validation and test images were never the same). Training was completed for each cycle once the cross-entropy error for the validation set was minimized using gradient descent, and the weights of the network were then used for the testing set.

The testing procedure involved computing the weighted sum of the 50-element test set using the weights from training, then again applying the softmax function. A simple max function was used to judge if the testing outputs matched the teaching signals, and to create the confusion matrix. However, since the weights were trained with six facial expressions, the max function was applied only among the four target emotions to create a four-way forced choice similar to what a human subject would have to perform. This is valid because the softmax function created outputs that were probabilities of each emotion being correct; thus, although the teaching signal was "all or nothing," the outputs were not. We note that when humans undergo the task of selecting among four target emotions, it is entirely possible that the emotion they perceive is not among the four options, and thus they may have to answer with their second or third choice. This is essentially what we have emulated with our network.

Results

Our model was able to fit the human data very well in several measures. Much of the data presented by Yu, Chai, & Chung (2011) is displayed in the form of confusion matrices, which pit the human responses against the actual targets. We used the same technique to display our data. Since Yu, Chai, & Chung (2011) presented their results on a poster, many of their figures lack numerical data. As such, much of our analysis will be dependent on comparing the visual presentations of data. The color spectrum of our confusion matrices were closely matched to that used by Yu, Chai, & Chung (2011) so that comparisons and conclusions can be made.

Performance on Unfiltered Images

Figures 4a and 4b show the confusion matrix for the unfiltered images given by Yu, Chai, & Chung (2011) and by our model, respectively. In addition to the hits, the columns show false alarms and rows depict misses.

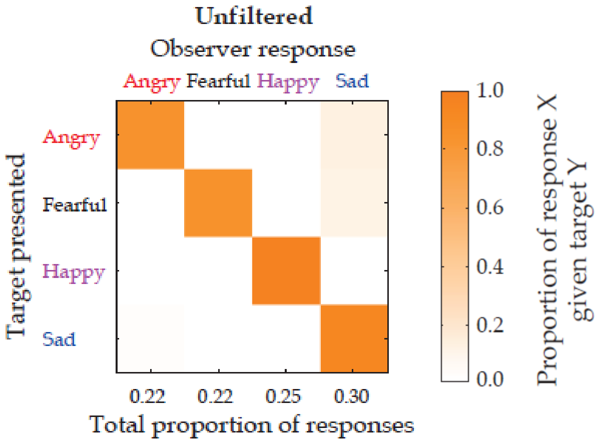


Figure 4a: Presented target vs. human responses (unfiltered) as presented in Yu, Chai, & Chung (2011).

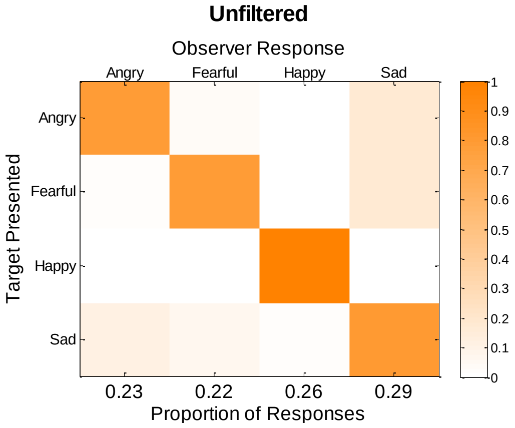


Figure 4b: Presented target vs. model responses, for unfiltered images.

Both the model and the human subjects demonstrated very good performance overall in recognizing the unfiltered images. The model also exhibited similar behavior as the human subjects in terms of having false alarms for sad faces when presented with Angry and Fearful faces. This is likely a characteristic of the closed-mouth sad faces in the NimStim set in general, since similar false alarms were present in Tottenham et al. (2009). The total proportion of responses was also similar between EMPATH and the human subjects, which suggests that the model was sensitive to many of the same facial features that the human subjects used for classification.

Performance for individual filters

The results for the individual filters demonstrate that recognition performance decreased as the filter orientations approached 90°. Figures 5a and 5b illustrate the performances of humans and of our model, respectively.

Both sets of confusion matrices distinctly show greater occurrences of misses and false alarms at orientations near the vertical; i.e. 60°, -60°, and 90°. Some other general trends can be drawn from the data. For both humans and the model, sad faces tended to draw more false alarms and misses, regardless of the filter condition. Angry expressions tended to lose their uniqueness as the filters neared vertical, resulting in many misses, and few hits and false alarms.

One informative visualization of recognition performance is plotting the d prime calculations for each filter, which indicates how strong a signal is in relation to surrounding noise (Abdi, 2010). Hence, the d prime calculation for each filtering condition is proportional to how recognizable the expression is with that filter. Figures 6a and 6b depict graphs of d primes for each filtering condition normalized to the d prime of the unfiltered images (higher d primes still correlate to higher recognition performance).

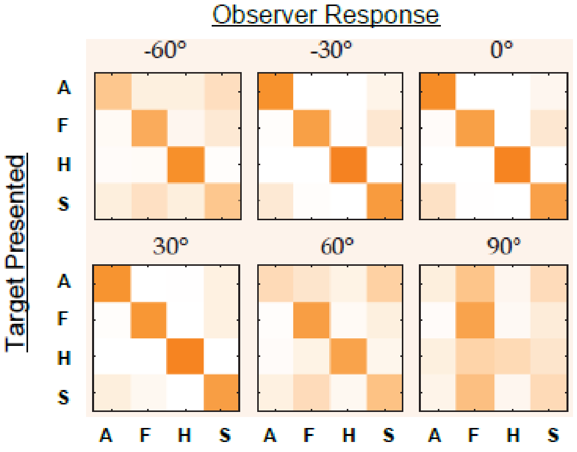


Figure 5a: Human performance for each filtering condition. From Yu, Chai, & Chung (2011).

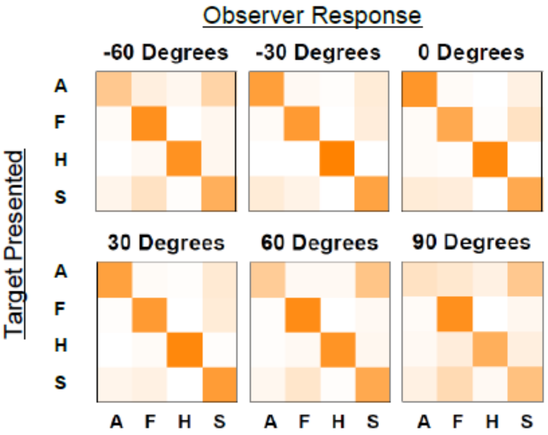


Figure 5b: EMPATH performances for each filtering condition.

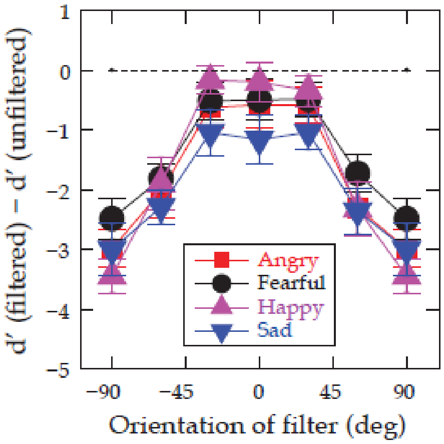


Figure 6a: Human data from Yu, Chai, & Chung (2011), showing d primes of each filtering condition, normalized to the unfiltered condition. Note that data for -90° was copied from data for 90°.

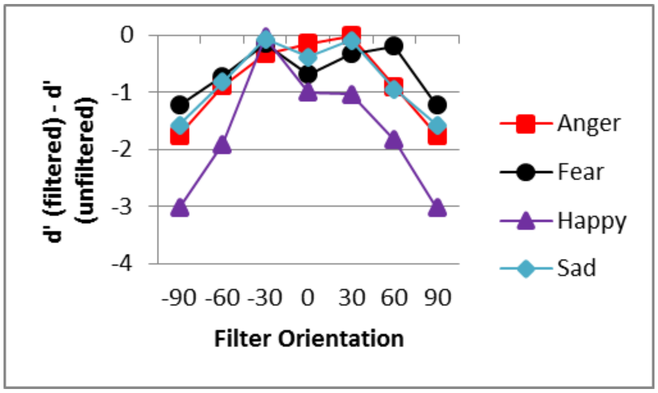


Figure 6b: Data from the EMPATH model showing d primes of each filter normalized to the unfiltered images.

Both d prime charts demonstrate lower recognition performance as the filters approached 90°. We note that results from the EMPATH model are not as symmetrical as those from the human data (e.g. the discrepancy between the values for -30 and 30 degrees for happiness, and the M-shaped graph for fear). This is likely due to the fact that the original images are not vertically balanced; i.e. the positive and negative filters each obstruct slightly different features of the expressions. The resulting images were likely different enough to confuse the network. It would be worthwhile in the future to explore this phenomenon of asymmetry, especially since it was not apparent in the human data.

As noted earlier, fearful faces were less affected by the filter orientations as the other three emotions. Both d prime charts show that fear was the most easily recognized at the 90° orientation. The earlier confusion matrices (Figures 5a and 5b) likewise depict a relatively steady percentage of hits for fearful faces. Much of this is attributed to the fact that both human observers and EMPATH exhibited a significant bias towards fear at the vertical orientations, which increased the occurrences of both hits and false alarms. Figure 7 illustrates EMPATH's high proportion of responses for fear at the vertical orientations. At the horizontal orientations, each emotion constituted close to 25% of the responses, but at the vertical orientation, responses in favor of fear approached 40%.

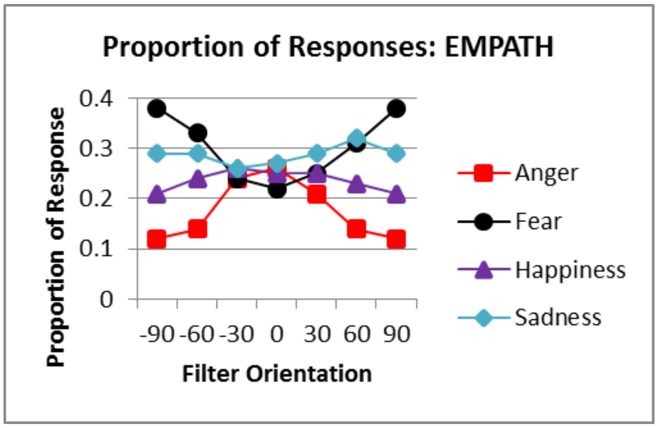


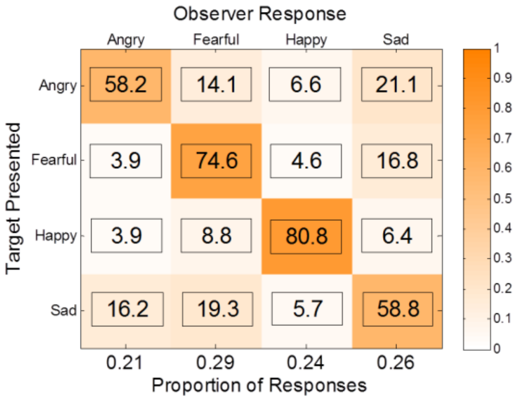
Figure 7: Total proportion of responses exhibited by EMPATH for each filter orientation.

Aggregate Performance of 6 filters

EMPATH's cumulative performance across all 6 filters is also similar to that of the human observers. Tables 1a and 1b show the overall performance of the human subjects and of EMPATH, respectively.

Firstly, the two tables show that with the exception of anger, EMPATH does significantly better in recognizing expressions than do the human subjects. Of course, the model can always be tailored to perform either better or worse, but we did not want to make adjustments just to suit these images. Secondly, the two tables depict many similar trends in the responses between the humans and the model. The overall proportion of responses shows that both the humans and the model were biased towards fear and sadness, and that both were biased against anger. In both cases, anger was the most difficult expression to recognize and also the most difficult with which to be confused, based on the overall percentage of responses. The human subjects and EMPATH also had difficulty recognizing sad faces, but there was a high false-alarm rate as well. The Spearman rank correlation between the two matrices was very good, at r = 0.976 (p < 0.001) for the complete matrices. Since we were also interested in the misses and false alarms, we also calculated the rank just for the off-diagonals, which was very similar at r = 0.942 (p < 0.001).

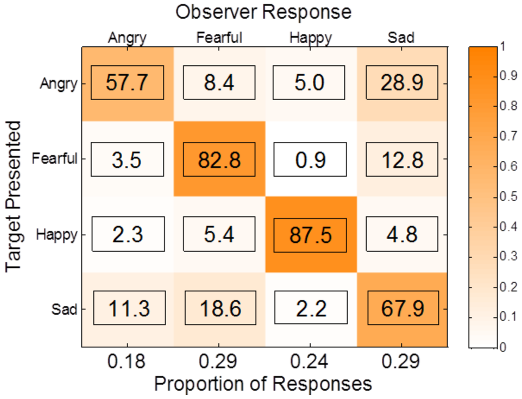
Table 1a: Aggregate performance of human subjects for all 6 filter orientations.



	Observer Response			
Target Presented	Angry	Fearful	Happy	Sad
Angry	58.2	14.1	6.6	21.1
Fearful	3.9	74.6	4.6	16.8
Happy	3.9	8.8	80.8	6.4
Sad	16.2	19.3	5.7	58.8
Proportion of Responses	0.21	0.29	0.24	0.26

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Table 1b: Aggregate performace of EMPATH for all 6 filter orientations



	Observer Response			
Target Presented	Angry	Fearful	Happy	Sad
Angry	57.7	8.4	5.0	28.9
Fearful	3.5	82.8	0.9	12.8
Happy	2.3	5.4	87.5	4.8
Sad	11.3	18.6	2.2	67.9
Proportion of Responses	0.18	0.29	0.24	0.29

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Discussion

The aim of this present experiment was to model the experiment conducted by Yu, Chai, & Chung (2011) and determine if their results can be replicated using a neural network that was not specifically tuned to their images or their data. Our results demonstrate a strong similarity to the pattern of human responses, particularly in showing that information for facial expressions lies primarily on the horizontal orientation, with the modest exception of fearful faces, which solicited heavy bias from both human observers and EMPATH as the orientation approached vertical. In particular, there was a very high proportion of hits and false alarms, suggesting that the vertical filter accentuated features in other expressions normally attributed to fear. Based on this data, it seems that much of the information for fear lies on the vertical, making it distinct from other expressions. It would be interesting to conduct further experiments with other image sets to determine if this phenomenon is a trait of the NimStim images or if it is more universal.

Some discussion regarding our use of d prime is warranted. The procedure used by Yu, Chai, & Chung (2011) to calculate d prime follows the standard guideline of $d' = Z_H - Z_{FA}$, where Z_H and Z_{FA} denote the inverse Gaussian distribution of hits and false alarms, respectively (Abdi, 2010). However, since this formula is typically used for two-way "Yes – No" tasks, the validity of using it for a four-way forced choice is debatable, since each emotion has one "Yes" response and three distinct "No" responses attached to it. Very little literature exist detailing d prime calculations for multiple-way forced choice scenarios, but Alexander (2006) described an easily-computed approximation to the original version in Green & Swets (1966). Based on that, we have recalculated our graph of d prime, which we depict in Figure 8. It should be noted that this approximation does not take false alarms into account. This resulted in a significantly higher d prime for fearful expressions, which were actually greater for filtering conditions near vertical than for unfiltered images. Given that this is an approximation, the validity may of course also be debated, but we nonetheless present both calculations. This serves as a prediction of how the Yu, Chai, & Chung (2011) data will look if it were analyzed in the same way.

Given EMPATH's demonstrated consistency in modeling human face recognition, another possible future experiment could be to determine which filtering orientations are ideal for recognition of each particular expression. It seems, based on this study, that the majority of expressions with the exception of fear would have an ideal filtering condition near the horizontal, but determining exact orientations would form testable hypotheses generated by the model.

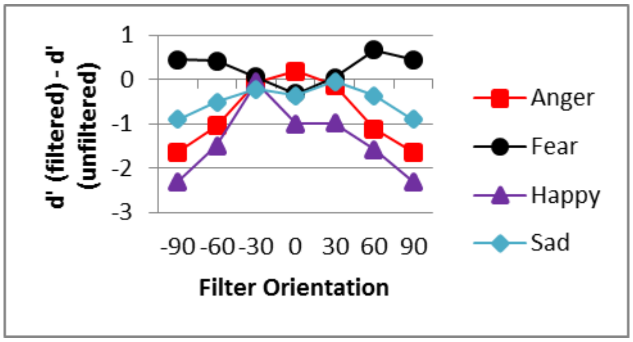


Figure 8: N-choice d prime approximations, following procedure outlined by Alexander (2006) and normalized to the unfiltered images.

Acknowledgements

We thank Dion Yu and Susana Chung for sharing their filtering code, and for providing us with ample clarification of their experiment. We thank Nim Tottenham for sharing the NimStim dataset with us. We also thank Gary's Unbelievable Research Unit (GURU) for providing feedback on this project, and Matthew Tong in particular for overseeing part of the reconstruction of EMPATH. This work was supported in part by NSF grant #SBE-0542013.

References

Abdi, H. (2010). Signal detection theory (SDT). In Peterson, P.L. Baker, E., & B. McGaw (Eds.). *International Encyclopedia of Education.* New York: Elsevier.

Alexander, J.R.M. (2006) An approximation to d' for n-alternative forced choice. University of Tasmania technical report, available at eprints.utas.edu.au.

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

Costen N.P., Parker D.M., & Craw I (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Percept Psychophys, 58*, 602–612.

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

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

Ekman, P., & Friesen, W. (1976). *Pictures of facial affect.* Palo Alto, CA: Consulting Psychologist Press.

Gold J., Bennett P.J., & Sekuler A.B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Res*, *39*, 3537–3560.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Nasanen R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Res*, *39*, 3824–3833.

Williams, N.R., Willenbockel, V. & Gauthier, I. (2009). Sensitivity to spatial frequency and orientation content is not specific to face perception. *Vision Res*, *49*, 2353–2362.

Tottenham N., Tanaka J.W., Leon A.C., McCarry T., Nurse M., Hare T.A., Marcus D.J., Westerlund A., Casey B.J., Nelson C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, *168*, 242-249.

Yu D., Chai A., & Chung S.T.L. (2011). *Orientation information in encoding facial expressions.* Poster presented at the Vision Sciences Society 2011 Annual Meeting, Naples, Florida.

Yu D. & Chung S.T.L. (2011). Critical orientation for face identification in central vision loss. *Optometry and Vision Science*, *88*, 724-732.