https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-113

Methods and measurements to compare men against machines

***Felix A. Wichmann*[1,2,3]*, David H. J. Janssen*[1]*, Robert Geirhos*[1]*, Guillermo Aguilar*[4,5]*, Heiko H. Schütt*[1,6]*, Marianne Maertens*[4,5]*, Matthias Bethge*[2,7]**

[1] Neural Information Processing Group, University of Tübingen, Germany; [2] Bernstein Center for Computational Neuroscience, Tübingen; [3] Max Planck Institute for Intelligent Systems, Tübingen, Germany; [4] Faculty of Computer Science and Electrical Engineering, TU Berlin, Germany; [5] Bernstein Center for Computational Neuroscience, Berlin; [6] Department of Psychology, University of Potsdam, Germany; [7] Centre for Integrative Neuroscience, University of Tübingen, Germany

Abstract

Recent advances in computational models in vision science have considerably furthered our understanding of human visual perception. At the same time, rapid advances in convolutional deep neural networks (DNNs) have resulted in computer vision models of object recognition which, for the first time, rival human object recognition. Furthermore, it has been suggested that DNNs may not only be successful models for computer vision, but may also be good computational models of the monkey and human visual systems. The advances in computational models in both vision science and computer vision pose two challenges in two different and independent domains: First, because the latest computational models have much higher predictive accuracy, and competing models may make similar predictions, we require more human data to be able to statistically distinguish between different models. Thus we would like to have methods to acquire trustworthy human behavioural data fast and easy. Second, we need challenging experiments to ascertain whether models show similar input-output behaviour only near "ceiling" performance, or whether their performance degrades similar to human performance: only then do we have strong evidence that models and human observers may be using similar features and processing strategies. In this paper we address both challenges.

Introduction

Successful visual perception constitutes a remarkable computational achievement, a complex inference in which we convert high-dimensional sensory input into meaning. As vision scientists we would like to understand the algorithms and computational principles used by the visual system when we perceive the world. Computational models of vision are essential tools in this endeavour, allowing and forcing us to precisely specify and subsequently test our hypothesized algorithms and computational architectures.

An early example of a ground-breaking—and both influential as well as inspirational—computational model in vision is the Reichardt-detector as a model of motion detection [1]. In recent years more complex models in vision have greatly expanded the scope of modelling in vision science because they are able to operate on arbitrary images, and are thus not limited to abstract and often one-dimensional parametric stimulus families (see, e.g. the successful early vision model of Goris and colleagues, limited to inputs specifying (putative) spatial frequency channel activities rather than image intensities [2]). Examples of successful models in vision science applicable to arbitrary images are the models of peripheral vision and crowding [3, 4], or the image-based model of early vision capable of predicting a large number of classic psychophysical findings developed in the Neural Information Processing group in Tübingen [5].

In computer vision the pace of the advances in the last few years has arguably been even faster: since the seminal work by Krizhevsky, Sutskever & Hinton [6], convolutional deep neural network (DNNs) models have proven superior to previous computer vision models in virtually all domains they have been applied to—and frequently not only by a small margin. DNN models of object recognition, e.g. rival human performance for the first time in history [7].

But even in the non-applied vision sciences DNNs have entered the limelight: it has recently been suggested that DNNs might not only be astounding tools for solving computer vision problems, but may also be good models for the neural architecture and algorithms of human core object recognition [8, 9, 10], possibly due to converging man and machine solutions to the same basic problem: "which objects are present in this scene?" [11]. Human observers are thought to achieve this feat via fast and presumably largely feedforward processing, allowing them to reliably identify objects in photographs of natural scenes in the central visual field within a single fixation in less than 200 ms [12, 13, 14, 15].

Obviously, we wholeheartedly welcome the recent advances in computational models of vision, both in vision science as well as in computer vision. Remarkable as they are, they pose two challenges in two different and independent domains, however:

First, because the latest computational models have a much improved prediction performance compared to previous ones, we are likely to require more human data to be able to statistically distinguish between different models. Thus we would like to have methods to acquire trustworthy human behavioural data easy and fast. Second, we need behaviourally challenging experiments to test the more successful models against human data. This is required to ascertain whether the latest computational models show similar input-output behaviour only for tasks for which they are near "ceiling" performance, or whether their performance degrades similar to human performance if challenged: only then do we have strong evidence that models and human observers may be using similar features and processing strategies.

Methods for the fast and agreeable acquisition of trustworthy human behavioural data

In general, better methods for measuring human performance includes improvements in experimental *stimuli* as well as experimental *protocols*. Both aspects receive continuous attention in the vision sciences. An example of the former is the *ei-*

dolon factory recently presented by Koenderink and colleagues [16]), which we believe will prove very useful to assess models of crowding as well as object and material property recognition. An example of the latter is the strict protocol to test metamerism for image-based models [17].

However, here we are seeking methodological improvements at an even more basic level, and we are interested in improvements allowing us to gather high-quality data faster and more intuitively. Ideally, we would like to find experimental paradigms suitable for experiments using naïve observers rather than highly-trained psychophysicists.

Quantifying human behaviour dates back to at least 1860, when the experimental physicist Gustav Theodor Fechner published *Die Elemente der Psychophysik* in which he argued to approach the mind using the rigorous measurement approach so successful in the natural sciences [18]. The year 1860 is now widely regarded not only as the year of birth of the scientific discipline of psychophysics, but as the beginning of the quantitative, scientific study of psychology.

Quantitative analysis of behaviour is limited to a triad of possible measurements [19]:

1. The open behavioural response—be it a response and its associated "correctness" or "subjective equality", or a judgement of appearance or magnitude.
2. The time it took to make the open behavioural response, i.e., the response or reaction time (RT).
3. The degree of belief in the accuracy of the response, i.e., the meta-cognitive feeling of certainty in the accuracy or appropriateness of one's response.

Even within the measurement of the open behavioural response there are two distinct traditions, however: One concerned with *thresholds* or *just noticeable differences* (JNDs), i.e. with measuring the minimal stimulus difference an observer can discriminate. The other one is instead concerned with the subjective magnitude of an observer's experience, e.g. with perceived brightness, length, size or depth, to name but a few. Some authors refer to the first tradition as a *sensory discrimination* tradition, and to the second tradition as one of *sensory judgement* [20]. Attempts to relate stimulus appearance to discriminability date back to the roots of psychophysics [18]. However, thus far the link has proven extraordinarily difficult to forge, and a unified psychophysical law is not yet in sight. For overviews and details of the debate the reader is referred to the work of Gescheider [21], Krueger [22] and Ross [23]).

To compare computational models to human vision, the open behavioural response of the JND-type is typically what models are assessed on: e.g. what is the percentage of correct object classifications for a model or a human observer? [1]

From a purely methodological point of view this reliance on JND-style experiments and data is reasonable: JND-style data, particularly if collected using the method of forced-choice, have been shown to yield the most reliable and robust estimates of human behaviour [24, 25, 19]. Forced-choice JND data are "good" data.

However, collecting forced-choice JND data is neither fast nor agreeable: First of all one typically requires dozens if not hundreds of trials to obtain a reliable and robust estimate of threshold (the JND). Second, being presented with two almost indistinguishable images next to each other or in quick succession, and being asked to indicate which one is the "target" or "signal" image is not intuitive; in our experience many observers do not find it particularly agreeable either. Typically observers require hundreds if not thousands of training trials before one can be certain that the measured threshold represents indeed the limit of an observer's visual system [25]. Thus within the field of psychophysics many experimenters rely on so-called *trained* observers, and conducting a psychophysical study is often a time-consuming undertaking.

Maximum likelihood difference scaling and stimulus discriminability

Contrary to JND-style methods, maximum likelihood difference scaling (MLDS) is a method for the estimation of perceptual (difference) scales based on the judgment of clearly visible or *supra-threshold* differences in stimulus appearance [26, 27]. Furthermore, the method can be used together with the method of triads, when observers are presented with three different stimuli and have to indicate which of two are more different to the third. In our experience both naïve as well as seasoned observers find the method of triads with supra-threshold stimuli very intuitive indeed. They require less training and find such experiments to be (almost) fun.

Recently MLDS has also been used successfully to estimate near-threshold discrimination performance in the context of the watercolour effect [29]. To achieve this link between JND-style stimulus discriminability and supra-threshold appearance the authors assumed a standard signal detection theory (SDT) [30] model with equal-variance Gaussian noise on the internal sensory scale (see [29] or [28] for details, as well as Figure 1, top panel).

Using MLDS as a psychophysical method for sensitivity estimation is potentially appealing, because, first, it is more intuitive and appealing to observers, and observers require less training prior to the commencement of the experiment proper. In addition, MLDS has been reported to need less data than forced-choice procedures [28].

Through computer simulations and a real psychophysical experiment in which observers' thresholds for slant-from-texture was measured, Aguilar and colleagues [28] showed that the thresholds recovered by both methods are indeed comparable, as shown in Figure 1, bottom panel. Particularly in the middle of the scale—typically in the middle of the stimulus range—both methods' point estimates are in good agreement.

Figure 1, bottom panel, shows a substantial difference, however, in the size of the confidence intervals around the estimated thresholds as returned by the respective software packages. The difference is particularly clear to see in the top panel of Figure 2: 2AFC confidence intervals are often a factor of two, three or four larger than those obtained from MLDS. The Bayesian confidence intervals—technically credible intervals—of the simulated forced-choice data were obtained using *psignifit 4*, a software thoroughly tested to calculate correct confidence intervals [31]. Thus we have no reason to doubt their adequate coverage, i.e. that, in a frequentist setting, approximately 95% of replications

[1] Except in the important case of perceptual image or movie quality assessment, when the *perceived* quality is of interest, very much in the sensory judgement tradition.

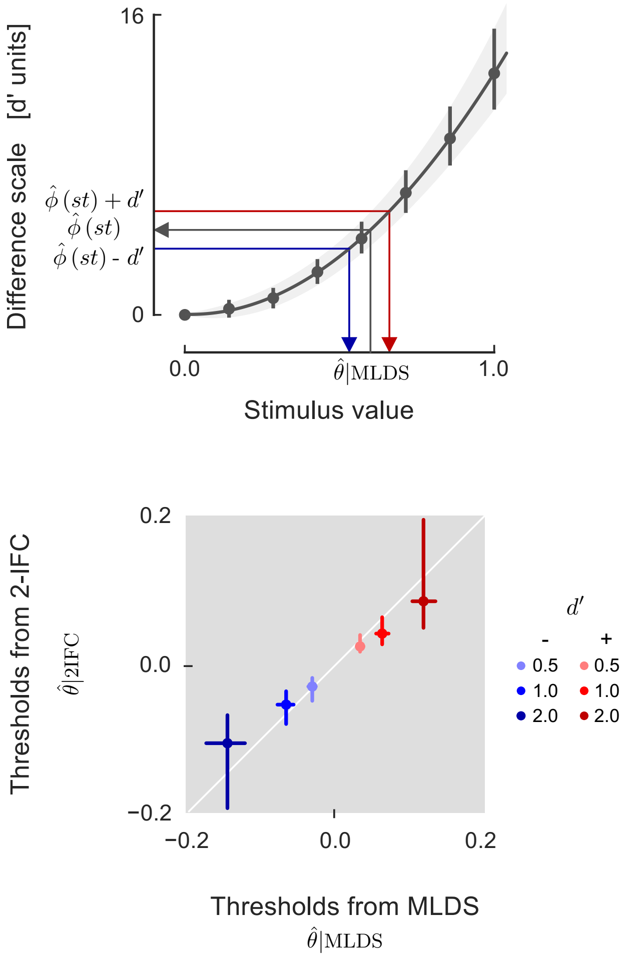


Figure 1. *Comparison of MLDS and forced-choice thresholds using simulated data.* **Top:** *Difference scale for a simulated MLDS experiment using an accelerating internal scale as found, e.g. for slant-from-texture. The procedure to read out thresholds is illustrated by arrows. Here we read out the threshold* ($\hat{\Theta}|MLDS$) *for a standard of 0.6 (vertical grey line) at a performance level of* $d' = 1$ *for comparisons above (red arrow) and below (blue arrow) the standard.* **Bottom:** *Thresholds derived with each methods are plotted against each other. They are expressed relative to the standard (st = 0.6) for comparisons above (red colours) and below (blue colours); error bars indicate 95 % confidence intervals as* returned by the respective software packages. *(Adapted from Figure 3 in [28])*

will fall within the 95% confidence interval. The bottom panel of Figure 2 shows the coverage of MLDS for three levels of internal noise (the internal noise is estimated by the MLDS package together with the scale). The dashed line in the figure indicates the correct coverage of 95%. Clearly, the confidence intervals returned by the MLDS package are too narrow, and thus do not achieve adequate coverage. Furthermore, coverage is neither independent of the amount of internal noise, nor constant along the scale: it is particularly poor towards the end where the internal scale was estimated to be shallow.

One should not forget that MLDS was not designed as a tool for sensitivity estimation—but it does remarkably well outside the bounds it was designed for. Aguilar and colleagues found it to return reasonably accurate threshold estimates from fewer trials than traditional forced-choice methods, at least for the tested slant-from-texture task. The confidence intervals around threshold should not be trusted, however, but for large scale comparisons between computational models and human observers there may well be scenarios where it is more important to obtain the thresholds from many naïve observers quickly.

Challenging methods for comparing DNNs and human observers

DNNs in computer vision are designed to accomplish high-level vision tasks, most notably object recognition, and they have undoubtedly proven their usefulness in the domain they were developed for.

Their success as algorithmic solutions for object recognition has generated substantial interest in DNNs within the vision science community. However, the usefulness of DNNs as models of human vision is not yet as clear. On the one hand, there is a growing number of studies reporting to find similarities between DNNs trained on object recognition and properties of the monkey or human visual system [9, 32, 10]. At the same time, however, there are, e.g., the well-known discrepancies as indicated by so-called adversarial examples. That is, it is possible to minimally perturb images such that they are misclassified by most DNNs but not by human observers [33, 34].

Thus we aim to obtain a better understanding of the similarities and differences in overt classification behaviour—and thus, very likely, computation—between DNNs and human vision. To this end we performed standard and straightforward object identification experiments with DNNs and human observers on exactly the same images under conditions favouring single-fixation, purely feed-forward processing to ensure a fair comparison between men and machines.

We chose to perform a contrast reduction experiment, because processing and perception of contrast is a fundamental and comparatively well understood aspect of human vision [35, 36, 37, 38, 39, 40]. Furthermore, we know that scene classification in animal and non-animal images is very robust to contrast reduction [41]. Thus we believe that comparing DNNs to human object recognition using contrast manipulations is an ideal task to to assess the degree of similarity between the tested DNNs and human observers, that is, to investigate to what degree the tested DNNs are good models for human vision.

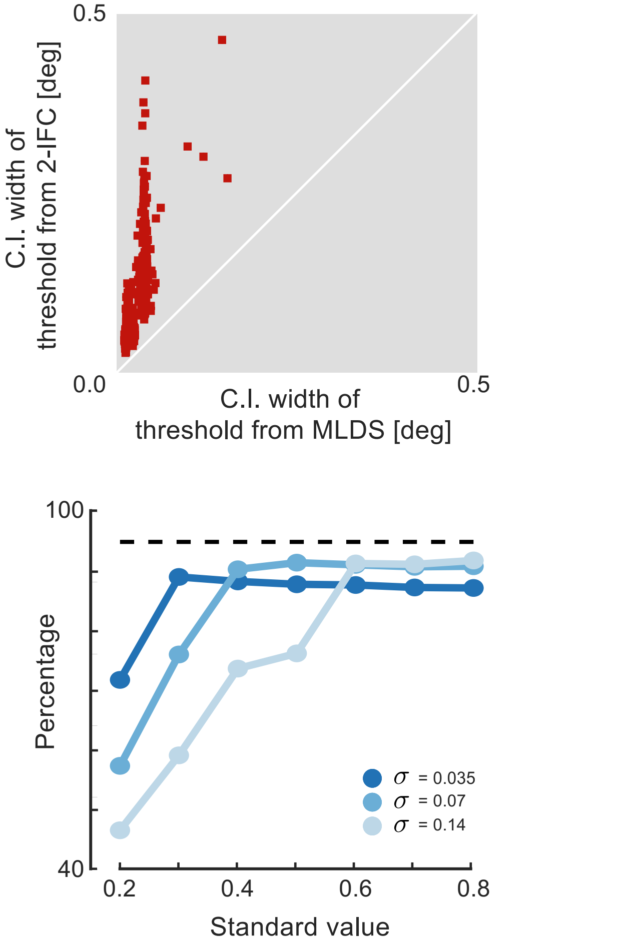


Figure 2. **Top:** *The width of confidence intervals derived from 2AFC and MLDS are plotted against each other for multiple simulations at one standard stimulus value (standard value 0.4).* **Bottom:** *Coverage of confidence intervals of MLDS at different standard levels, and for three different simulated noise levels (σ). Ideally, the coverage should be 95% (shown as a black dashed line) and independent of both σ and the standard value. (Adapted from Figure 5 in [28])*

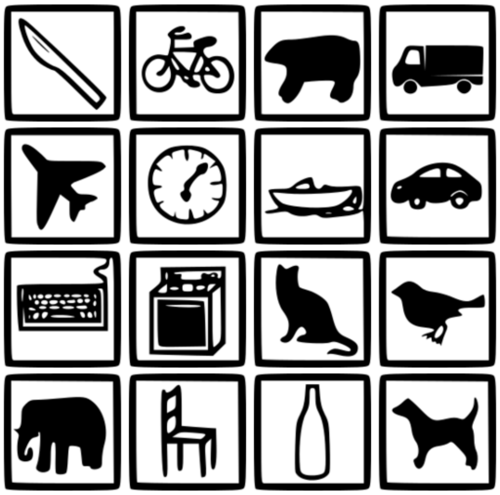


Figure 3. Response screen. Categories row-wise from top to bottom: knife, bicycle, bear, truck, airplane, clock, boat, car, keyboard, oven, cat, bird, elephant, chair, bottle, dog. Icons modified from the MS COCO website.

Methods

We used three DNNs for our comparisons between men and machines: AlexNet [6], GoogLeNet [42] and VGG-16 [43]. All three networks were specified within the Caffe framework [44] and acquired as a pre-trained model. VGG-16 was obtained from the Visual Geometry Group's website; AlexNet and GoogLeNet from the BLVC model zoo website. We reproduced the respective specified accuracies on the ILSVRC 2012 validation dataset in our setting.

The images serving as psychophysical stimuli were extracted from the training set of the ImageNet Large Scale Visual Recognition Challenge 2012 database [45]. This database contains millions of labeled images grouped into 1,000 very fine-grained categories (e.g., the database contains over a hundred different dog breeds). If human observers are asked to name objects, however, they most naturally categorize them into many fewer so-called basic or entry-level categories, e.g. *dog* rather than *German shepherd* [46]. The Microsoft COCO (MS COCO) database [47] is an image database structured according to 91 such entry-level categories, making it an excellent source of categories for an object recognition task. Thus for our experiments we fused the carefully selected entry-level categories in the MS COCO database with the large quantity of images in ImageNet. Using WordNet's *hypernym* relationship (x is a hypernym of y if y is a "kind of" x, e.g., *dog* is a hypernym of *German shepherd*), we mapped every ImageNet label to an entry-level category of MS COCO if there was such a hypernym relationship, retaining 16 clearly non-ambiguous categories with sufficiently many images within each category (see Figure 3 for a iconic representation of the 16 categories; the figure shows the icons used for the observers during the experiment).

In our psychophysical experiments all stimuli were presented on a VIEWPixx LCD monitor (VPixx Technologies, Saint-Bruno, Canada) in a dark chamber. The 22" monitor had a spatial resolution of 1920 × 1200 pixels and a refresh rate of 120 Hz. All stimuli were presented with a resolution of 256 × 256 pixels at the center of the screen, subtended an area of 3 × 3 degrees of visual angle at our viewing distance of 123 cm. Stimulus presentation and response recording were controlled using MATLAB (Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States) and the Psychophysics Toolbox extensions version 3.0.12 [48, 49] along with in-house routines. Responses were recorded using a standard computer mouse.

Our goal was to measure classification performance in man and DNN as a function of the type and amount of image manipulation—we show a manipulation of image contrast here, but used additional manipulations in our experiments (see [50]). We report data for the contrast-experiment using four contrast levels: 100%, 10%, 5% and 3% of the original contrast. In each trial the images were shown for only 200 ms, immediately followed by a full-contrast pink noise mask ($1/f$ spectral shape) of the same size. Participants had to choose one of 16 categories by clicking on a response screen (Figure 3), shown for 1500 ms. The surround of the screen was set to the mean grayscale value of all images in the dataset.

Our experimental protocol—short presentation times followed by a high contrast noise mask, fast-paced responding using a fixed temporal rhythm of the trial sequence not under the observer's control—was chosen to allow the fairest possible comparison between human behaviour and DNNs *as models of the human visual system for core object recognition*.

At each level we randomly chose 10 images per category from the pool of images without replacement for a total of 160 trials per contrast level per observer (i.e., no observer ever saw an image more than once. Within each category, all conditions were counterbalanced). Figure 4 shows, for illustration purposes, three images drawn randomly from the pool of images used in the experiment at various contrast levels.

Five observers took part in our experiment, one of them an author of this paper (RG). All participants except the author were either paid € 10 per hour for their participation or gained course credit. All observers were students of the University of Tübingen and reported normal or corrected-to-normal vision. Thus the human psychometric function shown in Figure 5 is based on $5x4x10x16 = 3200$ trials, and each confusion matrix in Figures 6 and 7 is based on 800 trials.

Results

Accuracies for the contrast-experiment ranged from 93% (VGG-16 and GoogLeNet) and $84-86\%$ (AlexNet and human average) for full contrast to near chance performance ($\frac{1}{16} = 6.25\%$) for 3% contrast. Figure 5 shows that AlexNet's and GoogLeNet's performance dropped rapidly with decreasing contrast, whereas VGG-16's decrease in performance for lower contrast levels was slower. Error bars in Figure 5 indicate the range of DNN accuracies resulting from seven repetitions on non-overlapping sets of images, with each run consisting of the same number of images per category and condition that human observers were exposed to. This serves as an estimate of the variability of DNN accuracies as a function of the images in ImageNet.

Figure 5 shows that human observers outperform all DNNs in the low contrast regime, despite that human observers have a lower performance at 100% contrast than VGG-16 and GoogLeNet. Thus a plot of the *relative* decrease of performance with decreasing contrast would show an even larger human contrast invariance as that exhibited by DNNs. (Error bars for human

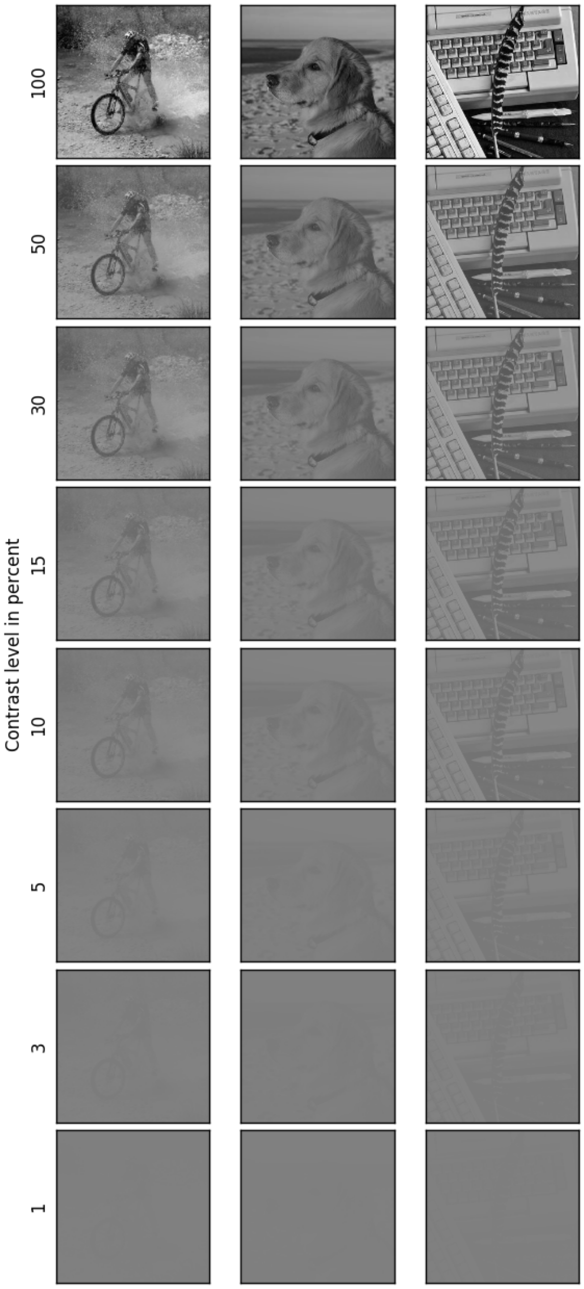


Figure 4. *Three example stimuli at various contrasts; experimental data reported in this paper used only four contrast levels, 100%, 10%, 5% and 3%. Because of the inaccuracies inherent in printing and displaying low contrast stimuli we show a number of additional contrast levels for illustrative purposes. Three images (categories bicycle, dog and keyboard) were drawn randomly from the pool of images used in the experiment.*

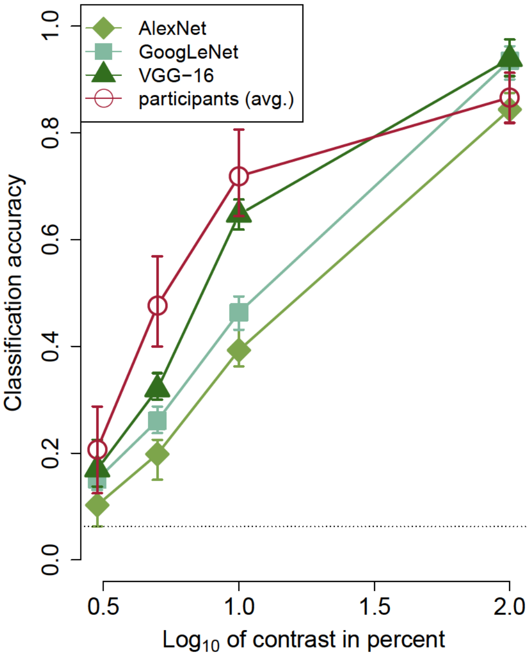
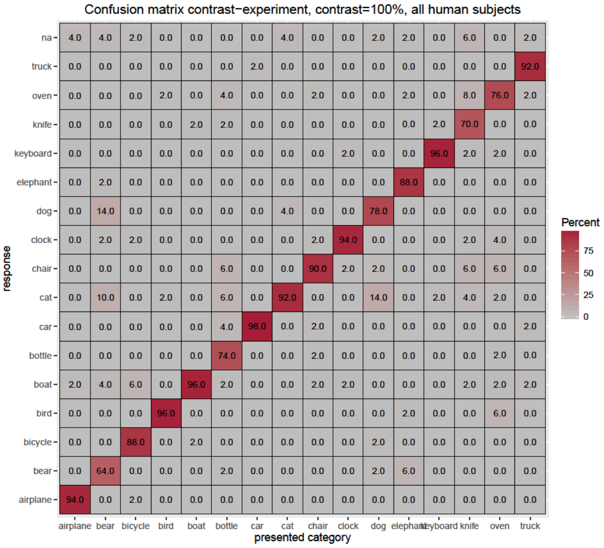


Figure 5. *Accuracy for our contrast-experiment. DNN performance is shown in greenish colours, average human performance in red on semi-logarithmic coordinates; see text for details.*



	airplane	bear	bicycle	bird	boat	bottle	car	cat	chair	clock	dog	elephant	keyboard	knife	oven	truck
na	4.0	4.0	2.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	2.0	2.0	0.0	6.0	0.0	2.0
truck	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	92.0
oven	0.0	0.0	0.0	2.0	0.0	4.0	0.0	0.0	2.0	0.0	0.0	2.0	0.0	8.0	76.0	2.0
knife	0.0	0.0	0.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	70.0	0.0	0.0
keyboard	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	96.0	2.0	2.0	0.0
elephant	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	88.0	0.0	0.0	0.0	0.0
dog	0.0	14.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	78.0	0.0	0.0	0.0	0.0	0.0
clock	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	2.0	94.0	0.0	0.0	0.0	2.0	4.0	0.0
chair	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	90.0	2.0	2.0	0.0	0.0	6.0	6.0	0.0
cat	0.0	10.0	0.0	2.0	0.0	6.0	0.0	92.0	0.0	0.0	14.0	0.0	2.0	4.0	2.0	0.0
car	0.0	0.0	0.0	0.0	0.0	4.0	98.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
bottle	0.0	0.0	0.0	0.0	0.0	74.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0
boat	2.0	4.0	6.0	0.0	96.0	2.0	0.0	0.0	2.0	2.0	0.0	0.0	0.0	2.0	2.0	2.0
bird	0.0	0.0	0.0	96.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	6.0	0.0
bicycle	0.0	0.0	88.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
bear	0.0	64.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0	6.0	0.0	0.0	0.0	0.0
airplane	94.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 6. *Average human confusion matrix for full-contrast images.*

performance show the range of data for all observers[2].)

Additionally we analyzed the distribution of errors over the 16 categories. Confusion matrices visualize systematic category-dependent error patterns, and are thus one way to gain insight into the categorization behaviour of biological or artificial classifiers. Figure 6 shows the average confusion matrix of our five human observers for the full (100%) contrast images. Correct object categorisation is shown on the main diagonal, ranging from 64% for *bears* to 98% correct for *cars* (average 86%, see Figure 5). Any cell off the main diagonal shows categorization errors; most notable are the 14% *dog* and 10% *cat* responses when the image shown contained in fact a *bear*, and the 14% *cat* responses when the image contained a *dog*[3]. On the whole the error pattern is “intuitively plausible”, showing that human observers mostly confused semantically and physically closely related animal categories with each other. Results for all DNNs at full contrast were qualitatively similar.

If we assumed—as some vision scientists do—that the tested DNNs are already good models for human ventral stream processing, we not only expect similar overall classification performance, but we expect human observers and DNNs to make errors on similar categories: they should use similar computations to arrive at their categorizations. Both predictions are born out in the data we have presented thus far in Figure 5 and discussed above in relation to Figure 6.

It should be noted, however, that the similarity in the confusion matrices between DNNs and human observers were calculated for near ceiling performance: All models and human observers have on average 84–93% of all responses on the main diagonal, so there are simply not many responses that *could* differ. Thus it may be more instructive to look at confusion matrices where less entries lie on the diagonal, i.e. for lower contrasts when categorization is more challenging.

Doing this, we observe that the similarity between human observers and DNNs progressively vanishes if the contrast is lowered: Figure 7 shows *confusion-difference* matrices, that is, the confusion matrix of a DNN—here VGG-16—minus the confusion matrix of average human performance at a given contrast level. We chose to show VGG-16 because of all the three DNNs tested, it seems to be the *most* humanlike (see Fig. 5); for AlexNet and GoogLeNet the differences would be even starker. Positive numbers, shown in green, indicate that VGG-16 responded more frequently in a given cell than human observers. Negative numbers in red show the opposite. Saturated colours thus indicate systematic differences between VGG-16 and human categorization performance; saturated colours off the main diagonal show systematic differences in the errors made by VGG-16 and human observers, saturated colours on the main diagonal show systematic differences in correct categorization. Inspection of Figure 7 shows that for contrasts of 10% VGG-16 begins to classify many of the presented images as *boats*, and by 3% it virtually always responds with *boat*, except for *clocks* and *bottles* which it still classifies reasonably correctly and better than human observers. Both other networks show a similar pattern to VGG-16, but they “home-in” on different categories: At low contrast—dense fog—

[2] Individual participants’ results were averaged, assuming that all observers have the same threshold, where we define threshold to mean the stimulus level necessary to obtain a performance equal to the arithmetic mean of lower and upper performance asymptotes. After fitting a psychometric function to each human observer’s individual data this appears warranted as individual thresholds showed only minimal divergence ($< 1.0\%$). The estimated average threshold for human observers was a contrast level of 5.11% with a 95% Bayesian credible interval of [4.46, 5.99%] [31].

[3] The entries in the confusion matrix are conditional probabilities, conditioned on the presented category; thus all columns sum to 100.

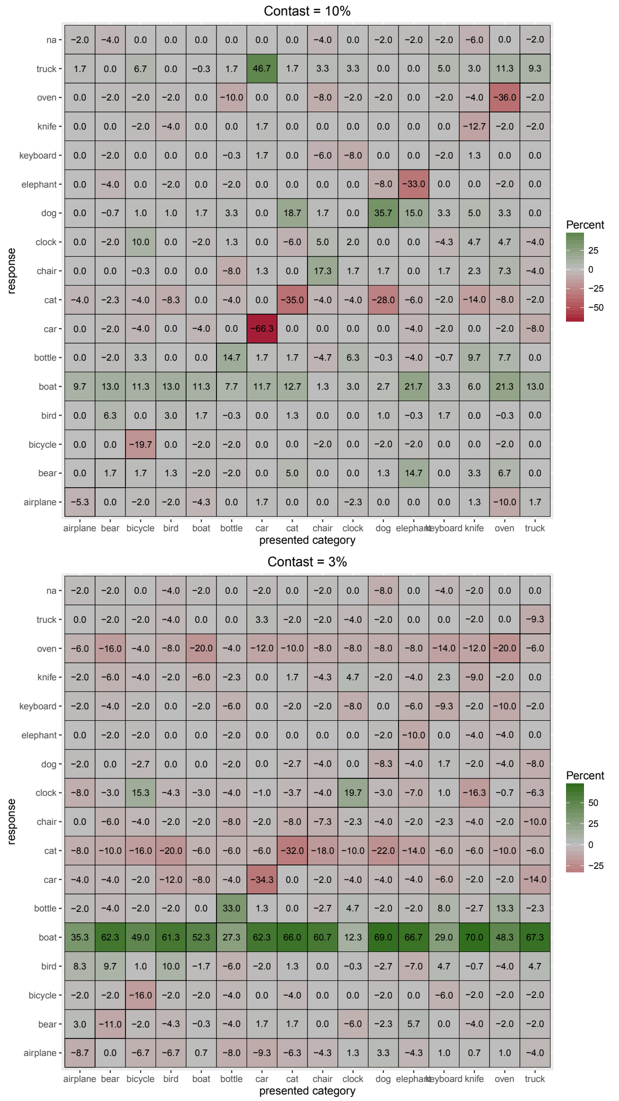


Figure 7. *Confusion difference matrices between VGG-16 and average human performance for 10% (top) and 3% (bottom) contrast images; see text for details.*

AlexNet mainly sees a world full of *bears*, GoogLeNet sees only *birds, bears* and *airplanes*. Contrary to such degenerate error patterns in DNNs for low contrast images, human observers show much more distributed and "reasonable"pattern of errors.

Discussion

DNNs and human observers exhibit roughly similar classification performance and similar confusion matrices for non-degraded ("easy") images under experimental conditions believed to result in single fixation, feedforward-only processing in human observers. If image contrast is reduced, however, we observe three effects: First, human observers' performance is more robust to contrast reductions. Second, for difficult low contrast object categorization all DNNs degenerate in a very non-human way: they begin to lump objects in one or very few categories only—a behaviour not shown by any of our human observers. Third, we find GoogLeNet and VGG-16 are equally good at categorising full-contrast images despite large architectural differences; however, for intermediate contrast images (10%) VGG-16 is clearly superior to GoogLeNet. Thus not only for comparisons of DNNs to human observers, but also for comparisons between different DNN architectures, it appears useful to include very challenging stimuli in one's benchmark dataset.

Implications for computer vision

Object recognition from ImageNet-like images as a computer vision problem is currently being almost considered "solved" because of the truly remarkable progress within the last five years in this area. As a result more and more researchers are turning their attention to even more difficult challenges such as learning from video sequences, 3D vision and unsupervised learning (to name but a few). Here we report evidence that despite excellent performance under "normal viewing conditions", all three investigated DNNs are yet to achieve human-level robustness under more difficult low contrast conditions. That human observers outperform DNNs when the signal gets weaker is also consistent with our more extensive exploration of additional image manipulations weakening the signal: added visual noise and eidolon-distortions [50]. This indicates that, from a computer vision perspective, DNNs do not yet show a similar level of robust object recognition under challenging conditions as that shown by the human visual system.

We do not think that the inferior contrast robustness of the three evaluated DNNs is insurmountable: Perhaps it would already be enough to include low contrast images into the training data to allow networks to acquire more contrast invariance and thus overcome the problem. In vision science, divisive (contrast) normalization [51, 52] is well-known, and part of almost all models of the early stages of the visual system (e.g. [38, 31]). Perhaps incorporating divisive normalization into DNNs as recently suggested by Ren and colleagues [53] may solve the problem in a more human-vision-like way than merely augmenting the training data. In fact, we speculate that contrast-normalization may well be the significant ingredient to help DNNs to exhibit robustness to contrast reductions.

An additional difference between current computational models and human observers is the latter's ability for meta-cognition, i.e. to notice when tasks are hard or when they are likely to fail—we mentioned this above as the third possible measure of human behaviour in the section on MLDS. This meta-cognitive feeling of certainty in the accuracy or appropriateness of one's response may enable human observers to switch their internal processing, e.g. to use different features, and thus to modify their behaviour. It is not inconceivable that such meta-cognitive abilities contribute to the diverging categorization behaviour between DNNs and human observers with decreasing image contrast, and as shown by the diverging confusion matrices with decreasing image contrast. Phrased positively, it may thus be beneficial to train machine learning algorithms to detect when they fail, or how sure they are about their own computations, in order to improve their robustness for applications.

In any case, we argue that precise *behavioural* comparisons between man and machine will advance our understanding of existing algorithmic differences between the two, and will enable us to build upon the insights gained in order to engineer more robust models.

Implications for human vision

As vision scientists we want to understand animate vision systems. Theory and computational models are essential if we ever hope to understand the complex inference made by our visual system, converting high-dimensional sensory input into meaning.

For object recognition, DNNs show the first successful algorithmic solution for this inference, a solution that only a few years ago seemed decades away. Thus it is perhaps not surprising that similarities found between DNNs and object recognition in human or monkey were greeted with enthusiasm by some vision scientists [9, 32, 10, 54]. We, too, are enthusiastic, but at the same time we think we need to carefully examine exactly what the similarities—and differences—are. In the 1990s during the previous wave of enthusiasm for neural networks, which in psychology went under the heading of "connectionism", all too often similarities were exaggerated and differences ignored. In 1991 Douglas and Martin criticised and warned against overgeneralisations, and pointed out that similarities between artificial and real neural networks were often superficial and more linguistic than substantial [55].

Obviously, models come at various levels, and in many flavours. The processing units in psychophysical models, e.g. correspond to "channels" and should not be thought of as representing single neurons. The linear-nonlinear-Poisson (LNP) cascade model is often useful to understand spiking patterns of neurons, but it does not include, and thus cannot explain, the dynamics of ion channels in the cell membrane. Clearly, this does not argue against psychophysical channel models or the LNP model being useful in some contexts. Useful models are not required to explain everything, and they need not be able to predict behaviour perfectly. What is important, however, is to specify exactly what the scope of one's model is, and what, in case of DNNs applied to vision sciences, putative similarities and correspondences in architecture, processing, features or receptive fields exactly refer to.

We show that three influential DNNs with very high object recognition performance for full contrast images are not as robust against contrast reductions as are human observers in terms of overt behaviour. Furthermore, human observers and DNNs do not appear to use similar algorithms at low contrast, as shown by the diverging confusion matrices at low image contrast. This, we argue, points to important differences between the tested DNNs and human vision: the difference increases with changing inputs, and there is thus not only a (less interesting) systematic, static offset between models and behaviour[4].

We think that DNNs could prove very useful to further our understanding of human vision: it is an exciting time for computational modelling of human vision. What our results do show, however, is that none of the three tested DNNs is yet an algorithmically equivalent model of the human ventral stream, not even for the putative feed-forward processing of core object recognition.

Conclusions

1. MLDS together with the method of triads represents a potentially useful method not only for appearance measurements for which it was designed for, but also for the fast and agreeable acquisition of trustworthy human threshold-type data—at least if the important caveats regarding the size of the returned confidence intervals are heeded.
2. AlexNet, GoogLeNet, VGG-16, and human observers exhibit roughly similar classification performance and similar confusion matrices for full contrast colour and black-and-white images under experimental conditions believed to result in single fixation, feedforward-only processing in human observers.
3. However, the similar performance for strong signals does not generalise to weak signals: human observers' performance is more robust to severe contrast reduction than that exhibited by AlexNet, GoogLeNet and VGG-16. Furthermore, at very low contrasts object categorization in all three tested DNNs degenerates in a non-human way: they predominantly categorize all objects in few categories only—a behaviour not shown by our human observers.
4. Evaluation methods going beyond prediction performance may help our progress: similarities in object recognition performance between models and human observers was shown to stem from different behaviour if analysed at a more fine-grained per category level (confusion matrices)—at least in the challenging, low contrast conditions.
5. We envisage that our specification of the failures, and their possible reasons—contrast normalization, metacognition—, as well as our newly collected dataset provide excellent opportunities to improve computational models in vision.

Acknowledgments

We would like to thank Uli Wannek for his scientific computing support and Silke Gramer for her administrative support. Funding was provided by the German Research Foundation (DFG; Sachbeihilfe Wi 2103/4-1 to FAW) and the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002).

References

[1] W. Reichardt, "Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensystems," *Zeitschrift für Naturforschung, Teil B*, vol. 12, pp. 447–457, 1957.

[2] R. L. Goris, T. Putzeys, J. Wagemans, and F. A. Wichmann, "A neural population model for visual pattern detection.," *Psychological Review*, vol. 120, no. 3, p. 472, 2013.

[3] B. J. Balas, L. Nakano, and R. Rosenholtz, "A summary-statistic representation in peripheral vision explains visual crowding," *Journal of Vision*, vol. 9, no. 12, pp. 13, 1–18, 2009.

[4] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature Neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[5] H. H. Schütt and F. A. Wichmann, "An image-based model for early visual processing [abstract]," *Journal of Vision*, vol. 16, p. 960, 2016.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in*

[4] It shows, furthermore, how important it is to explore the entire psychometric function from chance performance to best performance and not rely on a single threshold or point estimate only. Not that we claim this to be a novel insight: David Green writes about this already in 1960 (see p. 1199 [56]).

Neural Information Processing Systems, pp. 1097–1105, 2012.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

[8] D. L. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, pp. 356–365, 2016.

[9] D. L. 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*, vol. 111, no. 23, pp. 8619–8624, 2014.

[10] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing," *Annual Review of Vision Science*, vol. 1, no. 15, pp. 417–446, 2015.

[11] R. Dekel, "Human perception in computer vision," *ICLR*, vol. under review, 2017.

[12] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[13] M. C. Potter, "Short-term conceptual memory for pictures," *Journal of Experimental Psychology: Human Learning and Memory*, vol. 2, no. 5, p. 509, 1976.

[14] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[15] F. A. Wichmann, J. Drewes, P. Rosas, and K. R. Gegenfurtner, "Animal detection in natural scenes: Critical features revisited," *Journal of Vision*, vol. 10, no. 4:6, pp. 1–27, 2010.

[16] J. J. Koenderink, M. Valsecchi, A. J. van Doorn, J. Wagemans, and K. R. Gegenfurtner, "Eidolons: Novel stimuli for vision research," *Journal of Vision*, vol. in press, 2017.

[17] T. S. A. Wallis, M. Bethge, and F. A. Wichmann, "Testing models of peripheral encoding using metamerism in an oddity paradigm," *Journal of Vision*, vol. 16, no. 2, pp. 4, 1–30, 2016.

[18] G. T. Fechner, *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel, 1860.

[19] F. A. Wichmann and F. Jäkel, "Psychophysical methods," in *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. Volume 4: Methodology* (J. T. Wixted, ed.), p. (in press), John Wiley and Sons, 4th ed., 2017.

[20] D. Laming, "Psychophysics," in *International Encyclopedia of the Social and Behavioral Sciences* (N. J. Smelser and P. B. Baltes, eds.), vol. 18, pp. 12444–12448, Elsevier, 2001.

[21] G. A. Gescheider, "Psychophysical scaling," *Annual Review of Psychology*, vol. 39, no. 169-200, 1988.

[22] L. E. Krueger, "Reconciling Fechner and Stevens: Toward a unified psychophysical law.," *Behavioral and Brain Sciences*, vol. 12, no. 6, pp. 251–267, 1989.

[23] H. E. Ross, "On the possible relations between discriminability and apparent magnitude," *British Journal of Mathematical and Statistical Psychology*, vol. 50, pp. 187–203, 1997.

[24] H. R. Blackwell, "Studies of psychophysical methods for measuring visual thresholds," *Journal of the Optical Society of America*, vol. 42, pp. 606–616, 1952.

[25] F. Jäkel and F. A. Wichmann, "Spatial four-alternative forced-choice method is the preferred psychophysical method for nave observers," *Journal of Vision*, vol. 6, no. 11, pp. 1307–1322, 2006.

[26] L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *Journal of Vision*, vol. 3, no. 8, pp. 573–585, 2003.

[27] K. Knoblauch and L. T. Maloney, "MLDS: Maximum likelihood difference scaling in R," *Journal of Statistical Software*, vol. 25, pp. 1–26, 2010.

[28] G. Aguilar, F. A. Wichmann, and M. Maertens, "Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment," *Journal of Vision*, vol. in press, 2017.

[29] F. Devinck and K. Knoblauch, "A common signal detection model accounts for both perception and discrimination of the watercolor effect," *Journal of Vision*, vol. 12, no. 3, pp. 19:1–14, 2012.

[30] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*. Los Altos, California: Peninsula Publishing, 1988.

[31] H. H. Schütt, S. Harmeling, J. H. Macke, and F. A. Wichmann, "Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data," *Vision Research*, vol. 122, pp. 105–123, 2016.

[32] 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*, vol. 10, no. 12, 2014.

[33] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199, 2014.

[34] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv*, vol. 1412.6572, 2014.

[35] J. Nachmias and R. V. Sansbury, "Grating contrast: Discrimination may be better than detection," *Vision Research*, vol. 14, no. 10, pp. 1039–1042, 1974.

[36] M. A. Georgeson and G. D. Sullivan, "Contrast constancy: deblurring in human vision by spatial frequency channels," *Journal of Physiology*, vol. 252, pp. 627–656, 1975.

[37] F. A. A. Kingdom and P. Whittle, "Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing," *Vision Research*, vol. 36, no. 6, pp. 817–829, 1996.

[38] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *Journal of the Optical Society of America*, vol. 14, no. 9, pp. 2379–2391, 1997.

[39] F. A. Wichmann, *Some Aspects of Modelling Human Spatial Vision: Contrast Discrimination*. PhD thesis, The University of Oxford, 1999.

[40] C. M. Bird, G. B. Henning, and F. A. Wichmann, "Contrast discrimination with sinusoidal gratings of different spatial frequency," *Journal of the Optical Society of America A*, vol. 19, no. 7, pp. 1267–1273, 2002.

[41] F. A. Wichmann, D. I. Braun, and K. R. Gegenfurtner, "Phase noise and the classification of natural images," *Vision Research*, vol. 46, pp. 1520–1529, 2006.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, vol. 1409.1556, 2015.

[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, ACM, 2014.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and

L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[46] E. Rosch, “Principles of categorization,” in *Concepts: Core Readings* (E. Margolis and S. Laurence, eds.), pp. 189–206, 1999.

[47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *European Conference on Computer Vision*, pp. 740–755, Springer, 2015.

[48] D. H. Brainard, “The psychophysics toolbox,” *Spatial Vision*, vol. 10, pp. 433–436, 1997.

[49] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard, “What’s new in psychtoolbox-3,” *Perception*, vol. 36, no. 14, p. 1, 2007.

[50] R. Geirhos, D. H. J. Janssen, H. H. Schütt, M. Bethge, and F. A. Wichmann, “Comparing deep neural networks against humans: object recognition when the signal gets weaker,” *in preparation*, 2017.

[51] D. J. Heeger, “Normalization of cell responses in cat striate cortex,” *Visual Neuroscience*, vol. 9, pp. 181–197, 1992.

[52] M. Carandini and D. J. Heeger, “Normalization as a canonical neural computation,” *Nature reviews neuroscience*, vol. 13, no. 1, pp. 51–62, 2012.

[53] M. Ren, R. Liao, R. Urtasun, F. H. Sinz, and R. S. Zemel, “Normalizing the normalizers: comparing and extending network normalization schemes,” *arXiv*, vol. 1611:04520v1, 2016.

[54] J. Kubilius, S. Bracci, and H. P. Op de Beeck, “Deep neural networks as a computational model for human shape sensitivity,” *PLoS Computational Biology*, vol. 12, no. 4, p. e1004896, 2016.

[55] R. J. Douglas and K. A. C. Martin, “Opening the grey box,” *Trends in Neurosciences*, vol. 14, no. 7, pp. 286–293, 1991.

[56] D. M. Green, “Psychoacoustics and detection theory,” *Journal of the Acoustical Society of America*, vol. 32, no. 10, pp. 1189–1203, 1960.

Author Biography

Felix A. Wichmann received his BA (1994) and DPhil (1999) in Experimental Psychology from the University of Oxford. From 2007–2011 he held an associate professorship at the Technical University of Berlin, and since 2011 he is a full professor at the University of Tübingen and adjunct senior scientist at the Max Planck Insitute for Intelligent Systems. In his work he focuses on psychophysics as well as statistical and machine learning approaches to vision science.

David H. J. Janssen received his Bachelor of Psychology and Master of Cognitive Neuroscience from the University of Maastricht, and after a brief stint as data analyst at the Neurospin Center in Paris is now working on his PhD in Computational Neuroscience in Tübingen, Germany.

Robert Geirhos studied Cognitive Science at the University of Tübingen and the University of Glasgow. He is currently pursuing a Master’s degree in Computer Science at the University of Tübingen. In his work at Felix Wichmann’s Neural Information Processing Group he investigates synergies and differences between human vision and deep learning.

Guillermo Aguilar received his BS in Medical Sciences from the University of Chile (2007), his MS in Integrative Neuroscience from the Otto-von-Guericke-Universitt Magdeburg (2012), and he is currently finishing his PhD in visual perception at the Technical University of Berlin. His work has focused on the development and evaluation of psychophysical methods that aim to measure human visual perception, specifically in the domains of depth and lightness.

Heiko Schütt received his BSc in Psychology from the University of Gießen (2012) and his BSc in Mathematics and his MSc in Neural and Behavioural Sciences from the University of Tübingen (2014). Since then he has worked at at Felix Wichmann’s Neural Information Processing Group modelling early visual processing and eye movements.

Marianne Maertens is the head of an Emmy-Noether junior research group at the Faculty of Electrical Engineering and Computer Science at Technical University of Berlin. She received her PhD at the Max-Planck-Institute of Cognitive Neuroscience in Leipzig working on visual scene segmentation in human perception. Her current interest is the perception of material properties and the lightness of surfaces.

Matthias Bethge received his diploma in Physics from the University of Göttingen (1998) and his PhD in Computational Neuroscience from Bremen University (2003). Since then he has worked at the interface between machine learning and neuroscience on neural computations and representations for visual inference. In 2009 he joined the University of Tübingen as full professor and director of the Bernstein Center for Computational Neuroscience.