Feature-specificity in visual statistical summary processing

Harun Yörük¹ · Aysecan Boduroglu²

Published online: 3 January 2020
© The Psychonomic Society, Inc. 2020

Abstract

Visual statistical summary processing enables people to extract the average feature of a set of items rapidly and accurately. Previous studies have demonstrated independent mechanisms for summarizing low (e.g. color, orientation) and high-level (facial identity, emotion) visual information. However, no study to date has conclusively determined whether there are feature-specific summarization mechanisms for low-level features or whether there are low-level, feature agnostic summarization mechanisms. To address this issue, we asked participants to report either the average orientation or the average size from a set of lines where both features varied. Participants completed these tasks either in single-task or mixed-task conditions; in the latter, successful performance required extraction of both summaries concurrently. If there were feature-specific summarization mechanisms that could operate in parallel, then errors in mean size and mean orientation tasks should be independent, in both single and mixed task conditions. On the other hand, a central domain-general mechanism for low-level summarization would imply a correlation between errors for both features and greater error in the mixed than single task trials. In Experiment 1, we found that there was no correlation between the mean size and mean orientation errors and performance was similar across single and mixed-task conditions, suggesting that there may be independent summarization mechanisms for size and orientation features. To further test the feature-specificity account, in Experiment 2 and 3 (with mask), we manipulated the display duration to determine whether there were any differences in the summarization of earlier (orientation) vs. later (size) features. While these experiments replicated the pattern of results observed in Experiment 1, at shorter display durations, no differences emerged across features. We argue that our data is consistent with independent, multi-level feature-specific statistical summary mechanisms for low-level visual features.

Keywords visual perception · attention · Ensemble perception · Feature specific mechanisms

People can extract average information from object sets by visual statistical summary processing (also called ensemble perception) (for a review, see Alvarez, 2011; Whitney & Yamanashi Leib, 2018). Complementing the foveal high-resolution representation, ensemble perception allows viewers to experience the world in a holistic and detail-rich manner (Cohen, Dennett & Kanwisher, 2016). Previous research has shown that viewers can efficiently summarize various low-level features such as orientation (Dakin, 2001; Parkes, Lund, Angelucci, Solomon & Morgan, 2001), brightness (Bauer, 2009), size (Ariely, 2001; Yıldırım, Ögreden &

Boduroglu, 2018), color (Maule, Witzel & Franklin, 2014), position (Alvarez & Oliva, 2008), as well as higher-level features like facial identity (de Fockert & Wolfenstein, 2009) and facial emotion (Haberma & Whitney, 2007). Viewers do not only extract summaries of static visual displays, they can also summarize sequentially presented visual (Albrecht & Scholl, 2010; Hubert-Wallander & Boynton, 2015) and auditory information (Albrecht, Scholl, & Chun, 2012; Piazza, Sweeny, Wessel, Silver & Whitney, 2013). In addition, viewers can represent variance (Morgan, Chubb & Solomon, 2008; Khayat & Hochstein, 2018; Semizer & Boduroglu, 2013) and numerosity information (Utochkin & Vostrikov, 2017), as statistical summary information. Despite demonstrations of statistical summarizing across various domains, interestingly, there is no consensus on the mechanisms underlying this ability (e.g. Dubé & Sekuler, 2015; Whitney & Yamanashi Leib, 2018). While Haberman and colleagues suggested separate summarization mechanisms for higher level (e.g. facial identity) vs. lower level (e.g. size) features, their findings were

✉ Aysecan Boduroglu
aysecan.boduroglu@boun.edu.tr

¹ University of Houston, Texas, Houston, TX, USA

² Department of Psychology, Bogazici University, 34340 Bebek Istanbul, Turkey

inconclusive regarding the mechanisms supporting summarization of lower level features (Haberman, Brady & Alvarez, 2015). Specifically these findings ruled out a central domain-general summarization mechanism; nevertheless, they did not directly address the possibility of multiple, feature-specific summarization mechanisms. Actually, three separate findings, discussed in detail below, are suggestive of multiple feature specific summarization mechanisms. One, there are individual differences data that emphasize the independence of errors for the visual and spatial summarization tasks (Uner, Mutlutürk & Boduroglu, 2014). Two, there are findings showing that sequentially presented items are weighed differently when viewers are trying to estimate mean orientation versus size information (Hubert-Wallander & Boynton, 2015). Finally, there is evidence showing that there are no costs to summarizing concurrently, separate feature distributions of spatially segregated sets (e.g. Attarha & Moore, 2015).

Across three experiments, we specifically investigated whether there are multiple, domain-specific feature-based perceptual summarization mechanisms for low-level features. To do this, we took an experimental approach complemented by correlational analyses. Past research has shown that complementing behavioral experiments with an individual differences/correlational approach can be useful in identifying dissociations between constructs or mechanisms (e.g. Awh, Barton & Vogel, 2007; for reviews see Vogel & Awh, 2008; Wilmer 2008; Tulver, 2019). In all three experiments, we asked participants to extract the mean length (size) and/or mean orientation of a set of lines and report either one of these features. We compared error across blocks in which we manipulated whether participants had to extract the summary of one feature (single-task block) or of both features (mixed-task block); we also computed the correlations between mean length and mean orientation errors. As we explain in detail below, we predicted that if there are independently functioning domain-specific summarization mechanisms for low-level features, then they should be able to work in parallel, resulting in efficient concurrent summarization of different features and independence of errors across feature domains.

Domain-general vs. domain-specific mechanisms

Haberman, Brady and Alvarez (2015) proposed two possible mechanisms for how viewers extract visual summary information. One possibility is that there is a central, domain-general summary processor in the visual system that is responsible for averaging all types of information. The alternative possibility is that there are multi-level, domain-specific mechanisms for summarizing different visual properties, possibly at

different cortical levels (Whitney, Haberman & Sweeny, 2014). While the former view would require a significant correlation between the errors in summarizing different types of information, the latter mechanism would allow for independence of errors.

To determine whether there is a domain-general or domain-specific visual summarizing mechanism, Haberman, Brady and Alvarez (2015) compared summarizing performance using an individual differences approach across various visual properties. In their first experiment, they found that there was no correlation between orientation averaging and facial averaging tasks suggesting that statistical summary representations of high-level and low-level features may be governed by separate mechanisms. Subsequent experiments compared performance across various high-level (e.g. facial identity and facial emotion) and low-level (e.g. orientation and color) visual features. When viewers completed either two low-level or two high-level tasks, then there was a significant correlation in performance on the two tasks. On the other hand, when one task was from a low-level and the other one was from a high-level domain, there was no correlation between the two tasks. Thus, they concluded that statistical summary processing is not a uniform process, and that there are at least two separate and independent domain-specific summarizing mechanisms, one for low-level and another one for high-level visual information, specifically.

These findings were in line with claims made by Whitney, Haberman and Sweeny (2014) who argued that there might be multilevel processing mechanisms in the ventral and dorsal pathways for summarizing information rather than a single cortical area responsible for summarizing all types of visual properties. They argued that orientation, color, and brightness features might be summarized in the early cortical stages, while motion, position information, size and shape features might be summarized by separate mechanisms in the dorsal and ventral pathways, respectively. In addition, they argued that complex higher-level face and biological motion summaries might be processed later, after the convergence of ventral and dorsal pathways. Hubert-Wallander & Boynton (2015) provided some empirical support for these claims based on their comparison of how observers summarized sequentially presented sets. In separate experiments, observers reported the mean location or mean size of dots, the average motion direction of moving dot fields, or the mean facial expression. The differences in the temporal summarization profiles (i.e. the differential weighing of first as opposed to last items in the sequence in the reported mean summary) lead them to conclude that there may be distinct mechanisms involved in the summarization of location-based versus non-location based features.

Recent findings from our lab also supports the possibility that there may be separate mechanisms to summarize spatial (location-based) and visual information (Uner, Mutlutürk &

Boduroglu, 2014)¹. In two separate tasks, we asked participants to report the mean position of a set of colored squares (i.e. centroid) and the mean length of a set of randomly oriented lines. In both tasks, participants adjusted a probe to determine their response (i.e. drag to the centroid or extend the probe's length), allowing us to compute error in a continuous scale. We found that there was no relationship between error on the spatial centroid task and the visual length task ($r=.16$, $p=.16$).

In sum, the empirical research to date, has provided some evidence that there may be separate mechanisms responsible for the summarization of lower and higher-level complex visual features; some preliminary evidence suggests that this separation might extend to visual and spatial domains. However, except for the Hubert-Wallander & Boynton (2015) study, in the other studies, viewers summarized certain dimensions across separate tasks/blocks (e.g. Haberman et al., 2015; Uner et al., 2014). Then, researchers computed the correlation between the errors for different summarization tasks. For instance, Haberman et al. (2015), presented participants with four features/items in each display and asked them to summarize either higher level (e.g. facial identity and emotional expression), lower level (e.g. Gabor orientation and color of dots) or mixed (e.g. facial identity and Gabor orientation) feature pairs. While they found clear evidence for independence between higher level and lower level errors in summarization, they also found moderate significant correlations (ranging between .54–.73) between errors when tasks required the summarizing of two lower level features (for details see Table 1 in Haberman et al., 2015). It was not clear whether the correlation between lower level features emerged due to there being a low-level, domain general pooling mechanism or was due to an artifact of shared perceptual noise.

To address this issue, one possibility may be to utilize designs in which participants would need to summarize concurrently two low-level features from the same display as opposed to designs where summarization tasks are blocked by feature. Specifically, participants could be notified about the to-be-reported dimension after the offset of the study display via a post-cue. If there was a significant correlation across errors for different features under such a design, this could mean a number of possible things. A negative correlation would be suggestive of a trade-off between features, with summarization of one of the features possibly being prioritized over the other one; a positive correlation would suggest that the summarization relies on or is impacted by some

“shared” resource/noise. On the other hand, if under such a design, the errors were found to be independent (i.e. not correlated), this would suggest that the two summarization mechanisms are likely to be operating in parallel. Recent evidence demonstrating there are no costs to summarizing concurrently, *separate* feature dimensions of spatially segregated sets, suggest that there may indeed be feature-specific summarization mechanisms (e.g. for size and orientation: Attarha & Moore, 2015 for color and numerosity: Poltoratski & Xu, 2013; for size and numerosity: Utochkin & Vostrikov, 2017, but also see Emmanouil & Treisman, 2008).

In the present study, we specifically took the approach suggested above. To further test the domain-specificity argument, we asked viewers to summarize sets that consisted of heterogeneously sized lines presented at different orientations; depending on the condition, viewers either summarized length or orientation information. We chose these two dimensions for two reasons. One, there is evidence to suggest that both size and orientation may be summarized by mid-level texture processing mechanisms (e.g. Cain & Cain, 2018; Parkes et al., 2001); any evidence of independence of errors for the two summarization tasks would be a strong test for the domain-specific hypothesis. Two, using lines and asking about either mean size or mean orientation allowed us to use a single set and to equate physical properties of the presented stimuli for both summarization tasks and eliminate the need to present separate sets for each condition in a spatially segregated or a sequential fashion (e.g. Attarha & Moore, 2015). In this regard, the approach taken in the present study is different from earlier research because the to-be-summarized features belonged to the same items in a given set and participants reported the summary of either one of the features via a post-cue (mixed-task blocks). Successful performance on these mixed-task trials would require viewers to attend concurrently to both feature dimensions. If there is a domain-general summarization mechanism, then there may be costs to concurrent summarization, and consequently size and orientation errors may be correlated. On the other hand, if there are feature-specific summarization mechanisms, then summarization errors for these two low-level features should be independent.

In this study, in addition to the mixed task trials, we also included single-trial blocks, to investigate further the possible mechanisms of summarization. In these blocks, participants always reported either the mean size or the mean orientation of the viewed line set. Inclusion of the single-trial blocks enabled us to compare the error between single and mixed task conditions to test the domain specificity account. If there is a domain-general mechanism, then greater interference during the concurrent summarization in the mixed-task trials compared to the single-task trials is likely, possibly resulting in larger error in mixed as opposed to single task conditions. On the other hand, if there are independent mechanisms of

¹ As part of another project investigating the effects of action video game experience on ensemble perception, we compared action video-game players ($n = 23$), strategy video-game players ($n = 24$) and non-gamers ($n = 22$) on centroid (mean position) and mean size estimation tasks. Only for the AVGP group, there was a small yet significant relationship between the two domains, suggesting that these summarizing abilities may be benefitting from video-game related enhancements in visual processing skills ($r = .42$, $p = .05$).

Table 1. Bayesian ANOVA for size averaging errors

Model Comparison					
Models	P(M)	P(M data)	BF _M	BF ₀₁	error %
Null model (incl. subject)	0.200	0.752	12.128	1.000	
Duration	0.200	0.084	0.366	8.974	0.809
SingleMixed	0.200	0.145	0.678	5.188	0.903
Duration + SingleMixed	0.200	0.016	0.065	47.040	1.455
Duration + SingleMixed + Duration * SingleMixed	0.200	0.003	0.013	227.555	2.873

Note. All models include subject.

summarization for different low-level features, errors may be similar across single and mixed task conditions for each feature.

Experiment 1

Method

Participants We determined the necessary sample size for this and subsequent experiments by reviewing the sample sizes of recent experimental studies investigating the attentional modulation of ensemble perception (e.g. Attarha & Moore, 2015; Poltaratski & Xu, 2013). These studies had relatively small samples (approximating 10 participants per experiments), with similar number of trials to our experiments. We also reviewed recent studies that complement their behavioral findings with correlational analyses (e.g. Awh, Barton & Vogel, 2007; Scolari, Vogel & Awh, 2008; Vogel & Machizawa, 2004); in these studies, sample sizes were 20, 11, and 12, respectively. Based on these observations, in the first two experiments we collected data from 25 and 26 participants; in the third experiment, we used a slightly larger sample (40). Our post-hoc power analyses with GPower, revealed only a minimal increase when the sample size was increased to 100 (from .12 to .26). In addition, given that we were interested in showing support for a theoretically meaningful null hypothesis, we wanted to make sure that our experiments were not underpowered and our findings were not due to a Type II error. Therefore, we chose to also run Bayesian analyses to determine the support for the null hypotheses.

Twenty-five Bogazici University undergraduate students participated in the experiment, in return for course credit. We excluded data from 4 participants. Two of them had size and orientation averaging errors that were more than 3 standard deviations away from the group mean, another one because s/he was color blind and a final participant because s/he randomly responded throughout the experiment. Therefore, the following analyses were conducted on the data from 21 participants.

Materials and stimuli We programmed the experiment in E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA) and we ran it on a 17" monitor with screen resolution set to 1024x768 pixels (32x24 cm). Participants were sitting approximately 57 centimeters away from the computer screen. From that viewing distance, 1 cm was equal to 32 pixels and 1° visual angle.

Each trial (see Fig. 1) began with a green fixation cross, presented for 1500ms, which turned red for 500ms to indicate the beginning of a trial. We instructed the participants to fixate on the cross when it turned red. The study display consisted of 12 uniquely sized, randomly oriented lines and was presented for 200ms. Then, we presented the response screen. Participants had to adjust the length of the response probe for size averaging trials or rotate the orientation of the response probe for orientation averaging trials by using the left button of the mouse. When they finalized their response, they completed the trial by clicking to the right button of the mouse. In single-task blocks, we identified the relevant feature at the beginning of the block in the task instructions. In mixed-trial blocks, with the onset of the response probe, participants heard an auditory verbal cue from their headphones indicating the feature dimension that was supposed to be averaged (either orientation or size).

In each display, we presented 12 white lines on a gray background. To determine the orientation of the lines, we pseudo-randomly generated distributions within 60 and 90-degree intervals. For example, from a 60-degree interval, ranging between 25 and 85 degrees, we determined 12 different orientations.² The length of each line was determined randomly from a range of 24 and 152 pixels (.75° – 4.75° of visual angle) with the following constraints. No two lines were of the same length or orientation as the set mean and there were no repetitions of length or angular orientation within

² To determine the appropriate ranges of orientations, we conducted a pilot test (n=9) to see how display characteristics would influence viewers' orientation averaging performance. We tested orientation distributions chosen from 15, 30, 60, 90, 120-degree ranges. We found that orientation-averaging errors increased as the distribution range increased ($F(4,260) = 49.32, p < .001$). Therefore, we decided to use 60 and 90 degrees as orientation ranges to keep the task difficulty at an optimum level.

Fig. 1 Time course of a trial.

sets. We positioned the lines on a 5x4 invisible grid (760x608 pixels), with 3 locations randomly chosen from each row. We shifted the middle two rows of the grid by ± 16 pixels in order to prevent the endpoints of the lines touching the center of the display.

On the response screen, we presented viewers with either a blue or green response probe to distinguish the two types of trials, size and orientation averaging. For the size adjustment task, the length of the response probe was determined by adding/subtracting a random value between 21 and 27 pixels, to that of the half of the mean length of the set of lines. For the orientation rotation task, we determined the angle of the probe line by either adding to or subtracting a random number between 21 and 27 degrees from the mean angle of the set of lines. In mixed-task blocks, we presented an auditory verbal cue along with the color-coded response probe (green for orientation, blue for size). In both types of trials, the red fixation cross, indicated the beginning of a trial.

Procedure Before the actual experiment trials, we instructed the participants about the experiment procedure and they first completed a training session. To illustrate what the average size of a set of lines looked like we presented participants with example displays, which consisted of 2, 3, 4, 6, 8, 10, and 12 lines, sequentially. On the subsequent display following the example display, there was a red line indicating the mean length of the lines in the previous set. After that, participants completed 10 practice trials for the size-averaging task. During practice trials, we presented the displays for 1000 ms and we provided visual feedback by showing the correct response line in red. We followed the same procedure for the orientation-averaging task. After these training trials, participants also received a practice block with 16 trials, which was identical to the actual experimental trials. There were 360 trials in the actual experimental phase presented in 4 even

blocks. Half of the trials were single-task trials evenly split across size averaging and orientation averaging blocks (90 trials per feature, per block). The remaining 180 trials were mixed-task trials in which we randomly presented orientation and size-averaging tasks across two blocks. We counterbalanced the order of tasks and blocks across participants.

Results

We conducted the pairwise t-test and Pearson's correlation analyses using SPSS Version 25.0 (IBM Corp, 2017). For the Bayesian Factor analyses, we used the open source statistical program JASP (JASP Team, 2018). For each participant, for each task we calculated the mean errors on the size and orientation tasks in the single and mixed blocks. To determine whether there was an attentional cost of attending to both featural dimensions, for each summarization task, we first compared performance across single and mixed blocks. Pairwise t-tests showed that, there was no difference in errors in single ($M = 19.61$, $SD = 7.76$) versus mixed size trials ($M = 18.51$, $SD = 6.25$), ($t(20) = .70$, $p = .48$, Cohen's $d = .15$). Results from Bayesian t-tests provided further support for this null finding; we found moderate evidence in favor of the null hypothesis ($BF_{01} = 3.52$). There was also no difference between averaging errors for single ($M = 15.22$, $SD = 3.73$) and mixed orientation averaging tasks ($M = 14.88$, $SD = 2.83$), ($t(20) = .808$, $p = .43$, Cohen's $d = 0.10$). Bayesian analyses showed a moderate evidence in favor the null model ($BF_{01} = 3.28$), suggesting that observing no difference between the single and mixed orientation conditions were more likely than observing a difference (see Fig. 2).

To test for the independence of size and orientation averaging, we computed the correlations between the errors on each of these tasks, separately for the single and mixed-task

blocks. There was no significant correlation between size and orientation averaging errors in neither the single nor the mixed blocks ($r = -.08$, $p = .75$ and $r = -.11$, $p = .64$, respectively). Also, according to Bayesian Correlation Pairs analyses, there was a moderate evidence in favor of the Null hypothesis ($BF_{01} = 3.52$, $BF_{01} = 3.35$, for single and mixed task conditions, respectively), suggesting that finding independence between size and orientation errors was more likely than finding a relationship between them, for both single and mixed task conditions (see Fig. 3).

Discussion

There were two main findings from Experiment 1. One, there was no significant correlation between size and orientation errors in neither the single nor the mixed task conditions. Two, Experiment 1 also revealed similar levels of errors for both tasks across both single and mixed-task conditions suggesting that concurrent averaging of multiple summaries may be somewhat automatic. These findings, suggest that there may be independent mechanisms that support ensemble perception of these low-level features. Our findings demonstrating that participants can extract mean size and mean orientation information similarly across single task and mixed task trials is consistent with evidence showing that participants can extract separate feature summaries of spatially segregated sets in parallel (e.g. Attarha & Moore, 2015; Utochkin & Vostrikov, 2017, but also see Emmanouil & Treisman, 2008). However, we extend these findings by showing that viewers can summarize concurrently features that covary within a set (size and orientation of lines) without a cost. Utochkin & Vostrikov (2017) similarly reported that two features (size and numerosity) was averaged concurrently from a single set, however, in their study, within the display, there

were two spatially intermixed subsets denoted by color. Thus, viewers could have globally attended to the central region and utilize color-based grouping cues to extract the two summaries of the two subsets concurrently. In that regard, our study differs because the two features we asked participants to summarize belonged to the same items and the fact that viewers could summarize these featural distributions independently suggests that the summarization most likely happens before features are bound in object-file type representations (Treisman, 1996). The fact that line length variability did not influence errors in orientation summary further strengthens this possibility. Specifically, we categorized displays as having low, medium and high variance in line length and showed that orientation averaging errors did not vary as a function of variability in line length ($F(2, 180) = 2.82$, $p = .06$, $\eta_p^2 = .03$). These results suggesting there are independent summarization mechanisms for low-level features operating prior to object-file formation may be driven by the visual system's tendency to represent the incoming information as “loose bundles of features” as opposed to object files. Specifically, when the visual system encounters objects, the pattern of activations across a population of neurons may not be selective enough to ensure independent coding of each object. This may be particularly true when the objects share visual features and are presented simultaneously, as in our experimental displays (Treisman, 1999).

Experiment 2

Experiment 1 demonstrated that viewers could summarize independently size and orientation distributions. This finding is in line with the idea of feature-specific and domain-specific mechanisms for ensemble perception. However, the relatively

Fig. 2. (a) Mean size error across single and mixed task conditions. (b) Mean orientation error across single and mixed task conditions. Error bars indicate ± 2 standard error of the mean.

Fig. 3. Scatterplots depicting the relationship between errors in mean orientation and mean size in the a) single and b) mixed task conditions. Each data point represents a participant; the linear best fit is shown.

long display durations in Experiment 1 (200ms, unmasked) may have allowed viewers to utilize different mechanisms across the two conditions, possibly a feature-based mechanism for summarizing orientation and a separate system operating at the item level. It is known that orientation is processed early in the visual processing hierarchy, and size is processed further along the ventral stream (Whitney, Haberman & Sweeny, 2014), with there being dedicated groups of receptors to process different orientations but not size (Myczek & Simons, 2008). Instead, there may be a separate, focused attention-based, pooling mechanism that support the summarizing of size information as opposed to more feature-based mechanisms supporting summarization of orientation, motion and spatial frequency information (Simons & Myczek, 2008; but also see Ariely, 2008, Chong, Joo, Emmanouil & Treisman, 2008). If this is indeed the case, when viewers summarize sets having only viewed them for brief durations, pre-attentive processes (or processes driven by global attention) may only support the summarization of orientation information but not size information. Consequently, for briefly presented displays, errors for mean size trials may increase while errors for orientation summary trials may be immune to this manipulation. Furthermore, in mixed-task blocks, viewers may unknowingly prioritize orientation information during shorter presentation trials, at the cost of errors in size averaging. To test these possibilities, in Experiment 2, we examined the temporal independence of size and orientation summarization processes, by presenting displays for 50, 100 or 200ms.

Method

Participants A new group of 26 undergraduate Bogazici University students participated in this experiment in return for course credit. We excluded data of one participant because his/her orientation averaging errors were more than 3 standard deviations away from the group mean. We excluded the data

from another participant because s/he randomly responded throughout the experiment. Therefore, we conducted the analyses on the data from the remaining 24 participants.

Material, stimuli and procedure We used the same procedure and design as Experiment 2, except that we manipulated display durations so that displays for presented for either 50, 100, or 200 milliseconds. Display durations were intermixed throughout the trials of all blocks. The experiment consisted of 360 trials in total, with 180 mean size and 180 mean orientation trials in total. Participants completed half of these trials within the single task and the rest in the mixed task block. The display time manipulation was evenly split across conditions and blocks.

Results

For each participant, we calculated the errors for the size and orientation averaging tasks for each block type and display duration conditions (Figs. 4 and 5, respectively). We separately carried out a 2 (Block type: Single vs. Mixed) $\times$ 3 (Display Duration: 50, 100, 200 ms) repeated measures ANOVA, on size and orientation averaging errors. For the average size errors, results showed that, there was no significant effect of block type ($F(1,23) = 0.59, p = .81, \eta_p^2 = .003$), no significant effect of display duration ($F(2,22) = 2.64, p = .09, \eta_p^2 = .194$) and there was no significant interaction between the two variables, $F(2,22) = 3.14, p = .06, \eta_p^2 = .222$. Also, Bayesian ANOVA provided support for the Null model (see Table 1). A similar pattern emerged for orientation averaging errors, with neither main effect reaching significance, ($F(1,23) = 1.69, p = .21, \eta_p^2 = .069$), ($F(2,22) = .15, p = .87, \eta_p^2 = .013$), for block type and display durations, respectively. Bayesian ANOVA results were in favor of the Null model, with no main effects of display duration and task (see Table 2). For the orientation errors, there was a significant interaction between

Fig. 4. Mean size error across single and mixed conditions and display durations. Error bars indicate ± 2 standard error of the mean.

Block type X Display Duration $F(2,22) = 6.43$, $p < .05$, $\eta_p^2 = .369$.

As in Experiment 1, to directly test the feature-specificity claim, we also analyzed the relationship between participants' error on size and orientation averaging tasks, across all single-mixed conditions and durations. For single block trials, for all display durations, none of the correlations were significant (all $r_s < .24$, all $p_s > .25$). In addition, there was no significant correlation between participants' mixed block size and orientation averaging errors in none of the display durations (all $r_s < .13$, all $p_s > .59$).

Discussion

Experiment 2 replicated the overall pattern observed in Experiment 1, despite the reduction of display time to 50ms. Participants made similar levels of error on size and orientation averaging tasks across different display durations (50, 100, and 200 milliseconds); the errors across these two

Fig. 5. Mean orientation error across single and mixed conditions and display durations. Error bars indicate ± 2 standard error of the mean.

domains and across different task conditions were found to be independent. While previous research argued that attention-driven pooling mechanisms contributed to size averaging, we found no evidence to support this distinction between pre-attentive, featural and object-level attentional mechanisms (Myczek & Simons, 2008).

One potential problem with our design might have been the absence of a mask following the study display. While viewers can efficiently summarize unmasked displays at durations as short as 50 ms (e.g. Chong & Treisman, 2003) and even 30ms (Jacoby, Kamke & Mattingley), they need much longer durations to summarize masked displays (e.g. 100ms, Whiting & Oriet, 2011). By masking objects, the contributions of these objects were reduced on average size or orientation estimates (Jacoby et al., 2013, but also see Choo & Franconeri, 2010), suggesting that averaging processes go beyond the very early featural processing stages. In our case, the absence of masks following the display, may have let reentrant processes to contribute to the summarization process, and may have facilitated summary accuracy, particularly for the size task. To eliminate any processing advantage incurred by iconic memory (Sperling, 1960), in Experiment 3, we introduced a mask and replicated the experiment with the shortest display duration tested.

Experiment 3

In Experiment 3, to eliminate the potential confound from the iconic memory system and test whether both size and orientation averaging follow similar time courses, we introduced a mask subsequent to the presentation of the study set. Experiment 3 tested the shortest of the display durations tested in Experiment 2.

Method

Participants 40 undergraduate students at Bogazici University participated in this experiment in return for course credits. We excluded data from three participants for average errors exceeding three standard deviations of the group mean (2 participants for single orientation condition, and 1 participant for single orientation averaging condition errors). Therefore, we conducted the analyses on the data from 37 participants.

Material, stimuli and procedure The procedure and design of Experiment 3 was identical to Experiment 2, except that we only used the 50 milliseconds display durations for the presentation of object sets for all the trials, and a visual mask followed the display for 320 milliseconds before the presentation of the response screen. The visual mask was a black and white noise image covering the region area where we presented the displays. There were 360 trials in total.

Table 2. Bayesian ANOVA for orientation averaging errors

Model Comparison					
Models	P(M)	P(M data)	BF _M	BF ₀₁	error %
Null model (incl. subject)	0.200	0.595	5.873	1.000	
Duration	0.200	0.045	0.190	13.132	0.882
SingleMixed	0.200	0.242	1.275	2.461	2.657
Duration + SingleMixed	0.200	0.018	0.072	33.462	1.177
Duration + SingleMixed + Duration * SingleMixed	0.200	0.100	0.446	5.928	3.108

Note. All models include subject.

Results

As in the previous experiments, we calculated error for average size and orientation conditions, separately for the single and mixed trial blocks (See Fig. 6). For the size feature, there was no difference in the error between the single ($M = 19.03$, $SD = 6.57$) and mixed conditions ($M = 19.80$, $SD = 6.56$), $t(36) = -.79$, $p = .43$, Cohen's $d = 0.01$ (see Fig. 6). Also, the Bayesian analyses provided moderate evidence in favor of the null hypothesis ($BF_{01} = 4.22$). On the other hand, averaging errors for mixed orientation trials ($M = 16.97$, $SD = 3.75$) were significantly higher than single orientation averaging trials ($M = 15.87$, $SD = 2.42$), $t(36) = -2.457$, $p < .05$, Cohen's $d = 0.04$ (see Fig. 6). Bayes factor provided anecdotal evidence in favor of the alternative hypothesis ($BF_{10} = 2.44$), which suggests that the difference observed may not be sufficiently reliable to warrant further discussion.

As in the previous experiments, we computed Pearson's correlations between errors in orientation and size averaging for the single task and mixed task (see Fig. 7). None of the correlations reached significance ($r = -.21$, $p = .21$ and $r = .28$, $p = .09$) for single and mixed conditions, respectively). The Bayesian correlational analyses complemented these null findings; both correlations were more likely to favor the Null model with anecdotal evidence with no relation between

size and orientation averaging (single blocks: $BF_{01} = 2.30$, mixed blocks: $BF_{01} = 1.23$), errors than having a significant relation (single blocks: $BF_{10} = .43$, mixed blocks: $BF_{10} = .81$).

Discussion

Experiment 3 replicated the general pattern of results obtained in the first two experiments; the introduction of a mask subsequent to the display did not change the finding suggestive of a feature and domain-specific mechanism for averaging size and orientation information. While the results of the inferential analyses suggested that for orientation trials, errors were higher in the mixed-task blocks than in the single-task block, the Bayes factor provided only anecdotal or weak evidence for the likelihood of the alternative hypothesis over the null model. Together, these results suggest that the significant results obtained for orientation-averaging comparisons may have been due to the larger sample size in Experiment 3 compared to the previous experiments (Jarosz & Wiley, 2014). Consistent with the previous experiments, we again found that there was no significant relationship between viewers' errors on size and orientation averaging errors for single task blocks. There was a marginally significant positive correlation between viewers' errors in the mixed block; however, the results of the Bayesian analysis suggest that these correlations are

Fig. 6. (a) Mean size-averaging errors across single and mixed task conditions. (b) Mean orientation-averaging errors across single and mixed task conditions. Error bars indicate ± 2 standard error of the mean.

Fig. 7. Scatterplots depicting the relationship between errors in mean orientation and mean size in the a) single and b) mixed task conditions. Each data point represents a participant; the linear best fit is shown.

weak at best. Altogether, we replicated the general pattern of results from previous experiments, supporting independent feature and domain-specific processing mechanisms for size and orientation features.

General discussion

In this study, using a combined experimental and correlational approach, we investigated whether feature-specific mechanisms drive ensemble perception. Across three experiments, we investigated whether viewers can concurrently summarize size and orientation of a set of lines and whether the error in each task was independent. Previous research provided evidence for separate summarization mechanisms for higher versus lower level features (e.g. Haberman et al., 2015) and for features summarized along the dorsal and ventral pathway (e.g. Hubbert-Wallander & Boynton, 2015). However, the evidence to date was inconclusive as to whether there was any further specialization. To determine whether there are multiple feature-specific mechanisms for low-level features, we specifically targeted two features, size and orientation, for which mid-level texture based summarization mechanisms had been proposed (e.g. Cain & Cain, 2018; Parkes et al., 2001). The increased attentional demands in the mixed-task condition is more likely to affect the operation of domain-general summarization mechanisms that are agnostic to specific features (e.g. Bauer, 2017). In addition, if a shared mechanism were summarizing features, then we would expect a significant positive correlation between summarization errors for these features. However, a significant positive correlation would not necessarily mean a feature-agnostic mechanism is at work; such a pattern could emerge because of shared neural noise between different feature representations (e.g. Haberman et al., 2015). In contrast to the domain general mechanism, feature-specific summarization mechanisms, would allow for cost-free, concurrent processing of different features, especially under high

attentional load conditions (e.g. Attarha & Moore, 2015). Furthermore, independence of errors across different feature dimensions could only be consistent with the existence of multiple feature-specific mechanisms. In other words, null findings both at the accuracy level and at the correlation level would support the feature-specific view. Therefore, in our data analyses we took a Bayesian approach to evaluate the support for the null hypothesis in addition to the traditional NHST approach.

Across three experiments, we found that there were no attentional costs associated with concurrently averaging size and orientation features. The error for both features were similar across the single-task and mixed-task conditions. There were also no significant correlations between size and orientation errors in any of the conditions. Overall, our findings provide evidence for the view that there are multiple feature-specific operating at lower-levels of the processing hierarchy (Haberman, et al., 2015). We extend Hubert-Wallander & Boynton (2015)'s findings by demonstrating that there may be other feature-specific summarization mechanisms besides location vs. non-location ones. While our findings do not directly speak to the neural mechanisms underlying ensemble perception, our results are consistent with proposals speculating that there could be multilevel statistical summary processors in different regions of the brain (Whitney, Haberman & Sweeny, 2014).

While our results pointed to the feature-specificity of ensemble perception, we were not able to distinguish the temporal summarizing profiles for size and orientation. We had predicted orientation averaging to be equally efficient at shorter and longer durations, but size averaging to show greater errors at shorter display durations. We found no evidence for such a dissociation. This may have been partly due to our display duration manipulation not being sensitive enough; 50ms may have been a sufficient duration for both size and orientation summarization. Presenting stimuli for even shorter

display durations may reveal differences in summarization accuracy, possibly favoring encoding of mean orientation over mean size. There is also a related possibility that summarization in these two featural domains is differentially impacted by reentrant based mechanisms (e.g. Jacoby et al., 2013; Lamme & Roelfsema, 2000). In our particular case, it is also possible that early and late summarizing mechanisms supported orientation and size averaging, respectively (Myczek & Simons, 2008). In particular, an early feedforward mechanism may support orientation averaging while a late, reentrant-based mechanism that has been associated with object file formation and binding may support size-averaging (Treisman & Gelade, 1980; Kahneman, Treisman & Gibbs, 1992; Treisman, 2006; also see Bouvier & Treisman, 2010). This possibility is consistent with the pattern of results presented here but future research is necessary to specify contributions of feedforward and reentrant processing on ensemble perception.

The existence of multiple, feature-specific summarization mechanisms have also been supported by recent findings in rapid scene recognition. In addition to feature and domain-specific summarization mechanisms, there are conceptual summarization mechanisms. For instance, the visual system can accurately capture scene gist presented in RSVP paradigms (e.g. Potter, 1975; 1976). This ability to extract scene gist upon viewing images for as short as 40ms has been associated with summary mechanisms acting on global image characteristics (e.g. Oliva & Torralba, 2006). Critically, global ensemble texture information enables observers to determine the spatial layout and scene layout subsequently facilitates scene recognition (e.g. Sanocki, 2003; Ross & Oliva, 2010). Recently Brady, Shafer-Skelton & Alvarez (2017), also showed that high sensitivity to spatial texture information correlated with rapid scene recognition ability. They had participants complete a rapid scene recognition task along with a mean orientation task and a spatial ensemble task, requiring detecting changes to a grid of Gabor patches. Successful performance on this latter task required summarization at the texture level. Of all pairwise comparisons, the only significant correlation was between the spatial texture sensitivity and rapid scene recognition. These findings highlight that different summarization mechanisms can interact with one another to facilitate visual perception, and global ensemble texture representations facilitate scene recognition in rapid visual processing.

We believe that it may be no coincidence that feature-specific mechanisms drive both the capacity of visual short-term memory and ensemble perception (Haberman et al., 2015). There is both computational and empirical work demonstrating the interaction between ensemble and item representations, with the former scaffolding item representations (e.g. Brady & Tenenbaum, 2013). For instance, individual item representations in working memory are biased by ensemble representations (e.g. Brady & Alvarez, 2011), and

availability of spatial summary information was shown to facilitate item localization in working memory (Mutlutürk & Boduroglu, 2014). On the working memory end, there is considerable work suggesting that errors in memory for features such as orientation and color are independent (e.g. Bays, We & Husain, 2011) and there is selective decay for such features (e.g. Pasternak & Greenlee, 2005). These findings altogether underline the critical role played by feature-based mechanisms for both perception and working memory processes.

Acknowledgements This project is supported by a grant from The Scientific and Technological Research Council of Turkey (TÜBİTAK, 116K384) granted to A.B. We would like to thank Esra Mungan, Ahu Gökçe, Benjamin Tamber-Rosenau, Lindsay Santacroce, and Brandon Carlos for their feedbacks, and Emre Oral and Furkan Erdal for their assistance in data collection.

Open Practices Statement None of the experiments reported here were preregistered. The data or materials can be made available to interested parties. Please contact aysecan.boduroglu@boun.edu.tr.

References

- Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world extracting statistical summary representations over time. *Psychological Science*, 21(4), 560–567.
- Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Attention, Perception, & Psychophysics*, 74, 810–815.
- Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, 15(3), 122–131.
- Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4), 392–398.
- Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162.
- Ariely, D. (2008). Better than average? When can we say that subsampling of items is better than statistical summary representations? *Perception & Psychophysics*, 70(7), 1325–1326.
- Attarha, M., & Moore, C. M. (2015). The perceptual processing capacity of summary statistics between and within feature dimensions. *Journal of Vision*, 15(4), 1–17.
- Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, 18(7), 622–628.
- Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging?. *The Psychological Record*, 59(2), 171–185.
- Bauer, B. (2017). Perceptual averaging of line length: Effects of concurrent digit memory load. *Attention, Perception, & Psychophysics*, 79(8), 2510–2522.
- Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, 49(6), 1622–1631.
- Bouvier, S., & Treisman, A. (2010). Visual feature binding requires reentry. *Psychological Science*, 21(2), 200–204.
- Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3), 384–392.

- Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, 43(6), 1160–1176.
- Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, 120(1), 85–109.
- Cain, S., & Cain, M. (2018). A Texture Representation Account of Ensemble Perception. *Journal of Vision*, 18(10), 618–618.
- Chong, S. C., Joo, S. J., Emmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, 70(7), 1327–1334.
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4), 393–404.
- Choo, H., & Franconeri, S. L. (2010). Objects with reduced visibility still contribute to size averaging. *Attention, Perception, & Psychophysics*, 72(1), 86–99.
- Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience?. *Trends in Cognitive Sciences*, 20(5), 324–335.
- Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *JOSA A*, 18(5), 1016–1026.
- de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology*, 62(9), 1716–1722.
- Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory. *Journal of Vision*, 15(4):13, 1–13.
- Emmanouil, T. A., & Treisman, A. (2008). Dividing attention across feature dimensions in statistical processing of perceptual groups. *Perception & Psychophysics*, 70(6), 946–954.
- Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, 144(2), 432–446.
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17), R751–R753.
- Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision*, 15(4), 1–12.
- IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.
- Jacoby, O., Kamke, M. R., & Mattingley, J. B. (2013). Is the whole really more than the sum of its parts? Estimates of average size and orientation are susceptible to object substitution masking. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1), 233–2244.
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1), 2–9.
- JASP Team (2018). JASP (Version 0.9) Retrieved from <http://jasp-stats.org/>
- Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2), 175–219.
- Khayat, N., & Hochstein, S. (2018). Perceiving set mean and range: Automaticity and precision. *Journal of Vision*, 18(9):23, 1–14.
- Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11), 571–579.
- Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue. *JOSA A*, 31(4), A93–A102.
- Morgan, M., Chubb, C., & Solomon, J. A. (2008). A ‘dipper’ function for texture discrimination based on orientation variance. *Journal of Vision*, 8(11), 1–8.
- Mutlütürk, A., & Boduroglu, A. (2014). Effects of spatial configurations on the resolution of spatial working memory representations. *Attention, Perception & Psychophysics*, 76, 2276–2285.
- Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Attention, Perception, & Psychophysics*, 70(5), 772–788.
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155, 23–36.
- Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, 4(7), 739–744.
- Pasternak, T., & Greenlee, M. W. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, 6(2), 97–107.
- Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science*, 24(8), 1389–1397.
- Poltoratski, S., & Xu, Y. (2013). The association of color memory and the enumeration of multiple spatially overlapping sets. *Journal of Vision*, 13(8), 1–11.
- Potter, M. C. (1975). Meaning in visual search. *Science*, 187(4180), 965–966.
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509–522.
- Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, 10, 2–25.
- Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology*, 47(1), 43–86.
- Scolari, M., Vogel, E. K., & Awh, E. (2008). Perceptual expertise enhances the resolution but not the number of representations in working memory. *Psychonomic Bulletin & Review*, 15(1), 215–222.
- Semizer, Y., & Boduroglu, A. (November, 2013). Representation of variance and its impact on mean size estimation in visual short-term memory. *Poster presentation at the Annual Meeting of the Psychonomics Society, Toronto, CN.*
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11), 1–29.
- Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, 14(4–8), 411–443.
- Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, 6(2), 171–178.
- Treisman, A. (1999). Solutions to the binding problem: progress through controversy and convergence. *Neuron*, 24(1), 105–125.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Tulver, K. (2019). The factorial structure of individual differences in visual perception. *Consciousness and Cognition*, 73, 102762.
- Uner, O., Mutlütürk, A., & Boduroglu, A. (November, 2014). The impact of video game playing on the extraction of statistical summaries in visual displays. *Poster presentation at the Poster presentation at the Annual Meeting of the Psychonomics Society, Long Beach, CA.*
- Utochkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PLoS One*, 12(9), e0185452.
- Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, 428(6984), 748–51.
- Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, 17(2), 171–176.
- Whiting, B. F., & Oriet, C. (2011). Rapid averaging? Not so fast! *Psychonomic Bulletin & Review*, 18(3), 484–489.
- Whitney D., Haberman J., Sweeny T. D. (2014). From textures to crowds: Multiple levels of summary statistical perception. *Werner J. S.,*

- Chalupa L. M., (editors), *The New Visual Neurosciences*. Cambridge, MA: MIT Press.
- Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, 69, 105-129.
- Wilmer, J. B. (2008). How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. *Spatial Vision*, 21(6), 561-579.
- Yildirim, I., Öğreden, O. & Boduroglu, A. (2018). Impact of spatial grouping on mean size estimation. *Attention, Perception & Psychophysics*, 80(7), 1847-1862.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.