THE GRADUATE SCHOOL

We recommend that the dissertation prepared under our supervision by

SIDDHART SRIVATSAV RAJENDRAN

Entitled

Ensemble Coding In Color And Blur Perception

be accepted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Michael A. Webster, Advisor

Jennifer Hoy, Committee Member

Fang Jiang, Committee Member

Paul MacNeilage, Committee Member

Dennis Mathew, Graduate School Representative

David W. Zeh, Ph. D., Dean, Graduate School
August, 2018
Abstract

At any moment sensory systems are confronted with vast amounts of information, and therefore must represent information as efficiently as possible. In recent years, many studies have proposed ensemble coding as one strategy for forming a compact perceptual code. In a typical visual scene, both color and blur vary widely, and the average has important implications for representing visual norms (e.g. gray or focused) and visual function (e.g. to estimate the illuminant for color or the state of accommodation for blur). Yet whether and how the visual system uses an ensemble code for these features has received little attention. In this work, the properties of ensemble coding for the attributes of color and spatial blur were examined using behavioral and electroencephalogram (EEG) techniques. In a series of experiments, the sensitivity to this average in color and blur was studied using standard member identification and mean discrimination techniques used in summary statistics studies. The visual system has been reported to be biased towards aspects of a scene that are more meaningful to perception. In blur perception, a sharp/focused image is more relevant to vision than blurred images. In another set of experiments, adaptation to blur was studied to examine if the aftereffect is driven by the average blur level or biased towards blurred or sharper parts of the image. Finally, studies on color ensemble coding have focused on average color estimation in a visual scene. In the last set of experiments, aspects of color ensemble coding beyond the average were examined by probing the number of color levels that observers can distinguish in ensembles using both behavioral reaction times and an EEG oddball paradigm. In spite of important perceptual similarities, results from these experiments show that while color ensemble coding happens within opponent hues (e.g. separately for red and green), blur ensembles coding can happen even if an ensemble spans the whole blur spectrum from very sharp to
very blurred. This work also shows that ensemble perception is unaffected by specific properties of the
ensemble, or by whether or not an ensemble mimics natural variations of color and blur.
Introduction

The visual system encounters vast amounts of information over space and time, and thus the brain is confronted with the challenge of processing enormous volumes of information. However, neural processing faces strong limits on its resources like attention (Dux & Marois, 2009) and working memory (Luck & Vogel, 2013). To address these resource limits, the brain has been theorized to rely on principles of efficient coding to maximize its representational capacity. Efficient coding entails representing information in ways that removes redundancy and optimizes the responses of both individual neurons and neural networks.

One proposed efficient coding strategy used by the visual system is summary statistics extraction (Alvarez, 2011; Whitney & Leib, 2018). Extracting summary statistics of a scene involves representing a set of stimulus features according to its global parameters, like the average (Whitney & Leib, 2018) and variance (Haberman, Lee, & Whitney, 2015; Norman, Heywood, & Kentridge, 2015) of the distribution. This was first systematically studied by Ariely (2001), in an experiment showing that participants were better able to perceive the average of a set of different-sized circles than they could the individual circles. In a number of other reports following this study, ensemble perception has been shown to be robust and has been reported to occur for a number of stimulus parameters like orientation of simple lines (Norman et al., 2015); Norman et al 2015), direction of motion (Watamaniuk & McKee 1998), tone (Piazza et al, 2013), and biological motion (Sweeny et al 2013).
Representing the average and variance of an ensemble can have a number of potential advantages. For example, it gives a quick gist of the scene, affects decision making, estimates diversity in a scene, and even helps to ignore outliers (Haberman & Whitney, 2010). Ensemble coding appears to be an intrinsic property of sensory coding, because it develops early (Sweeny et al. 2015), and operates for abstract properties like lifelikeness (Leib et al. 2016). While the actual mechanisms of ensemble coding are intriguing in their own right, ensembles can also be used to study other aspects of neural processing. For example, Haberman and Whitney (2007) reported that ensemble coding behaved differently for upright, inverted and scrambled faces, pointing to a high-level or face-specific component of the averaging. Results like this thus help us understand both face perception and ensemble perception.

In spite of the vast range of stimuli for which ensemble coding has been reported, studies on how humans represent color ensembles have been limited, and there is almost no work examining how humans encode blur ensembles. Color and blur are two features of an image that are very important to the visual system, and also show large variations in a typical scene. Moreover, for both attributes the average value of a scene plays an important role. In color, the mean level is often taken to be a cue to the illuminant color, which is critical to factor out or discount in order to achieve color constancy (color percepts that are directly tied to the reflectance function of the surface; Foster 2011). Similarly, the average blur perceived in a scene is important because it plays a role in determining the process of accommodation. Accommodation is the ability of the visual system to change its focus so that the retinal image is sharp (Ciuffreda 1998). Blur is also important for making distance judgments in a scene. In fact,
the relationship between blur and the distance of objects from eye is so important, that it has its own internal consistencies (Sprague et al, 2016).

Importantly, in most natural scenes both color and blur can vary widely. In this case of blur, this results from the limited depth of focus of the eye’s optics. Due to limitations in focusing power of the eye, objects at only one distance can form a clear image on the retina. Objects in front of and behind the object of fixation are thus always blurred, raising the question of what the “average” blur is that an individual experiences in natural scenes.

Apart from being vital for object identification and location, color and blur have other representational similarities as well. For example, both can be modelled in a perceptual space centered on a neutral point. The neutral point in blur perception is the point of subjective focus, since deviations can appear either too blurred or too sharp (Elliot, Georgeson, and Webster 2011). Similarly, all colors are anchored by a neutral stimulus corresponding to gray (Webster and Leonard, 2008). There are many other examples of stimulus attributes that appear coded relative to a neutral point or norm. For example, models of face perception assume individual faces are coded by how they deviate from a neutral prototype or average face (Webster and MacLeod, 2011). Unlike color and blur, this “face space” has been studied extensively in the context of ensemble coding (Haberman and Whitney, 2007, 2009, Haberman et al 2015). Thus in a set of happy and sad faces, a relatively neutral face can be perceived as the average (Haberman and Whitney, 2009). Can the same be true for color and blur, where gray represents
the average of all hues in the plane, and the focused image represents the average of a set of heterogeneous features\(^1\).

Both color and blur are also cases where the visual system seems to constantly calibrate for the average value. In the case of color, as noted above the mean chromaticity of a scene is important to the visual system in the context of color constancy. Color constancy is the ability to assign colors to surfaces in spite of variations in spectral distribution and illumination of light sources. Estimating the average color in a scene can help with “correcting” for the dominant spectral variation in the ensemble and serve to preserve the color of an object (Foster, 2011). In fact, light adaptation, which involves at least some mean estimation, and which is known to occur in very early stages of visual processing, is known to play a significant role in achieving color constancy (D’Zmura & Lennie, 1986). Similarly, the visual system also adapts to the level of blur that it is usually exposed to (Webster and Marcos, 2016), and this adaptation plays a significant role in shaping the percept. For example, looking at a blurred image for just a few seconds can change the perception of a focused image (Webster, Georgeson, & Webster, 2002). Changes in perceived blur can result from either short term (Webster et al., 2002) or long term adaptation (Parkosadze et al., 2013). In the case of long term blur adaptation, the visual system is exposed to varying levels of blur within a scene, and representing the average blur could be potentially improve the efficiency of adaptation.

\(^1\) Due to the imperfections of the eye, an image on the retina is inherently blurred. Still, the perceived image is reported to be focused. This is because of active neural processes correcting for these imperfections (Elliott et al, 2011; Sawides et al 2011). Hence, though a sharp image might not be a plausible one for the optical system, a neural system can both blur or sharpen an image relative to focus (Webster and Marcos 2016). Consistent with the nature of these variations, the average of all the possible optical blur levels could be blurred, and the average of neural blur variations could be a focused image.
In summary, ensemble perception provides the neural visual system with a compact perceptual code for efficient coding. This process develops early, and occurs for a number of stimulus features. Color and blur are two important attributes of perception, and the scope of studying ensemble coding for these attributes has implications not only in understanding summary statistics representation and efficient coding, but also in understanding other properties of visual coding such as color constancy and blur adaptation. This dissertation includes four studies (as four chapters) which address different aspects of ensemble coding separately.

The first study ("Effects of Color Variance on Achromatic Settings") addresses the effect of stimulus variability on achromatic settings. Previous reports have shown that larger variance in colors in an ensemble make ensemble perception unreliable (Maule & Franklin, 2015; Webster, Kay, & Webster, 2014). The effect of color variance was examined using different color distributions, and the mechanisms of color averaging were studied by attempting to separate the different channels of color processing\(^2\) from one another.

In the second study("Ensemble coding of color contrasts") the mechanisms of color averaging were explored further by asking if gray holds a special status in the different attributes of color\(^3\).

\(^2\) Color vision is mediated by two stages early in the visual pathway. In the first stage there are the three types of photoreceptors (L, M, and S cones) each maximally sensitive to a different range of wavelengths. In the second stage, the color opponent channels are formed by combining the relative inputs from L and M cones to form the L/M channel and the relative signals from S and combined inputs of L and M cones are compared to form the S/LM channel. A third luminance channel combines the inputs of L and M signals.

\(^3\) Chromatic value for vision experiments are defined in a color space like the MacLeod-Boynton space (MacLeod and Boynton 1979). In this space, hue varies along the circumference, saturation along the horizontal axis, and lightness along the vertical axis.
– hue, saturation and lightness. Differences in ensemble coding strategies between the three attributes of color were also examined.

In the third set of experiments (“Discriminating color ensembles with different numbers of colors”), the ability to detect an increase in the number of colors between different ensembles was studied. Additionally, I explored the neural correlates of these ensemble discriminations by applying an EEG oddball paradigm.

Finally in the fourth study (“Perceiving the average blur in images”) includes three further experiments investigated ensemble coding in blur perception. The first experiment examined the ability to estimate the average blur in an image. The second explored this further by studying if ensemble blur coding operates differently for homogeneous (ensembles made of all blurred or all sharp images) and heterogeneous images (combination of blurred and sharp images). As in the case of color, the role of the norm in ensemble blur coding was also explored. The final experiment explored the aftereffects of adapting to blur ensembles.
Effects of Color Variance on Achromatic Settings

Abstract:
The average color in a scene is a potentially important cue to the illuminant and thus for color constancy, but it remains unknown how well and in what ways observers can estimate the mean chromaticity. We examined this by measuring the variability in “achromatic” settings for stimuli composed of different distributions of colors. The displays consisted of a 15 by 15 palette of colors shown on a gray background on a monitor, with each chip subtending 0.5 deg. Individual colors were randomly sampled from varying contrast ranges along the luminance, S and LM cardinal axes. Observers were instructed to adjust the chromaticity of the palette so that the mean was gray, with variability estimated from 20 or more repeated settings. This variability increased progressively with increasing contrast in the distributions, with large increases for chromatic contrast but also weak effects for added luminance contrast. Signals along the cardinal axes are relatively independent in many detection and discrimination tasks, but showed strong interference in the white estimates. For example, adding chromatic contrast along one of the chromatic axes increased variability in the white settings along both axes. This “cross-masking” and the effects of chromatic variance in general may occur because observers cannot explicitly perceive or represent the mean of a set of qualitatively different hues (e.g. that red and green hues average to gray), and thus may infer the mean only indirectly (e.g. from the relative saturation of different hues).

Introduction:
The stimulus that appears gray or achromatic is fundamental to color vision, for it is the norm relative to which all other colors are encoded [by the direction (hue) or distance (saturation) from gray]. Many studies have examined achromatic settings and how they depend on factors such as visual eccentricity (Beer & MacLeod, 2000; Webster & Leonard, 2008), the observer’s age (Wueger et al 2010; Werner & Schefrin, 1993) or the state of adaptation (Delahunt, Webster & Werner, 2004). In general, chromatic sensitivity is highest around the white point relative to other colors, consistent with the predictions for contrast masking in which white corresponds to zero contrast (Krauskopf & Karl, 1992; Switkes, Bradley, & De Valois, 1988). Achromatic settings have also played a central role in color constancy, as a measure of how the perceived colors of objects depend on the spectral characteristics of the illuminant (e.g. (Brainard & Wandell, 1992)). White settings tend to be more variable along bluish-yellowish directions compared to reddish-greenish axes (Bosten, Beer, & MacLeod, 2015; Chauhan et al., 2014; Werner & Schefrin, 1993). This blue-yellow bias has been attributed to weaker sensitivity or greater tolerance for variations in natural daylight spectra, which vary along a blue-yellow axis (McDermott & Webster, 2012).

Most studies of achromatic settings examine how or how well observers can judge the appearance of single surface. However, natural scenes often include a wide gamut of colors. It is thus important to understand the extent to which a scene as a whole can appear achromatic or unbiased. For example, this would be important for discounting the illuminant to achieve color
constancy, assuming a gray world or neutral set of reflectances. The ability to extract the mean of a set of elements has been repeatedly demonstrated in the context of “ensemble coding” or “summary statistics.” Ensemble coding is motivated by the observation that the visual system is often insensitive to the individual details comprising a scene, yet highly sensitive to the global statistics of these details such as their mean and variance (see (Whitney & Leib, 2018) for a review). This finding has been demonstrated for a variety of visual attributes including size (Ariely, 2001), orientation (Robitaille & Harris, 2011) and motion (Watamaniuk & Duchon, 1992). It has also been found for high level features such as facial expressions (Haberman & Whitney, 2007) and direction of gaze (Sweeny & Whitney, 2014). Recently, the ability to extract the mean of an ensemble has also been extended to audition, for example to represent the average of a set of tones (McDermott, Schemitsch, & Simoncelli, 2013; Piazza et al 2013). The ubiquitous nature of averaging in sensory processing suggests it confers fundamental advantages in perception, and in particular supports efficient representations of image properties (Alvarez, 2011; Whitney & Leib, 2018).

A number of studies have examined ensemble coding in color vision (e.g. (Chetverikov, Campana, & Kristjánsson, 2017; Maule, 2015, 2016; 2014; Webster, Kay, & Webster, 2014)). These have generally shown robust averaging for nearby colors, and weaker sensitivity as the color variance increases, but have not revealed how this averaging is related to basic mechanisms of color coding. Here we systematically probed how averaging and white percepts depend on different cardinal directions in color space (Krauskopf, Williams, & Heeley, 1982), and
the interactions between these directions. Specifically, our goal was to assess the relative independence of the cardinal chromatic axes in a color averaging task, and explore if the sensitivity of white settings for stimuli having different chromatic contrasts was separable based on these cardinal directions.

**General methods:**

**Participants:**
Participants included author SS and 12 students at University of Nevada, Reno who were unaware of the specific aims of the study. All observers had normal color vision as assessed by the Cambridge Colour Test and had normal visual acuity. The studies followed protocols consistent with the Declaration of Helsinki and were approved by University of Nevada, Reno’s Institutional Review Board, and participation was with informed consent.

**Stimuli:**
Figure 1 – Axes for chromatic contrast variations in different stimuli. The L/M and S axes are the cardinal directions of color space, and correspond to the 0-180 and 90-270 degrees in polar notation. The Blue-Yellow (BY) and Red-Green (RG) axes correspond to +45 and -45 degrees of rotation relative to the 0-180 axis.
Stimuli were displayed on calibrated NEC color monitor controlled by a Cambridge Research Systems ViSaGe graphics card. The gun luminances were linearized through calibration tables measured with a PhotoResearch PR 655 spectroradiometer. The test stimulus was composed of a 15x15 square array of color patches. Each uniform patch subtended 0.5 deg at the 110 cm viewing distance, and the array was shown centered in the 800 pixels by 600 pixels background of the monitor.

The test array and background (except where noted otherwise) had the same average luminance (30 cd/m2) and chromaticity (CIE 1931 x,y: 0.31, 0.316, equivalent to the chromaticity of Illuminant C). Luminance in the display was defined photometrically and not calibrated for individual observers. The colors of the individual array elements were specified by their LvsM, SvsLM, and luminance contrast relative to the mean, with contrasts scaled according to a version of the Macleod-Boynton (MacLeod & Boynton, 1979) color space (see Figure 1). The scaling was chosen to roughly equate sensitivity for the cardinal axes based on previous studies:

\[
\text{LvsM contrast} = (r_{mb} - 0.6568) \times 2754 \\
\text{SvsLM contrast} = (b_{mb} - 0.01825) \times 4099
\]

where \(r_{mb}\) and \(b_{mb}\) are the chromaticity coordinates of the stimuli in the Macleod-Boynton diagram, and the constants 0.6568, 0.01825 are the Macleod-Boynton coordinates of the background. For a given contrast level (e.g. 30), the contrasts of the 225 elements were chosen
to uniformly and equally sample 225 values along the contrast range (e.g. from -30 to 30), with
the spatial locations of the contrast levels randomized. The color of each patch displayed the
combination of the randomly assigned contrasts for the three cardinal axes. Different conditions
varied in 1) the contrast range of the elements; 2) the stimulus axis used to define the array; or
3) the type of background. The specific conditions are noted for each set of results below.

**General procedures:**

Observers viewed the display binocularly in an otherwise dark room. During an experimental
session they first adapted to the gray background for 1.5 sec, and then were shown the test
palette for 500 msec alternated with 1500 msec of the background. The mean of the array was
initially set to a random level, with the fixed contrast varied relative to the mean. For example, if
the specified contrast was 30 and the randomly selected mean was 10, then the elements
ranged in contrast from -20 to 40. Observers were instructed to adjust the mean chromaticity of
the array until it appeared gray, using two pairs of buttons on a handheld keypad to vary the
mean value along either the LvsM or SvsLM axis (Figure 2 bottom). The mean luminance
remained fixed throughout and could not be adjusted. The stimulus repeated until a setting was
selected by a button press, at which point the mean LvsM and SvsLM were recored and the
program advanced to the next stimulus. In a typical session participants were shown 11
different contrast levels, each shown 10 times in random order. Two separate sessions were
completed for each condition resulting in a total of 20 settings per stimulus. Results reported
are based on the standard deviation of the 20 repeated settings, averaged across the observers for each condition.

Figure 2 top – Examples of stimuli. Columns from left to right – Ensembles varying in luminance contrast only, ensembles varying in chromatic contrast only, ensembles varying in both luminance and chromatic contrasts. Stimuli with chromatic variations from top to bottom – chromatic contrasts along LM axis, along S axis, along “BY” axis, along “RG” axis.
Figure 2 bottom. Example of the task. The left array has 0 $S_{SvsLM}$ contrast and a fixed $L_{SvsLM}$ contrast with a mean chromaticity biased toward the +L pole. Observers adjusted both the $L_{SvsLM}$ and $S_{SvsLM}$ contrast so that the contrast distribution appeared to be centered on gray (right).

Result:

*Variations in luminance contrast:*

In the first case we examined the achromatic settings in palettes that had uniform (zero contrast) chromaticity, but varied in the level of luminance contrast. The luminance contrast was fixed at 0, 30, or 50, and observers adjusted the chromaticity until they appeared gray. Examples of the arrays are shown in Figure 2 top.
Figure 3 – Effect of luminance contrast on white settings. The standard deviation of white settings increased as the stimulus contrast increased showing that adding plain luminance noise to the stimulus made white settings harder.

Figure 3 shows plots of the resulting achromatic settings, again for stimuli that varied only in luminance contrast. In this and subsequent plots, the symbols show the mean standard deviation of the settings (averaged across the 20 trials and 8 observers (left), 4 observers (right)) as function of the array contrast, and the different lines correspond to the settings along the two chromatic axes. The left panel shows settings when observers adjusted the LvsM and SvsLM, while the right panel was for a case (discussed below) where the adjustment was instead along a “blue-yellow” and orthogonal “reddish-greenish” axis intermediate to the cardinal axes. In each panel the leftmost points correspond to a simple uniform square. For this condition observers simply had to adjust the square until it was indistinguishable from the background,
and the variance in the settings is near the nominal thresholds for discriminating a color change along either axis. It should thus be noted that in this case observers were not necessarily judging whether the square appeared achromatic since they may have simply been matching the background. However, a discriminable difference from a gray background at equiluminance does typically appear chromatic at the detection threshold (Webster, De Valois, & Switkes, 1990; Winkler et al. 2015).

Surprisingly, variance in the achromatic settings was increased when luminance contrast was added to the array. That is, even though the arrays remained chromatically uniform, judging the color was more difficult when the elements randomly varied in luminance. These effects were observed both for the mean settings as well as the settings for individual observers. Thus the increased variance is not simply because observers started to reliably differ in their chosen achromatic points, but because each observer was less reliable in their settings. A 2-way ANOVA (luminance contrast (0 vs 30 vs 50) vs response axis (LM vs S)) revealed a main effect of luminance contrast, but no main effect of response axis or a significant interaction between the groups. Posthoc comparisons showed that the difference in SD of white settings was between 0 vs 50 contrast levels (p = 0.013). A similar trend is apparent in the blue-yellow settings, though in this case the effect of luminance contrast did not reach significance.

**Variations in cardinal chromatic contrast:**

The next set of conditions varied the chromatic contrast along the LvsM or SvsLM axes, with contrast along one axis fixed at 0, 30, or 50 and the other fixed at 0. Because of the effects of
luminance variations noted above, the settings were also made for either 0 or 30 luminance contrast, for a total of 10 unique conditions. Examples of the palettes and the settings are shown in Figure 4. The leftmost points replicate the previous results for luminance contrast, with larger standard deviations in the settings when luminance contrast was added to the uniform chromatic array. For both axes, the variance in the settings also increased with increasing chromatic contrast. Importantly, this effect was similar whether the stimulus variance was along or orthogonal to the axis of the judgment. That is, the uncertainty in the LvsM achromatic setting was increased in a similar way whether the stimulus contrast was increased along the LvsM axis or the SvsLM axis. Finally, the results suggest potential differences in the
magnitude of the effect along each axis. However, this difference may merely reflect a difference in how contrasts along the two axes were scaled.

Figure 4 – White settings and simultaneous variations of luminance contrast and (a) LM contrast and (b) S contrast
A 3 way RMANOVA with 2 luminance contrast (0 vs 30) X 3 chromatic contrast (0 vs 30 vs 50) X 2 response axis (LM vs S) was done comparing the different groups, and this analysis showed significance in SD of white settings only between the chromatic contrast levels ($F(2,14) = 21.281, p < 0.001$). Posthoc pairwise comparisons between the chromatic contrast levels showed that the differences between the all the chromatic contrast levels was significant (0 vs 30, $p = 0.008$; 0 vs 50, $p = 0.006$; 30 vs 50, $p = 0.034$) with the SD for the chromatic contrast 50 being highest (mean SD = 7.017, +/- 1.12) followed by the SD for the 30 (5.576, +/- 0.961) and 0 (1.856, +/- 0.367) chromatic contrast levels.

Similar 3 way RMANOVA (2 luminance contrast (0 vs 30) X 3 chromatic contrast (0 vs 30 vs 50) X 2 response axis (LM vs S) for stimuli varying in S contrast showed main effect of chromatic contrast ($F(2,14) = 27.18, p<0.001$) with differences between the 0 vs 30 (0.002) and 0 vs 50 ($p = 0.002$) contrast levels being significant.

**Variations in blue-yellow chromatic contrast:**

As noted, a number of previous studies have shown that color discrimination and achromatic settings tend to be less reliable along the “blue-yellow” dimension of color space (Chauhan et al., 2014). Again, this effect has been attributed to the greater variance in natural lighting along the blue-yellow axis, which might reduce sensitivity to blue-yellow variations. We assessed whether a similar bias would occur for achromatic settings in the variegated color distributions. For this we repeated the same measurements, but for stimuli that were rotated 45 deg in the chromatic plane. One of these axes roughly corresponded to a blue-yellow hue variation, while the second, orthogonal axis corresponded to hues that varied from reddish-purple to green.
Examples of the palettes are shown in Figure 5. Unlike the preceding conditions, note that any differences in these stimuli cannot result from differences in how the LvsM and SvsLM axes are scaled, for both axes include the same LvsM and SvsLM contrasts, combined in opposite phase.

Figure 5 – White settings and simultaneous variations of luminance contrast and (a) BY contrast and (b) the corresponding orthogonal (RG) contrast
Here again, variability in the settings increased both with increasing luminance contrast and increasing chromatic contrast (Figure 7). Surprisingly, for our conditions the achromatic settings were if anything more reliable along the blue-yellow axis when the arrays were monochromatic. Moreover, when the stimulus varied in blue-yellow, settings also tended to be less noisy along the blue-yellow axis. However, the converse was not observed for stimulus variations along the reddish-greenish axis. Thus in general, did not reveal a weaker sensitivity to blue-yellow variations, but again confirmed that achromatic settings became less reliable as the stimulus variance increases. These effects were formally tested using a 3 way RMANOVA (2 luminance contrast (0 vs 30) X 3 chromatic contrast (0 vs 30 vs50) X 2 response axis (BY vs RG)). This analysis showed a significant main effect between the chromatic contrasts (F (2,6) = 14.36, p = 0.005) and no main effect of luminance contrast with the difference between 0 vs 50 being significant (p = 0.038).

Finally, stimuli varying in RG contrast in a 2 luminance contrast (0 vs 30) X 3 chromatic contrast (0 vs 30 vs50) X 2 response axis (BY vs PG) showed no main effect of chromatic or luminance contrasts.

**Variations in the background**

In the final conditions we assessed the effect of the background surrounding the arrays. As noted, in the preceding conditions the background was maintained at the gray average of the arrays, and thus provided a potential reference for setting the mean chromaticity of the array. To remove this cue, we repeated the settings for the cardinal axis contrasts, but on a black background. This increased the variability in the settings, although not significantly. (Figure 6).
As a second manipulation we introduced a gap between each element of the array. The gap subtended 0.5 deg and was again fixed at gray. An example of the array is shown in Figure 8 (inset). If the background served as a reference, then the added gap might provide a stronger local cue to the average chromaticity of the array. Alternatively, the gap might decrease sensitivity because it removed the potential cue for a mean bias in the color to reflect a uniform illuminant or transparent overlay. Instead, however, this condition did not alter the variability in the white settings.

These results were confirmed using a 3 way RMANOVA (3 stimulus type (luminance contrast only, LM & luminance contrast, S & luminance contrast), 3 background (gray, black, pixel), 2 response axis (LM vs S)), which showed no main effect of the stimulus ($F(2,10) = 2.18, p = .164$), or background ($F(2,10) = 2.17, p = .165$), or the response axis ($F(1,5) = 1.8, p = 0.238$).

**Discussion:**
To summarize, achromatic settings in our stimuli were strongly affected by the variance of the color distribution, becoming more variable as the contrast of the distribution increased. Surprisingly, this occurred even when the contrast in the distribution was along color directions that differed from the chromatic axis along which observers judged the achromatic balance. We consider the implications of these effects for how and how well the visual system might encode and represent the achromatic mean of a color distribution.

As noted in the Introduction, the average chromaticity in a scene can provide potential information about the illuminant and thus could provide a powerful cue to color constancy, though several authors have noted that its utility depends in part on an unbiased distribution of reflectances (the “gray world” assumption) (Foster, 2011; Smithson, 2005). Studies of color constancy have demonstrated mechanisms that do respond to mean changes in the stimulus. For example, the cones adapt to the time-averaged signals they are exposed to and thus could adjust to the mean by sampling the image with successive fixations (Chichilnisky & Wandell, 1995; D’Zmura & Lennie, 1986). However, this adaptation should actually work against detecting an average color bias since it acts to remove this bias. A uniform shift in the chromaticity (consistent with a global illuminant change) can also be readily distinguished from random changes in the
colors within a distribution (consistent with local reflectance changes) supporting relational color constancy (Foster & Nascimento, 1994). In principle, mechanisms tuned to lower spatial frequencies could also extract the average color of the scene by providing an estimate of the illuminant (Dixon & Shapiro, 2017).

As the latter example suggests, one strategy for estimating the mean could be to attend to the low spatial frequency component of the palettes, since this might correspond to the global illumination. There is evidence that the visual system can form layered representations of the illumination and surface color (Anderson & Winawer, 2005), and transparency effects have been widely studied (D’Zmura et al, 1997; Metelli, 1970).

Indeed, we specifically considered the possibility that observers might perceive the palettes as unbiased but with an overlying transparent layer of biased color. The control experiment with gaps added between the palette elements was designed to reduce this impression of transparency. However, this did not alter the pattern of settings, arguing against this interpretation.

Our principal finding is that the extraction of the average is more difficult as the variance in the distribution increases. One mechanism that could give rise to this effect for color is contrast masking. Contrast discrimination thresholds increase with increasing pedestal contrast for both luminance and chromatic contrast, consistent with a saturating nonlinearity in the contrast response (Legge & Foley, 1980; Switkes et al.,
1988). Thus the higher-variance patterns might reduce sensitivity to a chromatic bias. Contrast-dependent losses of sensitivity could also occur with contrast adaptation, which produces stronger losses for higher adapting contrasts (M. A. Webster & Mollon, 1994). Our stimuli involved brief and intermittent presentations and thus limited but did not completely control for potential adaptation effects.

However, a problem with both masking and adaptation accounts is that both are strongly color-selective. That is, masking or adaptation along the LvsM axis strongly affects sensitivity along the LvsM axis but not along the SvsLM axis or vice versa (Krauskopf & Karl, 1992; Li & Lennie, 1997; Sankeralli & Mullen, 1997). Instead, we found pronounced “cross-axis” interactions on the settings. Specifically, judging the LvsM balance was affected in similar ways whether the color distribution varied only along the LvsM axis or the SvsLM axis, and was also affected even when the distribution varied only in luminance. These cross-axis interactions argue against a simple nonlinearity as the basis for the effects, and suggest they may reflect a fundamentally different cause. Related to this, they may also reflect a fundamentally different judgment. In masking studies the task of the observers is to discriminate between two colors or color distributions, without regard to their color appearance. In contrast, our task required observers to directly judge the appearance.
The focus on appearance rather than discrimination in our task may make the achromatic settings more akin to ensemble coding. As we noted, ensemble coding of the mean has been demonstrated for a wide range of stimuli (Whitney & Leib, 2018), and like our color settings, becomes less precise with increasing variance in the ensemble (Maule & Franklin, 2015). The mechanisms mediating this representation are not well understood. However, our results suggest that the “averaging” does not occur independently within the cardinal mechanisms underlying early post-receptoral color coding, potentially pointing to higher stages of color processing. Moreover, we found little evidence for increased variability along the blue-yellow or daylight locus characteristic of achromatic settings (Bosten, Beer, & MacLeod, 2015; McDermott & Webster, 2012). This may be because the noise introduced by the color variations in the palette became more important in limiting performance than the intrinsic sensitivity to white, but at least for our limited settings this suggests that there do not appear to be any privileged axes for the averaging. How then was an average estimated?

An important issue for understanding the underlying processes is whether the achromatic settings reflected an explicit vs. implicit percept of gray in the stimulus – that is whether the visual system forms an actual representation of the mean of the distribution or only infers it indirectly. In the case of color – as well as potentially many of the other stimulus dimensions studied in ensemble coding – this issue remains
unresolved. However, it is worth noting that it is not intuitive that the mean of two complementary colors is gray; and in fact, when naïve observers are asked to select the complement of a color so that the mean is gray, their settings are highly variable (Webster et al, 2013). Thus at least anecdotally there seems to be little conscious access to the mean. Moreover, it is possible that observers were not directly sensing the mean hue, but instead gauging the relative contrast of different hues. For example, the white point might have been set by adjusting the reds and greens in the palette to appear equal in saturation. In fact some observers reported this strategy after testing. By this account, increasing the variance along other color directions might make these relative saturation judgments more difficult, because they introduced random hue variations in the stimuli to be compared.

Finally, we note that our settings were confined to the perception of unbiased color distributions, and very different factors may come into play when the average color of the stimulus strongly deviates from white. For example, a number of studies have found that in a variegated distribution, the more saturated colors tend to dominate the average perceived color (Anderson & Winawer, 2005; Kimura, 2018; Kuriki, 2004). This makes sense because in scenes with lighting-induced variation the purest colors may be interpreted as the most undiluted sample of the surface color. However, such effects again point to a role for higher-level inferences in the perception of the average color.
References:


https://doi.org/https://doi.org/10.1016/j.tics.2011.01.003


https://doi.org/https://doi.org/10.1016/S0042-6989(00)00151-6


*Psychological Science, 28*(10), 1510–1517.
https://doi.org/10.1177/0956797617713787

https://doi.org/https://doi.org/10.1016/0042-6989(94)00122-3


Ensemble coding of color contrasts

Abstract:

Ensemble coding has been demonstrated for many attributes including color, but the metrics on which this averaging is based remain uncertain. We examined ensemble percepts for colors that varied in contrast along color-opponent axes, to explore how averaging occurs across saturation within and between complementary hues. This was compared with ensemble percepts for luminance contrasts across increments and decrements. Ensembles were composed of 4 contrast levels along a saturation or lightness axis, displayed as 16 randomly positioned circles. Sets included different contrasts of the same hue (e.g. different saturations of red or green) or of complementary hues (different saturations of red and green), with the specific hue axis varied randomly across trials. Targets included the four displayed contrasts and five additional levels at intermediate or higher contrasts. On each trial, observers saw the ensemble for 0.5 sec and after a 1 sec delay responded whether a single displayed target was a member of the set. For sets with complementary colors (e.g. red and green), false alarms for gray targets were lower for chromatic contrasts than for luminance contrasts, even when this gray represented the mean chromaticity. In unimodal sets (e.g. red or green) false alarms fell precipitously at the gray boundary when the target had a different hue from the ensemble, while the fall off was more gradual for categorical differences in lightness ensembles. These effects suggest that averaging occurs primarily within rather than between hue (increments and decrements). Moreover, in
chromatically biased ensembles (e.g. high-contrast reds and low-contrast green) the perceived membership paralleled the asymmetry (e.g. fewer false alarms for high-contrast greens). This suggests that hue and saturation are not encoded as independent attributes, and instead is again consistent with averaging within each hue. Our results suggest that ensemble percepts of color are not based on a simple metric-like Euclidean distance, and that the overall mean of a color distribution with different hues may be available only implicitly, from the summary percepts of the different hues. Our findings also imply that “gray” forms a much stronger categorical boundary for chromatic differences than lightness differences.

Introduction:

In an ambiguous environment, the visual system can encode the gist of a scene, rather than individual items constituting a scene (Alvarez, 2011; Ariely, 2001). This ability, called ensemble coding, has been studied for a number of stimulus features including motion (Watamaniuk & Duchon, 1992), orientation (Solomon, 2010), and faces (Haberman, Harp, & Whitney, 2009; Haberman & Whitney, 2007) and even abstract properties like “lifelikeness” (Leib, Kosovicheva, & Whitney, 2016). In these cases observers show fine sensitivity to the mean of the stimulus array, and are more likely to think the mean level was present than one of the actual levels contributing to the average. Thus the visual system appears to be able to extract the average with high fidelity.

Here we examined ensemble coding in color vision. Estimating the average chromaticity of the scene could play an important role in processes like color constancy, for example to discount the
color of a global illuminant. Previous studies have examined ensemble coding for nearby hues, and have examined how sensitivity to the average varies with the range of the color differences (Maule & Franklin, 2015; Webster J, Kay, & Webster, 2014). Researchers have also tested for categorical biases in ensemble processing, for example to see if the perceived mean is shifted toward a category boundary (Maule, Witzel, & Franklin, 2014). In general however, it remains unknown how and how well the visual system could “compute” the average of a set of colors, and what rules determine this averaging.

Previous studies have focused primarily only on the dimension of hue, whereas color also varies along the dimensions of saturation and lightness. Incorporating these dimensions provides a richer test of the capacity for ensemble coding within color space. To explore the limits of this coding, we examined stimulus sets that varied in chromatic (saturation) or luminance (lightness) contrast relative to a neutral gray. Gray is special in color processing for a number of reasons. First, ensemble coding in color has generally not shown a categorical effect (Maule et al., 2014). However, the categories tested have been between adjacent hues (e.g. the blue-green boundary) and thus vary in the relative proportions of two hues (Chetverikov, Campana, & Kristjánsson, 2017; Maule & Franklin, 2015; Maule et al., 2014). Gray offers a stronger test of categorical coding because transitions through the gray point result in completely distinct complementary colors. Thus gray represents the strongest instantiation of a categorical color boundary. Second, gray represents a null or norm in color processing. Specifically, stimuli such as color (Webster & Leonard, 2008), faces (Webster & MacLeod, 2011) and blur (Elliott, Georgeson, & Webster, 2011) can be modelled in a perceptual space in which individual
variations of the stimulus appear to vary relative to a unique norm. The norm itself appears neutral, and the physical stimulus corresponding to the norm may elicit a response null within the encoding mechanism. Thus the norm has been theorized to hold a special perceptual status in visual coding (Webster, 2015). Third, it is not clear how ensemble percepts would incorporate this norm-based representation. For example, suppose an observer is exposed to a uniform set of hues that all have the same saturation. What would the perceived average of this set be? A metrical averaging of the chromaticities would yield an average of zero and thus the neutral gray. Yet this average would have a very different saturation than any of the color elements, and it is possible that observers might instead assume that the “average” had the same saturation as the elements, but differed in hue. That is, observers might instead compute the average independently for hue and saturation, and the nature of the ensemble percepts might therefore point to the representation of color at the level at which ensemble coding occurs.

Finally, color can vary not only in hue and saturation but also lightness. Gray is also a categorical boundary for lightness differences, since it demarcates the transition from increments to decrements. However, it is not clear whether the complements of light and dark behave in the same way as, for example, red and green. To explore these dimensions we compared ensemble coding along both the saturation and lightness dimensions.

**General Methods:**

**Participants:**
A total of 30 observers participated in the study, with different subsets tested in each of the conditions. All participants were recruited from the UNR student pool and were naïve to the specific aims of the experiment. Participants had normal color vision as assessed using the Cambridge Color Test, and gave informed consent following the protocols approved by the university’s IRB. Some observers participated in only one experiment while some participated in multiple sessions.

**Stimuli:**

Stimuli were presented on a CRT monitor controlled by a Cambridge Research Systems Visage graphics system. Chromaticities and luminances on the display were calibrated with a PR655 spectroradiometer. For all the conditions, ensembles were made of 16 randomly positioned circles arranged in a 4x4 irregular grid and were shown on a neutral gray background with the chromaticity of Illuminant C and a mean luminance of 20 cd/m². Depending on the condition, the individual circles either varied in their relative saturation or in their relative luminance. In the saturation condition, ensembles varied within a color category (e.g. reds) or between complementary hues (reds and green). In the luminance conditions, individual circles in an ensemble belonged to the same hue, but were either all darker or brighter relative to gray or were a combination of brighter and darker hues.
Procedure:

Figure 1 – Member identification task. Participants responded “yes” if they perceived the test as part of the ensemble, and “no” if they did not perceive the test as part of the ensemble.

Participants performed a member identification task where an ensemble of colors was briefly presented, and then followed by the presentation of a single target stimulus (Figure 1). Participants reported if the target was a part of the presented ensemble, and were given unlimited time to respond, though actual ensemble and targets were presented only briefly (500 ms) with a 1 sec interval. All ensembles were made of 4 contrast levels from a single contrast axis chosen at random hue angles in the color space. That is, on each trial the saturation of the
elements were fixed at a given level, while the hue angle of the ensemble could be along any
direction in the chromatic plane. The test target included the 4 contrast levels actually present
in the ensemble and an additional 5 contrast levels chosen from the same axis as the ensemble.
Each ensemble/target condition was shown in random order for a total of 20 repetitions, to
estimate the proportion of times a given target was perceived to have been a member of the
ensemble. If a contrast level was a part of the ensemble, and if participants reported it as a part
of the ensemble, it was considered a true correct response (TR). If a contrast level was not
present in the ensemble, but was still misreported as part of the ensemble, it was considered a
false alarm (FA).

![Opponent color space](image)

**Fig 2 – Opponent color space.**

**1. Ensembles in the cone opponent space:**

1.1 Method
The colors for the initial experiments were defined within a scaled version of the MacLeod-Boynton (MacLeod & Boynton, 1979) chromaticity diagram (representative figure – fig 2) where the scalings are based on previous measures designed to roughly equate threshold sensitivity along the LM and S cardinal axes. The specific conversion between the scaled space and the MB space is given below:

\[ \text{LvsM contrast} = (r_{mb} - 0.6568) \times 2754 \]

\[ \text{SvsLM contrast} = (b_{mb} - 0.01825) \times 4099 \]

where LvsM is the reported contrast level, \( r_{mb} \) and \( b_{mb} \) are the chromaticity coordinates of the stimuli in the MacLeod-Boynton color space.

The four ensembles tested in this condition are tabled below:

Table 1 – Ensembles used in experiment 1

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Contrast levels</th>
<th>Features of the ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-60, -15, 15, 60</td>
<td>Ensembles with gray average</td>
</tr>
<tr>
<td>2</td>
<td>-120, -30, 30, 120</td>
<td>Ensembles with gray average</td>
</tr>
<tr>
<td>3</td>
<td>-30, 0, 30, 120</td>
<td>Biased ensembles to test for hue/saturation interaction</td>
</tr>
<tr>
<td>4</td>
<td>0, 15, 30, 60</td>
<td>Categorical effects of gray</td>
</tr>
</tbody>
</table>
The metric average of the first two ensembles is gray. These sets were included to examine whether the gray would be misperceived as a member even though it did not share the "hue" of any of the actual elements. The third ensemble was designed to test the relative independence of hue and saturation in ensemble coding. For this ensemble, if a high saturation “red” is misreported as part of an ensemble made of low saturation “reds” and high saturation “greens”, such “illusory conjunctions” (Treisman & Gelade, 1980) could imply that ensemble coding processed the hue and saturation of the elements separately. The fourth ensemble was designed to examine if contrasts levels on the opposite side of gray (i.e. complementary to the ensemble elements) were misreported as part of the ensemble. This set was therefore included to test for categorical effects of the gray boundary.

1.2 Results and Discussion

(a)                                                                                   (b)
Figure 3 – Results for chromatic ensembles in the opponent color space. (a) Ensembles 1 and 2 show that the false alarm (FA) for gray was low even though it was the average chromaticity. (b) Ensemble 3 shows that ensemble percepts of hue and saturation are not dependent; Ensemble 4 shows that FA drops sharply at gray.

The results of this experiment are shown in fig 3. False alarm rates were low at zero contrast (gray) for ensembles 1 and 2, in which gray represented the mean chromaticity of the set. This shows that participants were able to correctly reject gray from an ensemble. The response pattern for ensemble 3 shows that the contrast levels actually present in the ensemble were the ones that had maximum membership responses. Hence, there was no evidence for illusory conjunctions between saturation and hue, suggesting that in an ensemble task hue and saturation coding are not represented independently. Further, for all of the ensembles, perceived membership quickly dropped for contrast levels outside the ensemble range. This pattern of correct rejection was similar for categorical boundary near gray (ensemble 4), and thus did not reveal a special or categorical effect of the gray boundary.

The results of these experiments help establish that a) ensemble coding of contrast does not reflect a simple metrical averaging of the contrasts; b) hue and saturation appear to be represented conjointly in ensemble coding, so that the average is not computed independently for the two attributes; and c) the falloff in false alarms is similar around the category boundary.
as for other contrasts, suggesting this boundary does not sharply delineate how the colors are averaged.

2. Effect of contrast scaling metrics:

The rejection rate for gray relative to the other intermediate non-member targets suggests that gray is not perceived as the average of an ensemble made of equal contrasts of red and green. That is, observers do not mistake a zero contrast stimulus for a member of an ensemble composed of visible color contrasts. However, this differential effect might also be explained through the non-linearities of the cone opponent space used to define the colors. For example, if there were a saturating nonlinearity in the contrast response, then higher contrasts would appear more similar to each other and gray would appear farther removed from the ensemble set. That is, the intermediate non-members might be perceptually more similar to the displayed set, and thus more likely to be misclassified. In fact, we attempted to select a roughly rational scaling in the ensembles to correct for this confound. However, it remains possible that this did not adequately equate the scaling, so that observers might still be susceptible to perceiving the gray as a member when the steps are instead perceptually uniform.

2.1 Methods

To address this issue we used a contrast scaling task in which a distribution of 9 contrast levels from -120 to +120 along the 0-180 axis of the opponent color space was displayed as a row of elements shown on the monitor. The ends of this series were shown fixed at -120 and +120 chromatic contrast and the center element at 0 contrast. Participants adjusted the remaining 6
intermediate contrast levels until they appeared to increase uniformly in saturation (Appendix I for further details). The results of this contrast metric task suggested that saturation was roughly linear, and thus there was no evidence for a significant nonlinearity in the contrast response on either side of gray (see Fig 11 in appendix I)

As a second test of the effect of the perceptual metric, we also measured ensemble coding for stimuli defined by constant steps in the CIELAB “uniform” color space. These experiments also varied the range of ensemble elements to further evaluate the status of gray in ensemble coding.

**Stimuli:**

As in the previous condition, three types of ensemble were tested in this condition as well. The specific contrast levels are listed in the table below. Note that in this case the units were chosen to correspond with the perceptual spacing predicted by CIELab

**Table 2 – Ensembles defined in the LAB space**

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Contrast Level</th>
<th>Feature of the ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>-30,-15,15,30</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>-45,-15,15,45</td>
<td>Symmetric ensembles with gray average</td>
</tr>
<tr>
<td>C3</td>
<td>-60,-30,30,60</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>-15,0,15,30</td>
<td></td>
</tr>
</tbody>
</table>
The task and experimental procedure were similar to the procedures of the first condition. However, in this case the background luminance was reduced (to 5 cd/m²) relative to the stimulus luminance (20 cd/m²), to prevent observers using the match between the gray target and background from affecting the response. That is, in this case the gray itself formed a clear change from the background similar to the other targets, and thus could not be classified simply on the basis of an “absence” of a stimulus. Up to four ensembles were tested in a given session lasting 40 minutes, and all ensembles were interleaved between multiple sessions. Membership reports were again based on 20 repeated settings for each target /ensemble combination, resulting in up to 720 trials.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C5</td>
<td>-30,-15,45,60</td>
<td>Asymmetric ensembles to test for illusory conjunctions between hue and saturation</td>
</tr>
<tr>
<td>C6</td>
<td>-30,0,30,60</td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>0,15,30,45</td>
<td>Ensembles that span and do not span the gray boundary to test for categorical effects at gray</td>
</tr>
<tr>
<td>C8</td>
<td>15,30,45,60</td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td>30,45,60,75</td>
<td></td>
</tr>
</tbody>
</table>
2.2 Results and Discussion

Figure 4 - Results for color ensembles with gray average. Ensembles C1:C3 vary along the saturation axis – Gray is correctly rejected from the ensembles which did not have gray.

The metric average of the ensembles C1:C3 was always gray. Figure 4 shows the membership ratings for these ensembles. For each, gray was less likely to be perceived as part of the set.

The false alarm (FA) for gray was compared with the correct response (TR) for the two extreme member of the ensemble. For all the ensembles, the FA was significantly lower than the TR for members of the ensemble as tested using a repeated measures analysis of variance (RMANOVA). These results have been summarized below:

Table 3 – Summary of analyses for ensembles C1:C3

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>F value</th>
<th>P value</th>
<th>Significant difference between contrast levels</th>
<th>Mean difference (Bonferroni adjusted p value)</th>
</tr>
</thead>
</table>
Figure 5 – Hue saturation independence in ensemble coding – The contrast levels marked with filled diamond were compared in the corresponding ensembles. Response for the member of the ensemble was higher than the response for the non-member showing that hue and saturation were relatively independent in ensemble coding.
The next set of colors included asymmetric ensembles (C4:C6) with higher contrasts for one of the hues and lower contrasts for the complementary hue, again to test whether hue and saturation were represented independently. The FA for a contrast level was always less than the TR for the corresponding contrast level that was a part of the ensemble (Figure 5). This was confirmed by paired sample t-tests comparing the two corresponding contrast levels of either side of gray for the 3 ensembles. The results of these analyses are tabled below:

Table 4 – Summary of analyses for ensembles C4:C6

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Contrast levels compared</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-15,0,15,30</td>
<td>-30 (65.5 ± 6.388) vs 30 (83.5 ± 3.9)</td>
<td>t (9) = -3.85</td>
<td>0.004</td>
</tr>
<tr>
<td>-30,-15,45,60</td>
<td>-45 (78.1 ± 5.1) vs 45 (94.1 ± 1.3)</td>
<td>t (9) = -2.88</td>
<td>0.028</td>
</tr>
<tr>
<td>-30,0,30,60</td>
<td>-60 (44 ± 6.5) vs 60 (79.5 ± 6.4)</td>
<td>t (9) = -6.6</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Figure 6 – Gray Categorical boundary for saturation. FA drop sharply for contrasts levels lower than that of the ensemble. The drop at zero contrast is not different from the drop at other chromatic contrast levels.

Finally, the remaining ensembles (C7:C9) were examined to again test for categorical differences at the gray boundary. In particular, we asked whether perceived membership fell more rapidly when the range of the ensemble included only one hue or a hue plus gray, compared to ensembles with the two complementary hues. For these, the drop was rapid for the contrasts immediately outside the set, and the decreased progressively for subsequent contrast levels (Figure 6). However, this pattern was similar near the gray boundary as at other contrast levels. Thus there was little evidence of a categorical boundary effect at gray.

Regardless of where the ensemble ended, FA for the first contrast level outside the ensemble was similar. This was again confirmed using a one way ANOVA
Table 5 – Summary of analyses for ensembles C7:C9

<table>
<thead>
<tr>
<th>Ensembles</th>
<th>Contrast Levels compared</th>
<th>F Value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,15,30,45</td>
<td>-15</td>
<td>F (2,24) = 1.02</td>
<td>0.376</td>
</tr>
<tr>
<td>15,30,45,60</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30,45,60,75</td>
<td>15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Luminance ensembles:

In the next set of experiments our aim was to compare the properties of ensemble coding for lightness variations vs. saturation variations. Specifically, we assessed whether observers might be more susceptible to misperceiving a gray when that gray corresponded to a neutral lightness rather than saturation. To test this, we conducted the same measurements but now for stimuli varying only in lightness.

3.1 Methods

In order to make this comparison, luminance and chromatic contrasts need to be on a comparable scale. Hence, we first asked participants to perform a contrast matching task between luminance and chromatic contrast, a task which can be performed reliably (Switkes & Crognale, 1999). In the experiment, 9 equally spaced contrasts ranging from -60 to 60 along the LM axis were displayed as circles in an upper row, and then luminance levels were adjusted in circles shown in a corresponding lower row until the lightness steps appeared the same.
magnitude as the chromatic steps. In this case only the central gray was fixed, and thus observers varied all eight of the other lightness levels. The result of this experiment showed that both luminance and chromatic contrasts progressed uniformly, and that a scale of -60 to 60 chromatic contrast was perceived as comparable to a scale of -50 to 50 luminance contrast. (Appendix II for more details). We therefore used this scaling to create the luminance ensembles as given in the following table:

i. **Luminance ensembles**

The ensembles used in this condition and the rationale for using them is tabled below:

Table 6 – Description of lightness ensembles

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Contrast levels</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>-30,-10,10,30</td>
<td>Ensembles with gray average</td>
</tr>
<tr>
<td>L2</td>
<td>-40,-20,20,40</td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td>-50,-30,-10,10</td>
<td>Asymmetric ensemble dark/light bias to test for illusory conjunctions</td>
</tr>
<tr>
<td>L4</td>
<td>-10,10,30,50</td>
<td></td>
</tr>
<tr>
<td>L5</td>
<td>0,-10,-20,-30</td>
<td></td>
</tr>
<tr>
<td>L6</td>
<td>-10,-20,-30,-40</td>
<td>Ensembles with dark bias that span and do not span the gray boundary</td>
</tr>
<tr>
<td>L7</td>
<td>-20,-30,-40,-50</td>
<td></td>
</tr>
<tr>
<td>L8</td>
<td>0,10,20,30</td>
<td></td>
</tr>
</tbody>
</table>
When presenting these ensembles, the background on the monitor was still a neutral gray, but the elements had a fixed chromatic contrast of 30, and again varied randomly in hue on each trial. Thus the set appeared as different lightness levels of a desaturated red or green, etc. The chromatic contrast was again added so that the neutral lightness level appeared distinct from the gray background.

3.1 Results and Discussion

![Figure 7 - Results for lightness ensembles with gray average. Ensembles L1, L2 vary along the luminance axis – Gray was included in this set even though it was not a part of the ensemble.](image)

As in the previous case, the FA for gray was compared with the responses for the members of the ensemble (L1, L2). For both sets, the FA for gray was marginally lower than the TR for the
extreme members, a difference that reached significance for the ensemble with higher variance (Figure 7). (see RMANOVA table below). However, proportion of false alarms for gray appeared markedly higher than for the comparable saturation ensembles ($t(43) = -4.76, p < 0.001; \text{mean difference} = 32.11 \pm 6.75$). Thus observers were more likely to misperceive a gray when it was part of the lightness set than the saturation set.

Table 7 – Summary of analyses L1:L2

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>F value</th>
<th>P value</th>
<th>Significant difference between contrast levels</th>
<th>Mean difference (Bonferroni adjusted p value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-30,-10,10,30</td>
<td>$F(2,16) = 3.35$</td>
<td>0.061</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>-40,-20,20,40</td>
<td>$F(2,18) = 4.47$</td>
<td>0.029</td>
<td>-40 vs 0</td>
<td>-$20 \pm 6.2 (p = 0.036)$</td>
</tr>
</tbody>
</table>
Figure 8 – Luminance ensembles with non-gray average. Contrast levels marked with filled diamond were compared for each ensemble. FA for the average was not significantly different from the response for members of the ensemble showing that luminance contrast varies from dark to bright in a continuum.

As in the color ensembles, paired sample t-tests were done to examine if the response for non-members was lower than the response for the members of L3 and L4. The true response for members of the ensembles was always greater than the FA for non-members showing that there was no evidence for illusory conjunctions (Figure 8).

Table 8 – Summary of analyses L3:L4

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>Contrast levels compared (mean)</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3: -50, -30, -10, 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4: -10, 10, 30, 50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 9 – Gray categorical boundary for luminance. FA drop more gradually compared to saturation. As in saturation, the drop at zero contrast is not different from the drop at other luminance contrasts.

The last set of conditions again probed the rate of fall off in false alarms around the gray category boundary. Figure 9 shows that there is a relatively gradual drop for luminance
contrasts outside the ensemble. Again, as in the previous case, there was no evidence for a steeper drop at the gray boundary. False alarms for the first contrast level outside the ensemble (L5:L10) was similar irrespective of where the ensemble ended, or whether it was a dark/ bright ensemble.

Table7 – Summary of analyses L5:L10

<table>
<thead>
<tr>
<th>Ensembles</th>
<th>Contrast level compared (mean)</th>
<th>F value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,-10,-20,-30</td>
<td>10 (68.75 ± 6.2)</td>
<td>Main effect of Ensembles F(2,42) = 2.66</td>
<td>p value = 0.082</td>
</tr>
<tr>
<td>-10,-20,-30,-40</td>
<td>0 (75 ± 6.2)</td>
<td>p value = 0.064</td>
<td></td>
</tr>
<tr>
<td>-20,-30,-40,-50</td>
<td>-10 (83.1 ± 6.2)</td>
<td>Main effect of bright vs dark F(1,42) = 3.62</td>
<td></td>
</tr>
<tr>
<td>0,10,20,30</td>
<td>-10 (60 ± 6.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10,20,30,40</td>
<td>0 (64.3 ± 6.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20,30,40,50</td>
<td>10 (73.75 ± 6.2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Finally, FA for the first contrast level outside the ensemble was significantly higher (F (1, 45) = 50.88; p value < 0.001) for the lightness ensembles (mean = 70.83 ± 3.7) than that of the saturation ensembles (33.62 ± 3.6) for all the ensembles. This confirmed that the drop in FA for lightness ensembles is more gradual than the drop for the color ensembles.

General discussion:
An important aspect of ensemble perception is the ability to estimate the average value of the parameter of interest. In most cases this is assumed to represent the metric average of the ensemble, though previous studies have shown that the averaging can for example exclude outliers in the ensembles (Haberman & Whitney, 2010). Based on the responses for contrast values within an ensemble, our findings with color show that color contrasts within a hue show evidence for such ensemble representation. These results are similar to other studies in ensemble color perception (Maule & Franklin, 2015; Maule et al., 2014). However, in our experiments, we aimed to study the extent to which this process could generalize across different stimulus categories, by focusing on averaging across the gray boundary. As noted in the Introduction, gray represents a unique categorical boundary in color perception, and thus might pose the greatest challenge to pooling signals across qualitatively different stimulus categories. Our results suggest that ensemble coding for color fails to generalize across complementary color categories. In particular, grays are less likely to be perceived as part of the ensemble, even though they represent the average stimulus and even though the gray is matched for the perceptual distance from the ensemble members. Thus at least in the extreme our results are inconsistent with a simple metrical averaging process underlying ensemble coding for color.

An alternative we explored for color coding was that the visual system might independently represent a color by its perceptual attributes of hue and saturation, and then average within each of these attributes. This scheme might lead to cross-attribute errors in the false alarms. However, we also did not find evidence for this representation. This suggests that even though
hue and saturation are perceptually separable they are not processed separately in ensemble percepts. Instead, our results are consistent with forming separate representations of summary statistics within each hue category.

Yet despite this, we did not find clear evidence for a categorical effect of the gray boundary. That is, false alarms were surprisingly no more likely when the outside element fell within the hue category or a separate category. It should be emphasized that these conditions represent the strongest categorical differences for color, since the two categories represent complementary colors. The failure to find strong categorical effects for these cases thus suggests that categorical effects in color ensemble coding are likely to be weak in general. This is consistent with the weak to absent categorical biases that have been found for other color tasks such as perceptual grouping (Webster & Kay, 2012).

Importantly, we observed different trends for variations in lightness levels. In this case, observers were much more likely to experience the neutral stimulus as a member of the set, even though the luminance and chromatic stimuli were matched for perceptual differences. Moreover, the false alarms for outliers faded more gradually with distance for lightness levels than for color. This raises the possibility that something like a metrical average is more likely to be computed for lightness than hue. Moreover, it points to intriguing asymmetries between luminance and chromatic processing. Opponency is considered a hallmark of color appearance and there are clear opposing differences between both complementary hues and complementary lightness levels as well as clear physiological substrates identified for these differences (Komban, Alonso, & Zaidi, 2011). Yet subjectively, it is perhaps easier to perceive
increments and decrements as part of a uniform continuum than two opposing colors such as red and green or blue and yellow. The ensemble coding differences are consistent with these subjective impressions, and suggest that the significance and signature of gray – a singularity at the center of color space – may depend importantly on whether the path through it varies in chromaticity or luminance.

**Conclusions:**

In summary, we conclude that ensembles percepts of color contrast are not based on simple metrics like Euclidean distance between the contrasts, and that these member judgements of contrast tend to be restricted to ensemble within rather than between hues. Also, we report qualitative differences in how lightness and saturation are represented in color processing based on the differences in member judgments between saturation and lightness ensembles.

**References:**


adaptation: data and multi-scale model. *Journal of Vision, 11*(2), 7-.
https://doi.org/10.1167/11.2.7


https://doi.org/https://doi.org/10.1016/j.cub.2007.06.039

https://doi.org/10.3758/APP.72.7.1825


https://doi.org/10.1364/JOSA.69.001183

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on


https://doi.org/10.1364/JOSAA.31.000A93


https://doi.org/https://doi.org/10.1016/S0042-6989(98)00219-3


https://doi.org/https://doi.org/10.1016/0042-6989(92)90036-I


https://doi.org/10.1364/JOSAA.31.00A283


Appendix I

Chromatic contrast scale metrics

Fig 10. Chromatic contrast scaling - 7 participants adjusted the intermediate contrast levels till the sequence appeared to progress uniformly.
Fig 11. Contrast scaling result - The progression from –120 to +120 appears to have a linear progression along the LM axis. Hence for all further experiments involving chromatic contrasts, the colors were defined in the CIELAB space. The LAB space is a uniform color space in which a unit distance on any two axes along an isoluminant plane has uniform chromatic change.
Appendix II

Luminance contrast scale metrics

In order to compare ensemble coding mechanisms for colors varying in their saturation and luminances, the relationship between luminance and chromatic contrast metrics need to be understood. The progression in luminance contrast was compared with the progression in chromatic contrast using a similar task described above. In this luminance contrast scaling task, participants were presented with two set of circles. The top set consisted to 9 circles uniformly varying in chromatic contrast from -60 to +60 and were fixed, while 9 test circles in bottom set could all be adjusted except the middle circle which had a fixed luminance that matched the luminance of the background. Participants were asked to adjust the luminance of all other test circles to match the chromatic contrast of the top set.
Fig 12. Luminance contrast scale to match the chromatic contrast progression. 4 participants adjusted the luminance contrast of the lower circles till they matched the chromatic contrast of the top row.

Figure 13. Participants set the relative luminance contrast of the darkest and brightest test circles at ~ 50% from that of the background. Also, there was an almost linear progression from the darkest to the brightest test circle. This made the progression in luminance contrast and chromatic contrast comparable, and more importantly, based on these settings, the minimum and the maximum of the test contrast levels was set at -40 and 40 for the luminance ensembles.
Discriminating color ensembles with different numbers of colors

Abstract:

Studies of ensemble coding have primarily focused on the visual system’s ability to estimate summary properties of an ensemble like the average and variance of the set of elements. However, the specific distribution defining the ensemble has received less attention. In this study, we examined the ability to discriminate between color ensembles that differed in the number of colors defining the set. This was explored behaviorally by asking participants to identify which of 4 color palettes included a larger number of color elements. The discrimination was also assessed neutrally by using an EEG – periodic visual stimulation “oddball” paradigm where the palette with an extra color was shown once per second in a stream of ensembles presented at 6 images per second. In both tasks, ensembles that differed from two to three colors were readily distinguished, yet performance fell rapidly when the set size was increased further. These results suggest that while the visual system may efficiently summarize the mean and variance of a set of colors preattentive sensitivity to the density of colors defining the ensemble is very restricted.

Introduction:

In 2015 a hand-woven sari was unveiled in India that contained 50,000 unique color swatches. This was a remarkable artistic achievement, but raises the question of how many colors the sari might appear to have. Human color vision is capable of discriminating between millions of colors when compared side by side (Foster & Nascimento, 1994). But most images are composed of
variegated colors interspersed across the scene. How well can we distinguish between variants of these color distributions?

One of the effective ways of coding large amounts of varying information is by representing an ensemble using its summary properties like the average value (Alvarez, 2011; Whitney & Leib, 2018). Color has shown some evidence for ensemble perception (Kimura, 2018; Maule & Franklin, 2015; J. Webster, Kay, & Webster, 2014). Apart from the average, the visual system can also represent the variance of the distribution (Norman, Heywood, & Kentridge, 2015), which can provide critical information such as the diversity or overall gamut of the ensemble (Haberman, Lee, & Whitney, 2015). The average and variance may be of special importance for coding color. For example, the mean chromaticity is a potential cue to the color of the illuminant (Foster, 2011) and also to the nature of the scene (e.g. whether it is lush green or arid yellow; Juricevic & Webster, 2009). Similarly, the variance provides cues to contexts such as fog or haze, and observers are very sensitive to the “correct” balance of color and luminance contrast in images (Mcdermott, Mulligan, & Webster, 2010). Vision may also encode higher-order statistics of scene colors. For example, Chetverikov et al, (2017) reported that observers can differentiate between a uniform vs. Gaussian distribution of colors. However, whether and how well the visual system encodes different statistics of color distributions remains poorly understood.

The perception of color ensembles has close connections to the perception of visual textures, which are also defined by the statistics and features of the elements forming them. Texture perception has been widely studied (Victor, Conte, & Chubb, 2017 for review), and shown to
depend on a small number of primitives (e.g. color, size, orientation) and how they are distributed in the image. Here again sensitivity is known to be high for some statistical differences (e.g. the average or variance of the orientations) while higher order statistics are difficult to resolve. These effects have also been demonstrated in a wealth of studies focused on visual search to ask which properties of scenes are or are not readily discriminated (Wolfe & Horowitz, 2004), as well as other works on change blindness to ask which differences we notice about the world (Simons & Rensink, 2005 for review). Together these paradigms have pointed to a general (but fuzzy) distinction between preattentive and post-attentive processing, which differ in whether the differences “pop out” and are thus automatically registered by the visual system, or require scrutiny and attention before they are recognized.

In this study we examined sensitivity to the number of colors in a set, for example whether they were composed of 3 colors or 4. The ability to estimate the number of items in a scene has also been widely studied in other contexts, and has given rise to the concept of a visual number sense (Dehaene, 2003; Guillaume et al., 2018). For example, humans are very good at detecting whether one image include a larger number of dots or higher density (Durgin, 2008). Numerosity discrimination judgements depend on several factors like the difference in the number of elements between the sets (Dehaene, 2003), relationship between the size of the each element and overall size of the set, perceptual cues such as pop-out, etc. (Guillaume M et al 2018). However, numerosity perception is in one way fundamentally different from judging the density if colors, because former probes “how many” elements there are while the latter varies “how many types” there are. Thus visual performance for the two tasks may be very
different. For example, Silva and Chubb (2014) studied sensitivity to different distributions of luminance levels in a texture. Remarkably, they showed that an appropriate set of as few as 3 lightness levels were indistinguishable for a texture with effectively continuous lightness gradient.

The present work extends the results of Chubb et al. to the color domain, to examine the ability to perceive the number and type of chromatic differences within a texture. The textures were palettes composed randomly colored squares chosen from a different number of colors. In some sets the individual colors differed in hue (with the same contrast or roughly equivalent saturation), while in others they differed in the number of saturation levels but had the same hue. The sets all had the same mean chromaticity (gray) and the same saturation (along the hue angle in the set) and thus differed only in the numerosity of colors. We then used a combination of behavioral and electrophysiological techniques to study sensitivity to the number of colors in the images.

The electrophysiological measures were included in order to specifically assess the rapid preattentive coding of differences in color number. A recent study objectively measured numerosity judgments in the visual system using electroencephalography (EEG) coupled with fast periodic visual presentation (FPVS) of stimuli (Guillaume et al., 2018; Park, 2018). Specifically, a periodic oddball technique was used to test if there are neural correlates to the number judgments. In this paradigm, a stream of similar images (e.g. same face identity) is presented at a particular frequency called the base frequency. Within the stream, an oddball image (e.g. – different facial identity) is presented at a slower frequency (the oddball frequency;
e.g. every 6\textsuperscript{th} image). If the oddball and base images are discriminated by the neural system, during frequency spectrum analysis, two simultaneous responses – one at the base frequency and the other at the oddball frequency can be extracted. This has been used in a number of studies (Liu-Shuang, Norcia, & Rossion, 2014; Retter & Rossion, 2016a; Rossion, 2014) to measure discriminability between object categories. In the EEG experiment, we used a similar technique to measure ensemble discrimination of color number.

**Methods:**

**Participants:**

9 observers participated in the behavioral experiment, and 10 participated in the EEG experiments. All of them were recruited from the University of Nevada student pool, and had normal or correct to normal visual acuity, and normal color vision assessed with the metropsis visual function assessment system and the Cambridge color test (CCT).

**Stimuli:**

Test stimulus was a 10X10 color ensemble made up of 2 to 6 individual colors. Two types of ensembles were used – ones that varied in number of colors and the ones that varied in number of saturation levels. This differentiation was made to explore differences between the way color numerosity varies along the hue and saturation dimensions. All stimuli were defined in a version of the Macloed-Boynton color space that was modified to scale the contrasts equally along the two cardinal cone opponent directions (Webster et al., 2000). The specific scalings are given below:
LvsM contrast = \((r_{mb} - 0.6568) \times 2754\)

SvsLM contrast = \((b_{mb} - 0.01825) \times 4099\)

where LvsM is the reported contrast level, \(r_{mb}\) and \(b_{mb}\) are the chromaticity coordinates of the stimuli in the MacLeod-Boynton color space.

**Chromatic ensembles:**

Colors were chosen by keeping the luminance and chromatic contrasts constant at 30 cd/m2 and 65 respectively, while varying hue angle such that the individual colors of the ensemble were always spaced equally apart in the color space. For example, in an ensemble made of two colors, the individual colors of the ensemble were made as distinguishable as possible by keeping angular distance between the two colors at 180 degrees. For a 3-color ensemble the hue angles were 120 degrees apart, and so on.

**Saturation ensembles:**

The saturation ensemble had individual saturation levels such that they were spaced equally apart along a single axis of the color space. For example, in an ensemble made of two colors, the individual colors were chosen from either side of gray, in an ensemble made of three colors, the individual saturation levels the two colors plus gray, and so on.

**Procedure:**

**Behavioral experiments:**
The ensembles were displayed on the four quadrants of a calibrated CRT monitor on a uniform color background. In any given trial, three quadrants had ensembles with the same number of colors, and one of the quadrants, chosen at random had an extra color. We call the three ensembles that were similar in composition as “base ensembles” and the one that looked different from the rest as the “odd ensemble”. When the base stimuli were color, the odd stimulus was also a color stimulus, and when the base stimuli were saturation stimuli, the odd one was also a saturation stimulus. Depending on the specific condition, the base stimuli were all made of the same set of colors (Color, Fixed (CF)) or saturation (saturation, fixed (SF)), or were made different set of colors (color, random (CR)) or saturation (saturation, random (SR)), all while maintain the number of colors in the ensemble. The odd ensemble for all these conditions had one color/saturation level more than the base. In order to test if responses are entirely driven by pop-out effects of the odd ensemble, we used a fifth condition (ensemble rotated (ER) condition) in which the base and odd ensemble all had same number of colors, but the odd ensemble had
Figure 1 – Task for behavioral experiment. In this example three of the base ensembles have only two colors, while the oddball ensemble is composed of 3 colors. In the condition shown the bae palettes were also constrained to sample from the same two hues, while in other conditions each was chosen from a random hue angle.

its constituent colors rotated 36 degrees relative to the angle of colors in the base ensemble. The angle of 36 degrees corresponded to the color difference that would occur between a base ensemble composed of 10 colors, and thus allowed us to test the sensitivity to the actual color differences in the palettes, separately from the number of colors in the palettes.

The number of colors in the base ensembles varied between 2 and 5, and thus the colors in the odd ensemble varied between 3 and 6. All four base colors were presented in the five test conditions (CF – Color, Fixed; CR - Color, Random; SF – Saturation, Fixed; SR - Saturation, Random; ER – Ensemble Rotated) twenty times in two different ways. In the first case, the
ensembles were presented briefly (250 ms), and this was used to measure the accuracy participants’ response. In the second case, the ensembles were presented for 5 secs, and participants freely viewed the stimuli to give a response. Responses from these longer stimulus duration trials were used to measure the reaction time (RT). At the end of the 5 secs, participants were asked to guess, and this value was used as the reaction time. As can be seen from the results (next section), no participant took more than 3 secs to make a response.

**Result:**

**Accuracy for discrimination color number:**

Accuracy for all the conditions was highest when the base ensembles were made of 2 colors and the odd ensembles were made of 3 colors. Accuracy responses fell sharply even when the base ensembles was made of 3 colors. Individual one way repeated measures analysis of variance (RMANOVA) for each of the test conditions with the four base colors as the repeated measure is given below:

<table>
<thead>
<tr>
<th></th>
<th>Color, Fixed (Fig 3a)</th>
<th>Color, Random (Fig 3a)</th>
<th>Rotated (Fig 4a)</th>
<th>Saturation, Fixed (Fig 3b)</th>
<th>Saturation, Random (Fig 3b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main Effect</strong></td>
<td>F(3,24) = 83.5</td>
<td>F(3,24) = 30.45</td>
<td>F(3,24) = 43.1</td>
<td>F(3,24) = 18.95</td>
<td>F(3,24) = 27.009</td>
</tr>
<tr>
<td></td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.001</td>
<td>p &lt; 0.01</td>
</tr>
<tr>
<td>2Color_base (mean accuracy ± SEM)</td>
<td>0.97 ± 0.02</td>
<td>0.64 ± 0.06</td>
<td>0.989 ± 0.01</td>
<td>0.739 ± 0.08</td>
<td>0.739 ± 0.06</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>3Color_base (mean accuracy ± SEM)</td>
<td>0.594 ± 0.06</td>
<td>0.33 ± 0.03</td>
<td>0.822 ± 0.06</td>
<td>0.267 ± 0.04</td>
<td>0.25 ± 0.046</td>
</tr>
<tr>
<td>4Color_base (mean accuracy ± SEM)</td>
<td>0.22 ± 0.04</td>
<td>0.25 ± 0.03</td>
<td>0.489 ± 0.09</td>
<td>0.25 ± 0.03</td>
<td>0.217 ± 0.019</td>
</tr>
<tr>
<td>5Color_base (mean accuracy ± SEM)</td>
<td>0.19 ± 0.03</td>
<td>0.2 ± 0.02</td>
<td>0.344 ± 0.06</td>
<td>0.278 ± 0.034</td>
<td>0.272 ± 0.034</td>
</tr>
<tr>
<td>Accuracy of response significant between</td>
<td>2Color_base &amp; 3Color_base (p=0.001); 3Color_base</td>
<td>2Color_base &amp; 3Color_base (p=0.012)</td>
<td>3Color_base &amp; 3Color_base (p = 0.03)</td>
<td>2Color_base &amp; 3Color_base (p = 0.013)</td>
<td>2Color_base &amp; 3Color_base (p=0.01)</td>
</tr>
</tbody>
</table>
Figure 2 – Accuracy of detecting the ensemble with more colors for (a) color ensembles (b) saturation ensembles
Figure 3 – Responses for the ensemble rotated condition. All the ensembles in this condition had the same number of colors, but one of them had colors rotated 36 degrees relative to the colors of other ensembles. (a) Drop in accuracy and (b) Increase in reaction time with increasing number of colors in the ensemble

Reaction times for detecting the ensemble with more colors:

Participants were given five seconds to find the ensemble that had the most number of colors, and the reaction time from this experiment was used for further analysis. Consistent with the accuracy measurements, the lowest reaction time was found for the fixed color ensembles, and reaction times were progressively slower for increasing colors in the ensembles.

Table 2 – Summary of analyses for reaction time:

<table>
<thead>
<tr>
<th>Color, Fixed</th>
<th>Color, Rotated (Fig)</th>
<th>Saturation,</th>
<th>Saturation,</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Fig 5a)</td>
<td>Random (Fig 5a)</td>
<td>4b)</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------</td>
<td>----------------</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Main Effect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F(1,8) =</td>
<td>F(3,24) =</td>
<td>F(1,8) =</td>
</tr>
<tr>
<td></td>
<td>12.97</td>
<td>5.997</td>
<td>26.5</td>
</tr>
<tr>
<td></td>
<td>p = 0.007</td>
<td>p = 0.003</td>
<td>p = 0.001</td>
</tr>
<tr>
<td><strong>2Color_base</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean RT (ms) ± SEM)</td>
<td>794.2 ± 79.2</td>
<td>1517 ± 196.6</td>
<td>706.9 ± 30.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>3Color_base</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean RT (ms) ± SEM)</td>
<td>1353.9 ± 184.9</td>
<td>2161.07 ± 314.8</td>
<td>989.6 ± 94.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>4Color_base</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean RT (ms) ± SEM)</td>
<td>2088.9 ± 317.2</td>
<td>2452.3 ± 438.2</td>
<td>1954.2 ± 272.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5Color_base</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(mean RT (ms) ± SEM)</td>
<td>2401.2 ± 451.3</td>
<td>2529.7 ± 459.2</td>
<td>2465.0 ± 337.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RT of response</strong></td>
<td>2Color_base &amp; 3Color_base</td>
<td>none</td>
<td>2Color_base &amp; 3Color_base (p = 0.026)</td>
</tr>
<tr>
<td><strong>significant</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
between (Bonferroni corrected p-value) 3Color_base & 4Color_base (p=0.047)  & 3Color_base & 4Color_base (p = 0.024); 4Color_base & 5Color_base (p = 0.038)

(a)  

(b)
Figure 4 – Increasing reaction time for increasing color number of base ensembles for (a) color ensembles (b) saturation ensembles

Fixed vs random conditions:
As noted, participants were tested in two conditions (fixed and random). In the fixed condition, all the base ensembles were made of the same colors, whereas in the random condition, all the base ensembles were made of different colors, still maintaining the number of colors. The random condition is a true numerosity task, since in the fixed condition observed could have based their responses on the differences in the chromaticities of the elements. To compare these, we used a two-way RMANOVA (ensemble type (fixed vs random) X base number (2-5)) for the color ensembles to compare performance. For the color ensembles, both the number of colors and the type of color ensemble (fixed/random) affected the accuracy and reaction time of the responses. As expected, participants were better in the fixed than the random condition. Also, there was a significant interaction between the types of ensemble and the base colors, presumably because performance on both tasks fell to chance at higher base levels. A similar 2 way RMANOVA (ensemble type (fixed vs random) X base number (2-5)) was done for the saturation ensembles, and it showed no main effect of base number or ensemble type and there was no interaction between the type of ensembles and the number of colors in the ensemble.

The summary of the 2 X 4 RMANOVAs are tabled below:

Table 3 – Comparison between fixed and random conditions:

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>RTs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Color number vs. color discrimination:

Based on the condition, participants were given explicit instructions to either find the ensemble with the most number of colors (for Color Fixed, Color Random, Saturation Fixed and Saturation Random), or to find the ensemble that was different from the rest (Color Rotated). Again, the latter was included as a control to determine if the discrimination was based on simply judging what the actual colors were in the array (rather than the number of colors). The Accuracy and RTs for the ensembles were compared with the rotated ensembles in two separate RMANOVAs, and it was found that the accuracy was higher for the rotated than for the random condition, and there was an interaction between the type of ensemble and the number of colors in the
ensemble, effectively confirming that the participant responses cannot be entirely explained by color discrimination alone. The details of the ANOVAs are tabled below:

Table 4 – Comparison between rotated and random condition:

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>RTs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main effect of random vs rotated</strong></td>
<td>F(3,24) = 30.9 p = 0.001</td>
<td>F(1,8) = 7.6 p = 0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Main effect of base ensemble</strong></td>
<td>F(3,24) = 63.8 p &lt; 0.001</td>
<td>F(1,8) = 20.23 p = 0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interaction</strong></td>
<td>F(3,24) = 7.4 p = 0.001</td>
<td>F(1,8) = 4.9 p = 0.057</td>
</tr>
</tbody>
</table>

**Discussion:**

In summary, the visual system’s ability to discriminate color number is poor for hue variations on an isoluminant plane. We found that participants were not able to distinguish an ensemble differing in the number of colors present for ensembles composed of 4 or more colors. Participants performed poorly and took longer to respond even while comparing ensembles made of 3 colors vs. 4 colors. The saturation ensembles were worse, wherein participants were unable to distinguish between 3 and 4 saturation levels (though we note that the saturation and hue steps were not equated, so this could partly reflect smaller step sizes for saturation). One reason for this poor performance could be that the odd ensemble always had just one color more than the base ensemble. This has been reported to be a factor affecting numerosity
judgments (Dehaene, 2003). Even so, participants performed more poorly in the rotated condition for ensembles made of 4 colors. This highlights that the visual system has poor sensitivity to color number.

**Fast periodic visual stimulation – Electroencephalography (FPVS-EEG Oddball):**

The behavioral experiments described above were compared with EEG experiments to probe the neural correlates for these results.

**Stimuli**

Unlike the stimuli for the behavioral experiments, the ensembles for the EEG experiments were shown in sequence one at a time. Three sets of ensembles were created using the same methods as in the behavioral experiment. However, to control for spatial artifacts in the palettes (e.g. clumping of the same color or mean color differences (e.g. too many of one color), for this experiment we also constrained the distributions so that the palettes had a similar number of each color. For example, for the two color condition we constrained the images to ones that did not deviate by more than 40-60% mixtures of the two colors.

**Procedure:**

Stimuli were presented on a calibrated LCD monitor with a refresh rate of 120Hz. The ensembles were presented one at a time at the center of the monitor in a dark room while participants performed a fixation task, and as the EEG signals were recorded using an EEG cap. Six ensembles were presented every second, and within this sequence five image were the base image while every 6th image was the odd ensemble. To ensure that the EEG measure could detect a change in the color number, we tested not only oddballs with one extra color but also larger
differences. For this the ensemble with 4 colors was compared to two odd ensembles – 5 and 6. The ratio between the 4 colors vs 6 colors was the same as the ratio between 2 vs 3 colors. Therefore, in the EEG part, 4 base colors and 5 odd colors were tested, and as in the behavioral part, both fixed and random conditions were tested. Images were presented in a square wave presentation mode as in a previous report (Retter, 2018) (Figure 6 for the task).

Figure 5 – EEG task. Stimuli were presented at 6Hz with every 6th image having more colors than the rest at 50% duty cycle. Participants pressed the space key whenever fixation changed to a square. Shown here is the 2vs3 random, color condition.

The trial details are tabled below:

<table>
<thead>
<tr>
<th>Event</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-sequence fixation cross</td>
<td>1-3 secs</td>
</tr>
<tr>
<td>Fade-in</td>
<td>2-4 secs</td>
</tr>
<tr>
<td>Testing sequence</td>
<td>50 secs</td>
</tr>
<tr>
<td>Fade-out</td>
<td>2-4 secs</td>
</tr>
<tr>
<td>Post – sequence fixation cross</td>
<td>1-3 secs</td>
</tr>
</tbody>
</table>
Participants were asked to press the space-bar every time the fixation changed from a square to a cross, and this was done to ensure fixation. The trial sequence was repeated thrice for every base vs odd ensemble comparison. This resulted in 30 trials per session (2vs3, 3vs4, 4vs5, 4vs6, 5vs6 – fixed and random, each repeated thrice), and a recording time of about 40 minutes.

**EEG acquisition:**

The BioSemi ActiveTwo EEG system containing 128 Ag-AgCl Active electrodes to collect EEG signals. The standard cap configuration around the 10/20 location was used (BioSemi B.V., Amsterdam Netherlands). The 10/20 location nomenclature was changed to the more conventional 10/5 nomenclature during analysis. The common mode sense (CMS) and direct right leg (DRL) electrodes were used as reference in the head cap. The offsets were maintained below 40 mV, and the signal was recorded at a sampling rate of 512 Hz. Vertical eye movements were recorded using two flat type Active electrodes placed above and below the right eye, and horizontal movements were recorded using another pair of electrodes placed on outer canthi of the either eyes.

**Analysis:**

Analysis was performed using Letswave 5, an EEG analysis toolbox for MATLAB R2013b (Mathworks, USA).

**Preprocessing:**

A Butterworth band-pass filter was applied to individual participant data with cutoff 0.1-100 Hz. Electrical noise was removed using a multinotch frequency filter at 60Hz (width 0.5Hz). Data was then segmented based on the 10 conditions in the experiment, and the fade-in/fade-out time of
each trial. The channels were then re-referenced to a common average of all the 128 channels, and individual image sequence trials were segmented based on the time of trial onset to 50 seconds coinciding with the time needed for a complete sequence presentation.

Results:

Frequency domain analysis:
A fast Fourier transform was applied to normalize the amplitude spectrum covering 0-256 Hz in the frequency domain. A baseline subtraction used by previous reports (Retter and Rossion 2016) was applied at each frequency, using the frequency of 30 neighboring bins. This resulted in the reduction of uncorrelated noise. This process was applied to each participant, and the overall grand-average amplitude and baseline-subtracted amplitude spectra across all participants was calculated for display (figure 6).

Responses to image presentation frequency of 6Hz and the differential response at 1 Hz occurred at their respective frequency and its harmonics. The unique harmonics were summed to obtain the total response to each tagged frequency. These summed responses were used to plot the topographical maps of individual harmonic and summed harmonic responses to display the data across all the channels. The response for the 1 Hz oddball frequency has been shown in figure 6.
Figure 6 – Grand averaged response topography across the back of head for each of the conditions. Isolated 1 Hz response for the ensemble with more colors than the base. Top panel – (a) for the fixed ensembles (b) for the random ensembles

Oddball response for the larger color number across all conditions:

Periodic responses were obtained for images at the presentation frequencies. As far as the oddball response was concerned, the response was expected to be maximal for the 2 vs 3 fixed and random conditions. Previous reports on measuring visual quantity discrimination have found responses mostly in the occipital regions (Guillaume 2018), and consistent with previous reports, responses to our stimuli were strongest over the occipital regions.

All EEG channels were grand averaged and the amplitude spectra were pooled. Z-scores were calculated on this amplitude spectrum for each discrete frequency bin (x) according to the
formula $Z = \frac{(x - \text{baseline mean})}{(\text{baseline standard deviation})}$. For z-score calculations, baseline was defined as the 20 frequency bins surrounding each target bin.

Oddball frequency harmonics up to the 9th including the first harmonic were considered for analysis. Harmonics which coincided with base stimulation frequency was excluded. Z-scores more than 1.64 ($p<0.05$) were considered significant according to the homologous criteria. There was a significant difference between the signal (oddball response) and noise for the 2 vs 3 color fixed and color random conditions. Z-scores for all other conditions were not significant.

Table 5 – Significant z-scores ROI – 128 channels:

<table>
<thead>
<tr>
<th></th>
<th>2vs3 Fixed</th>
<th>2vs3 Random</th>
<th>3vs4 Fixed</th>
<th>3vs4 Random</th>
<th>4vs5 Fixed</th>
<th>4vs5 Random</th>
<th>5vs6 Fixed</th>
<th>5vs6 Random</th>
<th>4vs6 Fixed</th>
<th>4vs6 Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.79</td>
<td>0.37</td>
<td>0.65</td>
<td>0.42</td>
<td>0.52</td>
<td>-0.4</td>
<td>0.67</td>
<td>0.29</td>
<td>-0.68</td>
<td>0.34</td>
</tr>
<tr>
<td>2</td>
<td>-0.82</td>
<td>0.89</td>
<td>-0.21</td>
<td>-1.09</td>
<td>-0.67</td>
<td>-0.48</td>
<td>1.08</td>
<td>-1.83</td>
<td>0.73</td>
<td>-0.02</td>
</tr>
<tr>
<td>3</td>
<td>-0.05</td>
<td>0.3</td>
<td>-0.59</td>
<td>0.78</td>
<td>-0.22</td>
<td>0.39</td>
<td>0.63</td>
<td>-1</td>
<td>0.72</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>-1.66</td>
<td>1.7</td>
<td>0.41</td>
<td>-0.07</td>
<td>-2.13</td>
<td>1.5</td>
<td>0.98</td>
<td>-1.07</td>
<td>0.42</td>
<td>-1.05</td>
</tr>
<tr>
<td>5</td>
<td>2.14</td>
<td>0.07</td>
<td>-0.39</td>
<td>0.79</td>
<td>0.35</td>
<td>0.44</td>
<td>2.95</td>
<td>-1.29</td>
<td>1.71</td>
<td>-0.69</td>
</tr>
<tr>
<td>6</td>
<td>2.37</td>
<td>3.92</td>
<td>2.19</td>
<td>1.61</td>
<td>-0.37</td>
<td>-0.97</td>
<td>-0.55</td>
<td>0.23</td>
<td>-1.29</td>
<td>0.39</td>
</tr>
<tr>
<td>7</td>
<td>1.37</td>
<td>0.35</td>
<td>0.41</td>
<td>-0.72</td>
<td>-1.26</td>
<td>1.03</td>
<td>0.89</td>
<td>1.45</td>
<td>0.73</td>
<td>1.09</td>
</tr>
<tr>
<td>8</td>
<td>-0.67</td>
<td>3.54</td>
<td>0.75</td>
<td>1.48</td>
<td>0.31</td>
<td>0.64</td>
<td>-0.13</td>
<td>-0.2</td>
<td>-0.17</td>
<td>-1.73</td>
</tr>
</tbody>
</table>

In a second case, the occipital region was isolated as the ROI, and z-scores of the 28 occipital electrodes was obtained. Again, as with the average of 128 channels, significant Z-score responses were obtained only for the 2 vs 3 fixed and random condition.

Table 6 – Significant z-scores ROI – 28 occipital channels:

<table>
<thead>
<tr>
<th></th>
<th>2vs3 Fixed</th>
<th>2vs3 Random</th>
<th>3vs4 Fixed</th>
<th>3vs4 Random</th>
<th>4vs5 Fixed</th>
<th>4vs5 Random</th>
<th>5vs6 Fixed</th>
<th>5vs6 Random</th>
<th>4vs6 Fixed</th>
<th>4vs6 Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.58</td>
<td>1.04</td>
<td>0.15</td>
<td>0.59</td>
<td>0.22</td>
<td>-0.13</td>
<td>0.58</td>
<td>-0.18</td>
<td>-0.39</td>
<td>0.17</td>
</tr>
<tr>
<td>2</td>
<td>-1.05</td>
<td>1.39</td>
<td>0.01</td>
<td>-0.01</td>
<td>-1.12</td>
<td>-0.01</td>
<td>1.4</td>
<td>-1.42</td>
<td>-0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>0.58</td>
<td>2.74</td>
<td>-0.9</td>
<td>-0.18</td>
<td>-0.05</td>
<td>0.25</td>
<td>-1.54</td>
<td>-0.34</td>
<td>-0.34</td>
<td>0.17</td>
</tr>
<tr>
<td>4</td>
<td>0.29</td>
<td>3.38</td>
<td>-0.54</td>
<td>-0.38</td>
<td>-1.36</td>
<td>1.21</td>
<td>-0.29</td>
<td>-0.34</td>
<td>0.61</td>
<td>-0.71</td>
</tr>
<tr>
<td>5</td>
<td>1.62</td>
<td>0.93</td>
<td>-0.63</td>
<td>0.47</td>
<td>0.08</td>
<td>-0.02</td>
<td>1.48</td>
<td>-1.79</td>
<td>0.85</td>
<td>-0.73</td>
</tr>
<tr>
<td>6</td>
<td>4.44</td>
<td>7.4</td>
<td>0.62</td>
<td>1.29</td>
<td>0.37</td>
<td>-0.83</td>
<td>-1.16</td>
<td>0.48</td>
<td>-0.95</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>2.25</td>
<td>1.51</td>
<td>0.74</td>
<td>-0.09</td>
<td>-0.7</td>
<td>1.02</td>
<td>1.1</td>
<td>0.72</td>
<td>0.3</td>
<td>1.29</td>
</tr>
<tr>
<td>8</td>
<td>1.53</td>
<td>5.35</td>
<td>0.16</td>
<td>0.09</td>
<td>0.12</td>
<td>0.33</td>
<td>0.18</td>
<td>-0.09</td>
<td>0.51</td>
<td>-1.82</td>
</tr>
</tbody>
</table>
Discussion:

Unlike the behavioral experiment, we used a control condition in the EEG comparing ensembles made of 4 colors with those that were made of 6 colors. This ratio between the number of colors was the same as the ratio between 2 color ensembles and 3 color ensembles. Yet, there were no significant response for this condition. In fact, the response for the 2 color vs 3 color condition itself was only weak. These results again show that the visual system’s ability to discriminate between color ensembles is weak.

General discussion:

Ensemble coding has proven to be a general process in sensory representations and is thought to allow efficient and rapid estimates of the “gist” of scenes by representing their summary properties such as the mean and variance of the stimulus distribution. Here we explored another statistic of the ensemble – the number of distinct elements composing the distribution. To summarize, both the behavioral and neural measures suggest that preattentive sensitivity to the number of colors in an ensemble drops rapidly when the number of colors increases beyond three. We first consider the potential bases for these effects and then consider the differences between the sensitivity as assessed by the psychophysical and EEG measurements.

As a feature color is one of the most powerful cues for perceptual grouping or for controlling visual salience (Wolfe & Horowitz, 2004). For example, a red fruit stands out easily from green foliage, and this is considered one of the primary advantages driving the evolution of primate color vision (Mollon, 1989). However, a unique color is difficult to detect when the chromaticity
lies within the distribution of background colors (Mcdermott et al., 2010). D’Zmura (1986) first systematically probed the conditions under which colors could be detected in a visual search task. He showed that targets popped out only when their chromaticity was linearly separable from the distractor colors. Specifically, a target that fall on a line in color space connecting two distractor chromaticities was difficult to find, while targets that could be separated from the background by a line were distinct. Our two-color stimuli were similarly defined by a single axis in color space, and thus when the third color was added the ensemble instead formed an area in the color plane that extended beyond the baseline axis. Thus these oddball ensembles should be readily discriminable. However, going from 3 to 4 colors instead resulted in a smaller expansion of the area from a triangle to a square, and this smaller excursion predicts a weaker perceptual difference between the patterns. By the time the base ensemble was composed of 5 colors then the ensembles appear to become indistinguishable. Chubb et al. found a similar result (personal communication) by showing that an ensemble formed by the 4 unique hues (red, green, blue and yellow) was not preattentively distinguishable from an ensemble made up of the 4 intermediate binary hues (i.e. orange, purple, yellow-green, and blue-green).

However, chromatic factors alone do not fully explain the results. First, we showed that observers could still discriminate the 2 vs 3 ensembles when the hue angles of each ensemble were randomized. Thus the oddball was not necessarily easily separable from the base colors collectively. Second, we did not explore the conditions for good discrimination more generally, because we only considered ensembles that differed by one color, except in the 4 vs 6 color condition in the EEG oddball task. In informal studies we examined ensembles with few vs many
colors, and the differences were obvious, but this was because of the spatial statistics of the palettes rather than their color statistics. Specifically, when an ensemble has only 2 colors there is a high incidence of neighboring elements with the same chromaticity, creating large uniform “clumps.” The texture scramble illusion of Chubb et al. (2014) controls for this clustering so that the 3 level ensemble has a similar spatial variegation to the multilevel ensembles. In our task, we did not attempt to control for this factor. It is likely that a second important factor in the drop in sensitivity was related to the fact that as the base number increases the spatial statistics start to asymptote so that there is, for example, little difference between the nearest neighbor statistics for 4 vs 5 colors compared to 2 vs 3. A useful extension of the present work would be to repeat the studies here but comparing ensembles that had 2 or 3 times the color differences, to see when performance again falls to chance. However, the point we emphasize here is that the discrimination we measured may be as much based on spatial differences as color differences.

As noted, both the behavioral and EEG measurements showed a general low sensitivity to the color density of the ensembles. Nevertheless, there were some differences between the two. The behavioral results showed that participants were able to discriminate ensembles made of three colors from those made of two colors both quickly and accurately. In the EEG experiment though, even the differential response between 2 colors and 3 colors was weak. There are a number of factors that could account for these differences.

*Exposure time:*
In the behavioral experiment, the stimulus was shown for 250 ms, while in the EEG experiment, images were presented at 6Hz frequency in a 50% duty cycle which limited image exposure time to 83 ms. Previous EEG experiments studying numerosity have used both longer (100 ms in sinusoidal stimulus presentation; (Guillaume et al., 2018)) and shorter (62.5 ms (Park, 2018)) exposures. One major difference between the stimuli in these studies and our study is that in all of these earlier studies a set of dots or other such similar shapes were used as stimuli, and numerosity judgments were based on an increase in the total number of elements in the set. In our study though, the stimulus always had the same number of elements, and only a subpart of the stimulus, namely the number of colors between the ensembles changed. This is more complex task than detecting an increase in the number of elements. Apparently detecting a change in the number of colors in the ensemble is a more difficult task and thus does not show similar high performance as numerosity judgments at these rapid rates.

Explicit vs implicit judgements:

Another major difference between the behavioral and EEG experiments is that the behavioral responses were based on explicit instructions to find the ensemble that had more colors than the rest. The EEG task was more passive where any difference in signal would be the result of implicit ensemble representation. Although this passive viewing task has produced differences in amplitude of signals between the base and oddball images (Guillaume et al., 2018; Retter et al., 2018; Retter & Rossion, 2016b), it is possible that the behavioral responses to the color ensembles were more likely to recruit attention and thus improve performance.
As we noted at the outset, this study was inspired by reports of a sari with 50,000 colors. What can we now conclude about the number of colors seen in it? Within the sari the colors varied in smooth gradients, and thus nearby checks involved only subtle color steps. Again observers are highly sensitive to these side by side comparisons, and this is the reason most electronic visual displays have a resolution of 8 bits or more per color, since coarser steps are readily apparent. But what of a random color distribution, or the sari crumpled on the floor? Our results suggest that the number of immediately apparent colors in the fabric would then be small, and probably much smaller than 50,000!

Reference:


Park, J. (2018). A neural basis for the visual sense of number and its development: A steady-


**Perceiving the average blur in images**

**Introduction:**

One of the strategies used by the visual system to process vast amounts of visual input efficiently is summary statistics extraction (Alvarez, 2011; Whitney & Leib, 2018). Extracting summary statistics of a scene involves representing a set of stimulus features using global statistical parameters, like the mean (Haberman & Whitney, 2007) and variance (Haberman, Lee, & Whitney, 2015; Solomon, Morgan, & Chubb, 2011) of the distribution. This was first systematically studied by Ariely (2001), and subsequently a number of stimuli have been reported to represent an ensemble using the average stimulus value (Whitney & Leib, 2018 for review). The consistent evidence for ensemble coding across diverse stimulus attributes and sensory modalities, from low-level to highly abstract (Leib, Kosovicheva, & Whitney, 2016) and from vision to hearing (Piazza et al., 2013), suggests this coding plays a central role in sensory processing.

In this study we examined the characteristics of ensemble coding in the perception of image blur. Blur is an important property of spatial vision, and typically shows large variations within a scene. Apart from being an important cue for accommodation (Ciuffreda, 1998), blur is also important for making distance judgements in a scene (Held et al., 2010). In fact, the relationship between blur and the distance of objects from eye is so important, that it has its own internal consistencies (Sprague et al., 2016), and violations of these can lead to strong distortions in size and distance judgments (Held et al., 2010). These variations in blur primarily occur due to the
limited depth of focus of the eye’s optics. Due to limitations in the focusing power of the eye, objects at only one distance can form a clear image on the retina. Objects in front of and behind the object of fixation are thus always blurred, raising the question of what is the “average” blur an individual experiences in natural scenes. In conditions like dichoptic viewing, the sharper features of an image dominate perception (Kompaniez et al., 2013; Radhakrishnan et al., 2015).

If summary statistics for blur are based on the perceived image blur, will the average estimation then be affected by such biases? In the first part of the study, we examined if the average blur of a scene can be extracted, and if so, whether there are any biases in estimating the average.

The visual system can represent the average not only of homogeneous sets (e.g. circle size; (Ariely, 2001)), but can also can represent the average of heterogeneous sets (Haberman, Harp, & Whitney, 2009). Faces provide a good example for studying heterogeneous sets because they vary in a continuum between opposite sub categories (ex-sad/happy) centered on a relatively neutral face which has equal proportions of the opposing sub categories (Haberman et al., 2009). Neural blur varies in a similar continuum where the perceived blur can vary between over-sharp to blurred with a focused image as the center or neutral point (Fig 1). This neutral point may represent a perceptual “norm” in blur processing (Elliott, Georgeson, & Webster, 2011; Radhakrishnan et al., 2015; Sawides et al., 2011; Webster, Georgeson, & Webster, 2002) where any image is perceived as blurred or focused with respect to its perceptual distance from the norm. In many cases, the norm also holds a special neural status. For example, the best-focus judgements in blur perception are not affected by adapting to the norm. Does this norm (a
focused image in this case) hold a special perceptual status in ensemble coding as well? In the second part of the study, we tested the role of the norm in the ensemble perception of a set of heterogeneous (both blurred and sharp) and homogeneous (either blurred or sharpened) ensembles.

Figure 1 – Optical vs Neural blur. Optical blur only increases relative to the focus, while neural blur can both sharpen and blur an image relative to a point of best focus

In the third part of the study, we examined ensemble adaptation in blur. Adaptation is also an inherent property of the visual system, and this adaptation plays a significant role in shaping blur percepts (Webster & Marcos, 2016). For example, looking at a blurred image for just a few seconds can change the perception of a focused image (Webster et al., 2002). Changes in perceived blur can result from either short term (Webster et al., 2002) or long term adaptation (Parkosadze et al., 2013; Sabesan & Yoon, 2010; Vinas, Sawides, de Gracia, & Marcos, 2012). In the case of long term blur adaptation under natural viewing conditions, the visual system is exposed to different levels of blur spatially and temporally, and representing the average blur could potentially improve the efficiency of adaptation. Here, we studied if the aftereffects of adapting to an ensemble of blur levels varying in space and time was driven by the
sharper/more focused parts of the ensembles, or were driven by the average blur level of the ensemble.

**General methods:**

**Participants:**
A total of 15 observers participated in the study with different subsets of participants in each of the conditions. All participants were recruited from the UNR student pool and were naïve to the specific aims of the experiment. Participants had normal or correct to normal visual acuity, and gave informed consent following the protocols approved by the university’s IRB.

**Apparatus:**
Stimuli were presented on a CRT Sony trinitron monitor with a resolution of 800 X 600 pixels driven by nVIDIA graphics card in a dark room under binocular, free viewing conditions.

1. **Estimating the average blur**

The ability to extract the average blur level in a stimulus with varying blur levels has not been systematically studied in the past, and hence the primary aim of this experiment was to establish if participants are able to estimate the average blur. This was tested using a simple mean discrimination task. If participants can represent the average blur of an ensemble, it is expected that the estimated average will be close to the physical average of the ensemble.

1.1. Stimuli:
Stimuli for these experiments were simple edges. The edges were defined using a 2D cumulative Gaussian function (CGF), in which the standard deviation (σ) of the CGF determined the amount
of edge blur. These edges were further windowed in a Gaussian, and the standard deviation of
the Gaussian window remained constant for all image blur levels.

1.2 Ensembles:

Ensembles were composed of 36 elements, and within each ensemble there were 19
consecutive blur levels distributed across the 36 elements. Three ensembles with low, medium
and high average edge blur (edge blur being defined by the standard deviation of the CGF, \( \sigma =
10 \) (range 1-19), \( \sigma =18 \) (range 9-27), \( \sigma =27 \) (range 18-36)) were used in this experiment, and the
blur levels for all of the ensembles had constant variance. These ensemble were created
dynamically as the experiment progressed. The spatial location of the individual elements was
randomly jittered over an area of 70 pixels on each trial, but it was ensured that the elements
never overlapped on any trial.

1.3 Procedure:

Five observers participated in this experiment, and performed the task from a distance of 2.5
meters from the monitor. At that test distance, the ensembles subtended an angle of 5 degrees
at the fovea. The ensembles were presented for 1000 ms following which a single edge with a
random blur level was presented. Participants adjusted the test blur level until it appeared to be
the average blur of the ensemble using a two alternate forced choice (too blurred/too sharp
than the perceptual average) staircase procedure (Figure 2). Average blur of the ensemble was
calculated by taking the mean of the last 8 reversals of the staircase. Each ensemble type was
tested four times in a single experimental session, and the order of testing was randomized for
each participant.
1.4 Results:

The estimated average blur was close to the actual average blur of the ensemble, irrespective of the blur level (Figure 3). This was confirmed using a one sample test comparing the actual mean with the subjective setting for all three blur levels (low average blur, t(4) = 0.1575, p = 0.8825; medium average blur, t(4) = -1.3638, p = 0.2443; high average blur, t(4) = -0.529, p = 0.6248).

Figure 3 – Results for the average estimation experiment. Estimated average blur of the ensemble was very close to the actual average blur of the ensemble.
1.5 Discussion:

The results of this experiment suggest that the mean blur from an ensemble can be extracted accurately and that the subjective mean is not biased towards the sharper parts of the ensemble. Single edges were used in the first experiment to make the stimulus simple. Even with the apparent simplicity of the stimulus, the exposure time had to be increased to 1000 ms in order to get reliable subjective responses. Longer exposure times have been used in the past albeit not so commonly to study ensemble coding in higher stimulus features like faces (Florey, Dakin, & Mareschal, 2017) and biological motion (Sweeny, Haroz, & Whitney, 2013). The possibility that this could have increased dependence on serial representation cannot be excluded. However, in our experiment, though the single edge itself was “simple”, the ensemble was made of 36 such elements. Generally, factors such as increased variance (Haberman et al., 2015) and increased number of elements (Maule & Franklin, 2015) could make summary statistics extraction harder. The relatively longer exposure time could have at least in part compensated for the complexity of the ensemble.

2. **Ensemble coding in blur**

The result of the previous experiment suggests that the average blur can be reliably extracted from a set of blurred edges and that the estimate of this average is unbiased. We next tested heterogeneous ensembles composed of both blurred and sharpened edges filtered by different amounts relative to focus. As noted in the Introduction, in blur coding this focused image also represents the norm. In this experiment, we studied the ability to form summary
representations across the blurred to sharpened continuum, and the role of the focus norm in this ensemble coding. Notably, the visual system is exposed to different levels of blur more often than different levels of sharpness in natural viewing conditions. Hence, the ensemble coding strategies for these two images sets (blurred sets vs sharp sets) could be different from one another. Using a third set of ensembles we also examined if there are asymmetries between the representation of blurred and sharp ensembles.

2.1. Stimuli:

Inserting a Gaussian edge profile to sharpen the edge is problematic, and rapidly leads to luminance levels that fall outside the gamut of the display. As an alternative, we instead varied blur and sharpness in this experiment by filtering the amplitude spectrum of a natural image (Webster et al., 2002). The texture (256x256 pixels) shown in figure 4 was blurred or sharpened by varying the slope of the amplitude spectrum. The original amplitude spectrum of the texture was rescaled in the frequency domain at each spatial frequency by an exponent of frequency in such a way that the resultant slope varied between -1.25 and +1.25 relative to the nominal unfiltered slope of -1. The images were again windowed by a 2D Gaussian to smooth the border
of the images. As in the previous case, the standard deviation of the Gaussian window remained constant for all the images.

2.2. Ensembles:

As in the previous experiment, ensembles were created dynamically during the experimental session. Each ensemble was composed of 12 elements chosen from 4 blur levels, and was presented at a constant eccentricity from a central fixation target, in order to prevent biases in perceived blur between foveal and peripheral viewing. The blur levels in the ensemble varied across the different trials and conditions. A total of 12 ensembles were tested across three sessions. The individual elements in the ensembles were made of all blurred, all sharp, or a combination of both blurred and sharp images.

The original image textures were cropped to 160X160 pixels and the window (standard deviation) of the Gaussian was set at 32. These specifications for the individual elements ensured there was no overlap between the elements while maintaining enough width of the Gaussian to appreciate the blur/sharpness of the individual elements.

2.3. Procedure:

Unlike the previous experiment where the ability to estimate the average was studied, in this
experiment the role of the focused image in ensemble coding was studied. Hence, the task was switched to a member identification instead of the mean estimation task used in the previous experiment. Nine participants were presented with an ensemble for 750 ms followed by a test stimulus for 500 ms, and the participants reported if the test blur level was a part of the ensemble (Figure 5). Nine test blur levels with their relative slope ranging in 0.25 steps from -1 to +1 (relative to the unfiltered image spectrum) were tested for each ensemble, and every blur level was tested 20 times for each ensemble. The nine test blur levels included the 4 blur levels that were shown in the ensemble plus an additional 5 intervening levels that were not present in the ensemble. It was ensured that the nine test blur levels were discriminable at the retinal periphery where it was presented.

Figure 5 – Member identification task for experiment 2

2.4. Results:

2.4.1. Heterogeneous ensembles with neutral, sharp, or blurred averages:

The settings in figure 6a were for ensembles that both had a focused image as their average. For both, the false alarm for the focused image was at least as high as the true response for the sharp and blurred extremes of the ensembles. A one-way (3 blur levels) repeated measures analysis of variance (ANOVA) for the first ensemble E1 (blur levels: -0.75, -0.25, 0.25, 0.75) confirmed that the false alarm (FA) for the focused image was significantly different (F(2,16) =
5.689, p = 0.014) from the true responses for the sharp (SR) or the blurred (BR) extreme levels of the ensemble. Further posthoc pairwise comparisons showed the false alarms for the focused image (E1 mean FA = 82.22 ± 4.65) was significantly higher (p = 0.006) than the responses for the sharp end of the ensemble (E1 mean SR = 56.11±5.386), but was not significantly different (p = 0.242) from the responses for the blurred end of the ensemble (E1 mean BR = 68.89 ± 7.628).

Similarly, for the ensemble E2 (blur levels: -1,-0.5,0.5,1), in spite of the wider range of blur levels, again a one-way (3 blur levels) repeated measures ANOVA showed that the false alarm for the focused image (67.78 ± 7.41) was not different (F(2,16) = 1.069, p = 0.367) from the true response rates for the blurred (62.78 ± 5.59) or sharp (53.33 ± 7.45) extremes of the ensemble.

(a) (b)

Figure 6 – Ensemble blur coding in heterogeneous ensembles. (a) - Ensembles E1 and E2 have the focused image as the average but differ in the range of slopes composing the ensembles. (b) – Ensembles E3 and E4 have a sharp and blurred average respectively. The average blur level was never a part of any of these ensembles, but was misreported as a member the ensemble a high proportion of times.

The two ensembles E3 (blur levels: -1, -0.5, 0, 0.5) and E4 (blur levels: -0.5, 0, 0.5, 1) shown in figure 6b are heavily biased towards either the sharper or blurrier elements, and consistent with the ensemble bias, have a sharp or blurred average physical blur. One way repeated measures ANOVA each of the ensembles (E3 & E4) comparing the FA for the ensemble mean and the two extreme blur levels of the ensemble showed significance (E3, F(2,16) = 6.350, p = 0.009; E4,
\( F(2, 16) = 5.907, p = 0.012 \). Nevertheless, posthoc pairwise comparisons between the false alarm (FA) for the average blur (E3 mean FA = 86.67 ± 4.48; E4 mean FA = 83.33 ± 4.86) and the true response rate for the sharp and blurred ends of the ensemble (E3 sharp mean = 53.33 ± 9.61; E3 blurred end mean = 72.22 ± 7.459; E4 sharp end mean = 63.889+/−8.53; E4 blurred end mean = 66.11+/−5.76) with Bonferroni correction showed that these differences did not reach significance. (E3 sharp end vs FA for ensemble mean, \( p = 0.065 \); E3 blurred end vs FA for ensemble mean = 0.288; E4 sharp end vs FA for ensemble mean = 0.061; E4 blurred end vs FA for ensemble mean = 1.00). Therefore, the mean blur of a heterogeneous ensemble was likely to be perceived as part of the ensemble, even when this mean corresponded to the neutral focus.

2.4.2. Categorical effects at the sharp/blurred boundary:

To study the nature of this drop in FA, ensembles E5:E12 (figure 7) were used. The starting blur level for ensembles E5 through E8 (ensembles with a sharp bias) were offset by 0.25 blur step with E5 starting at -1.25, and the 4 blur levels in the ensemble progressing uniformly in 0.25 steps. Ensembles E9 through E12 (ensembles with a blurred bias) were similar in composition to E5:E8, but started at 1.25. While some ensembles in this condition (E7, E8, E11, and E12) spanned the sharp/blurred boundary, others (E5, E6, E9, and E10) did not.

The false alarms (FA) for two consecutive blur levels outside the limits of the ensemble, and along the longer tail of the response curve was used to study if the FA at focus was different from that of other blur levels.
Categorical effects at focus – sharp ensembles: One way ANOVA for the first blur level outside the ensemble along the longer tail showed that the FA was not different between the four ensembles ($F(3,32) = 0.549, p = 0.653$). A similar analysis for the second blur level also showed no difference between the rate of FA for the four ensembles $F(3,32) = 0.42, p = 0.74)$. Note that one of the comparison points for both the analyses included the focus, yet there was no difference between the ensembles which included FA for focus and those that did not (figure 7 top).

Categorical effects at focus – blurred ensembles: One way ANOVA for the first blur level outside the blurred ensembles along the longer tail showed that the FA at focus was not different from the FA at other blur levels ($F(3,32) = 0.563, p = 0.643$). A similar analysis for the second blur level also showed a similar pattern of result $F(3,32) = 1.117, p = 0.357$ (figure 7 bottom).
Figure 7 – Tests for categorical effects and blur/sharp asymmetry in ensemble coding of blur.

The top four ensembles have a sharp bias and the bottom four have a blurred bias. The ensembles on the right span the blurred/sharp boundary and the ones on the left do not pass
the boundary. The pattern of drop in false alarms (shown by solid subparts in the plots) was compared among the blurred ensembles, sharp ensembles and between the blurred and sharp ensembles. The drop in false alarms was not significantly different between any of the groups. Collectively, these results suggest that the drop in false alarm was gradual for blur levels just outside the ensemble, and that this drop was not significantly different at the categorical boundary around the level of physical focus.

2.4.3. Asymmetries between ensemble coding of blurred and sharpened image sets:

In this section, the blurred and sharp ensembles were compared to test for asymmetries in their encoding.

The FA for the first blur level outside the ensembles along the longer tail was compared between the blurred and sharp ensembles using a two way ANOVA (2 blur vs sharp image set X 4 Ensembles E5:E12). This analysis showed that there was no difference between the FA for blurred and sharp ensembles $F(1,64) = 2.246$, $p = 0.139$. A similar analysis for the second blur level also showed that the there was no difference between the blurred and sharp ensembles $F(1,64) = 1.349$, $p = 0.250$.

2.5. Discussion:

Similar to the first experiment, we used a slightly unconventional exposure time in this experiment, but again the ensemble was presented in the periphery, and each ensemble had 12 elements. This could have made it harder for participants to encode every blur level individually,
and in fact participants consistently rated the task as being difficult. Consistent with this speculation, the pattern of results for all the ensembles (E1:E12) is typical of ensemble representation reported by others (Ariely, 2001; Maule et al 2014), where the mean of the ensemble has the highest false alarms, even when the mean was not actually a part of the ensemble. These results also suggest that despite the slightly longer exposure times used in our experiments, participants indeed depended on ensemble coding mechanisms rather than other perceptual cues like serial viewing.

Encoding heterogeneous distributions increases the functionality of ensemble coding, and one of the commonly studied stimuli in this context is faces. One of the representational similarities between faces and blur is that both can have a heterogeneous distribution around a neutral point. The evidence from this experiment suggests that a heterogeneous distribution of blur behaves the same way as faces where the neutral stimulus is perceived as the average of the heterogeneous set. Even when the average is not the neutral stimulus, a heterogeneous set with a slight sharp or blurred bias was represented using the average blur level, and this average blur level was the physical average of the blur levels in the ensemble.

Previous experiments in color have shown that there are no categorical biases in average hue estimation (Maule et al 2014). Although blur perception does not involve as many categories as color, the whole continuum from sharp to blur can be divided into at least three categories – blurred, focused and sharp. With the focused category being more relevant to spatial vision than others, it was expected that the absence of a focused image from an ensemble of blurred or sharp images would be more noticeable for the participants. Our results, however, show that
the drop in false alarm around the focused image was comparable to the drop in false alarms for other blur levels, suggesting that there was no categorical bias in ensemble blur coding. Finally, the ensemble coding plots of blurred and sharp ensembles almost mirror each other, and these effects suggest that, irrespective of the component blur levels of the ensemble, the mechanisms operate in a similar and robust manner.

3. Adaptation and average blur:

Perceived blur is affected by previous blur that the neural system has been exposed to, and a compact perceptual code for environmental blur can be achieved by averaging the blur within and between visual scenes. There is at least some evidence to think that the visual system can adapt to the mean of a set rather than the individual elements of a set (Corbett et al 2012), and in this experiment, we studied the adaptation aftereffects of an ensemble varying in blur across space and time.

3.1. Stimuli:

Stimuli were filtered image of the rock texture used in the previous experiment, only this time they were not windowed in a Gaussian, and instead were juxtaposed to pave the length and width of the screen.

3.2. Procedure:

All 3 participants performed the task from a distance of 150 cm, and adjusted a test image (composed of the array of texture images all with the same filtering) until it appeared focused
using a two-alternative forced choice (too blurred/too sharp) staircase procedure. Mean settings were based on the average of the last 7 reversals. These settings were repeated after the participants were adapted for 3 minutes to two types of adapting stimuli. In the first case, they adapted to a sequence of textures chosen from five different levels with a constant variance. The five blur levels were chosen such that their average was either sharp, focused, had low blur or high blur. The spatial and temporal distribution of each of the blur level was changed every 0.3 seconds. In the second case, they adapted to a texture with a single blur level which was the average of the ensembles in the previous case. In both cases, a topup adapt of 6 seconds was provided before every stimulus presentation. Participants repeated this three times in three separate sessions (figure 8).

![Figure 8 – Procedure for ensemble blur adaptation experiment](image)

3.3. Result:
Figure 9 shows that the shift in best focus judgments from baseline after adaptation. The size of the shift to the mean value and to the sequence of blur levels is very similar. This was confirmed using a two-way (Adapt Type - sequence vs average value X 4 Adapt Level) ANOVA which showed no main effect of the adapt type ($F(1, 2) = 0.000, p = 0.994$) and no interaction between the adapt type and adapt level ($F(3, 6) = 0.129, p = 0.939$). However, there was a main effect of the adapt level ($F(3, 6) = 183.66, p < 0.001$). Posthoc pairwise comparisons with Bonferroni corrections showed that the shift in best focus was significantly different between the sharp adapt and blur adapt conditions (sharp vs low blur, $p = 0.002$; sharp vs high blur, $p = 0.015$), and between focus adapt and high blur adapt ($p = 0.034$).

**Figure 9 – Similar pattern of aftereffects after adaptation to a sequence of blur levels**

(ensemble adapt) and to the average value of the sequence (average adapt).

3.4. Discussion:
This result is in line with previous reports of average adaptation (Corbett et al., 2012) where it has been shown that average values are adaptable quantities and have aftereffects that can significantly affect perception. However, we note that the same effects could be explained by the average effect of adaptation to each of the individual adapting levels. Either way, the results again suggest that the average controlling the adaptation remains very close to the physical average of the ensemble, and thus is not biased by the more focused or sharper elements in the ensemble.

General Discussion:

One of the consistent results from all of the above experiments is that the visual system’s is able to extract the mean blur from an ensemble of blur levels with high fidelity. In the first experiment, this was observed in spite of the complexity of the ensemble. In the second experiment this was observed by the high false alarms for the average values of the ensemble compared to the other blur levels in the ensemble. In the third experiment, this was observed by the similar pattern of aftereffects between adaptation to the average blur and the sequence of blur levels.

At least at closer viewing distances, vision has been reported to be more dependent on higher spatial frequencies for stimuli like faces (Shahangian & Oruc, 2014). This “sharp bias” can be seen in other conditions like dichoptic viewing (Kompaniez et al., 2013; Radhakrishnan et al., 2015) and while looking at hybrid images (Schyns & Oliva, 1999). When such sharp or even focused features are present in parts of an image, those features are more salient than the
blurred ones. For spatial variations in blur, the foveal image usually appears focused, and it is also perceptually more meaningful than other parts the visual field. Hence, it would not have been surprising if the estimated average had been biased towards the focused/sharp images. In experiments 1, in which stimuli were presented in the fovea and periphery, there was no sharp or focused bias in estimating the average blur.

Vision is generally thought to be less sensitive to variations in stimuli that mimic the variations in the natural environment. For example, the sensitivities for white settings along the Blue-Yellow axis of the color space are lower than the sensitivities for other axes (Bosten, Beer, & MacLeod, 2015), potentially because the natural environment has large color variations along the BY axis (Mcdermott, Mulligan, & Webster, 2010). However, there have been no reports examining the efficiency of ensemble coding for naturally occurring stimuli vs other stimuli. In blur perception, variations in blurriness are more common than variations in sharpness during natural viewing conditions. Nevertheless, there were no differences between the coding strategies for blurred or sharpened sets, either during ensemble representation or while adapting to the average blur/sharpness. These results show that ensemble coding mechanisms may be robust and operate in similar ways for familiar and novel stimulus attributes. It would be interesting to further test this hypothesis with other dimensions such as familiar vs. unfamiliar faces.

Finally, we found little evidence for categorical effects in ensemble coding for blur. This is in spite of the fact that the focus point reflects a clear norm and category boundary in the blur/sharp continuum. In this regard our results parallel findings we report for lightness variations (see chapter “Ensemble coding of color contrasts”) in which there is little evidence for
a categorical effect of the gray boundary separating increments and decrements. Like lightness, blur/sharpness appears to vary as a uniform continuum.

Conclusions:

We conclude that the average blur can be estimated from a scene reliably, and there are no biases towards sharper or more focused elements of the ensemble while estimating this average. This averaging is a robust process which is largely unaffected by the distribution of blur levels in the ensemble. Therefore, in our tasks, the blur variations from blurred to sharp seem to behave as a seamless continuum in which the focus level was not unique in the way it is represented.

References:


https://doi.org/10.1167/11.2.7


https://doi.org/10.1016/j.cub.2007.06.039

https://doi.org/10.1145/1731047.1731057

Kompaniez, E., Sawides, L., Marcos, S., & Webster, M. A. (2013). Adaptation to interocular


Sabesan, R., & Yoon, G. (2010). Neural Compensation for Long-term Asymmetric Optical Blur to


https://doi.org/10.1371/journal.pone.0046361


Summary

A typical visual scene is both colorful, and has wide blur variations in it. Nevertheless, these variations are qualitatively different from one another. In color, a heterogeneous set is more representative of natural color variations of a scene while in blur, a homogeneous set with varying levels of blurriness is more natural. In this sense, homogeneous and heterogeneous sets can have different applications in color and blur, and have to be compared with caution. Such a comparison is further restricted by other features of color and blur variations. For example, a unit step size in the color space and blur space cannot be compared directly.

However, there are other representational similarities between color and blur which have been explained earlier, and which can help us understand ensemble coding mechanisms in general. In the following sections, ensemble coding in blur and color are discussed in parallel. Here I summarize some of the main theoretical implications and unresolved questions arising from these studies.

Previous studies have reported that categorical effects are weak in color ensemble coding when the individual colors in the ensemble vary in hue (Maule, Witzel, & Franklin, 2014). In this work, both weak and strong categorical boundaries were tested. The categorical boundary of blur between blurred and over-sharp is relatively weaker than the strong categorical boundary of gray between complementary colors. This work has shown that for the attributes of color and blur, irrespective of the type of ensemble, or the strength of the categorical boundary, perceptual categories in color (eg – color labels based on appearance) and blur (eg –
blurred/sharp) do not seem to affect ensemble coding. Hence, the weak categorical effect observed for hue ensembles can be extended to saturation, lightness and blur as well.

The other similarity between color and blur in this report is the status of norm, and its interaction with average estimation. For both color and blur, the norm seems to play very little role in ensemble coding. In ensemble perception, the average of an ensemble is defined by the individual items. Yet, while computing the global nature of the ensemble, information on individual items is often lost. Reports have suggested that individual items are either processed and discarded, or the representation of individual items itself is noisy (Alvarez, 2011). Either way, once the average is estimated, the relevance of individual items to the ensemble, or the relevance of the average to the individual items becomes less important. The relationship between the perceptual norm and individual items on the other hand, is more mutual. In norm based coding, individual items are processed relative to its distance from the norm, and the recalibration of norm is based on long-term exposure to the properties of individual items. This difference between ensemble perception and norm based coding could delineate the norm and the average because they serve different purposes in visual coding. While the norm contributes to the processing of individual items, the average provides a quick compression of the scene. Statistics of this compression could contribute to recalibration of the norm, but the norm itself does not seem to affect estimation of the average. The lack of a special role for norm in ensemble perception, along with the lack of categorical effects for color and blur show that summary statistics extraction is an abstract function serving a highly specific role in perception, relatively immune to other aspects of visual coding.
A number of reports on color averaging have consistently shown that ensemble perception of color is less reliable for sets of colors showing large variance in hue (Maule & Franklin, 2015; Webster et al., 2014). Overall, this report suggests that other dimensions of color, and even blur operate in a similar manner. Apart from this commonality, there are certain distinctive differences in ensemble coding of color and blur, and even differences between the different aspects of color. Firstly, ensemble variance seems to affect hue and saturation ensembles differently from lightness and blur ensembles. The effect of color variance on ensemble perception was markedly similar between sets varying in hues and those varying in saturation. For saturation ensembles, the variance in the ensemble seems to be an important feature affecting response strategy. While ensemble perception for sets varying within a hue are more likely to depend on the explicit representation of the mean, ensemble perception for sets with colors from different hue categories appears to be less explicit. For ensembles varying in complementary hues, ensemble perception seems to depend on indirect inferences from other cues like relative saturation of the colors. This implicit inference for saturation ensembles is similar to previous report examining heterogeneous hue ensembles by Webster (2014). The representation of heterogeneous hue and saturation ensembles is in contrast to the representation of heterogeneous lightness and blur ensembles. Results from this work suggests that variations in lightness and blur appear to be in a continuum going from dark to bright and from blurred to over-sharp (or vice versa in each case) respectively. This continuous variation, probably provides a cue of “sameness” which serves to derive a more explicit measure of the mean value in the ensemble. It should be noted that for lightness and blur ensembles, the
degree of heterogeneity does affect subjective performance making responses less reliable. But as in the case of hue and saturation ensemble, the degree of heterogeneity in lightness and blur ensembles does not affect performance so much that the response strategy itself changes from an explicit estimation for homogeneous ensembles to implicit inference for heterogeneous ensembles.

Beyond mean, the visual system has been shown to be sensitive to certain other features of the ensemble such as the variance, and the nature of distribution (Chetverikov, Campana, & Kristjánsson, 2017). Here the sensitivity to the number of individual color levels in the ensemble was tested, and this appears to be poor. Again, there were differences between the explicit measurement of this quantity and implicit measurement suggesting that the subsampling of the number of color levels in the ensemble is a task requiring additional features like attention.

**Recommendations**

Although a number of studies have examined various aspects of ensemble coding, the actual neural mechanisms behind these processes still remain unknown. Reports like the current one identifying similarities and differences for various stimulus attributes in ensemble coding can progressively contribute towards understanding the dynamics of ensemble coding. A recent report has suggested that ensemble coding increases between 6-18 of age, and that at least for faces, coding of individual faces and coding of face ensembles can be dissociated from each other (Rhodes et al., 2017). Ensemble coding mechanisms can be tested in such young groups but with a history of sensory deprivation to examine if these mechanisms depend on a critical
age for development. Another recent report (Maule et al., 2017) has suggested that certain aspects of ensemble coding are affected for individuals with autism who are at a disadvantage of processing global information. These individuals also have been reported to have atypical sensory processing (Baum, Stevenson, & Wallace, 2015; Simmons et al., 2009). If in fact, the development of ensemble coding requires a critical age, what would be the real life disadvantage of not having ensemble coding? The current report, along with previous ones (eg - Corbett et al., 2012) have shown that average values from ensemble affect adaptation aftereffects. If individuals with poor ability to estimate the average are at a disadvantage, can plasticity mechanisms be exploited to train for efficient ensemble coding in sensory deprivations? Although these are specific questions, the general theme is that there are a number of issues in ensemble coding which could have far reaching implications, even as far as sensory plasticity.

Conclusions

This work has shown that ensemble coding mechanisms between color and blur are similar for homogeneous sets, but dissimilar for heterogeneous sets. Specific features of the ensemble like nature of color or blur distribution, its relationship to natural environmental variations, or even important aspects of perception like norm do not affect ensemble perception. Finally, the ability to discriminate between ensembles is poor, when the difference in number of colors between them is only one as measured through behavioral and EEG methods.
References:


