5-30-2017

Using Computational Models to Understand ASD Facial Expression Recognition Patterns

Irene L. Feng
Dartmouth College

Using Computational Models to Understand ASD Facial Expression Recognition Patterns

Irene Feng
Dartmouth College Computer Science
Technical Report TR2017-819

May 30, 2017

1 Abstract

Recent advances in computer vision have led to interest in studying how computer vision can simulate our own perception to better understand the intricacies of human neurobiology [1]. Researchers have made strides in computer vision to imitate many facets of human perception, such as object detection, character recognition, and face identification [45]. However, there have been fewer studies that try to model atypical human perception [4]. My thesis focuses specifically on individuals with Autism Spectrum Disorder (ASD) and their deficit in the facial expression recognition (FER) task. I built multiple computer vision models using hand-crafted features and also convolutional neural network architectures to explain the differences of facial expression recognition between typically developing (TD) individuals and individuals with ASD. The models I created that resembled varying levels of configural processing support the hypothesis that diminished configural processing contributes to the FER deficit in individuals with ASD. The models that resembled different areas focus do not support the hypothesis that eye-avoidance and therefore focus on the bottom half of the face contributes to the FER deficit in individuals with ASD.

2 Literature Review - FER in autism

2.1 FER patterns in autism

The ability to recognize various facial expressions is extremely important for social interaction. Individuals with ASD exhibit a variety of deficits in social behavior, including recognizing facial expressions. As of 2012, ASD is reported to affect 1 in 68 individuals [5]. Therefore, many studies have been devoted to discovering the reasons and exact nature of the FER deficits exhibited by people with ASD. In this section, I will review several studies that seek to uncover the nature of facial expression recognition patterns in autism.

Uljarevic and Hamilton conducted a meta-analysis summarizing 48 facial emotion studies comparing ASD to TD subjects [40]. Across these studies, there is a deficit in ASD accuracies compared to TD accuracies on the facial expression task that is robust to age and IQ. Accounting even for publication bias (the event in which studies with null findings are often not published), this correlation was still significant. In looking at specific emotions, there is no reliable difficulty in the recognition of happiness for those with ASD across the studies in the meta analysis. When mean effect size across other emotions are compared to happiness' mean effect, only the recognition of fear was significantly worse than the recognition of happiness.

Despite the conclusive reports from this meta-analysis, most studies report widely varying and even contradictory findings in autistic FER - some report overall deficits in expression [3, 28] but some do not report a deficit in any emotions tested [30, 31]. Some also report deficits in only a combination of negative-valence emotions (anger, sadness, disgust, and fear). Table 1 shows that many studies do not exactly agree on which emotions individuals with ASD have trouble recognizing. However, on the whole, more studies report that negative-valence emotions are more difficult for individuals to identify than non-negative emotions (happy, neutral, surprise). Other studies measured response time as well as accuracy and found a longer response latency during FER tasks [33, 3], while some did not [28, 30]. Other studies also examined the recognition of *intensity* of facial expression. One study found an FER impairment on emotions expressed at low intensities, in which the emotion is subtle and not exaggerated by the poser, but not at high intensities, in which the emotion is quite exaggerated [37]. Competing views are further discussed in another meta-analysis of autistic FER studies [32].

The widely varying conclusions from these studies may be due to demographic discrepancies of

Table 1: Summary of studies which conducted the emotion-labeling task comparing ASD and TD individuals. x's indicate that an FER deficit for that emotion was observed in the study. -'s indicate that the study did not measure this emotion. The total for each row shows the number of studies reporting a group deficit in that emotion over the number of studies that tested for this emotion.

Emotion/study	[34]	[41]	[35]	[37]	[28]	[29]1	[29] 2	[39]	Total	Percent
Anger	x	x				x	x	x	5/8	62.5%
Disgust		x	x		-			x	3/7	42.9%
Sad		x	x	x	x	x		x	6/8	75.0%
Fear		x	x	x		x			5/8	62.5%
Surprise					-		-		0/6	0%
Neutral	-					-	-	x	1/5	20.0%
Happy					x			x	2/8	25.0%

participants across studies in age and IQ, as well as different task types [32]. Therefore, the study of autistic FER in this paper will be narrowed in two ways. First, this paper will only test the recognition of the 6 basic emotions (anger, disgust, fear, happiness, sadness, and surprise) defined by Ekman [25] and the neutral expression. Second, this paper will restrict its focus to studies that perform the emotion labeling task, in which participants are given an image or video sequence as a stimulus, and then choose from a list of written emotions which one best matches the emotion expressed by the stimulus.

In order to understand face processing mechanisms in individuals with ASD in more detail, it is important not to look at just the accuracy and latency of certain intensities of emotions, but also error patterns. That is, we need to determine whether people with ASD have *distinct* patterns in which emotions tend to get confused. Wingenbach et al. cites that at high expression intensity, people with ASD mistake fear as surprise more than TD individuals, although this is a mistake made in both groups [37, 18]. The expression of fear and surprise indeed only differs by one small local feature of the face: the inner brow is lowered in fear expressions and not in surprise [37]. The other features in the eyes and mouth are mostly the same. On the contrary, other studies report this common confusion of surprise for fear in people with ASD but find that people with ASD did not mistake fear as surprise as often as TD individuals. Rather, individuals with ASD mistake fear as disgust [41] or anger [35] more than being mistaken as surprise. This discrepancy may be because the Wingenbach study used videos as stimuli. Therefore, TD individuals may be better at discriminating fear when it is in motion while ASD individuals may not benefit from motion context - nonetheless the studies agree that fear is an emotion that prompts atypical response in people with ASD.

Other confusions found in the literature are that individuals with ASD confused sad as neutral expressions [35, 28], sad as angry expressions, and angry as disgusted expressions [36], while TD individuals confuse the converse: disgusted as angry expressions [18]. They also confuse happy as neutral expressions and neutral faces as negatively-valanced emotions [28] more often than controls.

To further examine overall differences between FER in TD and ASD individuals, I aggregated results from several papers listed above - those specifically testing accuracies in the emotion labeling task [34, 41, 28, 35, 37, 29]. Figure 1a shows these weighted mean accuracies added from these respective papers. (See Appendix A for constructed emotion confusions for the studies.) According to this analysis, anger, disgust, fear, and sadness are harder for individuals with ASD to recognize than happy, neutral, and surprised expressions. The former expressions show at least a

10% accuracy deficit when comparing ASD to TD individuals, while the latter expressions have less than a 10% accuracy deficit. All expression deficits are statistically significant. Figure 1b shows that when grouped by valence, the group difference between ASD and TD individuals in accuracy across negative-valenced emotions is 13.51% (71.72% ASD to 87.39% TD) while the group difference across non-negative emotions (happy, surprise, neutral) is only 4.87% (87.39% ASD to 92.26% TD). This suggests that in emotion labeling tasks, a deficit in negative-valence emotions may be the main contributor to an overall deficit in facial expression recognition for people with ASD.

2.2 FER in autism theories

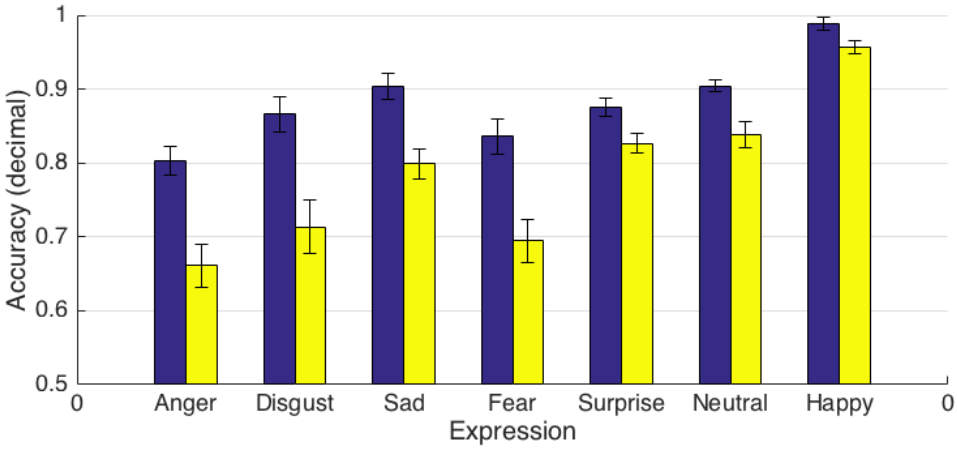
This section will cover theories in the literature that explain the FER deficit in individuals with ASD. Two theories covered in this paper that can potentially explain these deficits are the configural processing theory and the eye-avoidance theory.

2.2.1 Configural Processing

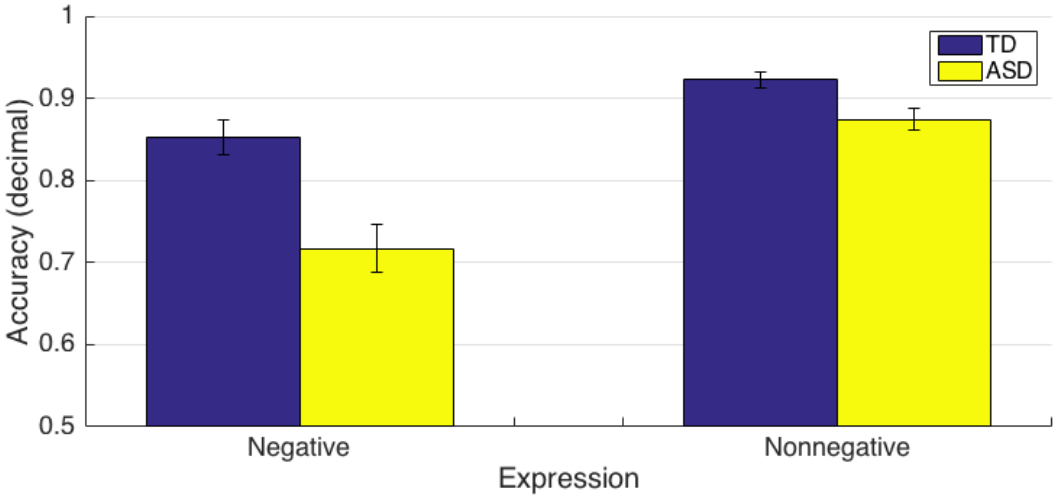
Processing of visual information has classically been segmented into featural processing and configural processing. In faces, featural processing refers to understanding the structure and shape of individual components of a face, independent of their relations to other features. On the other hand, configural processing (also referred to as holistic processing[1]) considers the relations among individual features, such as their spatial arrangement and relative shapes [20]. A seminal study by Tanaka and Farah revealed that a part of a face will be more easily recognized in the whole face than as an isolated part compared to recognition of the parts and wholes of other kinds of stimuli (inverted faces, scrambled faces, and houses) [22]. This suggests that configural processing is especially important in recognizing facial features, which in turn impacts facial expression recognition.

Many studies report that configural processing is not as prevalent in individuals with ASD than TD individuals. In various types of object recognition tasks, individuals with ASD disproportionately use local, feature-based processing, possibly to compensate for having diminished configural processing. For example, participants with ASD have better performance/faster reaction times on block design and the embedded figures task (EFT - recognizing a small geometric shape within a complex image), which require more local processing than global and holistic processing [32, 17]. People who score higher on the autism-spectrum quotient (AQ) have a lower accuracy than people with lower AQ scores on scene categorization, which requires more use of holistic processing than object categorization [19]. Furthermore, while individuals with ASD often perform *better* than TD individuals in (non-social) local visual search tasks, they often have difficulties with global visual tasks such as identifying the direction of motion of multiple dots [16]. In a study comparing ASD and TD performance of both identity and expression recognition (only happy, unhappy, fear, and anger), removing only the mouth and the brows from the pictures affects individuals with ASD much more than TD individuals, although they perform similarly when asking to discriminate full pictures of faces [14]. This indicates that individuals with ASD rely more on specific featural processing rather than holistic processing, especially in the facial expression recognition task.

It will be important to validate the hypothesis that atypical configural processing contributes to atypical FER in individuals with ASD. One behavioral marker to utilize is the inversion effect, i.e. the impaired ability to recognize upside-down faces versus upright faces. Studies have shown that the inversion effect occurs because inverted faces do not stimulate configural processes while upright faces do [21, 13]. An experiment showed that individuals with autism recognized the emotions (happy, unhappy, anger, fear) of inverted faces better than control participants [14]. Another experiment found individuals with ASD still performed worse on inverted faces than upright faces,



(a)



(b)

Figure 1: Weighted means for TD and ASD individuals in 6 FER studies: [34, 41, 28, 35, 37, 29]. (a) shows the weighted mean accuracies of each emotion were generated by multiplying the reported mean accuracies for each emotion and the total number of trials per emotion for each study, and then taking the sum of these numbers for each study divided by the total number of trials across all studies. Some studies did not test all seven expressions, and are accordingly left off when aggregating for that specific emotion. Weighted standard errors (SEs) are also shown, calculated in the same way as the means. (b) shows the weighted mean of emotion accuracies grouped by valence. Weighted standard errors (SEs) are also shown.

but they were not as impacted as controls [35]. These studies show that individuals with ASD have less of an inversion effect than controls, supporting the hypothesis that they use decreased configural processing in the FER task, if at all.

2.2.2 Eye-Avoidance/Amygdala Theory

Though the studies above suggest some sort of atypical configural processing in FER, individuals with ASD may also have featural processing deficits unrelated to configural processing. Some ASD researchers have presented the eye-avoidance hypothesis to account for the specific deficits in recognizing basic emotional expressions with negative valence. Several studies have reported that individuals with autism have reduced attention to the core features of the face, such as the eyes and nose, relative to typical individuals [15], and look at the mouth and other (cheek) regions more than the eye region in FER tasks [34].

Reduced attention to the eyes may lead to deficits in recognizing negatively-valanced emotions, as studies have demonstrated that emotions are differentially expressed on specific regions of the face. A study reported that TD individuals can recognize anger, fear, and sadness more from the top half of the face (recognizable-top expressions), happiness and disgust from the bottom half of the face (recognizable-bottom expressions), and surprise equally from its top and bottom sections [13]. (Note that disgust is a negative emotion, but is more recognizable from the bottom half of the face in this study. However, it is the only emotion in which its valence is not in line with the top/bottom half recognition results from the study above). Moreover, it is well-known in the literature that fear specifically is best recognized from the eyes [12], and results from the Dalton et al. study indicated that ASD children looked significantly less in the eye region when fear is presented [9].

These findings support that the eye-avoidance hypothesis, which claims that reduced attention to the eyes accounts for the specific deficits for people with ASD in recognizing negatively-valanced emotions. But why would this be so? Possibly, individuals with ASD find the mouth more perceptually salient than the eyes, suggesting a low-level (bottom-up) account of the FER deficit. However, it has been suggested that atypical eye-gaze patterns in ASD may reflect top-down differences of eye gaze, rather than bottom-up reasons [10]. More specifically, the specific deficits in recognizing anger, fear, and sadness may not be just explained by a perceptual low-level difference; there may be additional higher-level processing stages in facial expression recognition that impact this difference in focus. Researchers suggest that the brains of people with autism treat facial information differently, even when their visual focus is the same [59].

Some researchers have posited that the behavioral patterns of eye-avoidance in people with ASD are caused by atypical amygdala function in the brain during FER tasks. The amygdala functions in perceiving and controlling emotions, such as controlling aggression in events perceived as threatening [11]. When the amygdala is damaged, individuals have trouble with specifying fear as well as negative basic emotions in general in expression tasks [41].

ASD participants show decreased activation of the amygdala during the processing of negative emotional expressions [41], but at the same time, the magnitude of amygdala activation in people with ASD was found to be positively correlated with time spent looking at the eye region of the face [9]. Therefore, to explain these two observations, researchers posit that when viewing emotional eyes, the amygdala is stimulated more for people with ASD, triggering a fear response, and so eye-avoidance is habitually learned [60, 9]. This behavioral pattern then may contribute to the deficit of recognizing anger, fear, and sadness.

3 Literature Review - FER in Computer Vision

Now that both the specific differences in FER that individuals with ASD exhibit and theories that explain these differences have been covered, I review current computer vision methods that tackle facial expression recognition tasks. In the following methods section, I will explain how I used these existing computer vision methods to test the competing theories in the ASD literature.

3.1 Non-neural network methods

There are many machine learning methods adapted for automatic facial expression recognition (see [23], [24], and [58] for in-depth surveys of computer vision expression recognition methods). Many feature extractors use features that are geometric components such as the shapes of the facial features and the location of facial salient points. These features, called Action Units (AUs) correlate with the Facial Action Coding System (FACS), which encodes muscle groups in the face [27]. These AUs can be grouped together in order to classify facial emotions [27, 23].

Other feature extractors use appearance features representing the facial texture, including wrinkles, bulges, and furrows. There are two types of appearance features that are popularly used for facial expression recognition: local binary patterns and Gabor filters. A local binary pattern labels a neighborhood (of block size nxn, where n is an odd number) of pixels in a binary fashion, with the threshold value c as the center pixels value such that the label of pixel (x, y) in the black is determined by the ternary expression $(x, y) >= c\ ?\ 1 : 0$. The results of each neighborhood is considered as a p-bit binary number, in which p is the size of the neighborhood (if the block size is 3x3, p is 8, because there are 9 pixels in the block and 8 excluding the center pixel). Therefore, we can store all the local binary patterns of each neighborhood into a 2^p-bin histogram, which can then be seen as a type of texture encoder: it detects edges, spots, flat areas, etc [7].

Gabor filters are also used to extract appearance features for facial expression recognition. Gabor filters are complex sine waves restricted by a Gaussian window. The filters have varying frequency and orientations that also act as edge and texture detectors when convolved with an image. A group of Gabor filters with different frequencies and orientations convolved at one location in an image is called a Gabor jet [42]. A strong response (large amplitude) of a filter convolved at a certain location indicates an edge or texture in that location at that filter's frequency (size) and orientation. One advantage of Gabor filters is that the filter responses are robust to small changes in overall illumination [24, 64]. Importantly, Figure 2 shows that Gabor jets accurately model lower-level human and animal vision [64, 43].

Once features are extracted, Principal Component Analysis (PCA) is often used for further dimensionality reduction, as the facial features described above have very high dimensionality and can result in overfitting unless a very large dataset is available. PCA is a linear, unsupervised technique which finds a new orthogonal basis using eigenvalue decomposition. Each basis vector is called a principal component. Using any subset of the principal components that have the largest eigenvalues will capture the most variation out of any other same-sized subset of vectors, thus making PCA a useful tool to reduce the dimensionality of the data while preserving their (linear) relationships. PCA has been used as a dimensionality reduction technique for the FER task [64]. In one study, 60 principal components were chosen from each training image represented by 1470 Gabor filters (5 sizes at 6 orientations = 30 filter responses for each point on a 7x7 grid overlaid on each face) reached maximal generalization accuracy of 92% on the JAFFE dataset, and 75% on the POFA dataset (see Table 2 for a list of databases). This shows that PCA can be used to reduce dimensionality almost 25-fold. After feature selection, both Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs) have been used as discrete classifiers for facial expression

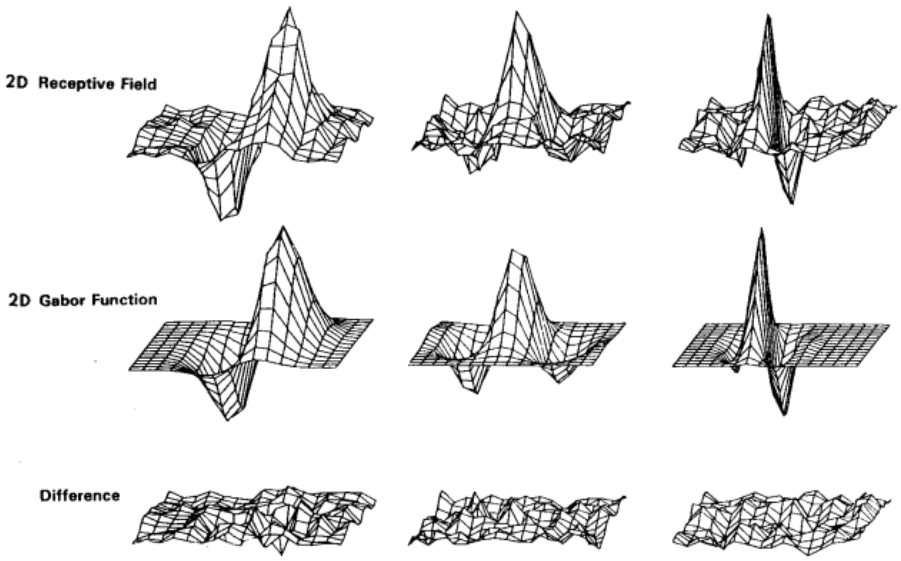


Figure 2: Reproduced from Figure 5 in [43]. Top row: illustrations of 2-D receptive field profiles in simple cells of the cat visual cortex. Middle row: best-fitting 2-D Gabor elementary function for each neuron. Bottom row: non-significant residual error of the fit.

recognition [64, 6, 57].

3.2 Deep convolutional neural networks

Gabor wavelets in feature extraction capture local variations of the inputs as well as some global features due to overlapping filters. However, these features are directly concatenated for final expression recognition, with no further mechanism to learn higher-level representations [8]. Deep neural networks can learn higher-level representations due to having many layers that are able to apply nonlinear functions on previous layers. Therefore, they would be interesting to use in building an ASD-like FER system because the causes of expression recognition deficits in ASD individuals may be primarily at the higher-level.

Any neural network with nonlinear activation functions is a universal function approximator: a machine learning model that can approximate any continuous function of the input (if allowed sufficient number of features). But historically, training those features was not efficient and it was very prone to overfitting, even with the advent of the backpropagation algorithm in 1986 [45]. In the 2010's, neural networks regained popularity in solving computer vision problems because of the emergence of the deep convolutional neural network (CNN) as well as large datasets to train them on. The deep CNN architecture has multiple convolutional layers, which applies a convolutional filter(s) across training images to detect features. This models low level processes in human vision by simulating the receptive fields of neurons. Then fully connected layers, which receive the outputs from the last of these convolutional layers as inputs, resemble higher level processes that occur in human and primate brains [45].

Calculating Gabor filters and applying a discriminative classifier over them to determine emotion is essentially a two-layer network: the Gabor filters act as a convolutional layer, and the discrete classifier (LDA or SVM in this case) acts as the fully connected layer. This shallow network aptly resembles the low-level processes of human vision, but may not resemble the high-level processes as adequately as deeper networks. Therefore, the ability of deep CNNs to model a task that requires

a complex interaction between low-level perception and high-level nonlinear relationships, such as those underlying face expression recognition, may render them a better model to simulate expression classification patterns in ASD than previously hand-crafted classification methods.

One deeper network that solves the facial expression task uses a deep convolutional neural network architecture with "AU-aware" layers as called by the authors. The output from a convolutional layer is used to learn a higher-representation by feeding it into 3 layers of Restricted Boltzmann Machines (RBMs). The final RBM layer is fed into a linear SVM classifier that classifies the six basic expressions defined by Ekman and Friesen as categories, which gets a 92.04% validation accuracy when using the CK+ database as the training and validation set and 74.76% when using MMI (see Table 2 for databases) [8].

Another network model consists of the same convolutional layers, followed by multi-layered structures called Inception layers [62]. Inception layers, coined in [66], increase the depth and width of networks while retaining the efficiency of computing and training dense layers. Both increasing the depth of networks by simply increasing the number of layers and increasing the width of networks by adding more parameters to certain layers may lead to overfitting and increased use of computational resources, so Inception layers are a solution to this problem. Instead of having only one-sized filter in a convolutional layer, an Inception layer concatenates the activation outputs of a group of varying-size convolutional filters (typically 1x1, 3x3, and 5x5), as well as an average or max pooling layer ([66] uses max pooling). This concatenation of layers increases the width of this Inception layer. This method of concatenating varying sizes of convolutional filters for feature selection is similar to the method of using Gabor jets from multiple Gabor filters at different frequencies and orientations. Moreover, the depth of the network is increased by adding 1x1 convolutional filters before the larger convolutional filters (3x3 and 5x5).[2] GoogLeNet has achieved extremely successful results in object detection tasks by utilizing multiple Inception layers in between convolution and fully connected layers [45, 66].

The Inception layers described in [62] take inputs from the traditional convolutional layers, and their output is fed to fully-connected layers. The resulting network has a 93.2% generalization accuracy on categorizing the 7 facial expressions when using part of the CK+ database as validation and 77.6% with part of the MMI dataset as validation (multiple datasets are used for training, see Table 3).

4 Methods

Returning to the motivations of my own study, I built computational models that simulate each of the theories about FER in autism. These models tested the following two hypotheses. (1) The configural processing hypothesis: diminished configural processing in individuals with ASD leads to a comparative deficit in identifying facial expressions, especially negative ones. (2) The eye-avoidance hypothesis: eye-avoidance in individuals with ASD contributes to a comparative deficit in identifying facial expressions, especially negative ones.

4.1 Dataset

An ideal dataset to test these hypotheses is a large enough database in the computer vision literature that has enough examples to train a neural network classifier, which is approximately 1 million images, but also is used in experiments testing facial expression recognition in individuals with ASD. I did not find a publicly available database that fit these criteria. Therefore, I created my own dataset from multiple facial expression databases to train on. Table 2 shows all the datasets I surveyed - the datasets I used in this paper are listed above the midline.

Table 2: Summary of facial expression databases surveyed. The column *Results of Human Data* is a summary of ASD deficits found in the previous column (*Human Data Available*). The aggregate dataset was created from images from datasets above the midline. Ones below the line are not included for various reasons: the dataset was too small (JAFFE), it did not have studies comparing ASD to TD individuals (RAFD and SFEW), or the images were labeled by AU but not expression (DISFA), or it needed payment (POFA).

Dataset	Media Type	Color	Size	# exemplars per emotion	# emotions	Human Data Available	Results of Human Data	Used in Computer Vision	Readily usable?
ADFES [48]	videos	color	720x576	22	10 (+ pride, contempt, embarrassment)	TD/ASD confusions [37]	Negative Valence, esp. Fear		
CK+ [49]	videos	grayscale	640x490	∼25-80	7 (- neutral, + contempt)	TD/ASD mean accuracies [34]	Only Anger	[62, 6, 7, 8]	Needs Cropping
Dartmouth[3]	photos	color	2340x2340	100	9 (+ calm, contempt)	TD confusions	No comparison to ASD		
FER2013 [50]	photos	grayscale	48x48	∼500-8900	7			[62]	
KDEF [52]	photos	color	562x762	140	7	TD/ASD mean accuracies [39, 41]	Overall [39], Negative [41]		
MMI [51]	videos	color	∼128128	∼590-1100	7			[62, 8]	Needs Cropping
JAFFE [65]	photos	grayscale	256x256	∼10	7			[64]	Needs Cropping
RAFD [53]	photos	color	681x1024	67	8 (+contempt)	ASD mean accuracies [38]	No comparison to TD		
SFEW [54]	photos	color	143x181	∼100	7			[62]	
DISFA [55]	videos	color	1024x768	27	7			[62]	Only AU
POFA (EFAS) [26]	photos	grayscale	720x576	∼13	7	TD/ASD confusions [35]			Needs purchase

4.1.1 non-CNN dataset

I created two separate datasets: the non-CNN dataset was used for training non-neural network models, and the CNN dataset was used for training deep CNNs. For the non-CNN dataset, I used images from the ADFES, CK+, Dartmouth, KDEF, and MMI databases. The images extracted from the CK+ and MMI databases were not uniformly aligned, so as a preprocessing step I cropped the faces using an automatic face-recognition method described in [2] (code can be found on OpenCV).

4.1.2 CNN dataset

For the CNN dataset, all datasets from the non-CNN dataset are included along with the FER2013 dataset. I used the FER2013 dataset only in training the neural networks because the dataset has the most images, but has the smallest resolution (48x48). This way, I could train the neural network with a dataset containing a considerable number (∼45,000) of images.[4]

I exclude FER2013 images from the non-CNN dataset because this dataset is noisy: the highest accuracy achieved on FER2013 is 71%, and even human performance is estimated to only be between 65% and 68% [56]. This is due to different orientations of faces (profile versus front-face orientation), partial faces, and mislabeled images. Also, with the exclusion of the FER2013 dataset, the images can be scaled to sizes of 128x128 as opposed to 48x48 to retain more information. Non-CNN methods do not need as many images to train, so without the FER2013 dataset, there is a total of 8024 images used for creating a non-CNN dataset. I resized all the images to 128x128 and 48x48 for the non-CNN and CNN datasets respectively, and converted them to grayscale. The number of images per each category is given in Table 4.

4.2 Non-CNN methods

In designing my non-CNN classifier, I chose to use appearance-based features over geometric features. Geometric features are difficult to extract in low resolutions [7]. Also, the salient facial points have to be manually located for extraction of geometric features [46]. Many of the datasets listed in Table 2 do not align their faces uniformly, and if they do, they can still fail to keep face positions consistent [23]. For example, if all photos are aligned in the eyes, that does not help in automatically locating the outline of the mouth because it can be in an open-o shape or a closed-line shape.

Additionally, I used Gabor filters instead of local binary patterns as appearance-based features. Although extracting LBPs is computationally faster and achieves just as good results as using Gabor filters [6, 7], I used Gabor filters because they model the visual cortex of humans, which is relevant to the question I am considering in this paper.

Most importantly, the Gabor jet representation of an image also captures human configural processing by simulating the overlap of the receptive fields (RFs) of neurons in human vision [67]. In a study by Xu et al., it was found that the interaction between local features and the contextual face background can be picked up by overlapping Gabor filters, especially by those with larger receptive fields [67].

Choosing the size and orientation of Gabor filters to use requires a number of design choices. We have to choose how many (and which) sizes and how many orientations of Gabor filters we want to take at each location, and how many locations of the facial image we will sample from. I followed the parameters in Lades et al. [42], in which a 2-D Gabor filter is modeled by the Gabor

Table 3: Number of images used for training the network in [62]. AN=Anger, DI=Disgust, FE=Fear, HA=Happy, NE=Neutral, SA=Sadness, SU=Surprise. There is a significantly unbalanced number of images in each class with more happy and neutral faces, and an undertraining of anger, fear, and sadness.

Database	AN	DI	FE	HA	NE	SA	SU	Total
Multipie	0	22696	0	47338	114305	0	19817	204156
MMI	1959	1517	1313	785	0	2169	1746	9489
CK+	45	59	25	69	0	28	83	309
DISFA	436	5326	4073	28404	48582	1024	1365	89210
FERA	1681	0	1467	1882	0	2115	0	7145
SFEW	104	81	90	112	98	92	86	663
FER2013	4953	547	5121	8989	6198	6077	4002	35887
Total	9178	30226	12089	87579	169183	11505	27099	346859

Table 4: Number of images for each dataset used in this paper. AN=Anger, DI=Disgust, FE=Fear, HA=Happy, NE=Neutral, SA=Sad, SU=Surprise. The NN set has extra samples of disgust images from the MMI and ADFES datasets and the number of happy examples in the FER2013 database was subsampled from 8989 to 6198 so that all emotions are less than 1.5 standard deviations from the mean number of exemplars per emotion ($\mu = 6435, \sigma = 889.44$). The non-CNN set excludes images from the FER2013 dataset, and keeps only some of the over-sampled disgust images.

CNN/non-CNN	Database	AN	DI	FE	HA	NE	SA	SU	Total
CNN	MMI	825	2365	725	1100	708	800	1075	7598
	CK+	45	59	25	69	0	28	83	309
	FER2013	4953	547	5121	6198	6198	6077	4002	33096
	KDEF	140	140	140	140	140	140	140	980
	ADFES	22	2222	22	22	22	22	21	2353
	Dartmouth	100	99	100	100	100	100	100	699
	Total	6085	5432	6133	7629	7168	7167	5421	**45035**
non-CNN	MMI	118	**194**	725	1100	708	800	1075	5427
	CK+	45	59	25	69	0	28	83	309
	KDEF	140	140	140	140	140	140	140	980
	ADFES	22	**478**	22	22	22	22	21	609
	Dartmouth	100	99	100	100	100	100	100	699
	Total	1132	970	1012	1431	970	1090	1419	**8024**

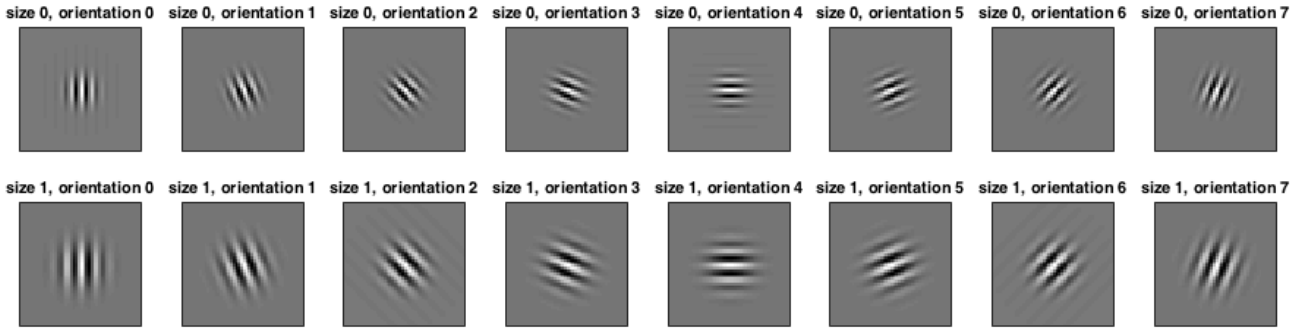


Figure 3: A sample of Gabor wavelets created by function $\Psi_k(\mathbf{x})$. Here, the wavelets are shrunk to fit the page, but $\mathbf{x}$ ranges from (-64,-64) to (63,63). $\mathbf{k}$ is determined from sizes $v = 0, 1$ (32 and 23 cycles per image), and $\omega = 0...7$ ($0, \frac{\pi}{8}, \frac{2\pi}{8}, ...\frac{7\pi}{8}$ radians).

function Ψ_k:

$$\Psi_k(\mathbf{x}) = \frac{\mathbf{k^2}}{\sigma^2} exp\left(\frac{-\mathbf{k^2 x^2}}{\sigma^2}\right)\left[exp(i\mathbf{k}\cdot\mathbf{x}) - exp\left(\frac{-\sigma^2}{2}\right)\right]$$

where $\mathbf{x}$ is a location in the 2D lattice over which the filter is defined and (0,0) indicates the center of the filter. $\mathbf{k}$ is a wave vector which defines the frequency and orientation of the Gaussian filter. The $exp(\frac{-\sigma^2}{2})$ term is subtracted to render the filters insensitive to illumination of the image. Following Lades et al., the hyperparameter $\sigma = 2\pi$. Forty Gabor filters are defined by $\mathbf{k} = (k_v \cos k_\omega, k_v \sin k_\omega)$, chosen from a range of 8 orientations and 5 frequencies. k_ω determines the orientation of the filter: $k_\omega = \omega * \frac{\pi}{8}$ for $\omega = 0, 1, 2, ...7$, which values correspond to $0, \frac{\pi}{8}, \frac{2\pi}{8}, ...\frac{7\pi}{8}$ radians. k_v determines the frequency: $k_v = \frac{\pi}{2*(\sqrt{2}^v)}$ for $v = 0...4$. The values of $v = 0, 1, 2, 3$ and 4 correspond approximately to 32, 23, 16, 11, and 8 cycles per image, respectively.

Figure 3 and Figure 4 both visualize Gabor filters in the spatial domain. For efficient calculation, I did not convolve the Gabor filters with the image but rather converted both the image and the Gabor filter specified by Ψ_k into the Fourier domain and multiplied them to get the filter response in the Fourier domain, and then performed inverse Fourier transform on the filter response. I used both the magnitude and the phase of the response (the absolute value and angle of the complex result, respectively) from each point on a grid of size 11x11.[5]

4.2.1 Dimensionality Reduction

Because using the magnitude and phase of an 11x11 grid of Gabor filter responses with (say) 5 different sizes and 8 different orientations of filters results in each facial image being represented by $2 * 11 * 11 * 5 * 8 = 9680$ features, and the dataset used for non-CNN feature extraction has around 8000 images to use for testing and training, PCA was performed after feature extraction to avoid overfitting.

The number of PCA components p was chosen by selecting the "elbow" point at which the proportion of variance captured by taking the first p components with the largest corresponding eigenvalues stops significantly increasing.

Figure 4: A sample image from the non-CNN image dataset with two jet filters overlaid in two positions. The filter on the higher grid point is the smallest size $v = 0$ (32 cycles/image), and orientation $\omega = 0$ (0 radians). The filter on the lower grid point has size $v = 2$ (16 cycles/image), and its orientation is $\omega = 5$ ($\frac{5\pi}{8}$). The grid is 11x11 with 11 pixels between each grid point.

4.2.2 Classifiers

As mentioned above, both Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs) have been used as discrete classifiers for facial expression recognition [64, 6, 57]. Therefore, both LDA and SVM models were used as classifiers to compare accuracies in this analysis.

Like PCA, LDA finds a new linear basis underlying the data, but in contrast to PCA, LDA is a supervised method that tries to model the difference between and within the classes of data in the new linear subspace. In the case of multi-class classification with c classes, LDA projects data onto a maximum $(c-1)$-dimensional linear subspace that minimizes within-class scatter of data points and maximizes between-class scatter [44]. To perform classification, a test x_t can be projected into this subspace as well as all of the class means. The projected class mean that has the smallest Euclidean distance from this test point is this point's predicted class.

SVM is another popular method of classification which determines a linear separating hyperplane between two classes with a maximal margin (called the geometric margin) from any point in the training data (kernel functions can be applied to create a nonlinear boundary, but in this analysis it is restricted to be linear). Both LDA and SVM classification is done in MATLAB (to create a multiclass SVM, the one-against-all method is used).

4.2.3 Cross-Validation

Although Lades et al. use 5 different sizes and 8 orientations [42] for feature extractors in FER, other papers use fewer sizes and orientations [64, 57]. To determine how many different filters (number of sizes and orientations) gives optimal performance for my dataset, 5-fold cross-validation was used to test the performance of the number of sizes (from 2 to 5 sizes, i.e. adding sizes starting from $v = 0$ [32 cycles/image], so that 2 sizes will have as the largest filter size $v = 1$ [23 cycles/image] and 5 sizes will have $v = 4$ [8 cycles per image] as the largest filter size) and number of different orientations (6 and 8, i.e. all angles from 0 to π radians, π not inclusive, with a spacing of $\frac{\pi}{6}$ and $\frac{\pi}{8}$, respectively) of the filters for both LDA and SVM classifiers. The entire training set (7221 images out of 8024, the other 803 images were used for a validation set) was randomized and split into 5 chunks. For each combination of a unique number of sizes and orientations, I created 5 separate models in which four chunks are used to train the model and the remaining chunk is used as the testing set to calculate the error for this model. The errors are averaged to get the mean cross-validation error for each unique combination of sizes and orientations. The number of sizes

Table 5: Comparison of cross-validation scores of PCA with LDA for whole faces testing different sizes and orientations. The number of sizes is selected from increasing values of v and ω starting at 0, e.g. 2 sizes: $v = 0, 1$; 3 sizes: $v = 0, 1, 2$; and 6 orientations: $\omega = 0, 1, ...5$. The size and orientation selection that had the highest accuracy is bolded.

sizes/orientations	6 orientations	8 orientations
2 sizes	0.7661	0.7858
3 sizes	0.7772	**0.7912**
4 sizes	0.7668	0.7883
5 sizes	0.7615	0.7761

and orientations with the lowest error was selected for both LDA and SVM models (both preceded by PCA).

Table 5 shows that 3 sizes (32, 23, and 16 cycles per image) and 8 orientations ($0, \frac{\pi}{8}, \frac{2\pi}{8}...\frac{7\pi}{8}$ radians) got the best score for the LDA+PCA model, with an 79.12% overall accuracy on the validation set with dimensions reduced down to 159 principal components. See Appendix B for the accuracy scores of the SVM model, as well as other hyperparameters tested.

4.2.4 Hypotheses applied to non-CNN methods

To test the configural processing hypothesis, I trained two different classifiers with different-sized Gabor filters. Because Gabor filters with larger, overlapping receptive fields play a more prominent role in configural processing than smaller Gabor filters, I trained one classifier that uses less configural processing by selecting the two smallest filters, and one classifier that uses more configural processing, which selects the smallest filter and the largest filter for classification. The classifier trained with filters having smaller receptive fields is called the local model, and the one trained with both the smallest and the largest frequency filters is called the global model. The results from the global model and local model should resemble TD and ASD responses for FER, respectively, to validate the configural processing hypothesis.

To test the eye-avoidance hypothesis, I tested whether more focus on the mouth rather than the eyes contributes to the FER patterns found in individuals with ASD. I simulated mouth and eye focus by spatially-blurring the top half and bottom half of faces, respectively. Additionally, a better simulation is to implement foveated vision with Gabor filters, in which different-sized Gabor filters will be taken relative to a given point of focus, explained later in the paper. I also trained separate computational models by training only on Gabor filters located in the top and bottom halves of the images in the non-CNN dataset. The results from the top-focus models and bottom-focus models should resemble TD and ASD responses for FER, respectively, to validate the eye-avoidance hypothesis.

4.3 CNN methods

The pre-trained neural network weights were acquired from [62]. Because the pre-trained network has been trained on an imbalanced dataset, there is a significant bias in the dataset towards predicting happy and neutral images (see Appendix C). Therefore, I fine-tuned this pre-trained neural network using the dataset I described in section 4.1.2, which is more balanced in the number of exemplars per class. I split the CNN dataset of 45k images into a validation set of 1,805 images, 4% of the entire dataset. Table 6 shows all the learning parameters I used for fine-tuning (these learning parameters were used for all subsequent trainings, unless otherwise specified). I fine-tuned

Table 6: Learning parameters used in caffe neural network library [63] for fine-tuning networks used in this study.

parameter	
base_lr	0.01
lr_policy	step
stepsize	320000
gamma	0.96
momentum	0.9
weight_decay	0.0002
max_iterations	1000000

the weights of the network for 1 million iterations, updating weights with a momentum of 0.9 and a base learning rate of 0.01, dropping the learning rate by a factor of gamma = 0.96 every 320k iterations. The overall accuracy of this network trained on whole faces is 75.46%.[6] See Appendix D for the performance graph over training iterations for this network.

Figure 5 shows the detailed confusion matrix of the fine-tuned network. Interestingly, the network does exhibit a confusion for fear as surprise, which is found in both TD and ASD individuals (and in my analysis, more prominent in individuals with ASD: see Appendix A). There is a confusion of anger as disgust which is found in ASD individuals, but there is almost no confusion for disgust as anger, which is found in TD individuals [18]. There is a bias towards neutral faces that is not present in either TD or ASD individuals, however. We can conclude that the way the neural network is categorizing emotions is more different from TD individuals compared to ASD individuals, but also has unique confusions of its own.

4.3.1 Hypotheses applied to CNNs

To test the configural processing hypothesis, I trained a deep convolutional neural network that is more robust to the inversion effect compared to the CNN above. By making a neural network more robust to the inversion effect, it will force the neural network to rely less on configural processing in performing FER on upright faces. This was done by taking the CNN I trained on the original dataset of facial expressions and generating another CNN by fine-tuning it on randomly inverted pictures. Then I tested both CNN's again on upright pictures to determine if these responses resemble TD and ASD responses for FER tasks.

To test the eye-avoidance hypothesis, I fine-tuned two pairs of CNNs. First, I trained a pair of models on two differently blurred imagesets. I also trained another pair of models on imagesets with the top or bottom half replaced with all black pixels to simulate the most rudimental method of testing top half/bottom half focus. I will compare the performance of the top model and bottom model for each pair to determine if these results resemble TD and ASD responses for FER tasks.

5 Analysis for Hypothesis 1: Configural Processing

5.1 Non-CNN: local vs global Gabor filters

I compared diminished configural processing with more use of configural processing by training local and global models in the non-CNN context, which differ by the size of Gabor filters used in feature extraction. The local model only selects the smallest 2 filters out of 3 sizes of filters in the LDA model, i.e. $v = 0, 1$ (approximately 32 and 23 cycles/image), while the global model only

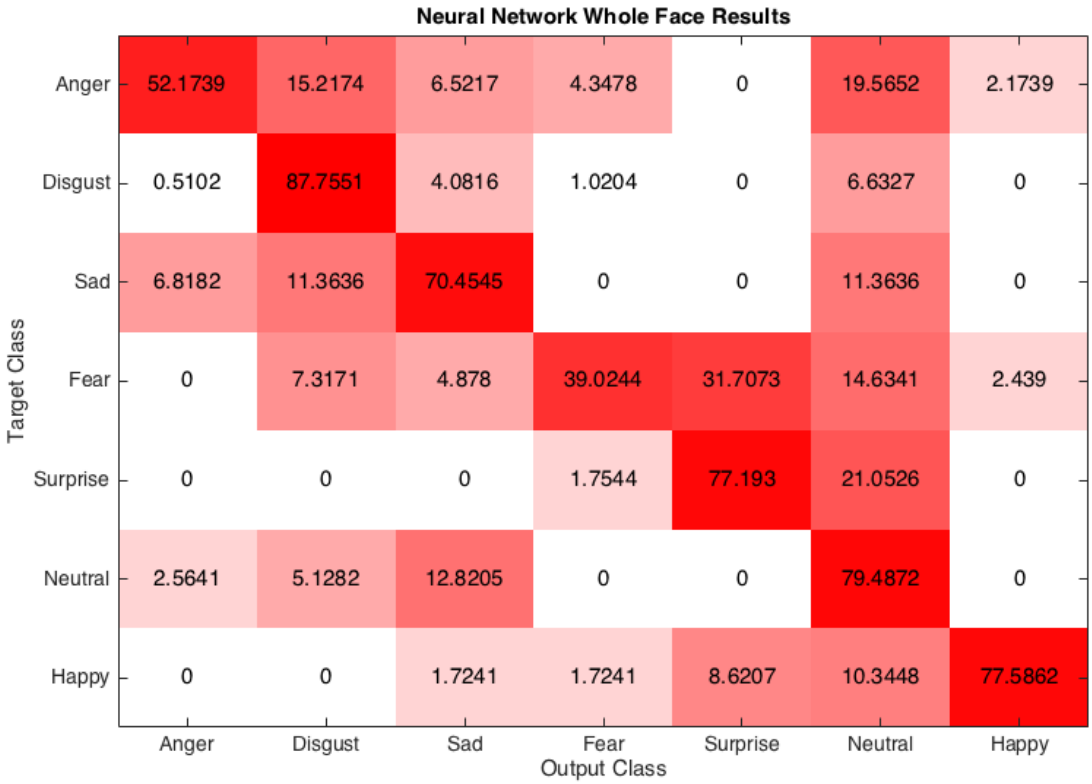


Figure 5: Confusion matrix on fine-tuned neural network on whole faces. The target class represents the stimuli expression, and the output class represents the network's prediction. For example, this network shows a confusion of predicting surprise for fear stimuli.

Table 7: Performance and 95% confidence intervals of the local and global LDA + PCA models. 159 components were taken for the models, accounting for 79.5% of the variance of the local filters and 81.5% of the variance of the global filters.

	PCA+LDA	95% CI
Global	0.7945	0.7649 - 0.8219
Local	0.7858	0.7558 - 0.8137
Local - Global	-0.0087	

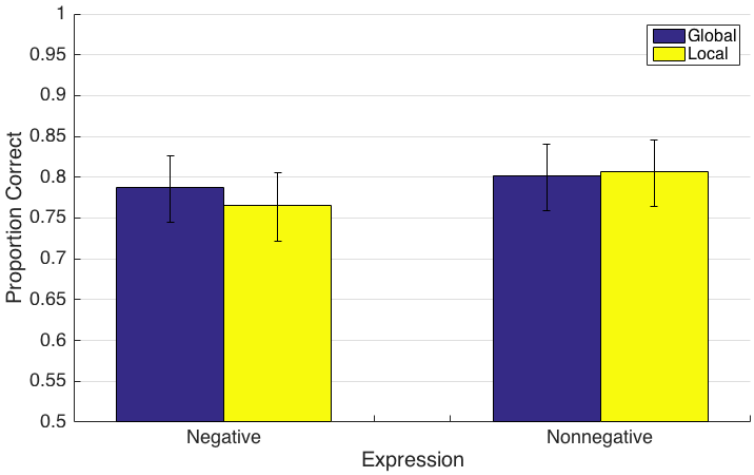


Figure 6: Comparing mean emotion accuracies grouped by valence between local and global models. 95% confidence intervals are graphed as error bars.

selects the smallest filter and the largest filter, i.e. $v = 0, 2$ (32 and 16 cycles/image). PCA reduced the number of features from 2*11*11*2*8 = 3872 features to 159 features, and LDA was performed on the training set in this reduced dimensional-space.

5.1.1 Results

Table 7 shows the overall performance of training these two separate models. There is a slight overall (<1%) deficit in the local modal versus the global model, but this difference is not significant, judging by the substantial overlap in the 95% confidence intervals. Figure 6 shows the mean accuracies for both negative and nonnegative emotions with 95% confidence intervals based on a binomial distribution on each mean. The difference between the local and global model in mean accuracy across negative-valenced emotions is 2.17% (76.57% local to 78.74% global), while the difference across non-negative emotions is only 0.51% (80.21.% local to 80.72% global). However, this deficit in negative-valenced emotions vs non-negative emotions is not statistically significant. (See Appendix E for the difference in confusions of local and global models.)

5.1.2 Discussion

Configural processing is a promising explanation in the FER deficit in ASD individuals. The deficits in discerning negative emotions in the literature are reflected when more local Gabor filters are used compared to more global Gabor filters. There is also substantially less of a deficit in the local model

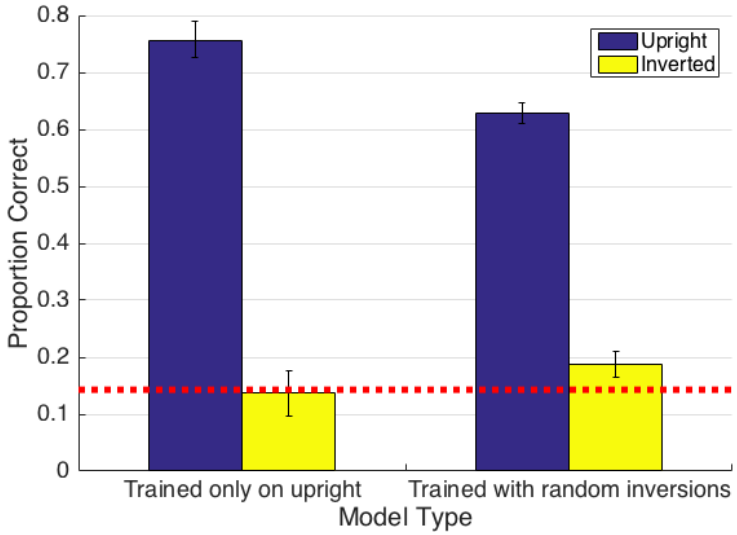


Figure 7: Bar plot of performance in testing two neural networks on upright and inverted faces: one only trained on upright faces, and one trained on randomly inverted ($p = 0.5$) faces. 95% confidence intervals are plotted. Chance performance (0.1428 correct) is plotted as a dotted red line.

(in fact, there is a slight advantage) for detecting nonnegative emotions. However, these results are not statistically significant.

5.2 CNN: inversion fine-tuning

In addition to using a non-CNN method to test the configural processing hypothesis, I trained a CNN that is more robust to the inversion effect, as individuals with ASD are. I took the CNN I previously fine-tuned on the original dataset of facial expressions, described in section 4.3, and generated another CNN by fine-tuning it on the same dataset but with randomly (probability 50%) inverted pictures. I compared each CNN's performance on both upright and inverted images. Then I tested this CNN on upright pictures to determine if these responses resemble ASD responses for the FER task.

5.2.1 Results

Figure 7 shows the difference in performance between the CNNs (the original fine-tuned CNN, and the one fine-tuned from it using randomly inverted faces) both on upright and inverted faces. The neural network trained only on upright faces has a 13.72% accuracy when tested on only inverted faces, which is right below chance (14.28%). Meanwhile, the neural network trained on randomly inverted faces has an accuracy of 18.83% when tested on only inverted faces. On upright faces, this neural network has an accuracy of 62.8%, compared to the 75.46% accuracy of the CNN previously trained in section 4.3. The upright accuracy deficit between the inverted-trained network and the upright-only-trained network is statistically significant.

Figure 8 shows the mean accuracy for both negative and nonnegative emotions between the two models tested on upright faces. For negative emotions, the upright model has a 15.11% higher accuracy than the inverted model (74.31% compared to a 59.20% accuracy). For nonnegative

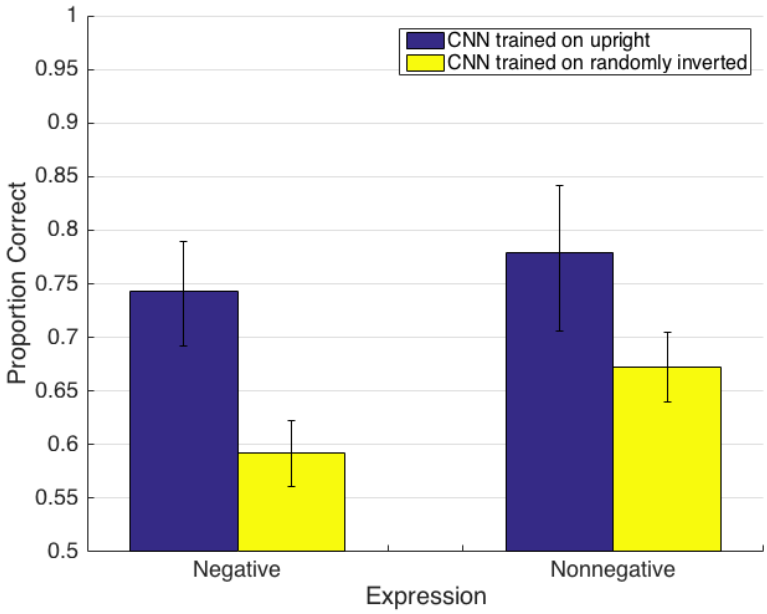


Figure 8: Comparing mean emotion accuracies grouped by valence between the CNN trained only on upright images and a CNN trained trained on randomly inverted (50% probability) faces.

emotions, the upright model has a 10.64% higher accuracy than the inverted model (77.92% to 67.28%). Both of these differences are statistically significant.

5.2.2 Discussion

As expected, training the neural network on random inversions does increase the accuracy of the neural network on inverted faces, but this difference is not statistically significant. Therefore, no further analysis was done on the inverted performance of either network. Training on inverted faces impaired the network's performance in classifying both negative and non-negative emotions on upright faces, as expected. The deficit in classifying negative emotions also is more pronounced compared to the deficit for nonnegative emotions, which agrees with the literature for the difference between ASD and TD individuals in identifying negative and nonnegative emotions.

5.3 Summary of all configural processing tests

Both non-CNN and CNN methods of simulating varying amounts of configural processing show promising evidence that diminished configural processing may account for the negative-valence FER deficit seen in individuals with ASD. However, the negative-valence FER deficit in the non-CNN method is not statistically significant, while the CNN method is.

Compared to a CNN only trained on upright images, a fine-tuning on randomized inverted facial expressions affected the CNN's performance in a way that matched the FER deficit patterns of individuals with ASD. However, this type of training might simply disadvantage a CNN's performance in general, because the accuracy in identifying inverted faces did not get significantly better. It would be interesting to design more ways to simulate a deficit in configural processing using a neural network that not only removes the use of global information, but better resembles the use of more local information, which could bring out the same FER patterns individuals with

ASD exhibit. A suggestion is to change the neural network architecture by removing a fully connected layer which is very dense and deals with the most high-level abstractions, and adding more convolutional layers instead.

6 Analysis for Hypothesis 2: Eye-Avoidance

6.1 Non-CNN: blurring, foveated, and halves tests

As covered in the literature, people with ASD place more focus on the mouth region than TD individuals. To test the eye-avoidance hypothesis that this mouth focus contributes to the negative-valence FER deficit in ASD, I performed three different non-neural network methods to train three different pairs of models that simulate focus on the top half/bottom half of the face. To model the different fixations on the top half and bottom half of the face, I blurred the imagesets to remove spatial detail from the respective halves of the images, created a foveated model of Gabor filters for feature selection, and also simply took Gabor filters only at top or bottom halves of the 11x11 grid overlaid on each image.

6.1.1 Blurring Results and Analysis

For the blurring analysis, I blurred the bottom and top halves of the non-CNN image set, creating two new image datasets, and then trained two separate LDA models (with PCA as a dimensionality reduction step) as before. To blur one half of each image, I convolved the image with a Gaussian filter of length $l = 15$ pixels using averaged-padding (padding with the image's mean intensity value). The one-dimensional filter can be described as

$$f(x, l) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-.5(l))^2}{2\sigma^2}}$$

evaluated at $x = 1, 2...l$. The two-dimensional filter is the cross-product of the one-dimensional filter with itself. I reconstructed the images to have spatially varying blur by applying filters with different σ values depending on the horizontal location within the image: for the bottom-blurred images, I split every image into 8 horizontal bars, each of height $128/8 = 16$ pixels. Then increasing values of $\sigma = \{2, 4, 6, 8\}$ were used starting from the 5th bar down to the 8th. Larger values of σ indicates a filter that has a stronger blurring effect. For the top-blurred images, I split the image into 8 horizontal bars again and then used the same increasing values of σ, but instead starting from the 4th bar up to the 1st. Figure 9 shows a sample of blurred images.

Table 8 shows the performance results for training Gabor filters on spatially blurred images. The images that were blurred on the top, which simulated bottom focus, were classified more accurately than the images that simulated top focus. This difference is not statistically significant by assessing the overlap of the 95% confidence intervals. The accuracy for the bottom focus model (78.21%) is very close to performance for LDA + PCA given whole faces (78.83%). Figure 10 shows the mean accuracy for both negative and nonnegative emotions between the two models tested on the half-blurred images. For negative emotions, the top focus model has a slightly lower (0.24%) accuracy than the bottom focus model (77.05% compared to a 77.29% accuracy). For nonnegative emotions, the top focus model has a 6.43% lower accuracy than the bottom focus model (72.75% to 79.18%). Both of these differences are not statistically significant.

These results are the opposite of what the eye-avoidance hypothesis would predict when comparing top half to bottom half accuracies. The bottom focus model actually performs better than

Figure 9: Top row: Images with spatially varying blur on the top half (called bottom focus), which models the focus predicted for ASD individuals. Bottom row: the same images with spatially varying blur on the bottom half (called top focus), which models the focus predicted for TD individuals.

Table 8: Performance and 95% confidence intervals for training Gabor filters on different sets of blurred images. 159 components were taken on the PCA step, accounting for 80% of the variance on the training set.

	PCA+LDA	95% CI
Top focus	0.7497	0.7182-0.7793
Bottom focus	0.7821	0.7519-0.8102
Bottom-Top	+.0324	

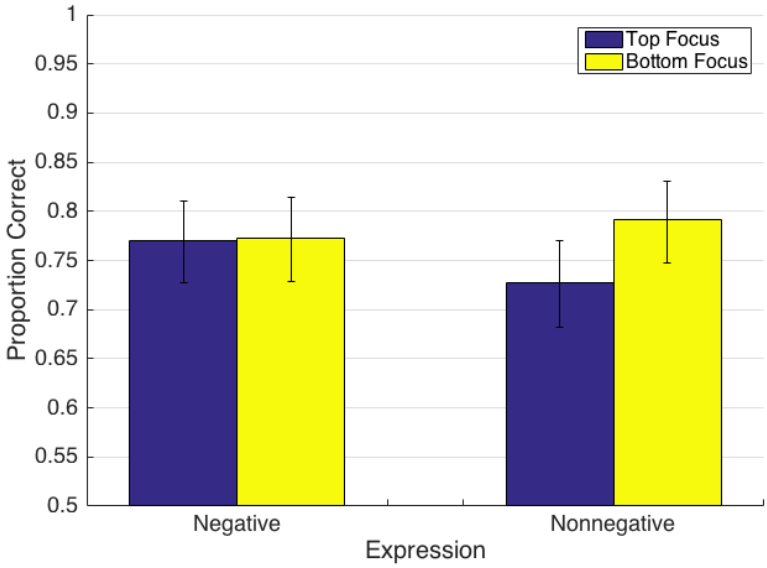


Figure 10: Comparing mean emotion accuracies grouped by valence between training LDA+PCA Gabor jet models on the top/bottom blurred imagesets. 95% confidence intervals are graphed as error bars.

top focus model. This could indicate that ASD individuals' focus on the bottom half of the face does not contribute to their FER deficit, but these results are not statistically significant.

Moreover, the bottom focus model may have higher performance than the top focus model due to the way the spatially varying blur filter was applied to the images. Because the spatial blur gets increasingly blurrier from the center of the images and the eyes in the images were often closer to the center than the mouth was, the eyes may not have been blurred enough while the mouth was sufficiently blurred. Therefore, the model trained on images in which the bottom half was blurred was given less information than the models that were trained on images in which the top half was blurred. Also, because blurring an image effectively reduces its image resolution, the resolution of the image in the regions of blur may have been too low to get responses even from the lowest-frequency (largest) Gabor filters. There is a possibility that effectively, no information from a blurred part of an image could be detected from the Gabor filters. Therefore, the blurred imagesets may not be fairly simulating a top-half focus and a bottom-half focus.

6.1.2 Foveated Results and Analysis

Therefore, for a second analysis, I trained foveated models of the Gabor jets. Foveated image processing attempts to simulate one aspect of human vision, in which focus points of an image will correspond to the center of the retina, the fovea, and therefore spatial resolution decreases as points in the image get more distant from the focus points. This is done by decreasing the resolution in peripheral regions of an image so that when human eyes focus on an image, they cannot distinguish between the original and the foveated versions of that image [61]. I chose one point (h, c) at which there is the most spatial detail - the most Gabor filters are taken at this point. Then, fewer Gabor filters capturing high spatial frequencies are taken at points increasingly distant from (h, c). This reflects the human visual system in which the point (h, c) is a point of focus. For a foveated model simulating top focus, a point is chosen at grid position $(h, c) = (3, 6)$: in an 11x11 grid, this point is centered in the x-direction and is one-fourth from the top of the image in the y-direction. I calculate the radial distance from this point for all other points in the 11x11 grid. I calculated how many filters I would take at each grid point. When choosing from n sizes of filters, the number of filter sizes taken at the point (x, y) is $f(x, y) = n - min(n - 1, \lfloor\sqrt{(x-h)^2 + (y-c)^2}\rfloor)$. The smallest filter size at this grid point is determined by $v_s = n - f(x, y)$.

Figure 11 illustrates the relationship between the radial distance of a point from the focus and the number and size of the filters (i.e. $f(x, y)$ and v_s) taken at that point. As the distance of (x, y) from (h, c) increases and $f(x, y)$ decreases, fewer filters are taken and the filters that are taken are only capturing the lowest spatial frequencies, which do not pick up details at high spatial frequencies. These filters are then concatenated as before to create the feature vector representing the image. For a foveated model simulating bottom focus, the foveated filters are selected the same exact way except the focus is chosen as $(h, c) = (9, 6)$.

Table 9 shows the performance results for training models with foveated filters with top and the bottom focus points. The bottom foveated model performs marginally worse than the top foveated model. The difference is not statistically significant. Figure 12 shows the mean accuracy for both negative and nonnegative emotions between the two models. For negative emotions, the top foveated model has a slightly higher (0.21%) accuracy than the bottom foveated model (69.81% compared to a 68.60% accuracy). For nonnegative emotions, the top foveated model has a slightly lower (0.26%) accuracy than the bottom foveated model (73.26% to 73.52%). Both of these differences are not statistically significant.

Although the top foveated model performs better than the bottom foveated model for negative emotions, which agrees with the literature in comparing TD and ASD individuals, these results are

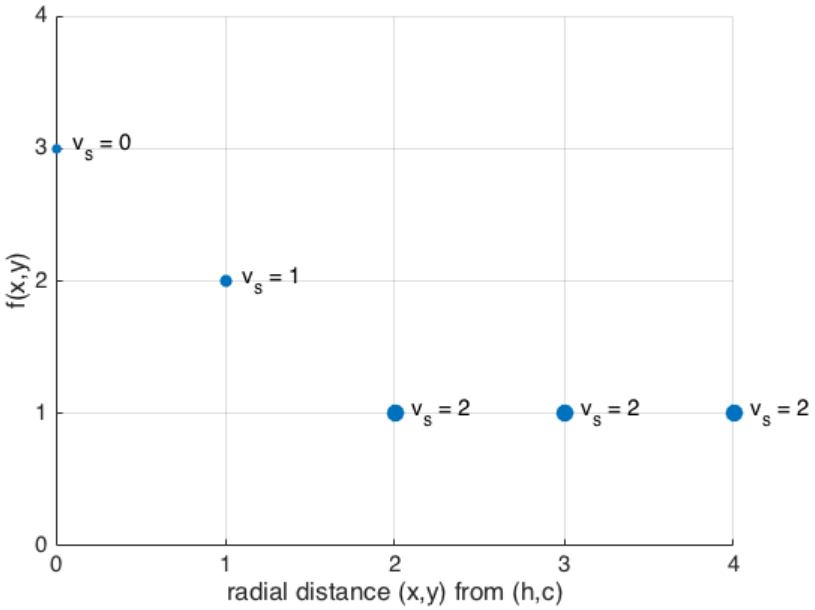


Figure 11: Number of Gabor filters in the foveated models taken out of $n = 3$ filters for a point (x, y). The number of filters taken is determined by the radial distance of the point (x, y) from the point of focus (h, c). v_s, the smallest-sized Gabor filter taken, is also labeled. The size of the points plotted also represents the relative size of the filters for easier visualization.

not statistically significant. Therefore, for my third analysis, I trained two different models to test the eye-avoidance hypothesis in the most rudimental way: one that selected Gabor filters taken only from the first 5 rows in the 11x11 grid, and one that selected Gabor filters from the last 5 rows.

6.1.3 Halves Results and Analysis

Table 10 shows overall performance results for taking half (5 out of 11 rows) of the Gabor jet results. Although there is an overall deficit for bottom trained as compared to top trained models, it is marginal and not statistically significant. Figure 13 shows the mean accuracy for both negative and nonnegative emotions between the two models. (See Appendix E for the difference in confusions of local and global models.) For negative emotions, the top half model has a slightly higher (1.69%) accuracy than the bottom half model (71.74% compared to a 70.05% accuracy). For nonnegative emotions, the top half model has a lower (1.28%) accuracy than the bottom half model (73.78% to 75.06%). Both of these differences are not statistically significant. Although the differences across valence accuracies of the top half and bottom half models are more pronounced than the foveated models, the differences are still not statistically significant.

6.1.4 Discussion on non-CNN methods for the eye-avoidance hypothesis

The results from all three non-CNN methods are inconclusive as to whether eye-avoidance contributes to the FER deficit found in individuals with ASD. As discussed before, blurring the images may inherently give an advantage to the bottom focus model over the top focus model. Therefore, the blurring method may not fairly simulate focus on the top and bottom half of a face. Although

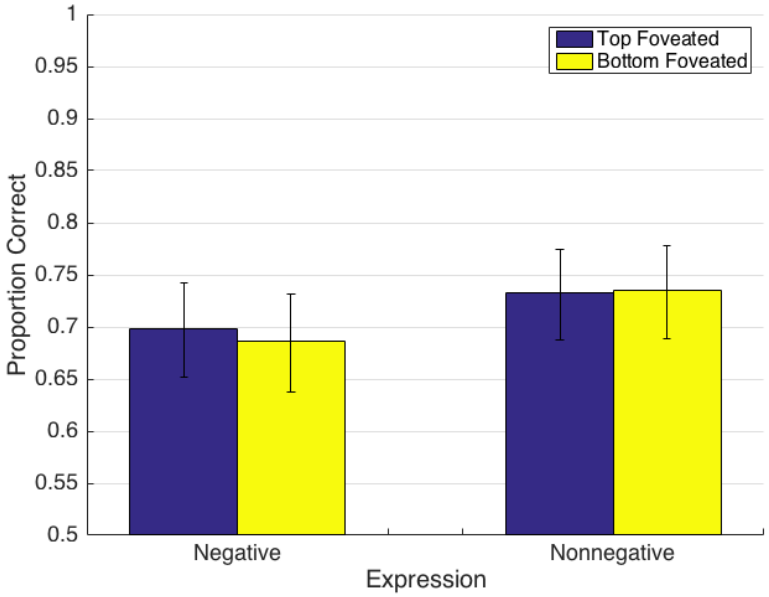


Figure 12: Comparing mean emotion accuracies grouped by valence between training LDA+PCA top-foveated and bottom-foveated models. 95% confidence intervals are graphed as error bars.

Table 9: Overall accuracy and 95% confidence intervals for training foveated models with PCA and LDA on 96 components, accounting for 79.5% of the variance.

	PCA+LDA	95% CI
Top Foveated	0.7148	0.6822-0.7458
Bottom Foveated	0.7098	0.6771-0.7410
Bottom-Top	-0.0050	

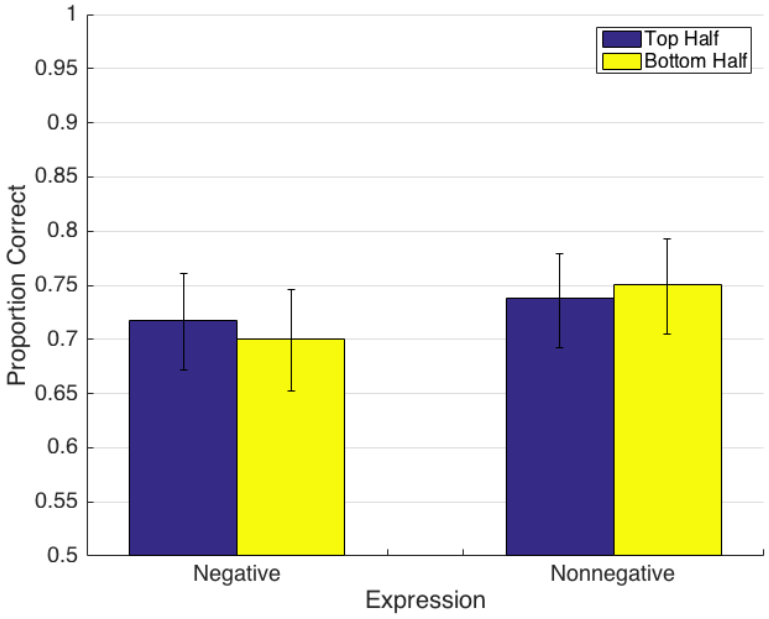


Figure 13: Comparing mean emotion accuracies grouped by valence between training LDA+PCA Gabor jet top-half and bottom-half models. 95% confidence intervals are graphed as error bars.

Table 10: Overall accuracy and 95% confidence intervals of the PCA+LDA models trained on Gabor filters taken at the top half and bottom half of the grid. 111 components were taken for the models, accounting for 81% of the variance.

	PCA+LDA	95% CI
Top Half	0.7273	0.6951-0.7578
Bottom Half	0.7248	0.6925-0.7554
Bottom-Top	-0.0025	

the foveated and halves methods may have been more fair than the blurring method, no significant differences emerged from comparison of the top/bottom models of these two methods. This may have occurred for two reasons: the resolution of the images was too low, and how the top half and bottom half of a face was determined does not exactly match the actual focus on the top half and bottom half of a face.

If the image resolution is too low, information with high spatial frequency is lost. The image resolution of the non-CNN dataset is 128x128 pixels, which is higher than our CNN dataset, but it still is low relative to many images collected in Table 2. Therefore, the models trained on the top halves of the images in all of these analyses may already be at a disadvantage, since the top half of faces contain more high-spatial frequency information that is important to facial expression recognition than the bottom half of faces, such as furrowing of different parts of the eyebrows and squinting and widening of the eyes [27]. Extracting Gabor filters from images of higher resolution may help confirm whether or not focus on the bottom half of the face more than the top half contributes to the FER deficit in individuals with ASD.

Also, what the *top half* and *bottom half* of a face refer to may not be accurately captured in my analyses. In the literature, the upper and lower part of the face refer to the eye and mouth areas, respectively, but the exact regions are not clearly given [34, 27]. Among studies testing FER on the upper and lower halves of the face, there are slight disagreements on where the dividing line is. The line dividing the halves is right at the bridge of the nose in one study [13], and right below the bridge of the nose in another [68]. As previously mentioned, due to the variability of the positions of the facial features in the dataset I used, I chose the top half and the bottom half in terms of image coordinates, which may have had confounding effects on my results. An alternate way to determine halves would be in terms of salient facial points of a face and create a facial grid from which Gabor filters could be extracted (instead of a rectangular grid). However, I have not yet found a robust method that can automatically extract these facial grids.

In the next analyses, I use convolutional neural networks to test the eye-avoidance hypothesis. We should continue to be cautious that image resolution as well as imprecision of the top half and bottom half can also impact the results of the neural network.

6.2 CNN: blurring, halves

I performed two different methods using CNNs to test the eye-avoidance hypothesis. I again trained a pair of models on two differently blurred imagesets. I also trained another pair of models on imagesets with the top or bottom half replaced with all black pixels to simulate the most rudimental method of testing top half/bottom half focus.

6.2.1 Blurring results and analysis

The same Gaussian filter described in section 6.1 (blurring for non-CNN dataset) was used to blur bottom halves and top halves of images for training the CNN, but the values of σ was reduced to $\sigma = \{1, 2, 3, 4\}$ with bar heights of $48/8 = 6$ pixels because the images to train the neural network are smaller. Table 11 shows the overall accuracies of two neural networks fine-tuned from the Mollahosseini network at all layers on blurred images. Like the non-CNN blurred analysis, the model trained on images in which the top half is blurred (bottom focus model) performs better (+4.16%) than the top focus model, but the difference is also not statistically significant due to the overlap of the 95% confidence intervals.

Figure 14b shows the mean accuracy for negative and nonnegative emotions between the two networks. For negative emotions, the top-focus network performs 11.01% worse than the bottom-

Table 11: Overall accuracy and 95% confidence intervals for fine-tuning CNNs on images with different spatial blurs. Training was done for 1 million iterations for each CNN.

	CNN	95% CI
Top Focus	0.6590	0.6251-0.6918
Bottom Focus	0.7004	0.6674-0.7319
Bottom-Top	-0.0416	

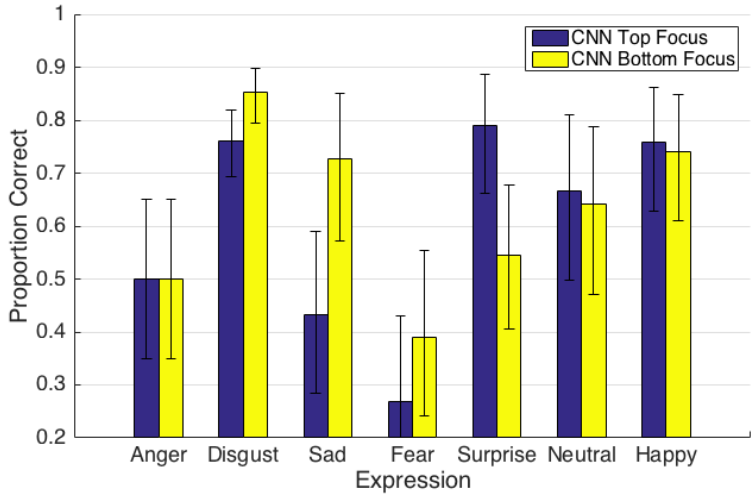
focus model (61.77% compared to a 72.78% accuracy). For nonnegative emotions, the top-focus model has a higher (10.38%) accuracy than the bottom-focus model (74.68% to 64.29%). The difference of the mean accuracies for the models is statistically significant only for negative emotions. Figure 14a shows the results separated by emotion, demonstrating that the main emotion driving the bottom-focus model's better performance on negatively-valenced emotions is sadness.

Like the non-CNN model trained on blurred images, the model trained on blurred images in which the bottom half is in focus performs overall better than the model trained on images in which the top half is in focus. Still, this trend is not statistically significant. However, the CNN method supports the same (though statistically insignificant in the non-CNN method) finding from the non-CNN method on blurred images: that the model trained on bottom-focus images did better on recognizing negative emotions.

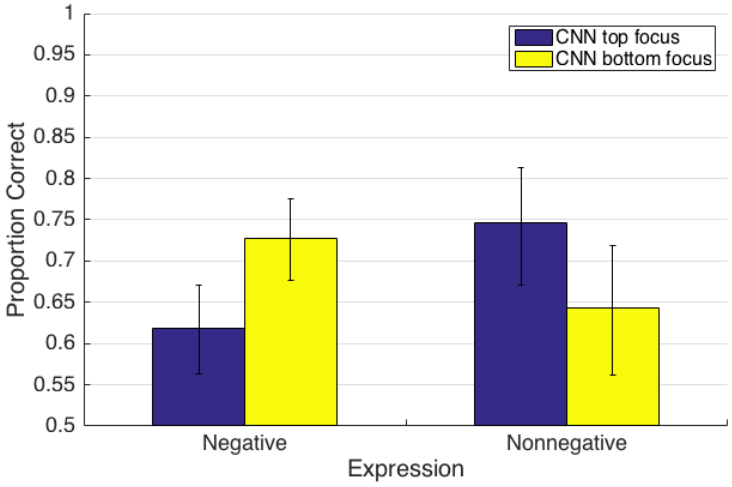
The inherent advantage of there being effectively less blurring in the eyes region than the mouth region when spatially blurring images could still be present in the CNN model as well as in the non-CNN Gabor jet model. If the resolution of the training images for the CNN was the same as the non-CNN dataset (128x128), it would be easier to determine what the effect of blurring is for the amount of useful information on the face by directly comparing neural network blurred performance to non-CNN blurred performance.

Leaving this possibility aside and assuming this way of spatial blurring is an accurate depiction of focus on top and bottom halves of faces then, these results do not support the eye-avoidance hypothesis. Focus on the bottom half of the face is more useful for identifying negative emotions than focus on the top half of the face. These results were similar to ones found in another computer vision study which also extracted Gabor filters from images on the CK+ database, some with the eye-area removed and some with the nose and mouth area removed. The model overall performed worse when the nose-and-mouth area was removed versus when the eyes area was removed [69]. More specifically, the mouth region being removed had more effect on the recognition accuracy rate of anger, fear, happiness and sadness, while the eyes region affected the recognition of disgust and surprise more: the same pictures used in the computer vision study was used in a human study and the same results were achieved. Figure 14a shows that this finding coincides with my own results, at least for sadness and surprise, the emotions of which the difference in accuracy are close to being statistically significant: sadness is better detected from the bottom half of the face, and surprise is better detected from the top half of the face. Therefore, these results do not exactly match what the eye-avoidance hypothesis suggests.

However, though my results (the ones close to statistical significance) agree with both the human and computer model FER results in [69], they do not agree with the human results in the Calder et al. [13], in which anger, fear, and sadness are more easily recognized from the top rather than from the bottom (mouth region). Also, disgust was found to be more of a bottom-recognizable emotion in [13], while in [69], it was more recognized from the eyes. This may be because [69] used images in which only the eye-area/mouth-area was blacked out, while [13] used halves of faces cropped at the bridge of the nose. Therefore, in order to align my study closer to one of these two studies,



(a)



(b)

Figure 14: Mean accuracies of neural networks fine-tuned from Mollahosseini network [62] on blurred images. 95% confidence intervals are graphed as error bars. (a) shows mean accuracies across emotions between CNNs trained on blurred imagesets. (b) compares the emotion accuracies grouped by valence.

Table 12: Overall accuracy and 95% confidence intervals for fine-tuning CNNs on images with either the top half or the bottom half of the image present. Training was done for 320k iterations for each CNN.

	CNN	95% CI
Top Half	0.3992	0.3551-0.4445
Bottom Half	0.4990	0.4534-0.5446
Bottom-Top	-0.0998	

I fine-tuned another pair of CNNs on images in which either the top or bottom half is all black, simulating the experiment done in [13].

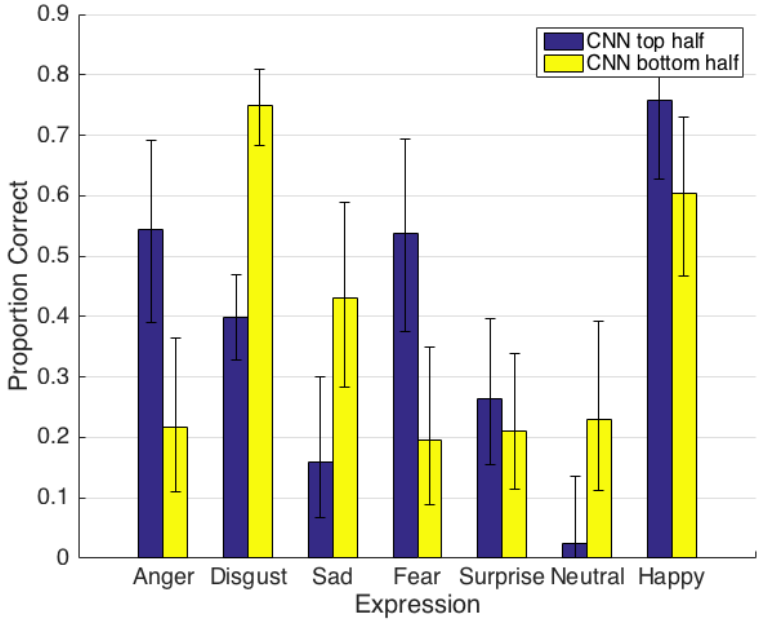
6.2.2 Halves results and analysis

I tested the eye-avoidance hypothesis in the most basic way via neural networks, similar to the non-CNN method of taking Gabor filters from certain rows in a grid. I created two new image sets in which the bottom and top halves of the images are all black. Two CNNs were fine-tuned from the weights of the pre-trained CNN, one trained on the top half (bottom half is black) of the images and one trained on the bottom half (top half is black). Training was done for 320k iterations.

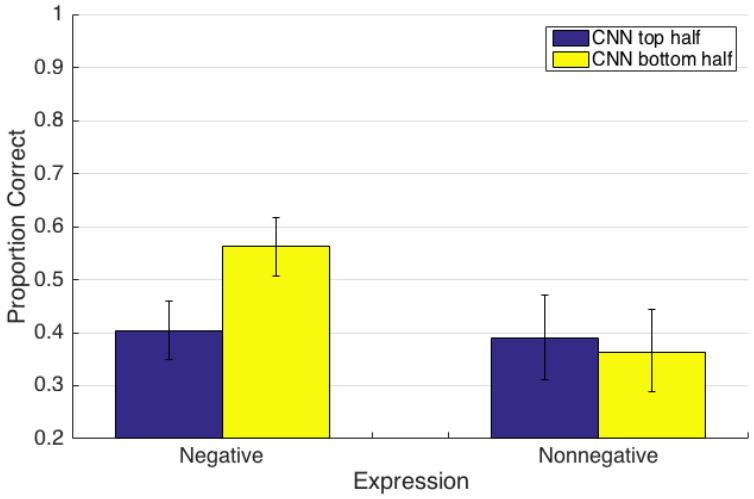
Table 12 shows that the overall performance of the network trained on the top half of images is significantly worse (9.98%) than the bottom half of images. Figure 15b shows the mean accuracy for both models grouped by negative and nonnegative emotions. The difference of the mean accuracies for the models is statistically significant only for negative emotions, in which the top half model has a 40.37% accuracy and the bottom half model has a 56.27% accuracy. However, Figure 15a shows that the CNN trained on top halves is significantly better at detecting anger and surprise, while the CNN trained on bottom halves is better at detecting disgust and sadness (sadness almost reaches significance). There is no significant difference in recognizing surprise and happiness between the two models. There is almost a significant difference in detecting neutral faces, in which the CNN trained on bottom halves performs better. Like the CNNs trained on blurred images, the CNN trained on the bottom half performed overall better on negative emotions than CNN trained on the top half. This is unlike the non-CNN models trained on only top halves and bottom halves of the Gabor jet grid, in which bottom models performed worse, although these results were statistically insignificant.

As expected, the results from fine-tuning the CNN on these imagesets more closely resembles the results from [13] than fine-tuning the CNN on the blurred imagesets. Figure 15a shows that anger and fear are top-recognizable, and disgust is bottom-recognizable. The only near-significant discrepancy is that sadness is bottom-recognizable in the CNN model. A reason for the discrepancy of sadness between my model and what is found in the literature may have to do with the fact that neutral was also a choice for classification while in [13], it was not. Sadness could be confused for neutral faces when only presenting the top half, and therefore may account for the relatively low accuracy on sadness for my model trained on the top halves of images compared to human studies.

The result that sadness is bottom-recognizable in this experiment contributes against the eye-avoidance hypothesis, then. On the whole too, there was no evidence supporting the eye-avoidance hypothesis. Firstly, although the CNN trained on top halves of images does perform marginally better on recognizing nonnegative emotions than the CNN trained on the bottom halves of images, this result is insignificant. Secondly and most substantially, the CNN trained on the top halves of images performs *worse* in recognizing negative emotions, and this result *is* significant.



(a)



(b)

Figure 15: Mean accuracies of neural networks fine-tuned from Mollahosseini network [62] on images showing the top half or the bottom half (the other half of the picture is blacked out). 95% confidence intervals are graphed as error bars. (a) shows mean accuracies across emotions between CNNs trained on half images. (b) compares the emotion accuracies grouped by valence.

6.3 Summary of all eye-avoidance tests

Although no group differences between the top and bottom models in the non-CNN eye-avoidance tests were statistically significant between valence, the CNN methods were statistically significant, and they generally do not support the eye-avoidance hypothesis. The models that simulated focus on the top half of the face actually performed worse on negative emotions than the models that simulated focus on the bottom half of the face.

However, there still may be information not captured by both the non-CNN and CNN methods that humans perceive. In this study, I only performed image manipulations on the CNN methods, simulating merely perceptual differences of individuals with ASD. I did not make any changes to the CNN architecture that would simulate a high-level change in the vision of individuals with ASD. Moreover, although the actual method of feature selection was changed rather than only making a perceptual-level image manipulation in the non-CNN foveated model, non-CNN methods probably do not have the capacity to model higher-order processing that is done in human facial expression recognition.

In addition, while individuals with ASD do focus more on the bottom half of faces, they do not exclusively focus on the bottom half of faces. Moreover, TD individuals do not exclusively focus on the top half of faces. Therefore, the comparisons of models focused on top halves of faces to those focused on the bottom halves of faces may not accurately represent the focus of TD and ASD individuals.

Another way to test the eye-avoidance hypothesis with an emphasis on high-level differences is to freeze the weights of the convolutional layers: this simulates that the low-level perceptual system of individuals with ASD is essentially the same as TD individuals. The only changes to the network would be made in the fully connected layers. Ideally, more research should be done in modeling the human vision system with neural networks in order to change the neural network architecture rather than change the perceptual inputs to CNNs.

7 Discussion

In this paper, I employed computer vision methods to more clearly understand the facial expression recognition patterns of individuals with ASD. The first hypothesis tested, the configural processing hypothesis, is promising in explaining the FER deficit in individuals with ASD. There are statistically significant deficits in identifying negative emotions and to a lesser extent, nonnegative emotions when forcing a neural network not to utilize higher-order abstractions by also using inverted face images in training. Because training on inverted faces disrupts configural processing, the CNN trained with randomly inverted faces models a vision system that has diminished configural processing. Therefore, because the performance of the CNN resembles the relative deficits of identifying nonnegative and negative emotions when comparing ASD and TD individuals, this supports that diminished configural processing contributes to the FER deficit for individuals with ASD.

On the other hand, the computer vision tests do not support the eye-avoidance hypothesis, which states that eye-avoidance contributes to the FER deficit in individuals with ASD. Although the foveated and halves tests using the non-CNN dataset had a larger deficit in identifying negative emotions for the bottom model compared to the top model, which matches the comparison of ASD to TD individuals, these results were not significant. Additionally, when using CNNs, comparing the bottom and top models showed the contrary: the bottom model performed significantly *better* on negative emotions than the top model. Therefore, because the relative accuracies of the CNNs do not resemble the relative deficits of identifying nonnegative and negative emotions when com-

paring ASD and TD individuals, this does not support that eye-avoidance contributes substantially to the FER deficit for individuals with ASD.

Limitations to these various methods should be discussed. Although studies have reported a larger inversion effect for faces on configuration tasks rather than on part-based matching tasks, a later study finds that inverted faces interfere equally with both local processing and configural processing [70]. The researchers conducting this study attribute the discrepancy between previous studies to the fact that the part task in the previous studies was easier than the configuration task, which may account for the relatively small inversion effect seen in the parts task. Therefore, the inversion effect affects both configural and part-based featural processing.

We should be cautious in asserting that training a CNN on inverted faces is exactly simulating a system with less configural processing, since it may be interfering with local processing as well. However, [13] suggests that configural processing in facial expression may differ from configural processing in facial identification. Therefore, it is possible that the inversion effect affects configural processing more than local feature processing in the domain of facial expression. More research should go into configural and local processing in the context of reading facial expressions.

To simulate a deficit in configural processing while preserving (and even heightening) local feature processing, one might create a neural network that not only removes the use of global information, but better resembles the use of more local information, which could bring out the same FER patterns individuals with ASD exhibit. It could be possible to do this by removing a fully connected layer in the neural network architecture and adding more convolutional layers instead.

There were also limitations in simulating a bottom focus and top focus on faces. As mentioned before, the image resolutions for both the CNN and non-CNN methods may be too low for information in the eyes to be picked up.[7] Also, the top and bottom half was determined by image coordinates and not face coordinates. Both of these could have contributed to the present findings against the eye-avoidance hypothesis.

My methods for testing the eye-avoidance hypothesis only made manipulations at the low-level (although the low-level differences can and do propagate to high-level differences, high-level differences were not explicitly implemented). Therefore, even if training on the top halves and bottom halves of faces may accurately represent the focus of TD and ASD individuals, respectively, the focus itself may not be the only difference between TD and ASD individuals that needs to be accounted for. Therefore, more research should be done in modeling the human vision system with neural networks in order to implement high-level differences by changing the neural network architecture rather than changing the perceptual inputs to the CNN.

Future studies will benefit from having a large labeled dataset of facial expressions that also has high resolution in order to utilize the advantage of neural networks' capacity for complexity. My research shows that choosing Gabor jets and then using a relatively simple dimensionality reduction technique (PCA) and discrete classifier (LDA) matches and even exceeds the performance of fine-tuning convolutional neural networks on a relatively small (<50k) dataset with low resolution. Therefore, a larger and higher resolution dataset may be needed for a CNN to exceed the performance of hand-crafted classification methods. A higher resolution dataset can also benefit the non-CNN methods as well by possibly magnifying the nonsignificant results found for both the configural processing and eye-avoidance hypotheses.

Also, more complex computer vision methods can be explored. For non-CNN methods, there are other tools to reduce the high-dimensionality of Gabor jet features rather than the relatively simple method of PCA, such as Adaboost, which has also been used in facial expression recognition [6]. It is a method of feature selection that is capable of nonlinear dimensionality reduction, and therefore may be more suited to accurately retain the complex relationships of Gabor features during dimensionality reduction than PCA. For CNN methods, video streams of dynamic facial

expressions can be collected and fed into recurrent neural networks for classification, which better matches real-world FER interactions.

It would also be interesting to test a combination of both hypotheses to see if the interaction of diminished configural processing with a bottom-half focus emulates FER patterns for people with ASD. The performance of the CNN models trained on halves showed that although not all negative emotions were harder to identify from the bottom half, anger and fear were. It is possible that an accuracy deficit for the other negative emotions can arise from the interaction of both diminished configural processing and a bottom focus. Therefore, the combination of diminished configural processing and eye-avoidance may better simulate the FER deficit in individuals with ASD than either of the mechanisms alone.

Acknowledgement

I am indebted to my thesis advisors Professor Emily Cooper and Professor Brad Duchaine for their guidance in every stage of designing my study. Thanks to Tim and Jonathan for reading and commenting on my work. I also thank Professor Mohammad Mahoor and Ali Mollahosseini at the University of Denver for providing me with the weights of their pre-trained neural network and further instruction on how to use it, as well as Professor Paul Whalen and Ali Mattek at Dartmouth College for providing me with a dataset of facial expressions used in their own studies.

Notes

[1]In a review, Maurer et al. [20] classifies holistic processing as one aspect of configural processing, in which holistic processing is specifically processing a bunch of features as one single unit, which does not include spatial relationships between individual features (the review defines this as second-order relations) but for the purposes of this paper, configural and holistic processing are both distinguished from local featural processing.

[2]While increasing the depth of the network, adding 1x1 convolutional filters as a layer before the other convolutional filters also crucially reduces the dimensionality of the output, which is the primary reason why they are used. See [66] for more details.

[3]This unpublished dataset is courtesy of the Whalen Lab at Dartmouth College.

[4]45k images is still too small to train a neural network from scratch, so in all analyses, I fine-tuned networks on the pre-trained neural network from Mollahosseini et al. [62].

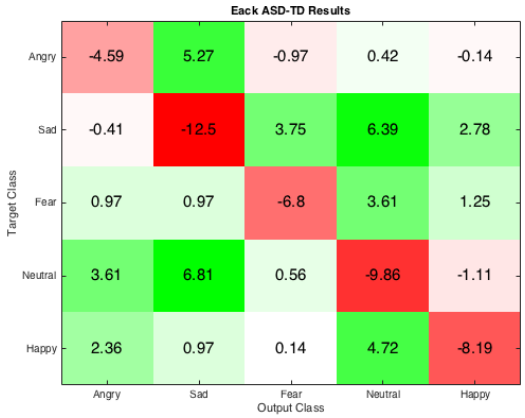
[5]Gabor jet filter extraction code modified from [47].

[6]Although the validation set for the CNN has 1805 images, the accuracies reported for this CNN and all other fine-tuned CNNs in this paper come from a subset which consists of all images in the validation set excluding FER2013 images totaling 481 images, making it more comparable to the nonCNN validation set.

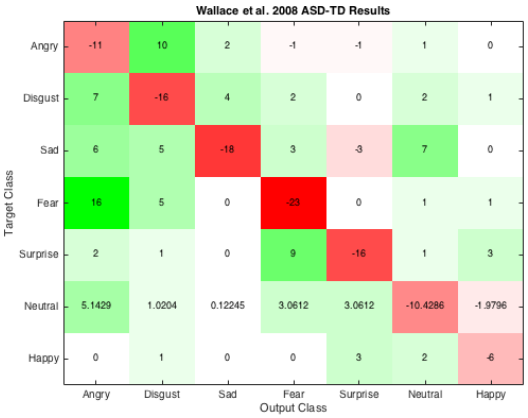
[7]Lades et al. used image resolutions of 128x128 which performed adequate face identification (recognition) [42]. But as mentioned previously, face identification and face expression recognition employ different processes in the brain, and so adequate FER may require more high-resolution information than face identification.

Appendices

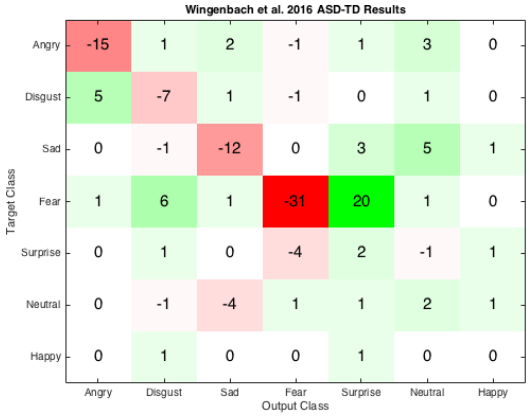
A Confusion matrices of facial expressions - literature



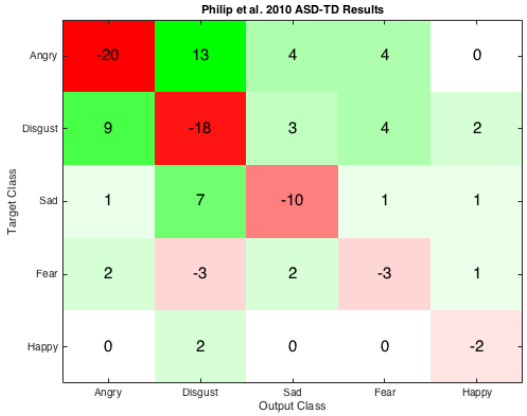
(a) ASD-TD Confusions from Eack et al. [28]



(b) From Wallace et al. [35]



(c) From Wingenbach et al. [37]



(d) From Philip et al. [29]

Figure 16: Confusions of the difference between ASD and TD accuracies (percentage correct) from various studies. Most cite an overall deficit in emotion recognition.

Figure 17 shows the weighted aggregation of confusions of the four studies in Figure 16. Comparing this aggregation to specific confusions cited in the literature, individuals with ASD do confuse fear for surprise more than TD individuals (4.25% more), which is in line with results from the Wingenbach study in Figure 16c. However, confusing fear as anger is more common as a behavior for ASD individuals (4.82 more than TD%), which affirms observations from the Wallace et al. study in Figure 16b. A substantial difference is that individuals with ASD confused disgust as anger (7.14%) more so than anger as disgust (4.29%), which does not entirely support findings from [36, 18], which state that individuals with ASD only confuse anger as disgust more than TD individuals. Another substantial difference is that ASD individuals confused anger as sadness (4.43%) more than TD individuals, which is consistent with findings from the Eack et al. study in Figure 16a.

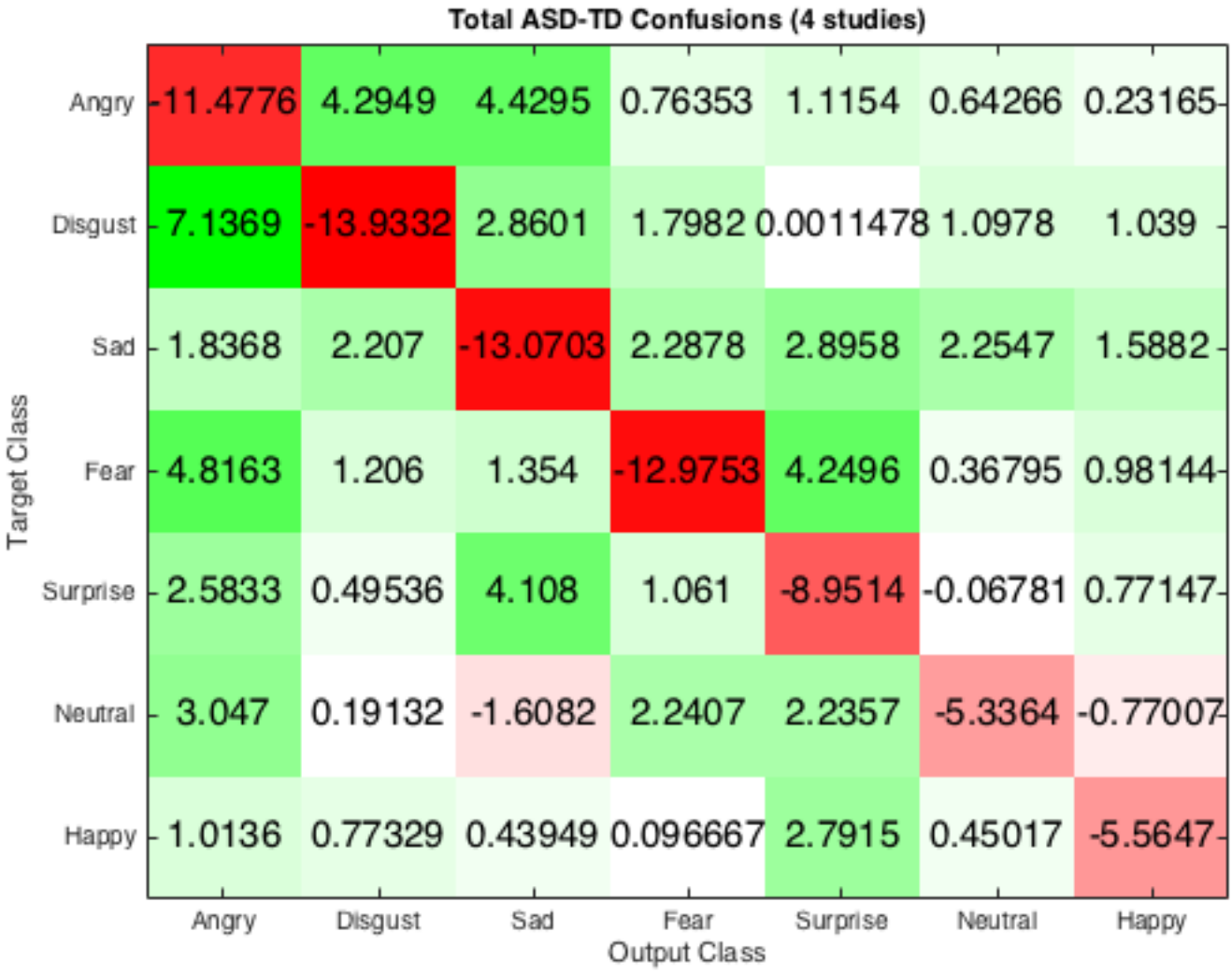


Figure 17: Summary of confusions in the emotion-labeling task comparing ASD and TD individuals across four studies in Figure 16. For each group in each study, the raw number of all output responses for each target stimuli were calculated. The number of responses were added by their respective group (ASD, TD) across all studies, then divided by the total number of target stimuli for each emotion to get the overall percentage % response distribution. The TD percentage distributions of confusions per emotion was subtracted from the ASD percentage distribution to get this resulting matrix.

Most, but not all, of these aggregated confusions agree with the results from the individual studies. Therefore, we can summarize the specific confusions of emotions that people with ASD confuse more than typically developing individuals: confusing sadness as neutral expressions, anger as sadness, anger as disgust, and disgust as anger.

B Non-CNN feature selection and classifier results

LDA+PCA had the best results for further feature selection after extracting Gabor filters. SVM+PCA results were lower than LDA+PCA results. Accuracy results are detailed in Table 13. This indicates that there is an incompatibility between PCA and SVM for facial expression classification, which is congruent with the literature [6]. In [6], SVM performance on features chosen from PCA was worse than both LDA performance on features chosen from PCA and SVM performance on all Gabor filters as features (no dimensionality reduction).

Table 13: Accuracy of SVM+PCA methods with various sizes and orientations of filters. None of the accuracies reached 70%, while LDA+PCA results for the same sizes and orientations were at least 70%.

sizes, orientations	6 orientations	8 orientations
3 sizes	0.6754	0.6931
4 sizes	0.5851	0.6129
5 sizes	0.5102	0.5528

Different values for some hyperparameters of the Gabor filters were also tested, such as spacing between different frequencies and maximum frequency. This was done on the LDA+PCA results, as this method is the fastest to train. This showed that the best spacing between Gabor frequencies should be half-octaves with the highest frequency being $\frac{\pi}{2}$ ($f = \sqrt{2}$, and $k_{max} = k_{v=0} = \frac{\pi}{2*f^0} = \frac{\pi}{2}$) instead of octaves ($f = 2$).

C Pre-trained neural network

The pre-trained neural network from the Mollahosseini paper was trained for 100k iterations using a polynomial learning rate policy. The accuracy on all of the CK+ images was 69.06%, while the FER results gave around a 40% accuracy. This is in line with the results from the paper which tested cross-database accuracies of CK+ and FER, which are 64.2% and 34.0%, respectively. I expected my testing accuracies to be higher than the cross-database results because some of the images were included in this training dataset. The confusion matrix for the CK+ results are shown in Figure 19. This reveals that the network shows biases towards happy, surprised, and angry faces. Therefore, I fine-tuned this network on my dataset and used this fine-tuned network as the basic working network for the rest of my hypotheses.

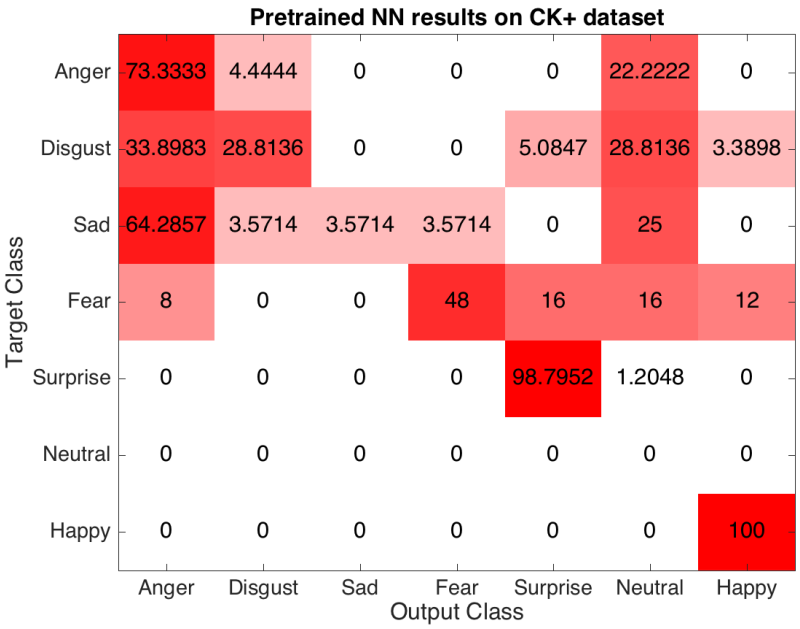


Figure 18: Confusion matrix for the pre-trained neural network from [62], tested on CK+ dataset. The neural network is exceptionally good at identifying happy and surprised expressions, but very poor at identifying sad, disgust, and fear expressions.

D Training graph for fine-tuned neural network

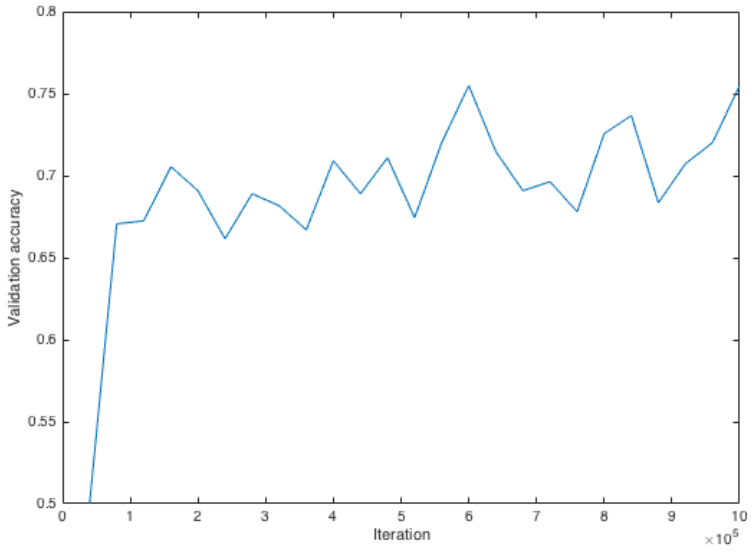


Figure 19: Training graph for fine-tuned neural network on whole faces. Trained for 1 million iterations.

E Confusion matrices for computer vision models

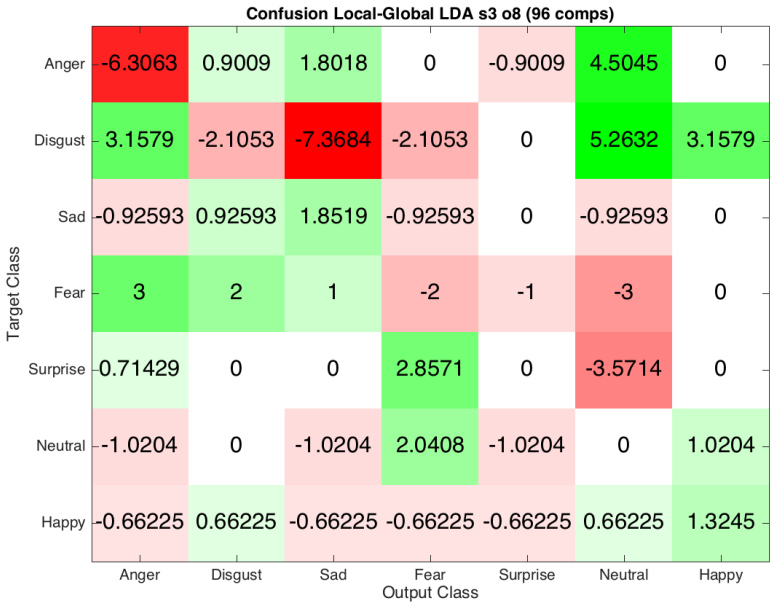


Figure 20: Difference confusion matrix for local-global training with PCA+LDA. The percent distribution of confusions of the local model was subtracted by the percent distributions of the global model's confusions.

Figure 20 shows that there is a deficit in the local model for identifying anger, disgust, and fear, compared to the global model. Figure 21 shows the confusion matrix for training on halves of Gabor filters does show that the bottom model is worse in recognizing disgust, fear, and sadness. However, it is better at recognizing anger, which is contrary both to results on top/bottom recognizability of emotions in TD individuals [13] and to the impaired ability of individuals with ASD to determine anger compared to TD individuals. We can conclude that focus on the bottom half of a picture may have a slight influence in ASD FER patterns, but it does not completely account for all ASD classification patterns. In fact, almost all of the observations in terms of confusions yield opposite results from the literature: the bottom-trained half is *less* likely to confuse anger as sadness, sadness as neutral, disgust as anger, and fear as anger than the top-trained half: all of which ASD individuals are *more* likely to make these confusions. Sadness seems to be the most effected expression when dividing the top half and bottom half.

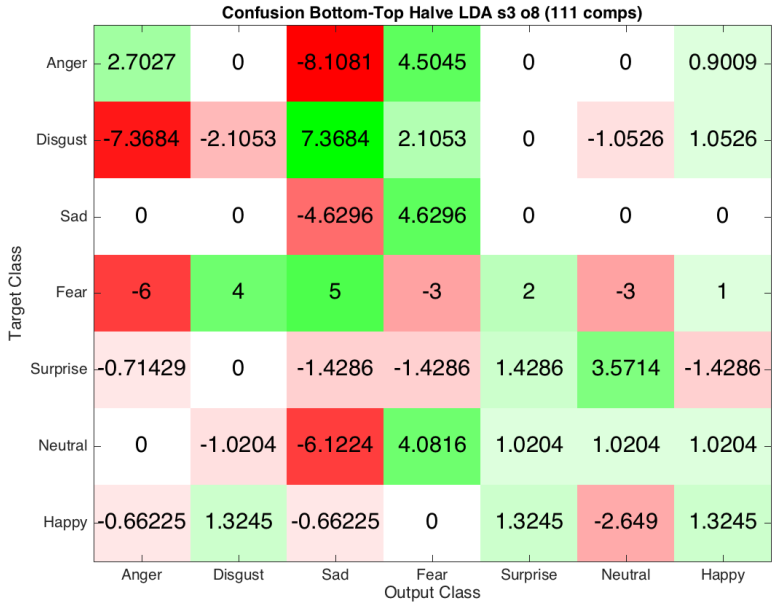


Figure 21: Difference confusion matrix comparing models trained on Gabor filters at the top half and bottom half of the grid, with PCA+LDA. The percent distribution of confusions of the bottom model was subtracted by the percent distributions of the top model's confusions.

References

[1] Yamins, D.L. and DiCarlo, J.J. 2016. Eight open questions in the computational modeling of higher sensory cortex. Current Opinion in Neurobiology, 37: 114-120.

[2] Viola, P. and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition. Proceedings of the 2001 IEEE Computer Society Conference, (1): I-511.

[3] Loth, E., Garrido, L., Watson, E., Duff, A., Ahmad, J. & Duchaine B. Facial expression recognition as a candidate biological risk marker for autism spectrum disorders: how frequent and severe are deficits? In press.

[4] Farah, M.J. and McClelland, J. 1991. A computational model of semantic memory impairment: modality specificity and emergent category specificity. Journal of Experimental Psychology: General, 120(4): 339.

[5] Christensen, D.L. 2016. Prevalence and characteristics of autism spectrum disorder among children aged 8 years- autism and developmental disabilities monitoring network, 11 sites, United States, 2012. Morbidity and Mortality Weekly Report. Surveillance Summaries: 65.

[6] Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. & Movellan, J. 2005. Recognizing facial expression: machine learning and application to spontaneous behavior. Computer Vision and Pattern Recognition, 2005. IEEE Computer Society Conference, 2:568-573.

[7] Shan, C., Gong, S. & McOwan, P.W. 2009. Facial expression recognition based on local binary patterns: a comprehensive study. Image and Vision Computing, 27(6): 803-816.

[8] Liu, M., Li, S., Shan, S. & Chen, X. 2013. Au-aware deep networks for facial expression recognition. Automatic Face and Gesture Recognition, 2013 10th IEEE International Conference and Workshops: 1-6.

[9] Dalton, K.M., Nacewicz, B.M., Johnstone, T., Schaefer, H.S., Gernsbacher, M.A., & Goldsmith, H.H. 2005. Gaze fixation and the neural circuitry of face processing in autism. Nature Neuroscience, 8(4): 519-526.

[10] Neumann, D., Spezio, M.L., Piven, J. & Adolphs, R. 2006. Looking you in the mouth: atypical gaze in autism resulting from impaired top-down modulation. Social Cognitive and Affective Neuroscience, 1(3): 194-202.

[11] Whalen, P.J. and Phelps, E.A. 2009. The human amygdala. New York, Guilford Press.

[12] Whalen, P.J., Kagan, J., Cook, R.G., Davis, F.C., Kim, H. & Polis, S. 2004. Human amygdala responsivity to masked fearful eye whites. Science, 306: 2061.

[13] Calder, A.J., Young, A.W., Keane, J. & Dean, M. 2000. Configural information in facial expression perception. Journal of Experimental Psychology: Human Perception and Performance, 26(2): 527.

[14] Hobson, R.P., Ouston, J. & Lee, A. 1988. What's in a face? The case of autism. British Journal of Psychology, 79(4): 441-453.

[15] Dawson, G., Webb, S.J. & McPartland, J. 2005. Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. Developmental Neuropsychology, 27(3): 403-424.

[16] Simmons, D.R., Robertson, A.E., McKay, L.S., Toal, E., McAleer, P. & Pollick, F.E. 2009. Vision in autism spectrum disorders. Vision Research, 49: 2705-2739.

[17] Rudra, A., Ram, J.R., Loucas, T., Belmonte, M.K. & Chakrabarti, B. 2016. Bengali translation and characterisation of four cognitive and trait measures for autism spectrum conditions in India. Molecular Autism, 7(1): 50.

[18] Jones, C.R., Pickles, A., Falcaro, M., Marsden, A.J., Happ, F., Scott, S.K., Sauter, D., Tregay, J., Phillips, R.J., Baird, G. & Simonoff, E. 2011. A multimodal approach to emotion recognition ability in autism spectrum disorders. Journal of Child Psychology and Psychiatry, 52(3): 275-285.

[19] Vanmarcke, S. and Wagemans, J. 2017. Priming facial gender and emotional valence: the influence of spatial frequency on face perception in ASD. Journal of Autism and Developmental Disorders: 1-20.

[20] Maurer, D., Le Grand, R. & Mondloch, C.J. 2002. The many faces of configural processing. Trends in Cognitive Sciences, 6(6): 255-260.

[21] LeGrand, R., Mondloch, C.J., Maurer, D. & Brent, H.P. 2001. Neuroperception: early visual experience and face processing. Nature, 410(6831): 890.

[22] Tanaka, J.W. and Farah, M.J. 1993. Parts and wholes in face recognition. The Quarterly Journal of Experimental Psychology, 46(2): 225-245.

[23] Zeng, Z., Pantic, M., Roisman, G.I. & Huang, T.S. 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE transactions on Pattern Analysis and Machine Intelligence, 31(1): 39-58.

[24] Fasel, B. and Luettin, J. 2003. Automatic facial expression analysis: a survey. Pattern Recognition, 36(1): 259-275.

[25] Ekman, P., Friesen, W.V. & Ellsworth, P. 1972. Emotion in the human face: guidelines for research and an integration of findings. Pergamon Press.

[26] Ekman, P. and Friesen W.V. 1976. Pictures of facial affect. Human Interaction Laboratory, University of California Medical Center.

[27] Ekman, P. and Friesen W.V. 1978. The facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press.

[28] Eack, S.M., Mazefsky, C.A. & Minshew, N.J. 2015. Misinterpretation of facial expressions of emotion in verbal adults with autism spectrum disorder. Autism, 19(3): 308-315.

[29] Philip, R.C.M., Whalley, H.C., Stanfield, A.C., Sprengelmeyer, R., Santos, I.M., Young, A.W. & Atkinson, A.P. 2010. Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. Psychological Medicine, 40(11): 1919-29.

[30] Tracy, J.L., Robins, R.W., Schriber, R.A. & Solomon, M. 2011. Is emotion recognition impaired in individuals with autism spectrum disorders? Journal of Autism and Developmental Disorders, 41(1): 102-109.

[31] Castelli, F. 2005. Understanding emotions from standardized facial expressions in autism and normal development. Autism, 9(4): 428-449.

[32] Harms, M.B., Martin, A. & Wallace, G.L. 2010. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. Neuropsychology Review, 20(3): 290-322.

[33] Capps, L., Nurit, Y. & Marian S. 1992. Understanding of simple and complex emotions in nonretarded children with autism. Journal of Child Psychology and Psychiatry, 33(7): 1169-1182.

[34] Bal, E., Harden E., Lamb D., Van Hecke A,V., Denver J.W. & Porges, S.W. 2010. Emotion recognition in children with autism spectrum disorders: relations to eye gaze and autonomic state. Journal of Autism and Developmental Disorders, 40(4): 358-370.

[35] Wallace, S., Coleman, M. & Bailey, A. 2008. An investigation of basic facial expression recognition in autism spectrum disorders. Cognition & Emotion, 22(7): 1353-1380.

[36] Wallace, G.L., Case, L.K. & Harms, M.B. 2011. Diminished sensitivity to sad facial expressions in high functioning Autism Spectrum Disorders is associated with symptomatology and adaptive functioning. Journal of Autism and Developmental Disorders, 41(11): 1475.

[37] Wingenbach, T.S., Ashwin, C. & Brosnan, M., 2016. Diminished sensitivity and specificity at recognising facial emotional expressions of varying intensity underlie emotion-specific recognition deficits in autism spectrum disorders. Research in Autism Spectrum Disorders, 34: 52-61.

[38] Serret, S., Hun, S., Iakimova, G., Lozada, J., Anastassova, M., Santos, A., Vesperini, S. & Askenazy, F. 2014. Facing the challenge of teaching emotions to individuals with low-and high-functioning autism using a new Serious game: a pilot study. Molecular Autism, 5(1): 37.

[39] Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B. & Hoekstra, R.A. 2013. Empathy and emotion recognition in people with autism, first-degree relatives, and controls. Neuropsychologia, 51(1): 98-105.

[40] Uljarevic, M. and Hamilton, A. 2013. Recognition of emotions in autism: a formal meta-analysis. Journal of Autism and Developmental Disorders, 43(7):1517-1526.

[41] Ashwin, C., Chapman, E., Colle, L. & Baron-Cohen, S. 2006. Impaired recognition of negative basic emotions in autism: a test of the amygdala theory. Social Neuroscience, 1(3-4): 349-363.

[42] Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.P. & Konen, W. 1993. Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers, 42(3): 300-311.

[43] Daugman, J. 1988. Complete discrete 2D Gabor transform by neural networks for image analysis and compression. IEEE Transactions on Acoustics, Speech, and Signal Precessing, 36(7): 1169-1179.

[44] Balakrishnama, S. and Ganapathiraju, A. 1998. Linear discriminant analysis-a brief tutorial. Institute for Signal and Information Processing, 18.

[45] Kriegeskorte, N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual Review of Vision Science, 1: 417-446.

[46] Pantic, M. and Patras, I. 2006. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36(2): 433-449.

[47] Margalit, E., Biederman, I., Herald, S.B., Yue, X. & von der Malsburg, C. 2016. An applet for the Gabor scaling of the differences between complex stimuli. Attention, Perception, & Psychophysics, 78(8): 2298-2306.

[48] Van der Schalk, J., Hawk, S.T., Fischer, A.H. & Doosje, B.J. 2011. Moving faces, looking places: the Amsterdam Dynamic Facial Expressions Set (ADFES). Emotion, 11: 907-920.

[49] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. & Matthews, I. 2010. The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops: 94-101.

[50] Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B. & Zhou, Y. 2013. Challenges in representation learning: a report on three machine learning contests. International Conference on Neural Information Processing: 117-124.

[51] Valstar, M.F. & Pantic, M. 2010. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. Proceedings of the 3rd International Workshop on EMOTION, Corpora for Research on Emotion and Affect: 65.

[52] Lundqvist, D., Flykt, A. & Ohman, A. 1998. The Karolinska Directed Emotional Faces (KDEF). Department of Clinical Neuroscience, Psychology section, Karolinska Institute, Stockholm, Sweden.

[53] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T. & van Knippenberg, A. 2010. Presentation and validation of the Radboud Faces Database. Cognition & Emotion, 24(8): 1377-1388.

[54] Dhall A., Goecke R., Lucey S. & Gedeon, T. 2011. Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. 2011 IEEE International Conference.

[55] Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P. & Cohn, J.F. 2013. DISFA: a spontaneous facial action intensity database. IEEE Transactions on Affective Computing, 4(2):151-160.

[56] Tang, Y. 2013. Deep learning using support vector machines. CoRR, abs/1306.0239,2.

[57] Lekshmi, P.V. and Sasikumar M. 2009. Analysis of facial expression using Gabor and SVM. International Journal of Recent Trends in Engineering, 1(2).

[58] Tian, Y.L., Kanade T. & Cohn J.F. 2005. Facial Expression Analysis. Handbook of Face Recognition, Springer: 247-276.

[59] Spezio, M.L., Adolphs, R., Hurley, R.S.E. & Piven, J. 2007. Atypical use of facial information in high-functioning autism. Journal of Autism and Developmental Disorders, 37(5): 929-39.

[60] Tanaka, J.W., Wolf, J.M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L. & Kaiser, M. D. 2012. The perception and identification of facial emotions in individuals with autism spectrum disorders using the Let's Face It! Emotion Skills Battery. Journal of Child Psychology and Psychiatry, 53(12): 1259-1267.

[61] Wang, Z. and Bovik, A.C., 2001. Embedded foveation image coding. IEEE Transactions on Image Processing, 10(10): 1397-1410.

[62] Mollahosseini, A., Chan D. & Mahoor, M.H. 2016. Going deeper in facial expression recognition using deep neural networks. IEEE Winter Conference on Applications of Computer Vision (WACV): 1-10.

[63] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S. & Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM International Conference on Multimedia: 675-678.

[64] Lyons, M. J., Budynek, J. & Akamatsu, S. 1999. Automatic classification of single facial images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(12): 1357-1362.

[65] Lyons, M., Akamatsu, S., Kamachi, M. & Gyoba, J., 1998, April. Coding facial expressions with Gabor wavelets. Automatic Face and Gesture Recognition, 1998. Proceedings of the 3rd IEEE International Conference: 200-205.

[66] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. 2015. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 1-9.

[67] Xu, X., Biederman, I. & Shah, M.P. 2014. A neurocomputational account of the face configural effect. Journal of Vision, 14(8).

[68] Bassili, J.N. 1979. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology, 37(11): 2049.

[69] Kotsia, I., Buciu, I. & Pitas, I. 2008. An analysis of facial expression recognition under partial facial image occlusion. Image and Vision Computing, 26(7): 1052-1067.

[70] Yovel, G. and Kanwisher, N. 2004. Face perception: domain specific, not process specific. Neuron, 44(5): 889-898.