Do computational models differ systematically from human object perception?

R. T. Pramod
Department of Electrical Communication Engineering
& Centre for Neuroscience,
Indian Institute of Science, Bangalore, India.
E-mail: pramodrt@ece.iisc.ernet.in

S. P. Arun
Centre for Neuroscience,
Indian Institute of Science, Bangalore, India.
E-mail: sparun@cns.iisc.ernet.in

Abstract

Recent advances in neural networks have revolutionized computer vision, but these algorithms are still outperformed by humans. Could this performance gap be due to systematic differences between object representations in humans and machines? To answer this question we collected a large dataset of 26,675 perceived dissimilarity measurements from 2,801 visual objects across 269 human subjects, and used this dataset to train and test leading computational models. The best model (a combination of all models) accounted for 68% of the explainable variance. Importantly, all computational models showed systematic deviations from perception: (1) They underestimated perceptual distances between objects with symmetry or large area differences; (2) They overestimated perceptual distances between objects with shared features. Our results reveal critical elements missing in computer vision algorithms and point to explicit encoding of these properties in higher visual areas in the brain.

1. Introduction

What is a good man but a bad man's teacher?
What is a bad man but a good man's job?
-- *Tao Te Ching* [1]

Convolutional or deep neural networks have revolutionized computer vision with their impressive performance on object classification tasks [2-4]. This performance (~75% correct on top-1 classification), while impressive compared to older algorithms, is nowhere close to humans, whose ability is so high that it is rarely measured and often taken as ground-truth [5]. This performance gap raises the intriguing possibility that human vision can be used to "teach" machines to do better. Conversely, understanding the specific deficiencies in machine algorithms can elucidate the specialized computations performed by the brain [6, 7]. The obvious approach would be to compare humans and machines on a classification task, but this involves an additional classification process that complicates any inference about the underlying features.

A simpler alternative would be to compare distances between objects in feature space. For a machine algorithm this involves calculating the distance between the corresponding feature vectors. In humans, these distances can be measured experimentally in behavior [8-10] or in distinct brain regions [11, 12]. This has permitted the detailed comparison of object representations in the brain with existing computer vision algorithms, which has yielded several insights. First, object representations in early visual areas are explained best by Gabor filters [13], which is not surprising given the well-known orientation selectivity of early visual areas. Second, object representations in higher visual areas (that are crucial for object recognition) in both human and monkey, are explained well by later layers in convolutional neural networks [12, 14-16], as well as by other computational models such as SIFT [11] and other hierarchical models such as HMAX [12]. Thus, the consensus view is that deep neural networks have object representations that are highly similar to those in the brain. But this predicts that these networks should perform as well as humans, yet they clearly do not.

This apparent contradiction could have arisen for two reasons. First, most of these comparisons are based on ~50-100 natural objects differing in many features. This could have produced a large correlation even if the object representations were entirely different. Second, there may be systematic patterns in the match between brain representations and computational models: it could be worse for certain types of images and better for others. These issues have never been investigated.

Here, we systematically compared object representations in humans with a number of computational models. To this end, we collected a large dataset of 26,675 perceived dissimilarity measurements from 2,801 visual objects across 269 human subjects. To measure perceived dissimilarity in humans, we asked subjects to locate the oddball in an array containing one object among multiple instances of the other. The reciprocal of the visual search time was taken as an estimate of perceived dissimilarity. This measure behaves like a mathematical distance [10], shows linear summation across multiple features [17], explains rapid visual categorization [18] and is strongly correlated with subjective dissimilarity ratings [19]. We used this dataset to train and test a number of popular computational

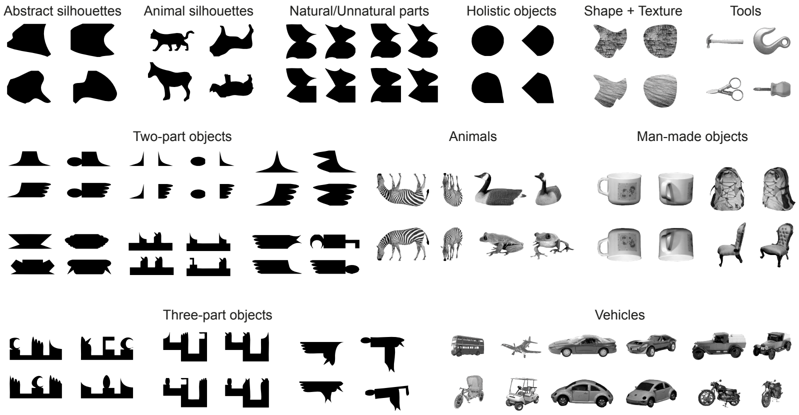

Figure 1 – Examples of the 2,801 objects used in the study, which ranged from simple and abstract silhouettes to complex/grayscale natural objects.

models. The best individual model was a convolutional neural network, but it was outperformed by a combination of all the individual models. Importantly, however, this model explained only a part of the explainable variance, with systematic patterns in its residual error that offer insights into both human and machine vision.

2. Methods

2.1 Perceived dissimilarity measurements in humans

We collected a total of 26,675 perceived dissimilarity measurements from 2,801 objects across 269 human subjects. We measured perceived dissimilarities between isolated objects rather than natural scenes because the latter may produce variations in attended location and consequently in the feature representation. The dataset was compiled from 32 distinct visual search experiments typically with little or no overlap between subjects. Most experiments involved measurements of pairwise dissimilarity between all possible pairs of a fixed set of objects. Hence, we used two types of objects in the dataset – natural objects and silhouettes. The natural objects consisted of objects drawn from various natural categories such as animals, tools, vehicles. In some experiments, there were two views of each object – a profile (sideways) view and an oblique view corresponding to an in-depth rotation of 45 degrees. The silhouette shapes comprised both abstract and animate silhouettes. In some experiments, silhouette images were obtained by combining 7 possible parts on either end of a stem in all possible ways to get a total of 49 objects.

The fact that the dataset was compiled from experiments performed on different subjects raises the question of whether the measurements are a valid estimate of the perceptual distances within a single subject. We believe they are valid for the following reasons: First, the dataset is extremely consistent when subjects are separated into random groups, implying that it would be even more consistent if it were feasible to collect data within a single subject. Second, in a separate experiment, we measured perceptual distances for a random subset of 400 image pairs in a separate group of 4 subjects. These perceptual distances were strongly correlated with the distances in the original dataset ($r = 0.80$, $p < 0.00005$).

All participants were aged 20-30 years, had normal or corrected-to-normal vision and were naïve to the purpose of the experiments. All of them gave written informed consent to an experimental protocol approved by the Institutional Human Ethics Committee. In each experiment, the subject was seated ~60 cm from a computer monitor controlled by custom programs based on Psychtoolbox [20] in MATLAB. Each trial began with the appearance of a fixation cross for 500ms followed by a 4 x 4 search array. The search array contained one oddball target item among multiple identical distracters (Figure 2A). All items in the array were jittered in position to prevent alignment cues from driving the search. Subjects were instructed to respond as quickly and as accurately as possible to indicate the side on which the oddball target was present using a key press (Z for left and M for right). If the subjects failed to make any response within 10s of the onset of the search array, the trial was aborted. All aborted or incorrect trials were randomly repeated later in the task. For each pair of objects, subjects performed between 2-8 correct trials (depending on the experiment) on which their response time was recorded. The reciprocal of the average search time taken by subjects to make a key press was taken as an estimate of perceived dissimilarity between the oddball target and the distracter.

2.2. Computer vision models

We tested a total of 19 popular computational models for object vision. These models fell roughly into five groups: pixel-based, boundary-based, feature-based, image statistics-based and network models. Each model is described briefly below and in greater detail in the Supplementary Material (Supplementary Section S1). For most models, we accessed a feature vector for each image and calculated the Euclidean distance between two feature vectors as the distance. For models specified in terms of a distance metric rather than feature vectors (e.g. SSIM), we calculated the pairwise distances directly.

Pixel based models (n = 2). We tested two models that use a pixel-based representation. In the sum-of-squared error model (SSE), each pixel is a feature and the distance is computed as the sum-of-squared differences across pixels after shifting the images to obtain the best match. The coarse footprint model (CFP) was previously used to explain perceived and neural dissimilarity, and calculates the pixel-wise difference between two images after they undergo size-dependent blurring [21].

Boundary-based models (n = 4). These models only represent the boundary of an object and not its internal detail. We therefore reduced each object into a silhouette for the purposes of these models. The curvature scale space model (CSS) was used previously for planar curve matching, and uses local curvature zero-crossings as features. The curvature-length (CL) and tangent-angle-length (TAL) models represent a contour either using the curvature or the tangent angle at regular intervals along its length. The Fourier descriptor (FD) model represents the contour using the Fourier transform coefficients of a closed contour represented in the complex plane [22]. It has previously been used to study perceived dissimilarity in humans [23].

Feature-based models (n = 7). These are the most common type of computational models, and involve extracting specific features from an image that are then used for vision tasks. The Gabor filterbank (GABOR) model represents the image as coefficients of a wavelet pyramid and has been used previously to model early visual cortex [13]. The Geometric Blur (GB) model computes local image properties at selected interest points and calculates the distance using all pairwise comparisons of interest points. This model has also been used widely in object detection [24, 25]. The scale invariant feature transform (SIFT) is an extremely popular computer vision algorithm for describing and matching local image features [26]. It finds a set of distinctive interest points across a set of images, and represents each image by the histogram of interest points of each type present in the image. The Histogram of Oriented Gradients (HOG) model is another widely used feature descriptor in which the image is broken down into overlapping blocks spanning the entire image each containing a histogram of orientations [27]. Scene gist (GIST) is a variant of the Gabor filter bank in which each image is represented as a weighted sum of Gabor filters. Scene Gist is different from Gabor wavelet representation in that it uses Gabor wavelet filters on non-overlapping windows of the image. Fourier phase (FPH) and Fourier power (FP) are the phase and power of the 2-d Fourier transform of the image.

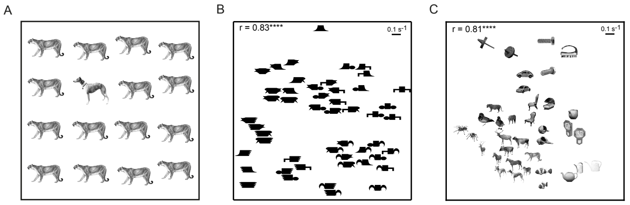

Figure 2 – Visual search dissimilarities (A) Example visual search array with target and distracters arranged in a 4x4 array. In the actual experiment, all the elements of the search array were presented on a black background with a red vertical line down the middle of the display; (B) 2D embedding of a set of abstract objects. The embedding is obtained using Multidimensional scaling on search dissimilarities (1/RT). The r-value indicates the agreement between search dissimilarities and embedded distances (**** is $p < 0.00005$); (C) Similar plot as in (B) for a set of natural objects.

Statistical models (n = 2). We tested two models that represent images using their statistical properties. Texture Synthesis (TSYN) is a popular texture synthesis/analysis algorithm where a texture image is represented by a set of statistics (central moments, range of pixel intensities and correlations) calculated on the wavelet filtered image [28]. The Structural Similarity Index (SSIM) is used to measure similarity between distorted images using the mean and covariance of pixel intensities [29].

Network models (n = 4). We tested four models that use hierarchical or network models inspired by biological vision. For each network model, the output of each unit in a specific layer of the network is concatenated into a single feature vector. The Jarrett (Jar) model has one stage of random filters followed by divisive normalization, output nonlinearity and pooling. This model uses random filters generated using white noise with no learning and was shown to achieve competitive performance on object classification tasks [30]. HMAX is a popular biologically inspired model that uses cascades of linear summation and nonlinear pooling (max) operations to build selectivity and invariance [31]. We used the outputs of complex cell (C1) layer model units. The V1 model is a simple model for area V1 in the early visual cortex containing Gabor filters with input and output divisive normalization and pooling [32]. The Convolutional Neural Network (CNN) model was an implementation of a deep neural network (VGG-16) trained on object localization and classification that shows the best performance to date [4].

2.3. Model evaluation

Some models (e.g. CNN) are already optimized for object classification, and therefore we were interested in knowing how well they match the perceptual data directly. In other cases (e.g. Fourier power), the model is not optimized in any way, yet a weighted sum of its features may explain the perceptual data. We therefore tested each model in two ways: First, we calculated the direct correlation between the distances of each model and the observed perceptual distances without any explicit fitting to the perceptual data. Second, we fit each model to the data by weighting its features to obtain the best match to the perceptual data. To avoid overfitting, we used a standard cross-validation procedure: we trained each model using 80% of the data and then tested it on the remaining 20%. This cross-validation procedure was repeated 10 times to obtain average model performance. To equate the number of free parameters across all models, we performed a dimensionality reduction on the feature vectors for each model using principal component analysis (PCA). Specifically, we compiled the feature vectors corresponding to all 2,801 objects in the dataset and performed a PCA. We then projected each feature vector onto the first 100 principal components to obtain a 100-dimensional feature vector for that object. We then asked whether the observed perceptual distances could be explained as a weighted sum of distances along each of these principal components. Specifically, if $x_1 = [x_{11}\ x_{12}\ x_{13} \ldots\ x_{1n}]$ and $x_2 = [x_{21}\ x_{22}\ x_{23} \ldots\ x_{2n}]$ are the n-dimensional feature vectors corresponding to two images (n = 100), then our model predicts the observed distance y_{12} between these two images to be:

$$y_{12} = w_1|x_{11}-x_{21}| + w_2|x_{12}-x_{22}| + \ldots + w_n|x_{1n}-x_{2n}| \qquad (1)$$

where w_1, w_2 etc. represent the importance or contribution of that particular principal component to the overall perceptual distance. Writing down these equations for all 26,675 observed distances results in the matrix equation $\mathbf{y} = \mathbf{Xw}$ that can be solved using simple linear regression. We obtained qualitatively similar results on changing the number of principal components available to each model.

2.4. Combination models

In addition to analyzing each model separately, we asked whether combining all models would yield an even better model. We tested two possible combined models. In the first model (comb1), we concatenated the z-scored features across 15 models (among the excluded four models, the SSIM model does not use explicit features, CSS and GB have too few features to calculate 100 principal components and V1 model has too many features (~10^6) to calculate the principal components). We then obtained a 100-dimensional reduced feature vector using PCA. We then repeated the procedures described above to characterize model performance. In the second model (comb2), we tried to predict the observed distances as a weighted sum of the net distances of each model. This resulted in a matrix equation of the form $\mathbf{y} = \mathbf{Xb}$ where $\mathbf{y}$ is a 26,675 x 1 vector containing observed distances, $\mathbf{X}$ is a 26,675 x 19 matrix containing distances predicted by each of the 19 models (without fitting to the data) and $\mathbf{b}$ is an unknown 19 x 1 weight vector that represents the relative contributions of each model to the observed distances.

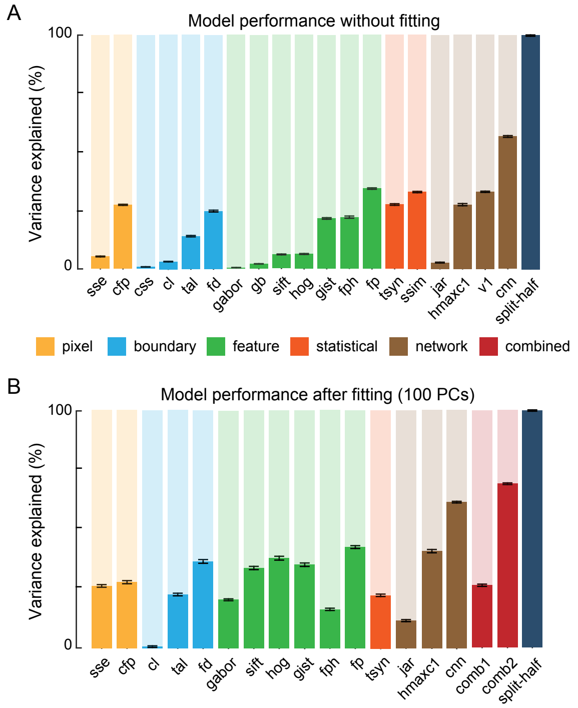

Figure 3 – Model performance on human perceptual data. (A) Bar plot of explainable variance explained by all models with un-weighted feature distances. Models are color coded by category. (B) Similar plot as in (A) for weighted PC model distances.

2.5. Model quality of fit

A simple measure of the quality of fit for a model is the squared-correlation (R^2) measure which represents the percentage of total variance explained by the model. However, this number alone is not meaningful because it does not capture the "explainable variance" or intrinsic reliability of the data. To estimate the explainable variance, we first separated the subjects into two random groups and calculated the perceptual distances separately. We then reasoned that the upper limit for any model would be the degree to which two random groups of subjects might be correlated. However there is a small issue because the split-half correlation obtained in this

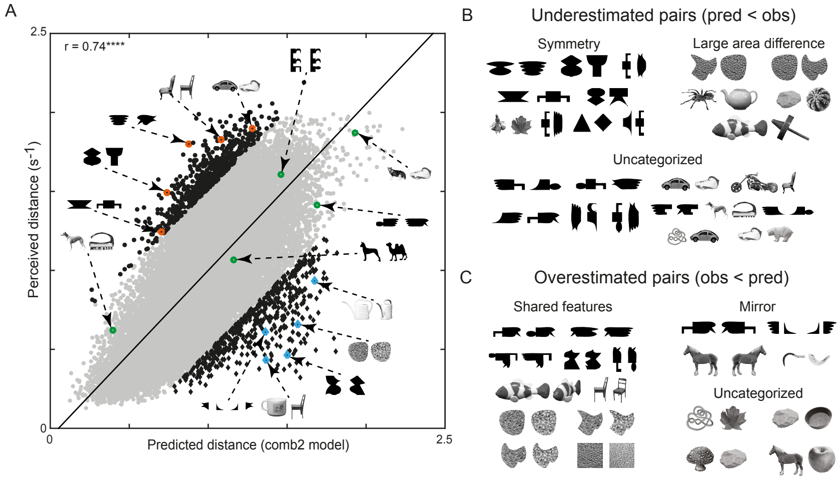

Figure 4 − Best model performance. (A) Plot of the observed perceptual distances versus the distances predicted by the best model (*comb2*) for all 26,675 pairs. Object pairs whose dissimilarity is underestimated by the model (residual error more than 1 standard deviation above the mean) are shown as *filled black circles* with example pairs highlighted in *orange*. Pairs whose dissimilarity is overestimated by the model (residual error less than 1 standard deviation below the mean) are shown as *filled black diamonds* with example pairs highlighted in *blue*. Pairs whose dissimilarity is explained by the model (residual error within 1 standard deviation of the mean) are shown as *gray circles* with example pairs highlighted in *green*. ; (B) Examples of over-estimated pairs of objects; (C) Examples of under-estimated pairs of objects.

manner (by comparing two halves of the subjects) may underestimate the true reliability of the data (which is based on all subjects). We accordingly corrected this split-half correlation using a Spearman-Brown correction. The corrected split-half correlation, which represents the upper limit on any model performance, is given by $rc = 2r/(1+r)$, where r is the split-half correlation and *rc* is the corrected correlation. To obtain a single composite number representing the percentage of explainable variance explained by each model, we divided the squared-correlation of the model with the observed data by the square of the corrected split-half correlation.

For all models with the same number of degrees of freedom, model performance can be compared directly using the measure of explainable variance explained. This is true in all models except *comb2* with 19 free parameters compared to other models with 100 free parameters. In this case, *comb2* is expected to perform worse by default due to its fewer free parameters. However its performance was in fact better than all other models despite their greater complexity. This obviates the need for detailed model comparisons using quality-of-fit measures that take model complexity into account.

2.6. Strength of symmetry

We quantified the strength of symmetry of an image by calculating the degree to which the two halves of the image are mirror images of each other. For a symmetric object, the pixel-wise difference between its halves mirrored about the axis of symmetry will be zero. Thus the strength of symmetry about the vertical axis (S_v) for an image A can be written as:

$$S_v = 1 - \frac{\Sigma\, abs(\, A - flipv(A))}{\Sigma\, abs(\, A + flipv(A))} \qquad (2)$$

where *flipv(A)* represents the image mirrored about the vertical axis, and abs() is the absolute value, and the summation is taken over all pixels. This measure ranges from 0 when the image and its vertical mirror reflection do not overlap at all, to 1 when the image and its vertical mirror reflection are identical in every pixel i.e. the image is symmetric. We calculated the strength of symmetry about the horizontal axis (S_h) in an analogous fashion.

For each image pair, we calculated the strength of symmetry about the vertical axis averaged across both objects in the pair, and likewise the average strength of symmetry about the horizontal axis across both objects. We then took the overall strength of symmetry for the image pair as the larger of the vertical and horizontal symmetry measures.

3. Results

To compare object representations in humans and computational algorithms, we collected a large dataset of perceptual dissimilarity measurements from human subjects using visual search. This dataset consisted of 26,675 perceived dissimilarity measurements made from 2,801 objects across 269 human subjects. The objects consisted of abstract and familiar silhouettes as well as natural objects (Figure 1). We tested only a small subset of all possible object pairs both due to experimental constraints as well as to avoid testing completely dissimilar objects (e.g. natural versus silhouette) that would saturate the range. The dissimilarity between each pair of objects was obtained using visual search: subjects had to locate the oddball item in a field of identical items, as illustrated in Figure 2A. We used the reciprocal of search time as an estimate of the perceived dissimilarity (see Introduction).

Subjects were extremely consistent in their performance as indicated by a highly significant correlation between the perceived dissimilarity of the 26,675 object pairs obtained from two random groups of subjects ($r = 0.84$, $p < 0.00005$). To visualize these dissimilarity measurements graphically, we performed multidimensional scaling to embed these distances for two subsets of the objects tested (Figure 2B,C). In these plots, nearby images represent hard searches. Both plots showed interesting patterns: in Figure 2B, silhouette objects that share parts are close together, as are objects that are vertical mirror images. In Figure 2C, profile and oblique views of natural objects are close together, indicative of pose or viewpoint invariance. These plots show that perceptual distances between objects are highly systematic and structured.

Next we asked whether the perceptual distances between objects can be explained using computational models. As a first step, we asked whether the distances between objects in each computational model (without any fitting to the data) are correlated with the perceptual data. We expressed each model's variance explained (R^2) relative to the explainable variance in the data (see Methods). All models showed a significant positive correlation with the perceptual data, but the convolutional neural network (CNN) was the best model ($r = 0.68$, $p < 0.00005$; Figure 3A). It explained 55.1% of the explainable variance in the data.

Although some of the tested models, such as the CNN, have been optimized for vision tasks, other models (such as Fourier power) are not optimized at all. We therefore investigated the ability of each model to fit the data by allowing it to prioritize its features to fit the data. For instance, in the case of Fourier power, we asked whether a weighted sum of Fourier power differences between two images – where the weights reflect the importance of each spatial frequency – could explain the perceptual data. However this cannot be done directly on all the parameters of each model, because models varied widely in the number of features (from a minimum of 6 features in the CSS model to 940,800 features for the V1 model). We therefore reduced the dimensionality of each model by projecting the feature vectors of each model along its first 100 principal components, and then asked whether a weighted sum of these PCA-based features could predict the perceptual data. Note that four models could not be fit using their principal components (CSS, GB, SSIM and V1; see Methods). We fit the perceptual data to each model using a standard cross-validation approach: 80% of the data was used for training and 20% for testing. Model performance after fitting using 100 PCs is shown in Figure 3B. As expected the performance of all models improved after fitting to the perceptual data. The best individual model was still the CNN model, which explained 62.6% of the explainable variance ($r = 0.72$, $p < 0.00005$; Figure 3B).

We then asked whether combining all models in some way would produce an even better fit to the data. We explored this possibility in two ways: First, we concatenated the features of each model and projected these features onto their first 100 principal components, and fit this model (denoted as *comb1*) as before. Second, we asked whether a weighted sum of the individual model distances (without fitting) could predict the perceptual data – we denote this model as *comb2*. The performance of these two combined models is also shown in Figure 3B. It can be seen that the comb1 model has a performance that is actually worse than some of the individual models. This could be because the concatenated features may contain more irrelevant variations that are captured by the principal components. In contrast, the *comb2* model had the best match to the perceptual data compared to all other models (variance explained = 68.1%; $r = 0.74$, $p < 0.00005$; Figure 3B). Note that the performance of the *comb2* model (with only 19 parameters) is better than any single individual model fit to the data with 100 free parameters based on their principal components.

3.1 Does the best model show systematic residual errors?

To investigate the performance of the *comb2* model in greater detail, we plotted the observed perceptual distances against the predicted distances (Figure 4A). For each image pair, we calculated the residual error as the signed difference between the observed distance and predicted distance. We then asked whether the residual errors varied consistently with the nature of the image pair. To this end, we examined all image pairs whose residual error was one standard deviation above or below model predictions (Figure 4A). This revealed many interesting patterns. Image pairs whose dissimilarity was underestimated by the model (i.e. predicted < observed) usually contained symmetric objects or pairs in which one object occupied a large area compared to the other (Figure 4B). Image pairs whose dissimilarity was overestimated by the model (predicted > observed) tended to be objects that shared features (Figure 4C): these could be objects with shared contours, texture or shape. The model tended to overestimate the dissimilarity even between two views

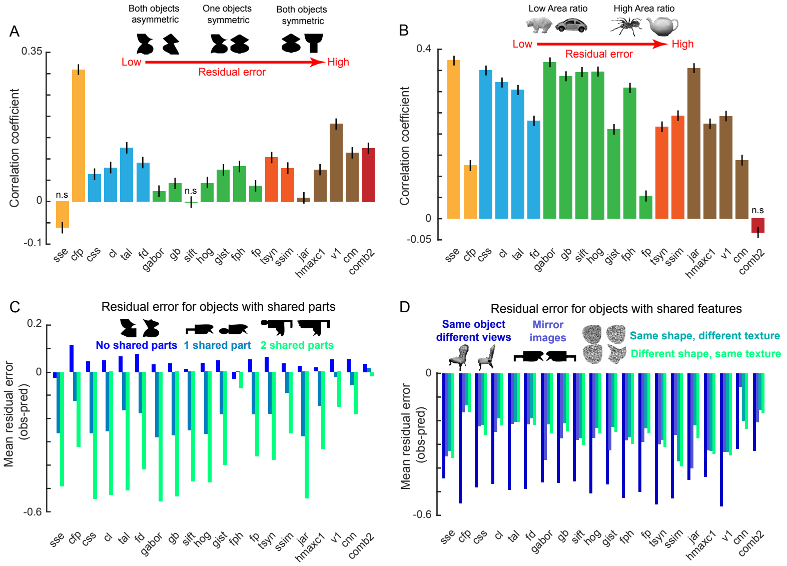

Figure 5 – Patterns of residual errors across models. (A) Correlation between symmetry strength and residual error across object pairs for each model. Error bars indicate bootstrap-derived estimates of standard deviation (i.e. by repeatedly calculating the correlation on image pairs sampled randomly with replacement ten times). All correlations are significant with p < 0.005. Non-significant correlations are indicated as "n.s"; (B) Correlation between area ratio and residual error across object pairs for each model; (C) Average residual error across image pairs with zero, one or two shared parts; (D) Average residual error for objects pairs related by view, mirror reflection, shape and texture.

of a single object, and mirror images (which can be construed as a 180-degree rotation in depth).

To confirm that these residual error patterns are not intrinsic to the data, we analyzed the residual error patterns obtained when data from one half of the subjects was used as a predictor for the other half. This yielded qualitatively different error patterns. Thus, the systematic residual error patterns present between models and data are not intrinsic to the data itself.

Because of the large number of image pairs it is quite possible that image pairs with small residuals also have the same properties as detailed above. Also it is possible that the combined model has these residual error patterns but individual models do not. To assess these possibilities, we quantified each of the above observations to assess the degree to which they modulate residual error. For each model, we calculated the residual error between the observed perceptual data and the predicted distances (without fitting the model explicitly to the data for individual models). We present these results for simplicity but obtained qualitatively similar results even when we used the models to fit the data based on their 100 principal components (Supplementary Section S2).

3.2. Symmetric objects are more distinct in perception than models

To quantify our observation that computational models underestimate the dissimilarity between symmetric objects, we note that symmetric objects by definition contain mirrored halves about their axis of symmetry. We therefore devised a measure of symmetry for each object pair (see Methods). We then asked whether this strength of symmetry co-varies with the residual error of each model across all image pairs. A positive correlation would confirm our observation that, as the objects in a pair become more symmetric, the residual error becomes larger – in other words, symmetric objects are more distinct in perception than in models. The correlation for all models between symmetry strength and residual error is shown in Figure 5A. All models including *comb2* showed a significant positive correlation. The only two exceptions to this trend were the SSE and SIFT models. We conclude that computational models underestimate the dissimilarity between symmetric objects. Even in perception, we have shown that symmetric objects are more distinct than expected due to local part differences [19].

3.3. Objects with large area differences are more distinct in perception than models

To quantify our observation that objects with large area differences are more distinct in perception, we calculated for each image pair the ratio of the area of the larger object to area of the smaller object. For all individual models, the area ratio had a significant positive correlation with residual error (Figure 5B). However the *comb2* model showed no significant correlation with area ratio, because it tended to assign positive and negative weights across models, which reduced its dependence on area ratio (r = -0.03, p = 0.08). Nonetheless, the fact that the area ratio correlates with the residual error indicates that objects with large area differences are more distinct in perception than in models.

3.4. Objects with shared features are more similar in perception than in models

We observed that objects that share parts tend to be more similar in perception than in the combined model (*comb2*) (Figure 4). We tested several ways in which objects could share features. First, we measured the average residual error for pair of objects that shared parts at one or two locations across two objects. Across all models, the residual error was large and negative for objects that shared two parts, smaller but still negative for objects sharing a single part and near-zero for objects with no shared parts (Figure 5C). Thus, objects with shared parts are more similar in perception compared to models. Second, we considered objects that share features because they are two views of the same object, or because they are mirror images of each other. The residual error was consistently negative for these objects as well (Figure 5D). Finally, this pattern was true even for objects that differed in texture but not shape (i.e. shared shape) and for objects that differed in shape but not texture (i.e. shared texture) (Figure 5D). Together these observations indicate that objects with shared features are more similar in perception than according to all computational models.

To confirm that these results (3.2-3.4) are not specific to the validation procedures used here, we repeated our analysis by training the *comb2* model using a leave-one-experiment out procedure (Supplementary Section S3) as well as on individual experiments (Supplementary Section S4) – and obtained similar results.

4. Discussion

Here, we systematically compared object representations in humans with a number of computational models. Our main finding is that all computational models show similar patterns of deviation from human perception. Importantly these deviations occur for specific types of images with identifiable properties, allowing us to make qualitative inferences about the missing elements in computational models. We observed two types of deviations: First, symmetric objects and objects with large differences in area are more distinctive in perception than predicted by all computational models. Second, objects that share features, objects in multiple views and mirror images are more similar in perception compared to all computational models. We propose that incorporating these properties into existing computational models should improve their performance.

We have found that convolutional neural networks outperform nearly all other computational models in explaining perceptual data. This is consistent with recent findings that these networks explain neural dissimilarities in higher visual areas [12, 14-16]. However we have gone further to show that these networks – and even all computational models – show systematic deviations from perception. The fact that all computational models, despite their extremely diverse formulations, exhibit the same systematic differences from perception indicates that they all lack specific properties that must be explicitly incorporated. It also suggests that these deviations are potentially properties that are explicitly computed by higher visual areas in the brain.

We have found that symmetry makes objects more distinct than predicted by computational models. Although symmetry is a salient percept that has been localized to higher visual areas [33, 34], we have shown for the first time that symmetry has a specific consequence for the ability of computational models to predict perceptual dissimilarity. We have also shown recently that symmetry also makes objects more distinct than expected in perception itself [19]. Our finding that mirror images and multiple views of an object are more similar in perception agrees with recent findings that these properties are encoded in high level visual areas [35,36]. Our finding that mirror confusion and view invariance in perception are not explained by convolutional neural networks is somewhat surprising considering that they show pose-invariant classification and are trained on vertical mirror images. It implies that view invariance may be implemented differently by these networks compared to perception. Finally, we speculate that there may well be other systematic patterns in the perceptual data that are not explained by computational models. Our dataset is made available publicly to facilitate their discovery as well as for benchmarking other models. (*http://www.cns.iisc.ernet.in/~sparun/searchmodels/*)

To conclude, our results constitute the first systematic analysis of how computational models deviate from perception. A direct implication of our results is that incorporating these deviations into computational models should improve their performance on classification tasks. This may not be straightforward for several reasons: First, we have compared object representations for isolated objects whereas real-world scenes contain typically many other objects as well as contextual scene information. Second, properties such as view invariance, mirror confusion, symmetry may be incorporated in several ways into models. Therefore, a thorough investigation may be required to confirm whether these properties do indeed improve classification. Nonetheless our results are an important first step towards understanding the performance gap between machine vision and human vision.

References

[1] S. Mitchell. *Tao Te Ching: A new English version*. Harper Collins, 1988.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105, 2012.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77 (1-3): 157–173, 2008.

[6] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[7] D.D. Cox and T. Dean. Neural networks and neuroscience-inspired computer vision. *Current Biology*, 24: R921–9, 2014.

[8] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2: 4, 2008.

[9] N. Kriegeskorte and M. Mur. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, 3: 245, 2012.

[10] S. P. Arun. Turning visual search time on its head. *Vision Research*, 74: 86–92, 2012.

[11] D. D. Leeds, D. A. Seibert, J. A. Pyles, and M. J. Tarr. Comparing visual representations across human fmri and computational vision. *Journal of Vision*, 13 (13): 1–27, 2013.

[12] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10 (11): e1003915, 2014.

[13] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452: 352–355, 2008.

[14] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. Dicarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences U. S. A.*, 111: 8619–24, 2014.

[15] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10 (12): e1003963, 2014.

[16] U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35 (27): 10005–10014, 2015.

[17] R. T. Pramod and S. P. Arun. Features in visual search combine linearly. *Journal of Vision*, 14 (4): 1–20, 2014.

[18] K. Mohan and S. P. Arun. Similarity relations in visual search predict rapid visual categorization. *Journal of Vision*, 12 (11): 1–24, 2012.

[19] R. T. Pramod and S. P. Arun. Object attributes combine additively in visual search. *Journal of vision*, 16(5), 8-8.

[20] D. H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10: 433–436, 1997.

[21] S. P. Arun and C. R. Olson. Global image dissimilarity in macaque inferotemporal cortex predicts human visual search efficiency. *Journal of Neuroscience*, 30 (4): 1258–1269, 2010.

[22] C. T. Zahn and R. Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, 100 (3): 269–281, 1972.

[23] J. M. Cortese and B. P. Dyre. Perceptual similarity of shapes generated from Fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, 22 (1): 133–143, 1996.

[24] A. C. Berg and J. Malik. Geometric blur for template matching. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001,* 1:l-607, 2001.

[25] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005.* 1:26-33, 2005.

[26] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60 (2): 91–110, 2004.

[27] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. Proceedings of the 2005 IEEE Computer Society Conference on,* 1:886:893, 2005.

[28] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40 (1): 49–71, 2000.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13 (4): 600–612, 2004.

[30] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *Computer Vision, 2009 IEEE 12th International Conference on*, 2146–2153, 2009.

[31] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2 (11): 1019–25, 1999.

[32] N. Pinto, D. D. Cox, and J. J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Computational Biology*, 4 (1): e27, 2008.

[33] M. Bertamini and A. D. J. Makin. Brain activity in response to visual symmetry. *Symmetry*, 6 (4): 975–996, 2014.

[34] S. Bona, A. Herbert, C. Toneatto, J. Silvanto, and Z. Cattaneo. The causal role of the lateral occipital complex in visual mirror symmetry detection and grouping: an fMRI-guided TMS study. *Cortex*, 51: 46–55, 2014.

[35] J. E. Rollenhagen and C. R. Olson. Low-frequency oscillations arising from competitive interactions between visual stimuli in macaque inferotemporal cortex. *Journal of Neurophysiology*, 94 (5): 3368–3387, 2005.

[36] N. A. R. Murty and S. P. Arun. Dynamics of 3d view invariance in monkey inferotemporal cortex. *Journal of Neurophysiology*, 113 (7): 2180–2194, 2015.