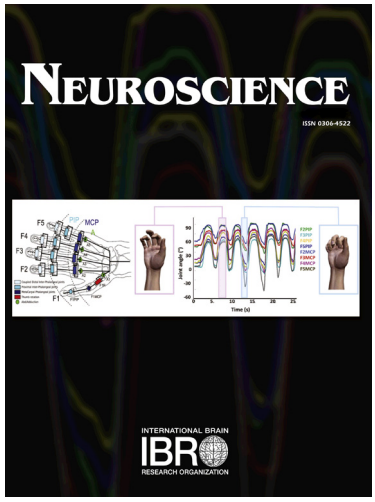
Accepted Manuscript



Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition

Hamid Karimi-Rouzbahani, Nasour Bagheri, Reza Ebrahimpour

PII: S0306-4522(17)30141-0
DOI: http://dx.doi.org/10.1016/j.neuroscience.2017.02.050
Reference: NSC 17639

To appear in: *Neuroscience*

Received Date: 16 December 2016
Revised Date: 17 February 2017
Accepted Date: 21 February 2017

Please cite this article as: H. Karimi-Rouzbahani, N. Bagheri, R. Ebrahimpour, Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition, *Neuroscience* (2017), doi: http://dx.doi.org/10.1016/j.neuroscience.2017.02.050

Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition

Hamid Karimi-Rouzbahani[1,2], Nasour Bagheri[1], Reza Ebrahimpour[2,3,4*]

[1]Department of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran

[2]School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

[3]Cognitive Science Research lab., Department of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran

[4]Institute for Advanced Technologies, Shahid Rajaee Teacher Training University, Tehran, Iran

*To whom correspondence should be addressed.

Email: rebrahimpour@srttu.edu,

Post office Box: 16785-163,

Lab address: http://ccvlab.ir/

Acknowledgments:

This work was supported partially by Shahid Rajaee Teacher Training University under contract number 30318. We are also grateful to Karim Rajaei for his helpful comments and generously editing this manuscript.

Abstract- Humans perform object recognition effortlessly and accurately. However, it is unknown how the visual system copes with variations in objects' appearance and the environmental conditions. Previous studies have suggested that affine variations such as size and position are compensated for in the feed-forward sweep of visual information processing while feedback signals are needed for precise recognition when encountering non-affine variations such as pose and lighting. Yet, no empirical data exist to support this suggestion. We systematically investigated the impact of the above-mentioned affine and non-affine variations on the categorization performance of the feed-forward mechanisms of the human brain. For that purpose, we designed a backward-masking behavioral categorization paradigm as well as a passive viewing EEG recording experiment. On a set of varying stimuli, we found that the feed-forward visual pathways contributed more dominantly to the compensation of variations in size and position compared to lighting and pose. This was reflected in both the amplitude and the latency of the category separability indices obtained from the EEG signals. Using a feed-forward computational model of the ventral visual stream, we also confirmed a more dominant role for the feed-forward visual mechanisms of the brain in the compensation of affine variations. Taken together, our experimental results support the theory that non-affine variations such as pose and lighting may need top-down feedback information from higher areas such as IT and PFC for precise object recognition.

Keywords: invariant object recognition, feed-forward vision, psychophysics, EEG, computational model

Introduction

Primates can accurately perform object categorization in fractions of a second (Thorpe et al., 1996; Fabre-Thorpe et al., 1998), despite substantial variations in objects' size, position, pose and the environmental lighting conditions. It has been suggested that, rapid object categorization is likely to be feed-forward (Riesenhuber and Poggio, 1999; VanRullen, 2007; DiCarlo and Cox, 2007; Afraz et al., 2014), and that more complex stimulus processing is achieved by feedback projections from higher visual areas. The latter include situations in which the target objects are in clutter (Hupe et al., 1998; Bullier, 2001; Lamme et al., 1998), occluded and in low contrast (Wyatte et al., 2012). There are also studies suggesting that object representations which are robust to variations in size and position are mainly constructed in a hardwired feed-forward manner in the visual pathways (Serre et al., 2005; Serre et al., 2007b), whereas for non-affine variations such as pose and lighting the brain needs to activate its top-down information resources at higher areas such as IT and PFC to yield invariant object representations

(Bullier, 2001; Serre et al., 2005). However, these hypotheses lack supporting empirical data, which provided the motivation for the current study.

A set of behavioral studies have addressed the impact of individual variations on categorization performance. These include the studies reporting that changes in objects' size (Jolicoeur, 1987; Peissig et al., 2006; Zoccolan et al., 2009), position (see (Kravitz et al., 2008) for a review), pose (Edelman, 1995; Troje and Bulthoff, 1996), and the lighting conditions (Braje et al., 1998) of the environment exert a considerable influence on categorization performance (i.e. on both accuracy and time). However, no conclusions could be made on the contributions from the feed-forward and/or feedback mechanisms in these studies, since those studies have put no constraints on the feed-forward sweep or the feedback processing of visual information. To separate the contribution of the feed-forward/feedback pathways of information, the backward masking strategy has been frequently used and revealed to be highly effective in blocking the influence of feedback resources on categorization (Pollen, 1999; Lamme and Roelfsema, 2000; Serre et al., 2007a). A recent study, which investigated the feed-forward categorization using backward masking, suggested that the impact of variations on categorization is relative to the level of the applied variation (Ghodrati et al., 2014). However, since the evaluated variations were combined in that work, the relative impact of individual variations on feed-forward performance remains unknown. In the current study, we aimed to find possible differences between the mentioned variations in feed-forward object categorization to understand whether there is a potential need for feedback information when encountering some specific variations rather than others. To do this, an image set was generated in which 3D object models underwent parametrically controlled variations in lighting, pose, size and position independently from one another. The image set was used in a behavioral object categorization set-up with a backward masking protocol. A short stimulus presentation time was chosen to reduce the probability of the integration of top-down with the ongoing bottom-up visual information, so that the observed results can be associated with the feed-forward visual mechanisms.

Although useful in the study of human categorization performance, the behavioral experiments may be influenced by undesirable effects such as decision-related cognitive processes and response-related motor actions inherent in such experiments. To avoid these effects, we also designed a passive EEG recording experiment to gain access to the brain correlates for our behavioral observations. We expected to see differences between variations, since it was previously reported that the activity levels of IT neurons in non-human primates were highly modulated by object variations (Ashbridge et al.,

2000; Freiwald and Tsao, 2010; Booth and Rolls, 1998; Desimone et al., 1984; Vogels and Biederman, 2002; Hung et al., 2005), and that such modulations were simply decodable from whole-brain MEG/EEG data when objects underwent variations in position (Carlson et al., 2011; Karimi-Rouzbahani et al., 2017) and size (Isik et al., 2014; Karimi-Rouzbahani et al., 2017). Our goal was to reduce the intervention of the top-down signals from higher visual areas in categorization. Therefore, contrary to our behavioral experiment, the subjects performed a category-irrelevant color-matching task while their whole-brain EEG signals were recorded. We used a shorter stimulus presentation time compared to previous studies (Carlson et al., 2011; Isik et al., 2014), to avoid potential impacts of feedback information from higher visual areas and to confine the task to feed-forward visual processing. A new analysis method was proposed in this work, which used 'Dunn' clustering index (Dunn, 1973) to explore the representational space of the EEG signals. The method helped in explaining the behavioral observations in time and space and provided several key advantages to traditional decoding approaches (Carlson et al., 2011; Isik et al., 2014; Hung et al., 2005). Isik et al., (2014) compared the dynamics of the appearance of size- and position-invariant representations. Results showed that the processing of size preceded position in time. Here we argue that such comparisons could have been biased since no attempt was made to equalize the separability of size- and position-affected images in the pixel space, nor was the potential bias removed from the representational results in the brain space. To avoid such problems, here we defined a modulation index to provide an unbiased comparison between the four different variations in the representational space.

Finally, a hierarchically-organized feed-forward computational model was used as a 'proof of existence' to provide support that a feed-forward structure seems to be enough to explain the behavioral as well as the EEG observations. The model was selected based on several recent studies supporting its brain plausibility from both the performance as well as the representational aspects (Khaligh-Razavi and Kriegskorte, 2014; Yamins et al., 2014; Cadieu et al., 2014).

Experimental procedures

Stimulus set

An object image set was generated in which 3D object models underwent variations in size, position, pose and lighting. The image set included sixteen distinct object exemplars (freely downloaded from http://tf3dm.com/), which were categorized into the groups of 'animals', 'cars', 'faces' and 'planes' (Fig.

1a). To apply the parametrically controlled variations, Blender software was used (https://www.blender.org/). The size, position and pose of the objects as well as the lighting conditions of the 3D space were altered in different conditions. In the size conditions, the objects were resized so as to cover approximately from 5k to 250k pixels in the pixel space in 9 linear steps. This ranged approximately from 2 to 13.5 degrees of visual angle when the images were presented on the screen in the psychophysical and EEG experiments (Fig. 1b, third row from top). In position conditions, objects were put at different circular radii from the image center to provide different levels of eccentricity from the fovea. This led to 9 steps of position conditions ranging from 0.8 to 7.7 degrees of visual angle into the periphery in the experiments (Fig. 1b, forth row). The variation in pose was applied by rotating the objects around their X, Y and Z Cartesian axes simultaneously in steps of 45 degrees. This led to a total of 8 conditions ranging from 0 to 360 degrees of orientation (Fig. 1b, second row). Size, position and pose conditions shared a default condition which is shown only once in Fig. 1b, highlighted by the orange box. However, this condition (i.e. which shows the objects in 0 pose orientation, 5.8 degrees of size and 0 degree of position eccentricity) is considered in the evaluation of pose, size and position conditions. A uniform light source was used in the three above-mentioned variations which had almost no influence on the objects as they underwent those variations. However, the uniform light source was replaced by a pointing light source in different lighting conditions and was directed to the objects at the same distance but from different angles to generate the 9 lighting conditions: top left, top, top right, right, bottom right, bottom, bottom left, left and front (Fig. 1b, first row). A unique gray-background, 512-by-512 pixel image was generated from each object-exemplar-variation-condition making a total of 560 images in the image set (i.e. 16 exemplars in 35 conditions).

Behavioral experiment

A psychophysical experiment was designed to study the behavior-level feed-forward rapid object categorization performance under different object variations. Twelve human subjects (mean age 21, five females) volunteered in the categorization experiment which lasted for about 30 minutes in four blocks of 140 trials with 5 minutes of rest time in between the blocks. Subjects had normal or corrected-to-normal vision and were seated in a dimly lit room 60 cm away from the monitor and the visual stimuli covered between 2 to 13 degrees of visual angle depending on the size of the stimulus (i.e. as defined in Fig. 1). Matlab PsychoToolbox (Brainard, 1997) was used for stimulus presentation and response recording. Images were presented to different subjects in random orders.

The stimulus followed a 500-ms fixation point and remained on the screen for 25 ms. A random noise mask followed the subsequent 20-ms inter-stimulus interval (ISI), and stayed on the screen for 100 ms. Following the random noise mask, subjects were supposed to make a response indicating the stimulus's category by pressing one of the four predefined keys on the keyboard which randomly changed from subject to subject. Fig. 2a shows the experimental paradigm in more detail. This set-up has been shown to be very effective in avoiding the contributions from most feedback signals in rapid object categorization tasks (Lamme and Roelfsema, 2000; Serre et al., 2007). Subjects were asked to respond as fast and accurately as possible. Subjects' answers and the corresponding response times (as measured from the stimulus onset) were used in the analyses of the behavioral data. The subjects were acquainted with the behavioral task in a training phase performed immediately before the main task.

EEG experiment

In order to provide electrophysiological correlates for the behavioral observations, an EEG paradigm was designed with a few changes to the behavioral paradigm: contrary to the behavioral experiment in which the main image set was presented to the subjects, a sub-sampled version of the image set (Fig. 1, images in the orange and red boxes) was used which included 96 distinct images each of which was presented six times to each subject to increase the signal-to-noise ratio. In the EEG experiment, the subjects' task was passive and designed to neutralize the category-related decision-making processes in the representational space (Fig. 2b). Therefore, a color-matching task was employed aimed at keeping the subjects fully attentive during the experiment. The experimental paradigm is shown in Fig. 2b. There was a fixation point on the center of the screen for 200 ms which accompanied the two following stimuli each of which were presented for 25 ms with a 1200-ms ISI. The fixation point changed color to either red or green when it appeared with the stimuli and would/would not change color from the first stimulus to the second. The task of the subjects was to report whether the fixation point changed color or not by pressing one of the two corresponding keys on the keyboard. Eleven human subjects (mean age 22, three females) volunteered in the EEG recording experiment. Other aspects of the EEG paradigm resembled the behavioral paradigm.

EEG recording and preprocessing

EEG signals were recorded using eWave, a 32-channel 1000-sample per second amplifier designed by ScienceBeam (http://www.sciencebeam.com/) and preprocessed in Matlab software (http://www.mathworks.com/). Among all possible choices for the filtering ranges reported in the EEG-

/MEG-based object decoding literature (e.g. 1-30 Hz (Taghizadeh-Sarabi et al., 2015), 0.1-150 Hz (Behroozi et al., 2015) and 2-100 Hz (Isik et al., 2014)) we chose the 0.1-100 Hz range for band-pass filtering since this range provided the highest Dunn indices in representational analysis, meaning that this range probably covered the most informative range of the EEG signals in object representation. We also notch-filtered the signals at 50 Hz to avoid mains noise. The band-pass and notch filters were linear phase FIR filters as implemented in EEGLAB (Delorme and Makeig, 2004). A separate set of analyses were done with 1-100 Hz band-passed signals to ensure no bias in the time course of the signals as a result of high-pass filtering (Widmann and Schröger, 2012), which revealed no significant effect on the results. Eye-blinks and movement artifacts were removed using Independent Component Analysis (ICA) as implemented in EEGLAB. To find the disrupting ICA components, and to plot the scalp maps, we used respectively the ADJUST (Mognon et al., 2011) and ERPLAB (Lopez-Calderon and Luck, 2014) plug-ins.

Data statistics

On average, 97.25% of the trials (6162 out of 6336 trials for the eleven subjects, s.d. = 2.1%) passed the artifact removal procedure and were used in the representational analyses. After the noise and artifact removal, stimulus-aligned channel activities were extracted by epoching the signals in the window from 200 ms pre-stimulus to 800 ms post-stimulus. Channel activities were averaged using a 5-ms moving-average filter to increase the signal-to-noise ratio. This window was chosen so, as it was not too coarse to lose the timing information and not too fine to be affected by the noise. The signal epoching step resulted in a 3-dimensional data matrix 'X' for each subject with 31 rows (i.e. number of channels), 200 columns (i.e. number of samples in the -200 to +800-ms window relative to the stimulus) and 576 layers (i.e. number of trials for a subject with no removed trials). The values in the matrix were the measured activities on the EEG electrodes (i.e. voltage values in microvolts) which were either positive or negative relative to the reference electrode.

To reach the representational Dunn index values for each time point in the analyses (i.e. results of Figs. 5, 6 and 7), we used the data values from the mentioned data matrix on every time sample (i.e. every matrix column). On every time sample (i.e. each column of the 'X' data matrix), to measure the separability between the four categories, one activity point was obtained in the 31-dimensional space (Fig 3). For instance, to find the corresponding point for the 8th trial at the +210 ms post stimulus, we used the 31 values in all rows of the 82th column (from left to right) of the 8th layer of the matrix. The 31 dimensions referred to the recording channels (electrodes) of the EEG amplifier. After finding all the

576 points corresponding to the representation of the four categories in the time step, the Dunn index was calculated using equation (1).

Representational analysis

Multivariate pattern classification of electrophysiological signals, also referred to as neural decoding, has recently been highly popularized for representational analyses in monkey and human studies (Carlson et al., 2011; Isik et al., 2014; Hung et al., 2005). Here we proposed using clustering evaluation indices as a substitute for decoding as they provide several key advantages:

- Clustering indices avoid many hyper-parameters which have to be determined in most decoding schemes each of which may influence the results. These include the train/test data sampling strategy, the choice of the cross-validation method to avoid over-fitting of the classifier, the type of the classifier, etc.
- Clustering indices, while being simple in formulation, directly capture the inter- and intra-class distances of individual exemplars while the decoding schemes provide an implicit measure of how good/bad the samples are distributed in the representational space.
- They avoid the ceiling effect which occurs when decoding totally-separable category clusters.
- Although there might be very rare cases in which two sets of category representations provide equal clustering indices, many more cases may be found where the decoding rates are equal while the category representations are positioned differently in the representational space.

Therefore, we used the 'Dunn' clustering index (Dunn, 1973) to explore the representational spaces obtained from the EEG signals as well as the computational model. The clustering evaluation indices have been generally used as metrics for evaluating clustering algorithms (Davies and Bouldin, 1979; Rousseeuw, 1987). However, using the definition below the Dunn index can measure the separability of category clusters in the representational space:

$$Dunn\ index = \frac{min_{1 \le i < j \le n} d(i,j)}{max_{1 \le k \le n} d'(k)} \qquad (1)$$

where $d(i,j)$ stands for the distance between classes i and j, and $d'(k)$ represents the intra-class distance for class k. The inter-class distance, $d(i,j)$, between the two classes may be any type of distance measure, such as the distance between the centroids of the classes as in this study. Similarly,

the intra-class distance $d'(k)$ may be measured in different ways. In this study, $d'(k)$ was the maximal distance between any pairs of elements in class k. Here, the four classes were animals, cars, faces and planes and the Euclidean distance was the distance measure.

As it is obvious from (1), since the Dunn index does not take into account the dimensionality of the representational space and changes in proportion to the inter-class and intra-class distances of category clusters, it can reflect the separability of category clusters in different representational spaces such as in pixel, EEG or computational model spaces. Taking advantage of these characteristics, we used the Dunn index to measure the improvement made to the category representations as they traveled from the pixel space to the EEG or the model spaces using the modulation index:

$$Modulation\ index = \frac{DI_2 - DI_1}{DI_2 + DI_1} \quad (2)$$

where DI_1 and DI_2 refer to the Dunn indices calculated in the first and the second spaces, respectively. In this work, the first space is always the pixel space and the second space is either the EEG or the model space. The modulation index could range from -1 to +1. Positive and negative values for the modulation index reflect respectively an increase and decrease in the category separability when going from the pixel to the EEG/model space.

Computational model

A hierarchical feed-forward computational model was used here to provide insight into the hierarchical strategy used by the visual system to yield invariance to variations. The model, known as 'AlexNet', has been recently developed based on the structure of the traditional convolutional neural networks and surpassed the highest performing machines on the ImageNet Large Scale Visual Categorization Challenge (ILSVRC) (http://www.image-net.org/). More importantly for this study, the model has provided highly precise predictions of the object representations observed at area IT of human (Khaligh-Razavi and Kriegskorte, 2014) and non-human (Yamins et al., 2014) subjects. The model, which is an 8-layer stack of feature extractors, comprised of alternating blocks of mathematical operations such as convolution, regularization, normalization and max-pooling and provided invariance to variations as image representations passed the consecutive model layers. The model was developed by Krizevsky et al., (2012) and provided in Matlab code by Vedaldi and Lenc, (2015) (available at http://www.vlfeat.org/matconvnet/). Using a gradient descent algorithm, the model was trained to discriminate between 1000 object categories from the ImageNet image sets which included the object

categories of the current study as well. To test whether the model accounted for the observations in the EEG representational analyses, it was applied to the same four-class image set and then the same representational analyses (i.e. Dunn and Modulation indices) were conducted on the representations provided by the model layers.

To compare the model's accuracy with that of humans, a maximum correlation classifier was applied to the object representational vectors obtained from the model's seventh layer as in (3). Among all more complex classifiers, the maximum correlation classifier was chosen since it was simple and influenced by the representational space more easily. The classifier is formulated as in (3):

$$i^* = argmax_i[corr(x^*, \bar{x}_i)] \quad (3)$$

where x^* refers to the test object's representational vector obtained from the model's seventh layer and $\bar{x}_i$ is calculated by averaging class i representational vectors from the training set images. x^* belongs to class i^* if $\bar{x}_i$ is the most correlated average representational vector from all training classes to x^*. To avoid ceiling of performance, only 50% of the model representations were provided to the classifier for training and the remaining 50% for testing. The training/testing sets were randomly resampled 200 times from the object representational set to avoid the possible noise in the data. The 200 repetitions were randomly subsampled to twelve samples to avoid any potential bias when comparing the significance results with those of the twelve human subjects. Although the 50% choice for the size of the training set underestimated the model's performance, it was of no concern since the categorization patterns of the model were to be compared with those of humans and not their absolute values.

Results

As stated earlier, our primary goal was to measure any possible differences between the two classes of variations on a feed-forward object categorization task. The classes included affine variations of position and size and supposedly more difficult (non-affine) variations of pose and lighting.

Humans' object categorization accuracy as well as their response times were obtained from the behavioral experiment (Fig. 2a). The response times were averaged across the correctly answered trials which took the subjects longer than 200 ms and shorter than 3000 ms to respond (i.e. this included more than 99% of trials). This time span was chosen since other correctly answered trials were considered to be either the outcome of an unintentionally button press or some other undesirable

cognitive processes which intervened (e.g. prior information coming from higher brain areas) in feed-forward categorization. The significance of the difference between all possible pairs of variations was evaluated using Wilcoxon's signed-rank test and the results were summarized in cross-variation significance matrices below the bar graphs in Fig. 3. To obtain each element of the cross-variation significance matrix, the performance/response-time vectors for a pair of variations were provided to the Wilcoxon's test and a probability measure was obtained and reported in p-value in the cross-variation matrix. These performance/response-time vectors consisted of 12 values, each of which was obtained from one human subject or a single model run.

On average, the participants responded correctly in 93.07% of the trials (±25.4%). The average correct response time was 832 ms across the subjects (±343 ms). Lighting, position, size and pose appeared to have the lowest to highest influences on the categorization accuracy with the mean accuracy of 96.56%, 95.5%, 94.09% and 84.81%, respectively (Fig. 3a, blue bars). Except for the lighting-position pair, other pairs of variations revealed significant differences. This is interesting that pose showed a profound impact on the categorization accuracy. This was reflected in the longer average response time of categorization for pose (883 ms) than for other variations (780, 786 and 803 ms for lighting, size and position, respectively). These results provide proof for imperfect categorization accuracy in humans under different object variations and environmental conditions. Although against a few studies supporting invariant object categorization in humans in specific paradigms (Guyonneau et al., 2006; Corballis et al., 1978), these results are on par with a big set of studies supporting the dependency of categorization on variations (Jolicoeur, 1987; Peissig et al., 2006; Zoccolan et al., 2009; Kravitz et al., 2008; Edelman, 1995; Troje and Bulthoff, 1996; Braje et al., 1998). In addition, Fig. 3a suggests significant differences between variations: pose significantly affected human performance in both accuracy and speed while lighting had the least impact on categorization. This can be explained by the mental alignment which was suggested to be needed for the compensation of pose (Hamm and McMullen, 1998), rather than other variations such as position and lighting. The results for the sub-sampled version of the main image set, namely 'the EEG image set' provided the same results (Fig. 3b) as the main image set. The lowest performance was still observed for pose, whereas size increased its influence on accuracy and speed leading to a lower performance compared to the results in Fig. 3a. Size no longer resulted in a significantly higher accuracy compared to pose. This was also reflected in the response times of size which showed an insignificant difference with pose in Fig. 3b.

We next sought to see whether the computational model is consistent with our empirical results. The application of the computational model on the same image set provided the same patterns of behavior as humans for different variations (Fig. 3a and b red bars): the highest average categorization rates were for lighting and position with 84.14% and 84.07%, respectively, and insignificantly and significantly lower categorization rates were observed for size and pose with 81.91% and 70.27%, on the main image set. A human-like decline in the model's accuracy was also observed for the size variation when investigating results of the EEG image set (Fig. 3b). Therefore, the model appears to have adopted a human-like strategy in the categorization of objects, at least at the behavioral level. This was interesting since no human data had been used to train the model parameters, and it had been trained to maximize the recognition accuracy on a different image set (Krizhevsky et al., 2012).

To provide an insight into the conditions which were the simplest and hardest in the categorization task, we next resolved the variations into the constituent conditions (Fig. 4a-d). Very high categorization rates were achieved in all lighting conditions (Fig. 4a), which was not significantly different between any possible pairs of lighting conditions (i.e. $p > 0.1$, cross-condition matrices were calculated over 12-element vectors of accuracy in humans over different conditions, $n = 12$), not even in the front lighting condition (condition 9). This means that the brain could compensate for all lighting conditions with no difficulty, which is at odds with the results of previous work which observed significant modulations of performance in different lighting conditions in face recognition (Braje et al., 1998). The disparity of the results may be explained by the different mechanisms involved in the recognition of faces and objects (Freiwald and Tsao, 2010) and the differences in the experimental paradigm.

The categorization performance under pose variation (Fig. 4b), however, showed a significant decline ($p < 0.05$) in the middle conditions (i.e. conditions 3 to 7) where the objects underwent between 120 to 260 degrees of in-depth orientation. This result concurs with the suggestion that mental rotation becomes more difficult as the objects undergo higher levels of rotation in time-limited tasks (Hamm and McMullen, 1998). Expectedly, the small size conditions (i.e. conditions 1 and 2) suffered from a significant decline of categorization accuracy ($p < 0.05$), which may be explained by the paucity of information provided to the subjects as a result of object shrinkage (Fig. 4c). The smaller amount of information in the small object sizes is explained by the reduced number of retinal photoreceptors evoked by the smaller-sized object. Objects' position proved to be very influential on performance when the objects were positioned more peripherally relative to the image center ($p < 0.1$ when comparing

conditions 1 to 6, Fig. 4d), which is in accordance with a big set of results reviewed in (Kravitz et al., 2008).

The position effect is explained by the lower density of photoreceptors in the peripheral retina causing less visual acuity compared to the foveal regions (Curcio et al., 1987). As opposed to the variation in pose which yielded a U-like performance curve, object categorization under the variations of size and position did not show any evidence for a reference-based object transformation in the brain (Hamm and McMullen, 1998).

The computational model showed a highly correlated categorization pattern with human subjects (Pearson's linear correlation, $r = 0.93$, $p < 10^{-15}$, $n = 35$). The condition-wise correlation between the humans and the computational model is shown in Fig. 4e. Human-model correlation coefficients indicated significant positive correlations for pose and size (with $p < 0.01$, $n = 8$ and $p < 10^{-6}$, $n = 9$ respectively), barely significant correlation for position ($p < 0.1$, $n = 9$) and insignificant correlation for lighting ($p = 0.82$, $n = 9$). Hence, the condition-wise correlation results supported the human-plausibility of the computational model. However, no conclusions can be made on the brain-plausibility of the model in the middle layers since the reported performances were calculated using the model's 7th layer representations.

Next we asked whether the model implemented a hierarchical solution to reach the human-plausible performances, as it was expected to be the case in humans. To investigate this matter, the correlation was calculated between the classification results at different model layers and the categorization accuracy of the human subjects (Fig. 4f). The human-model correlation increased as the object representations traveled from the pixel space to the input model layers and continued up to the last model layer. Only the two final model layers showed significant correlations with humans ($P < 0.001$, $n = 35$, red curve) and the other layers were only insignificantly correlated with the human ($p > 0.07$). In light of these results, the model has yielded human level categorization in a systematically hierarchical manner and not by accident. This not only added confidence to the brain-plausibility of the model, but also made the model a proper candidate for analyzing the hierarchical structure of the recorded EEG signals in the following analyses.

The results provided thus far facilitated the comparison between different variations at the behavioral level. These results proposed that, lighting and position were the two simpler variations compared to pose and size in the current categorization task. However, since the results were obtained

from the highest level of abstraction namely 'behavior', it failed to provide a detailed description of the differences inherent in the variations. Moreover, the evaluation of categorization performance at the behavioral level, although useful, may not be precise enough since it involves other cognitive processes such as decision making as well as motor actions like finger movements, each of which can be biased towards one of the experimental conditions. Therefore, we evaluated the brain representations which drive the behavioral outputs using the above-explained EEG recording paradigm and the representational analysis methodologies.

The following results for the EEG and the computational model are obtained from the EEG image set (as indicated in colored boxes in Fig. 1). Fig. 5 shows temporally resolved Dunn indices for the pool of variations and each variation separately. As mentioned earlier, the average number of trials which passed the artifact removal was 560.1 (s.d. = 9.4) for each subject, on which the following representational analyses were made. The shaded error areas and bars in Figs. 5 and 6 reflect the standard error across subjects. As Fig. 5a shows, the pooled Dunn index rose to significance at 98 ms ($p < 0.05$ as determined by Wilcoxon's signed-rank test with $n = 11$; the Dunn indices were considered significant at the time points when they rose significantly above the average Dunn value in the last 200 ms pre-stimulus window) and remained significant until 338 ms post-stimulus onset. This improvement was a result of changes in the categorical representations, with category clusters increasing their inter-category distances while each of them experienced a decrease in intra-category distances. This was referred to as 'untangling' of category representations (DiCarlo and Cox, 2007). The untangling of representations peaked at 186 ms post-stimulus and remained significant for as long as 240 ms. This timing dynamics is highly consistent with most object representational studies which reported the time course of visual processing in humans (Kaneshiro et al., 2015; Cichy et al., 2014). It should be noted that, the reported Dunn index is highly affected by the number of samples in the stimulus set (i.e. brain representational data points); in most cases it decreases when the number of samples increases in clusters. That is the reason why, unlike the decoding schemes, we did not decide on how high the obtained Dunn index was, and just compared that with a pre-stimulus window (i.e. the 200-ms pre-stimulus window here) to investigate significant positive increases.

The Dunn indices for individual variations were obtained as well (Fig. 5b). Differences were observed in the amplitudes as well as the temporal dynamics between the variations. A shorter rising latency (i.e. latency was defined as the first post-stimulus significant time point) was observed for lighting and size compared to pose and position (mean 83, 92, 95 and 106 ms respectively for lighting,

size, position and pose, Fig. 5c). The cross-variation significance matrices, shown in Fig. 5c, were calculated using 11-element vectors which contained the latency/peak times calculated for the Dunn indices obtained from individual EEG subjects. The rising latency was significantly shorter for lighting than for pose and position (with $p < 0.05$ and $p < 0.001$ with $n = 11$, respectively), which concurs with the results of the behavioral experiment with a significantly longer response time for pose compared to lighting ($p < 0.001$, Fig. 3b). Results in Fig. 5 also confirmed the earlier decoding time observed previously for size than for position (Isik et al., 2014), and highly correlated with the reported delays in the appearance of size, in-depth rotation and position-invariant representations in the human brain (Liu et al., 2003). Ritchie et al. (2015), suggested that the brain uses the optimal decoding time (i.e. the peak time of the Dunn index value here) to decide on the category of the perceived object; therefore, the behavioral response times should most probably be predicted by the peak Dunn index time for each variation. The peak untangling times are shown in Fig. 5c for each variation, which indicate a disadvantage for pose compared to other variations. These results, although not as correlated with the behavioral response times as the latency times, concurred with the study reporting a correlation between the peak decoding and the categorization times (Ritchie et al., 2015). To the best knowledge of the authors, the dynamics of variations in lighting and pose have never been pitted against other more-studied variations such as size and position, which is covered in the present study. This work, while confirming previous observations on the precedence of size to position (Isik et al., 2014), suggests that variations in lighting and pose are respectively the least and the most influential variations on the untangling time in object categorization.

We next investigated the amplitudes of the EEG Dunn indices to find brain correlates for the performance patterns observed in the behavioral experiment. In fact, we made the assumption that the object categorization accuracy was determined by how untangled the object representations of different categories were in the brain representational space. Possible correlations between the EEG Dunn indices and the behavioral measurements would not only support this assumption, but could also validate the index to be a proper metric for measuring category separability in the EEG space. Since the Dunn indices for different variations were time-variant values, direct comparisons between the EEG Dunn indices and the scalar behavioral accuracy values would fail to reveal EEG's dynamical behavior. Accordingly, the 50- 400 ms post-stimulus span was divided into 7 non-overlapping windows. This span was chosen since it covered the time-window in which all variation indices remained positive relative to their baseline values and could change their ranking when compared. This span was also considered by many to cover most of the brain's feed-forward visual processing activities (Thorpe, 1996; Isik et al.,

2014; Kaneshiro et al., 2015). The Dunn indices for variations were averaged in the 7 mentioned windows and pit against one another in Fig. 6a.

A higher Dunn index value was observed for lighting through all time-windows (Fig. 6a). Ascending Dunn indices were observed in the first half of the windows (i.e. 50 to 250 ms) which became descending in the later windows (i.e. the windows from 250 to 400 ms) for all variations except for position. Position revealed a more complex behavior; it remained lower than other variations in most preceding windows (i.e. the windows from 100 to 300 ms), outperformed pose and size and approached lighting in the final windows (i.e. 300 to 400 ms windows). The order of the Dunn indices for different variations showed the highest correlations with the behavioral patterns of accuracy in the last two time windows where lighting and position dominated size and pose (as it is in other following windows until the 650-700 ms window, data not shown). These results suggest a possibly different untangling process involved in the enhancement of position-affected representations which needs more processing time to take effect compared to the other variations under study. This is interesting that the latency in the appearance of position-affected representations is compensated for at the behavioral level by an earlier behavioral response time, when compared to other variations. This might be explained by the hypothesis that, contrary to the variations in pose and lighting, the variation in object position is less expected to influence the semantics of object categories (Kravitz et al., 2008). Therefore, a semantic labeling procedure seems to be absent when associating the representations to categorical decisions under the variation of position (Kravitz et al., 2008), resulting in an earlier response time.

One important question that the current study is trying to answer is: which variation is simpler and which is more complex for the feed-forward mechanisms of the visual system? The answer to this question can highly constrain the feed-forward/feedback models of the visual system. Although the reported behavioral patterns of performance or the EEG Dunn indices for variations might seem to suffice to answer this question at first sight, this comparison is unfair since it overlooks the fact that the input image set might have been biased towards one of the variations under study (e.g. lighting or position). Therefore, in order to conclude correctly, we calculated the Dunn indices for all variations in the input space (i.e. pixel space), and used them to calculate the improvement made to each variation by the brain using the modulation index defined in (2). The pixel-space Dunn indices are shown in Fig. 6a, and the resultant modulation indices in Fig. 6b. Variations in lighting, pose, size and position had the highest to lowest Dunn indices in the input space, with 0.2872, 0.2127, 0.2037 and 0.1950, respectively. Therefore, the pixel-space Dunn indices showed a high level of bias towards lighting. This proposes that

the observed higher EEG Dunn indices for the lighting compared to other variations in Fig. 6a might be the result of this bias.

Using the introduced modulation index, now it was possible to compare the variations under study. As shown in Fig. 6b, position, size, pose and lighting showed the highest to lowest modulation indices respectively for almost all time windows. This order was violated in the 50 to 250 ms windows with size being modulated more than position. This result suggested that the feed-forward paths have played a more dominant role at compensating for variations in position and size compared to lighting and pose. The observed order for the amplitude and the timing of the variations under study, while reappraising previous observations on the timing of object categorization (Freiwald and Tsao, 2010; Liu et al., 2009, Goddard et al., 2016), concurs with the suggestion that some variations such as position and size may be compensated for by hard-wired mechanisms (e.g. feed-forward mechanisms as they were dominant in the current recordings) while other variations such as pose and lighting may need auxiliary top-down mechanisms (e.g. feedback mechanisms as they were highly suppressed in this study) to involve in order to achieve an acceptable recognition performance (Bullier, 2001; Serre et al., 2005).

To inspect the contribution of different brain areas in the reported untangling results over the scalp, the Dunn indices were recalculated using the signals from individual electrodes for the pool of variations as well as for individual variations. In other words, in this analysis only a single electrode was considered each time to calculate the Dunn indices and the final scalp maps were generated by the superposition of individual electrodes at their relative locations. The generated scalp maps for the pool of variations and individual variations are shown at different time windows in Fig. 7.

Dunn indices declined significantly in Fig. 7, compared to the previously-reported values in Fig. 5, as a result of calculating in one dimensional space (i.e. single channel). The mean Dunn indices from individual electrodes ranged from 0.0292 (significantly higher than the last 200 ms prior to stimulus onset using Wilcoxon's signed-rank test, $p < 0.005$, $n = 11$) in anterior locations (consisting of electrodes F7, F3, Fz, F4, F8, AFz, FP1, FP2 and FPz) to 0.0419 ($p < 0.005$) in posterior locations (consisting of electrodes P7, P3, Pz, P4, P8, POz, O1, O2 and Oz) in the first 400 ms window after the stimulus onset on the pooled variation (Fig. 7a) which was significantly ($p < 0.005$) higher compared to anterior locations in the same window. The Dunn index started to rise in the 50-100 ms window and became significant ($p < 0.05$) in the occipital area (O1, O2 and Oz) in the 100-150 ms window. This observation replicated the appearance of low level feature differences among categories, particularly for the face category, found

at occipital areas at around the same time (i.e. the P1 component in ERP analysis) (Itier and Taylor, 2004; Rossion and Caharel, 2011).

The higher untangling indices were then observed on the whole posterior area in 150-200 ms along with activities at frontal areas (electrodes AFz, FP1, FP2 and FPz). The temporal dynamics of the Dunn scalp maps in Fig. 7a, though calculated methodologically different from the one used in ERP analyses, were in agreement with previously observed ERP results in humans which reported the start of category-detection signatures at around 150 ms which lasted until approximately 350 ms post-stimulus (Thorpe et al., 1996). However in Thorpe et al. (1996), the category-detection signatures were more localized in frontal areas from 150 to 200 ms and did not totally shift to posterior areas until 325 ms, while in the current study the most informative regions were the occipital, frontal, lateral occipital and parietal areas in the 150-350 ms post-stimulus window. It should be noted that, several paradigm-related variables such as the type and demand of the task and the visual stimuli can affect the distribution of the activity over the scalp; as it is the basis for all EEG analyses. It is suggested (Thorpe et al., 1996) that, the initial signature activity which was observed in the frontal areas was provoked by the go/no-go task to suppress the motor actions in no-go trials (Sasaki et al., 1998). Therefore, the absence of a strong frontal activity in the current study, which was performed under a passive paradigm, as it was the case in Kaneshiro et al., (2015), comes as no surprise. On the other hand, the observed weak untangling activities at the frontal areas in the 150-300 ms windows, might be provoked by the color-matching task in the present study, an interpretation supported by the studies reporting frontal activities specific to decision-making and motor actions at around the same latency (Thierry et al., 2012). The untangling activity then moved towards the lateral-occipital (i.e. P7 and P8) and temporal locations in the 200-250 ms window and started to fade away after 350 ms post-stimulus. These results are on par with recent findings which showed the occipital and lateral-occipital electrodes were the most influential channels in reflecting category-related information in a passive recording paradigm (Kaneshiro et al., 2015).

Fig. 7b shows the Dunn scalp maps which were normalized to the maximum observed Dunn index calculated across all variations. Although noisier here compared to the results in Fig. 7a as a result of fewer number of images, several observations were made for individual variations. For lighting, the average Dunn indices observed at the occipital locations (O1, O2, Oz) were significantly ($p < 0.05$, $n = 11$) higher than those found at parietal locations (POz and Pz) in the 100-250 ms windows, while the opposite was true for the other variations with significantly higher average Dunn indices at POz and Pz

locations. This result, while suggesting the significant role of the primary visual areas in the representation of the low-level image properties, provided evidence for the involvement of parietal areas in the mental rotation, size transformation and position representation (Muthukumaraswamy et al., 2003; Sereno and Lehky, 2011; Freud et al., 2016). It was previously shown that the latency of the ERP components such as N1/N170, which were suggested to be involved in face and object processing, vary with the type and the amount of applied variation (Itier and Taylor, 2004; Muthukumaraswamy et al., 2003). Therefore, the high separability indices observed in the 300-350 ms window at the parietal and central locations in size and position scalp maps, might be explained by the time needed to reach invariant representations for more difficult conditions in those variations, as it was the case in (Muthukumaraswamy et al., 2003). The observation of high Dunn indices in the parietal areas of the scalp maps is well supported by several recent studies showing the role of these areas in shape and category representations (Freud et al., 2016; Sereno and Lehky, 2011; Rishel et al., 2013; Swaminathan and Freedman, 2012).

Although useful for a coarse inspection of the visual processing in the brain, the proposed Dunn scalp maps failed to clearly present the expected feed-forward mechanisms involved in processing of the EEG signals. As a solution, we used the above-mentioned computational model to unveil the hierarchical structure underlying the EEG observations. Results indicated that a hierarchically organized feed-forward structure reasonably similar to the computational model must be underlying the observed EEG effects shown in Figs. 6 and 7 (data not shown). The hierarchical correlation between the untangling indices of the EEG and the computational model, proposed a potential hierarchical similarity between the strategy adopted by the humans and the model in the compensation of variations. To examine this, we calculated the Dunn and modulation indices at the output of each model layer for each variation (Fig. 8). The model's Dunn indices (Fig. 8a) were highly similar to the Dunn indices calculated from the EEG experiment (Fig. 6a): highest Dunn indices were obtained for lighting throughout the model layers compared to other variations which concurred with the EEG results through all time windows. Mounting Dunn indices were observed for position when going from the first to the last model layer with the EEG counterparts in early and late time-windows, respectively. Interestingly, in the model as well as the EEG results, the final untangling results (the results from the last model layer and the last time window) were highly correlated with the pattern of behavioral performance shown in Fig. 3b: higher Dunn indices for lighting and position as there were higher categorization performance for lighting and position at the behavioral level. Although one-to-one comparisons should not be made between the results from the model layers and the EEG time windows as they may represent different time scales, a bilateral

interpretation can result in a high-resolution spatio-temporal insight into how 'invariance' is achieved in the feed-forward visual system.

Position was the variation which underwent the highest amount of improvement going from the pixel space to the model space in all layers (Fig. 8b) as it was in the EEG (Fig. 6b). The last two model layers showed higher correlations with the EEG results: lighting showed the lowest and position the highest values of modulation indices. There are abundant studies reporting that the neural representations of objects reach the top layers of the ventral visual stream (i.e. V4 and IT) in less than 200 ms post-stimulus (Ashbridge et al., 2003; Desimone et al., 1984; Hung et al., 2005). Accordingly, it is not surprising to see high correlations between the outputs from the final model layers and the EEG windows after the 50-150 ms post-stimulus (Fig. 6b), since those layers have been shown to strongly imitate V4 and IT level representations of the brain (Khaligh-Razavi and Kriegskorte, 2014; Yamins et al., 2014). Regarding these results, the model seemed to mimic the representational untangling and modulation results observed in the EEG experiments. Together, the modulation results from the EEG experiment and the computational model provided evidence for a higher improvement made to variations in position and size compared to variations in lighting and pose. The EEG results also showed precedence and tardiness respectively in the appearance of lighting- and position-affected representations (Fig. 5b). This may be supported by the early and late rises of the Dunn indices in the beginning and final model layers respectively for the variations in lighting and position (Fig. 8a).

Following the computational evaluation of different variations which provided support for the brain-plausibility of the computational model, we used the model to generalize from the EEG image set to the main image set to examine whether the modulations observed for the variations would remain the same or whether they depended on the number of samples in the image set. The modulation indices obtained from the main image set are summarized in Table 1.

A big advantage for position was observed (Table 1) in all layers (as in Fig. 8), along with a continuous improvement for size and a disadvantage for pose. These results, while reconfirming the advantage for position and size compared to lighting and pose on a more complex image set with 592 sample images (compared to the EEG image set with only 96 samples), showed that the observed differences between the variations (Fig. 5 and Fig. 8) were not affected by the number of conditions but caused by the solutions provided by the feed-forward visual mechanisms to compensate for those variations.

Discussion

Feed-forward mechanisms are responsible for affine variations

We found that variations in pose and lighting caused more difficulty for the feed-forward visual mechanisms than variations in size and position. It was also shown that variations influenced both the level and the temporal dynamics of the category separability. The relative patterns of categorization performance measured in human behavior were in more agreement with the separability of category representations at higher visual areas than in lower areas. Moreover, the differences observed between the studied variations in the brain representations could be explained by a hierarchically organized feed-forward model of the visual system. Our results suggest that the feed-forward visual pathways show stronger impact on the compensation of affine variations. When encountering non-affine variations, however, those pathways fail to form representations with enough separability among categories. Even though there was no previous empirical evidence to support these findings, they can be interpreted in light of several theoretical studies (Bullier, 2001; Serre et al., 2005). In fact, the main contribution of the current study is to provide behavioral, electrophysiological and computational evidence supporting the hypotheses made in (Bullier, 2001; Serre et al., 2005) on the ineffectiveness of the feed-forward mechanisms when encountering complex variations.

Although several investigations reported the impact of variations on object categorization (Jolicoeur, 1987; Peissig et al., 2006; Zoccolan et al., 2009; Kravitz et al., 2008; Edelman, 1995; Troje and Bulthoff, 1996; Braje et al., 1998), they made no attempt to break up the contribution of the feed-forward or the feedback mechanisms in the task. Yet, using a backward masking paradigm along with a very short stimulus presentation time, we tried to neutralize the feedback and/or attentional effects (Chikkerur et al., 2010; Milner, 1974; Afraz et al., 2014) which could intervene in the behavioral experiment. As there are massive mechanisms of recurrent processing within the cortical pathways (Felleman and Van Essen, 1991), it is worth noting that by feedback, we mean the long back-projections (e.g. from higher visual areas such as IT to V1) and not the short recurrent loops (e.g. the horizontal connections within areas or between adjacent areas such as V1-V2). According to this definition, and the effectiveness of the backward masking on long feedback connections, it is inferred that the results of the behavioral experiment are still contributed to by the short recurrent circuitry. However, no suggestions have ever been made on the explicit role of these circuitries in the compensation of

variations. In the EEG experiment, we believe that the considerations such as the passive paradigm, as well as the randomly-ordered images with a short presentation time (25 ms) have resulted in the domination of the feed-forward mechanisms in the recording. This was reflected in the brain activities appearing mainly between 80 to 110 ms post-stimulus. In addition, the brain-plausible computational model provided a feed-forward explanation for the observations made in the EEG recording, supporting the domination of the feed-forward effects in the signals.

Dunn or modulation index, which one better unveils the differences between variations?

As it was explained in the Representational analysis section followed by the results in Figs 6 and 8, we used a modulation index (equation (2)) to remove the bias from the comparison made between variations. This index removed the input bias from representational separability index (i.e. Dunn index) and provided a measure showing the improvement made to the category representations as they travelled from the pixel space to the brain space. More specifically, the Dunn indices extracted from the EEG (Fig. 6a) and the model (Fig. 8a) representations were used to explain the behavioral results (Fig. 3b) as they revealed highly similar patterns to the behavioral results, while the modulation indices were used to answer the main question of this study: which variations were compensated for by the feed-forward visual processing mechanisms? As it was shown in Fig. 6a, in the pixel space, the category representations were more separable under the lighting variation compared to position. If we had compared the EEG Dunn indices in Fig. 6a, we would have concluded that lighting was the simplest variation among the variations under study, as this was done in (Isik et al., 2014). However, the modulation results in Fig. 6b showed that lighting was the least improved variation. That is why we only compared the variations based on their corresponding modulation indices and not their Dunn indices.

But, where do the observed differences across variations come from?

Robustness to variations in size and position was proposed to be achieved as the object representations pass layers of simple and complex cells in the ventral visual stream (Serre et al., 2005; Hong et al., 2016). Simple cells are suggested to provide generalization across objects (Poggio and Bizzi, 2004) while the complex cells cause invariance to position and size, by pooling the outputs of several simple cells with the same selectivity but at different positions and scales using a max-like operation (Riesenhuber and Poggio, 1999). While invariance to size and position appears to be hard-wired in the visual pathways especially at IT and not specific to particular objects (Tanaka, 1996), many IT neurons have shown to be highly influenced by the object poses which the subjects have experienced (Logothetis

et al., 1995). In light of these findings, it was suggested that variations such as lighting and pose needed prior visual experience with specific lighting and viewing conditions of the target objects in recognition, while the compensation of affine transformations of objects such as position and size, which are common to all objects, are hard-wired in the visual system (Serre et al., 2005). Results of the present study provided proof that the feed-forward mechanisms alone fail to enhance object representations under variations of pose and lighting as much as they do when encountering variations in size and position. However, the relative contribution of the feedback mechanisms to the compensation of non-affine variations still remains unknown. Future studies need to be conducted to unveil the answer to this query. For that purpose, causality methodologies such as the one proposed in (Goddard et al., 2016) can be utilized to separate the contribution of the feed-forward and feedback mechanisms in the representational analyses.

Model layers depict information beyond the EEG signals

The computational model was used to explain the behavioral as well as the electrophysiological observations. Although the model could predict the human behavior (Fig. 4e) and the EEG patterns accurately, the one-to-one correspondence between the model layers and reasonable EEG time windows appears far-reaching (i.e. compare Fig. 6 and Fig. 8). In other words, as the final model layer is supposed to imitate IT-level representations, and the visual information reaches the highest visual areas (e.g. IT) at around 100 ms post-stimulus, we expected the model's final layer to correlate maximally with the EEG signals at around the same time. However, the model's final layer becomes maximally correlated with the EEG signals at around 350 ms post-stimulus (data not shown). There can be two reasons behind this disparity. First is the low spatial resolution in EEG recording. The EEG channels provide weighted average representations from their nearby sources of activity, which means that each channel incorporates activities belonging to different brain areas including non-visual locations. This not only concealed the true dynamics of different brain areas, but also formed new dynamics for the whole-brain activity which could be totally distracting (i.e. this might be the case here). The second and more reasonable interpretation can lie in the time-invariant nature of the model. In other words, since the model lacked 'time' in its structure, it could have become maximally correlated with the EEG signals after a shorter or longer latency. The model was not developed to imitate the timing of visual processing but to reach a performance comparable to the highest performing machines in object recognition. To this end, what is really interesting is that the model's patterns of activity highly correlated with their EEG

counterparts in a layer-wise, time-locked manner (data not shown), and extracted the concealed information in the EEG signals in support for a feed-forward hierarchical structure of object processing.

Future work

The image set which was generated for the current study allowed us to evaluate the human and model performance under a single parametrically controlled variation without being affected by other variations. However, there are some potential improvements which can be made to the image set in future studies. This include a larger number of category exemplars in order to increase the generality of the image set. High- and low-level image variations can also be incorporated into the image set and evaluated for their impacts on feed-forward/feedback mechanisms. Variations in contrast, texture, clutter and object deformation can be of high interest in that regard. Although the impact of the feedback signals have been reported in cluttered images for figure-ground segregation (Hupe et al., 1998), in low-contrast images for representational enhancement (Hupe et al., 1998; Wyatte et al., 2012) and in occluded object categorization (Wyatte et al., 2012), not enough empirical findings exist to confine the contribution of the feed-forward/feedback mechanisms in invariant object recognition. This opens up the opportunity for future studies in this area.

References

Afraz A, Yamins DL, DiCarlo JJ (2014) Neural Mechanisms Underlying Visual Object Recognition. Cold Spring Harb Symp Quant Biol 107:79-99.

Ashbridge E, Perret DI, Oram MW, Jellema T (2000) Effect of image orientation and size on object categorization: responses of single units in the macaque monkey temporal cortex. Cognitive Neuropsych 17:13–34.

Behroozi M, Daliri MR, Shekarchi B (2015) EEG phase patterns reflect the representation of semantic categories of objects. Med Biol Eng Comput 54:205-221.

Booth MCA, Rolls ET (1998) View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb Cortex 8:510–523.

Brainard DH (1997) The psychophysics toolbox. Spatial Vision 10:433–436. doi: 10.1163/ 156856897X00357 PMID: 9176952

Braje WE, Kersten D, Tarr MJ, Troje NF (1998) Illumination effects in face recognition. Psychobiology 26:371-380.

Bullier J (2001) Inegrated model of visual processing. Brain Res Rev 36:96-107.

Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate it cortex for core visual object categorization. Plos Comput Biol, 10.

Carlson TA, Hogendoorn H, Kanai R, Mesik J, Turret J (2011) High temporal resolution decoding of object position and category. J Vision 11.

Chikkerur S, Serre T, Tan C, Poggio T (2010) What and where: a bayesian inference theory of attention. Vis Res 50:2233–2247.

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nat Neuroscience 17:455–462. doi: 10.1038/nn.3635 PMID: 24464044

Corballis MC, Zbrodoff NJ, Shetzer LI, Butler PB (1978) Decisions about identity and orientation of rotated letters and digits. Memory and Cognition 6:98–107.

Curcio CA, Sloan KR, Packer O, Hendrickson AE, Kalina RE (1987) Distribution of cones in human and monkey retina: individual variability and radial asymmetry. Science 236:579–582.

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE T Pattern Anal 1:224–227. doi:10.1109/TPAMI.1979.4766909.

Delorme A, Makeig S (2004) EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Meth 134:9–21.

Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neuroscience 4:2051-2062.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341.

Dunn JC (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. J Cybernetics 3:32–57. doi:10.1080/01969727308546046.

Edelman S (1995) Class similarity and viewpoint invariance in the categorization of 3-D objects. Biol Cybern 72:207–220.

Fabre-Thorpe M, Richard G, Thorpe SJ (1998) Rapid categorization of natural images by rhesus monkeys. Neuroreport 9:303–308.

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47.

Freiwald WA, Tsao DY (2010) Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science 330:845–851.

Freud E, Plaut DC, Behrmann M (2016) What is happening in the dorsal visual pathway. Trends Cogn Sci 20:773–784.

Ghodrati M, Farzmahdi A, Rajaei K, Ebrahimpour R, Khaligh-Razavi SM (2014) Feedforward object-vision models only tolerate small image variations compared to human. Front Comput Neurosci 8:74.

Goddard E, Carlson TA, Dermody N, Woolgar A (2016) Representational dynamics of object recognition: Feedforward and feedback information flows. NeuroImage 128:385–397.

Guyonneau R, Kirchner H, Thorpe SJ (2006) Animals roll around the clock: The rotation invariance of ultrarapid visual processing. J Vision 6:1008–1017.

Hamm JP, McMullen PA (1998) Effects of orientation on the identification of rotated objects depend on the level of identity. J Exp Psychol Hum Percept Perform 24:413–426.

Hong H, Yamins DKL, Majaj NJ, Dicarlo JJ (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neuroscience 19:613-622. doi: 10.1038/nn.4247

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310:863-866.

Hupe JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J (1998) Cortical feedback improves categorization between figure and background by V1, V2 and V3 neurons. Nature 394:784-787.

Isik L, Meyers EM, Leibo JZ, Poggio T (2014) The dynamics of invariant object categorization in the human visual system. J Neurophysiol 111:91-102.

Itier RJ, Taylor MJ (2004) N170 or N1? Spatiotemporal differences between object and face processing using ERPs. Cereb Cortex 14:132-142.

Jolicoeur P (1987) A size-congruency effect in memory for visual shape. Memory and Cognition 15:531–543. doi:10.3758/bf03198388

Karimi-Rouzbahani H, Bagheri N, Ebrahimpour R (2017) Average activity, but not variability, is the dominant factor in the representation of object categories in the brain. Neuroscience In press.

Kaneshiro B, Perreau Guimaraes M, Kim H-S, Norcia AM, Suppes P (2015) A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. Plos one 10. doi:10.1371/journal.pone.0135697

Khaligh-Razavi SM, Kriegskorte N (2014) Deep supervised, but not unsupervised, models may explain it cortical representation. Plos Comput Biol 10.

Kravitz DJ, Vinson LD, Baker CI (2008) How position dependent is visual object categorization? Trends Cogn Sci 12:114–122. doi:10.1016/j.tics.2007. 12.006

Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet: classification with deep convolutional neural networks. Adv Neur in 25:1106–1114.

Lamme VAF, Roelfsema PR (2000) The distict modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23:571-579.

Lamme V, Super H, Spekreijse H (1998) Feed-forward, horizontal, feedback processing in the visual cortex. Curr Opin Neurobiol 8:529–535.

Liu H, Agam Y, Madsen J R, Kreiman G (2009) Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. Neuron 62:281–290.

Logothetis NK, Pauls J, Poggio T (1995) Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5:552–563.

Lopez-Calderon J and Luck SJ (2014) ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front Hum Neurosci 8.

Mognon A, Jovicich J, Bruzzone L, Buiatti M (2011) ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. Psychophysiology 48:229-240.

Milner PM (1974) A model for visual shape recognition. Psychol Rev 81:521–535.

Muthukumaraswamy SD, Johnson BW, Hamm JP (2003) A high-density ERP comparison of mental rotation and mental size transformation. Brain Cognition 52:271–280.

Peissig JJ, Kirkpatrick K, Young ME, Wasserman EA, Biederman I (2006) Effects of varying stimulus size on object categorization in pigeons. J Exp Psychol Anim B 32:419–430. doi:10.1037/0097-7403.32.4.419

Poggio T, Bizzi E (2004) Generalization in vision and motor control. Nature 431:768–774.

Pollen DA (1999) On the neural correlates of visual perception. Cereb Cortex 9:4–19.

Rishel CA, Huang G, Freedman DJ (2013) Independent category and spatial encoding in parietal cortex. Neuron 77:969–979.

Ritchie JB, Tovar DA, Carlson TA (2015) Emerging object representations in the visual system predict reaction times for categorization. Plos Comput Biol 11. doi:10.1371/ journal.pcbi.1004316

Riesenhuber M, Poggio T (1999) Hierarchical models of object categorization in cortex. Nat Neuroscience 2: 1019–1025.

Rossion B, Caharel S (2011) ERP evidence for the speed of face categorization in the human brain: Disentangling the contribution of low-level visual cues from face perception. Vision Res 51:1297-1311. doi:10.1016/j.visres.2011.04.003

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. doi:10.1016/0377-0427(87)90125-7.

Sasaki K, Gemba H, Nambu A, Matsuzaki R (1993) No-go activity in the frontal association cortex of human subjects. Neurosci Res 18:249-252.

Sereno AB, Lehky SR (2011) Population coding of visual space: comparison of spatial representations in dorsal and ventral pathways. Front Comput Neurosci 4.

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005) A theory of object categorization: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. CBCL Paper #259/AI Memo.

Serre T, Oliva A, Poggio T (2007a) A feedforward architecture accounts for rapid categorization. P Natl Acad Sci USA 104:6424–6429.

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T (2007b) Robust object recognition with cortex-like mechanisms. IEEE T Pattern Anal 29:411-426.

Swaminathan SK, Freedman DJ (2012) Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nat Neuroscience 15:315–320.

Taghizadeh-Sarabi M, Daliri MR, Niksirat KS (2015) Decoding objects of basic categories from electroencephalographic signals using wavelet transform and support vector machine. Brain Topogr 28:33-46.

Tanaka K (1996) Inferotemporal cortex and object vision. Annu Rev Neurosci 19:109–139.

Thierry AM, Glowinski J, Goldman-Rakic PS (2012) Motor and Cognitive Functions of the Prefrontal Cortex. Springer.

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520-522. doi: 10.1038/381520a0 PMID: 8632824

Troje NF, Bulthoff HH (1996) Face categorization under varying poses: the role of texture and shape. Vision Res 36:1761–1771.

VanRullen R, (2007) The power of the feed-forward sweep. Adv Cogn Psychol 3:167-176.

Vedaldi A, Lenc K (2015) MatConvNet-convolutional neural networks for MATLAB. arXiv:1412.4564 [cs.CV].

Vogels R, Biederman I (2002) Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. Cereb Cortex 12:756–766.

Widmann A, Schröger E (2012) Filter effects and filter artifacts in the analysis of electrophysiological data. Front Psych 3.

Wyatte D, Curran T, O'Reilly R (2012) The limits of feed-forward vision: recurrent processing promotes robust object recognition when objects are degraded. J Cognitive Neurosci 11:2248-2261.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, Dicarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. P Natl Acad Sci USA 111:8619–8624. doi:10.1073/ pnas.1403112111.

Zoccolan D, Oertelt N, DiCarlo JJ, Cox DD (2009) A rodent model for the study of invariant visual object categorization. P Natl Acad Sci USA 106:8748–8753.

Figure captions:

Fig. 1 The image set. **(A)** Images show animal, car, face and plane categories and their constituent exemplars, which are separated by the blue lines. The exemplar images are chosen from the frontal light source condition. **(B)** The conditions that each category exemplar in **(A)** underwent (here only the conditions for the first animal exemplar are shown). Exemplar images are shown larger than used in experiments in lighting and pose variations for better illustration. The bottom table shows the condition's number which is referred to in the analyses. Rows from top show lighting, pose, size and position conditions, respectively. The conditions which are used in the EEG experiment are indicated in orange and red boxes (i.e. a total of 8 exemplars in 12 conditions). Information about each condition is provided below it.

Fig. 2 Paradigms used in the behavioral **(A)** and EEG **(B)** experiments.

Fig. 3 Categorization accuracy for the humans and the model as well as human response times. Human- and model-related data are in blue-green and red-yellow color spectra, respectively. **(A)** Top row, from left to right, presents human object categorization accuracy, model object categorization accuracy and human correct response times for each variation on the main dataset; bottom row, indicates the cross-variation significance matrices in the same order. **(B)** The same figures as in **(A)** but for the EEG dataset. Different colors of the cross-variation significance matrices (as explained by the color bars) present levels of significance obtained from Wilcoxon's signed-rank test. Error bars indicate the standard errors across subjects for humans and across simulation runs for the model.

Fig. 4 Categorization accuracy under different conditions for the humans and the model. The categorization accuracy for different conditions of variations in lighting **(A)** pose **(B)** size **(C)** and position **(D)** for the humans (blue line) and the computational model (red line). The corresponding cross-conditional significance matrices are provided below each graph in blue-green and red-yellow spectra for the humans and the computational model, respectively. The color-coded significance values for cross-condition matrices are the same as in Fig. 3. The scatter plot in **(E)** provides human-model correlation coefficients in different variation conditions as well as their pooled result. Blue, black, green and red spots show the data points for the lighting, pose, size and position conditions, respectively. The gray line shows the best fist-order fit to the pooled result. **(F)** Shows the correlation coefficient and correlation significance of the pixel/model space with those of humans. Error bars show the standard errors across the subjects for the humans and across classification runs for the model results.

Fig. 5 Temporally-resolved Dunn indices as metrics for the evaluation of the EEG representational space. **(A)** The stimulus-aligned Dunn index when the variations are pooled. **(B)** The Dunn indices for each variation separately. The baseline has increased in **(B)** as a result of fewer images in the variations compared to their pool. **(C)** The peak and the latency times of the Dunn indices for each variation, which

are averaged across subjects along with corresponding cross-variation significance matrices. Shaded error areas and error bars represent the standard errors across the subjects. The circles below each graph in **(A)** and **(B)** indicate the times when the Dunn indices were significantly above the average Dunn index in the last 200 ms pre-stimulus window ($p < 0.05$), which is shown by the horizontal line in each plot. The zero-aligned vertical lines indicate the time of the stimulus onset.

Fig. 6 Time-windowed Dunn and modulation indices for the input and EEG spaces. **(A)** The Dunn indices for the input and different windows of the EEG. **(B)** The modulation indices (as calculated using equation (2)) corresponding to each time plot in **(A)**. The cross-variation significance matrix for the modulation indices at each time window is plotted below it. Error bars show standard errors across the subjects.

Fig. 7 Single-electrode Dunn topographic maps resolved in 50 ms windows. **(A)** The Dunn index of the EEG dataset calculated on overlapping windows for individual electrodes. Colors indicate the Dunn index values. The numbers below each scalp map indicate the millisecond time span in which the index was calculated. **(B)** The same as in **(A)** but for each variation of the EEG dataset. Rows from top show the results for variations in lighting, pose, size and position, respectively.

Fig. 8 Dunn and modulation indices at the output of each model layer. **(A)** The Dunn indices for each variation are shown with different colors. **(B)** The modulation indices corresponding to **(A)**.

Figures:

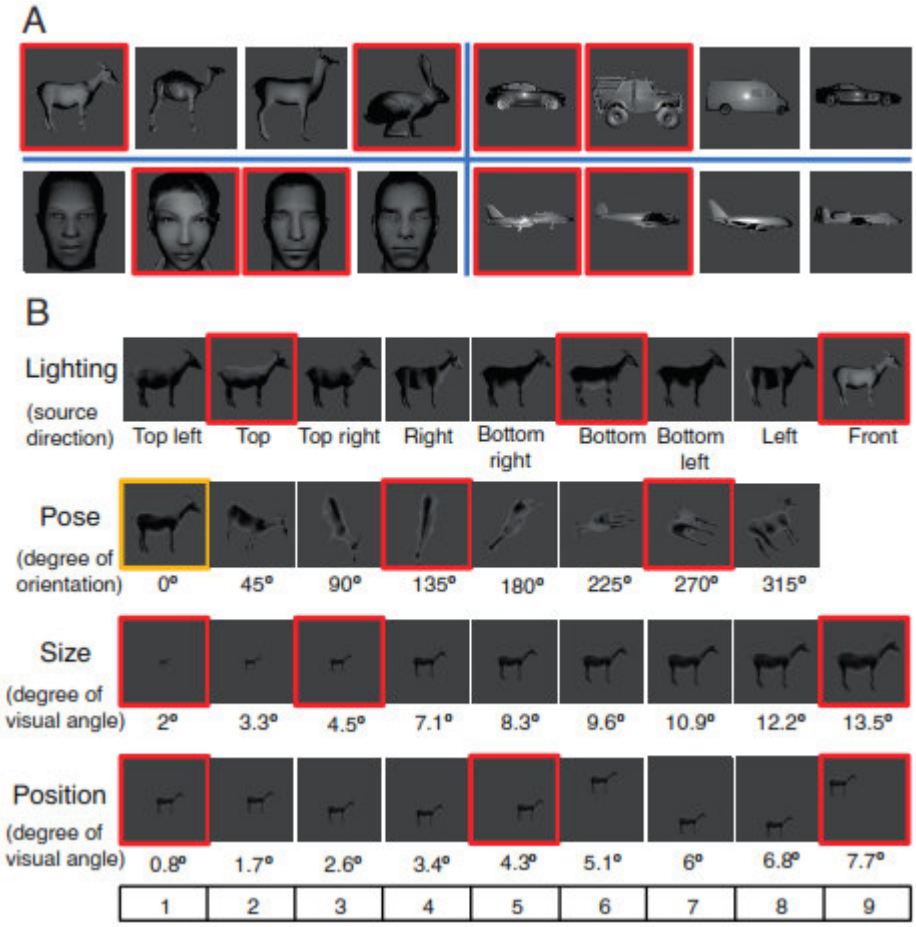


Figure 1.

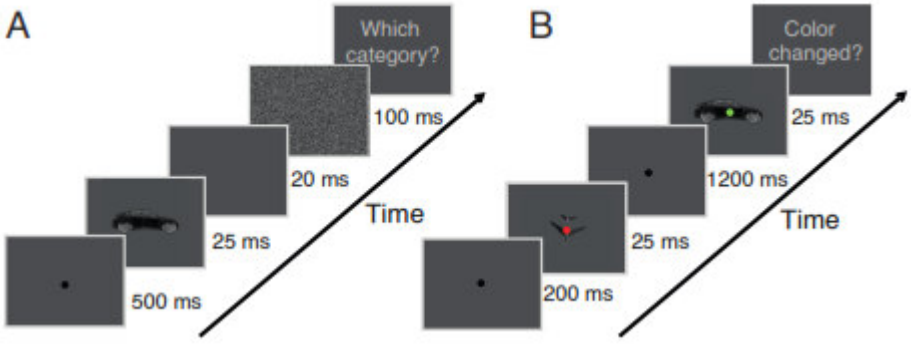


Figure 2.

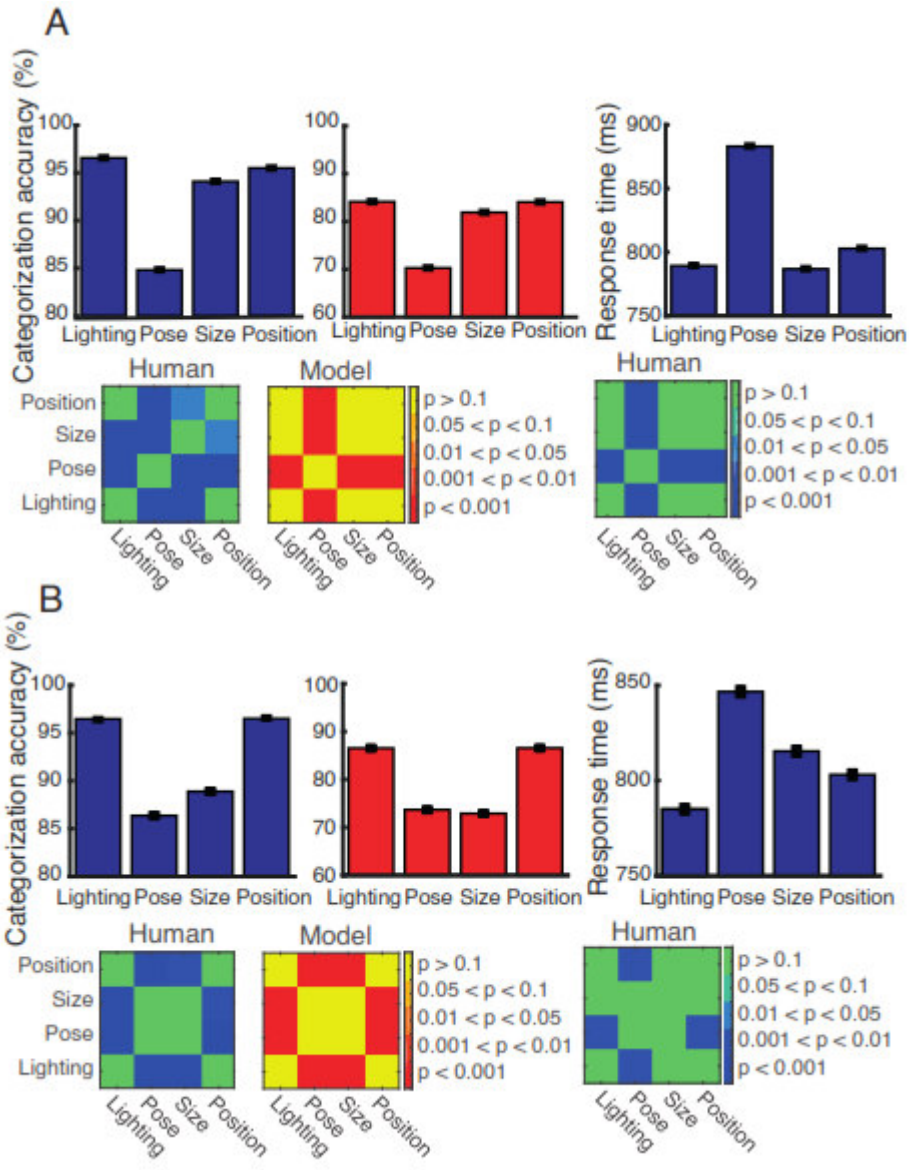


Figure 3.

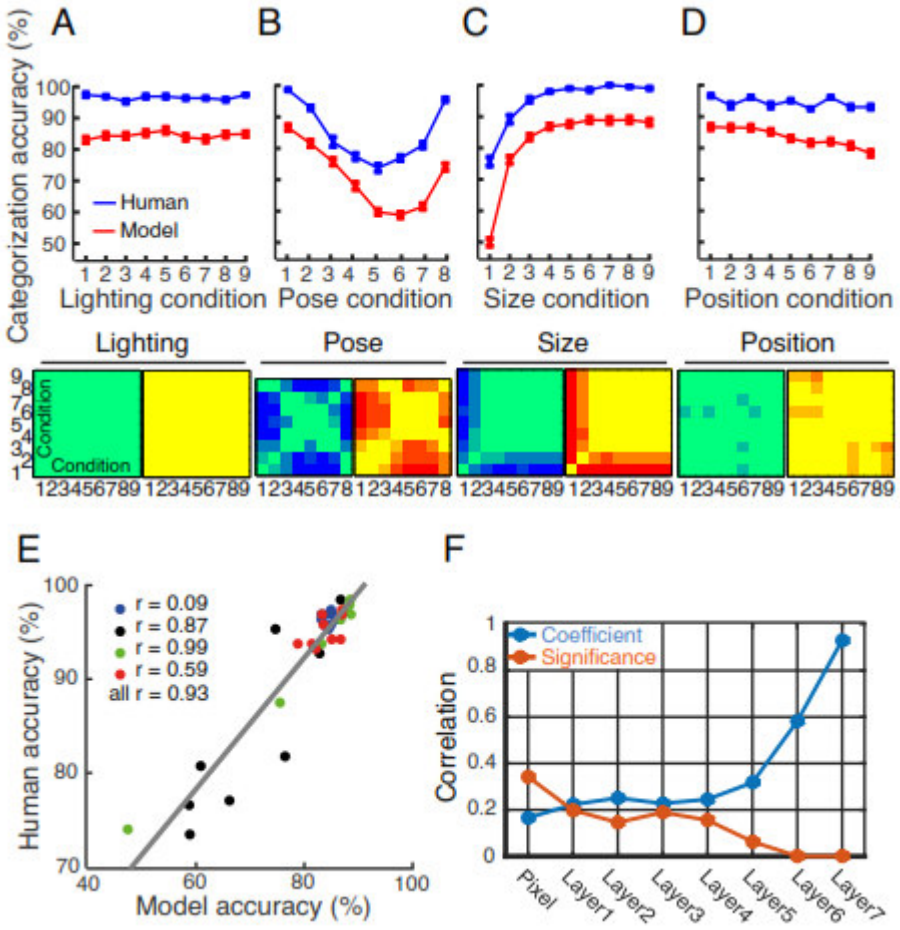


Figure 4.

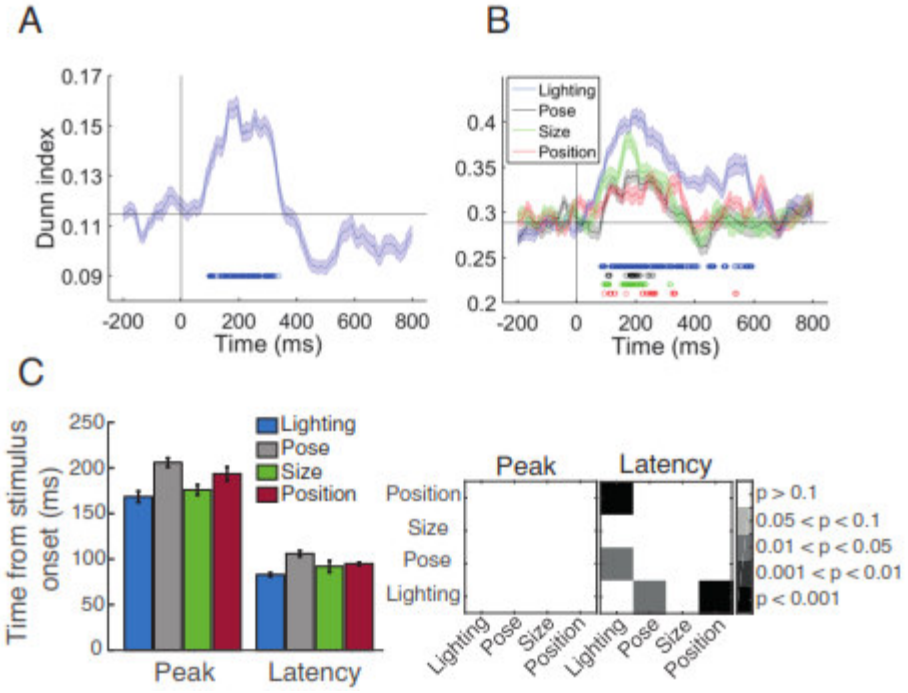


Figure 5.

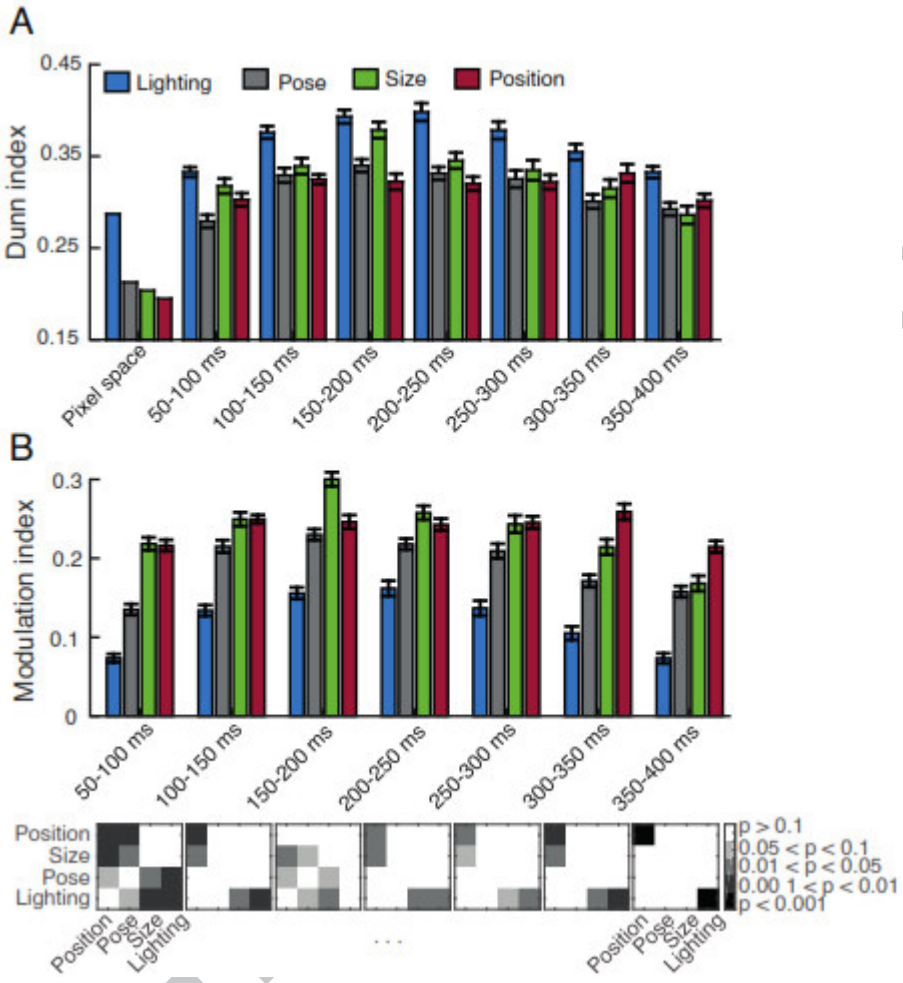


Figure 6.

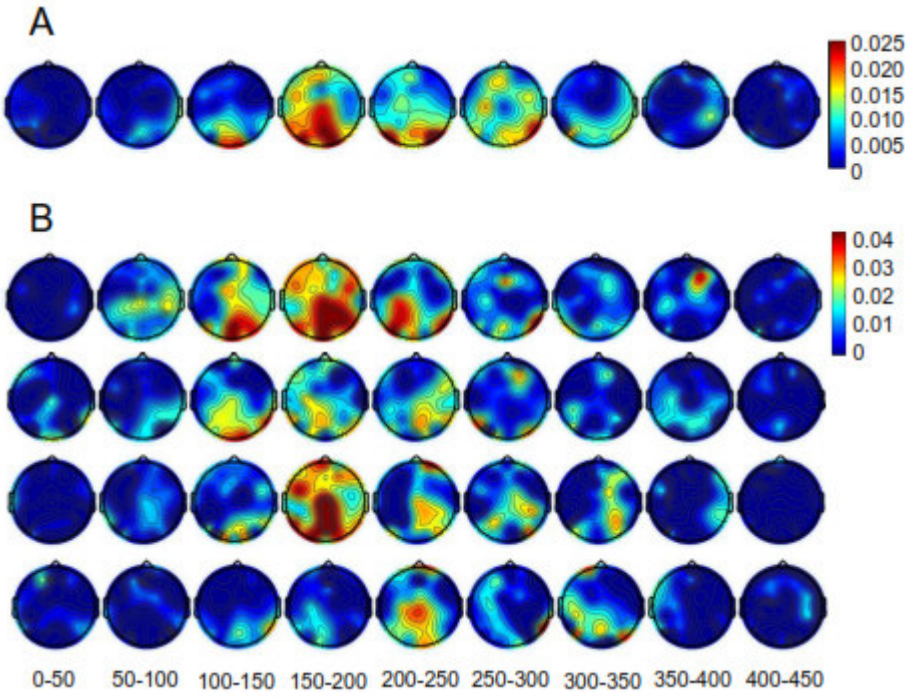


Figure 7.

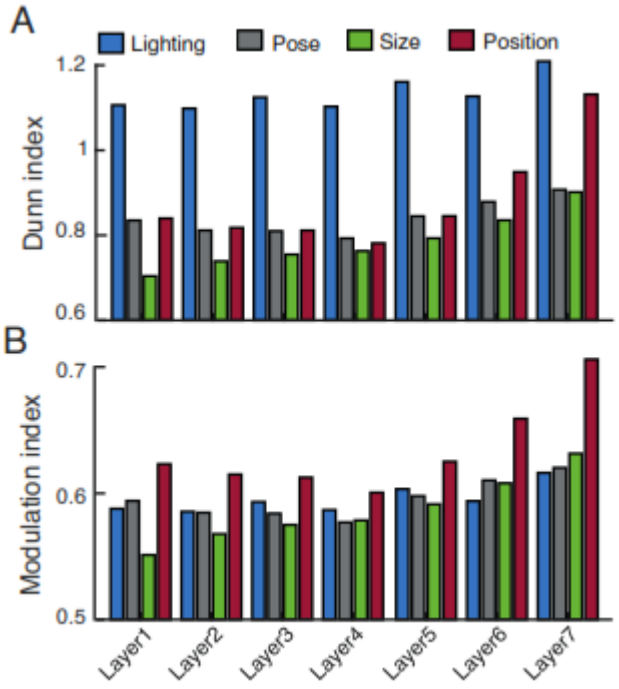


Figure 8.

Table 1. Modulation indices at the output of model layers for each variation on the main image set.

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7
Lighting	0.752	0.760	0.762	0.764	0.765	0.761	0.772
Pose	0.692	0.696	0.695	0.695	0.717	0.711	0.716
Size	0.724	0.740	0.739	0.737	0.753	0.768	0.782
Position	0.789	0.790	0.791	0.787	0.806	0.832	0.859

Highlights of the research "***Hard-wired feed-forward visual mechanisms of the brain compensate for affine variations in object recognition***" by Hamid Karimi-Rouzbahani et al.

- Feed-forward visual pathways contribute more dominantly to the compensation of affine variations such as size and position
- This is reflected in both the amplitude and latency of the category separability indices obtained from the EEG recording
- A hierarchically organized feed-forward neural network can explain these findings
- Non-affine variations may need top-down feedback information from higher visual areas for precise object recognition