NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

As a manuscript

Igor Utochkin

FEATURES, OBJECTS, AND ENSEMBLES IN VISUAL PERCEPTION AND MEMORY

Summary
of the dissertation for the degree of
Doctor of Science in Psychology

Moscow, 2021

The dissertation was prepared at the Laboratory for Cognitive Research, National Research University Higher School of Economics.

Seven published articles were selected for the defense:

1. Utochkin, I.S., & Brady, T.F. (2020) Independent storage of different features of real-world objects in long-term memory. *Journal of Experimental Psychology: General, 149*(3), 530-549. DOI: 10.1037/xge0000664
2. Khvostov, V.A., & Utochkin, I.S. (2019). Independent and parallel visual processing of ensemble statistics: Evidence from dual tasks. *Journal of Vision, 19* (8): 3, 1-18. DOI: 10.1167/19.9.3
3. Tiurina, N.A., & Utochkin, I.S. (2019). Ensemble perception in depth: Correct distance-size rescaling of multiple objects before averaging. *Journal of Experimental Psychology: General, 148* (4), 728-738. DOI: 10.1037/xge0000485
4. Utochkin, I.S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision, 15* (8), 1-14. DOI: 10.1167/15.4.8
5. Utochkin, I.S., & Yurevich, M.A. (2016). Similarity and heterogeneity effects in visual search are mediated by "segmentability". *Journal of Experimental Psychology: Human Perception and Performance,42* (7), 995-1007. DOI: 10.1037/xhp0000203
6. Utochkin, I.S., Khvostov, V.A., & Stakina Y.M. (2018). Continuous to discrete: Ensemble-based segmentation in the perception of multiple feature conjunctions. *Cognition, 179,* 178-191. DOI: 10.1016/j.cognition.2018.06.016
7. Utochkin, I.S., & Brady, T.F. (2020b). Individual representations in visual working memory inherit ensemble properties. *Journal of Experimental Psychology: Human Perception and Performance, 46* (5), 458–473. DOI: 10.1037/xhp0000727

The results are also published in the following articles on this topic:

8. Khvostov, V.A., Lukashevich, A.O., & Utochkin, I.S. (2021). Spatially intermixed objects of different categories are parsed automatically. *Scientific Reports, 11:* 377, 1-8. DOI: 10.1038/s41598-020-79828-4

9. Markov, Y.A., Utochkin, I. S., & Brady T. F. (2021). Real-world objects are not stored in holistic representations in visual working memory. *Journal of Vision, 21*(3): 18, 1–24. DOI: 10.1167/jov.21.3.18.
10. Utochkin, I.S., Khvostov, V.A., & Wolfe J.M. (2020). Categorical grouping is not required for guided conjunction search. *Journal of Vision, 20*(8): 30, 1-22. DOI: 10.1167/jov.20.8.30
11. Iakovlev, A.U., & Utochkin I.S. (in press). Roles of saliency and set size in ensemble averaging. *Attention, Perception, & Psychophysics.* DOI: 10.3758/s13414-020-02089-w
12. Im, H.Y., Tiurina, N.A., & Utochkin, I.S. (in press). An explicit investigation of the roles that feature distributions play in rapid visual categorization. *Attention, Perception, & Psychophysics.* DOI: 10.3758/s13414-020-02046-7
13. Brady, T.F., & Utochkin, I.S. (2019). Entities also require relational coding and binding (commentary on Bastin et al.). *Behavioral and Brain Sciences, 42,* E285. DOI: 10.1017/S0140525X19001924
14. Markov, Y.A., Tiurina, N.A., & Utochkin, I.S. (2019). Different features are stored independently in visual working memory but mediated by object-based representations. *Acta Psychologica, 197,* 52-63. DOI: 10.1016/j.actpsy.2019.05.003
15. Markov, Y.M., Tiurina, N.A., Stakina, Y.M., & Utochkin, I.S. (2017). The capacity and precision of visual working memory for objects and ensembles. *Psychology. Journal of HSE, 14* (4), 735-756. DOI: 10.17323/1813-8918-2017-4-735-755
16. 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. DOI: 10.1371/journal.pone.0185452
17. Zou, B., Utochkin, I.S., Liu, Y., & Wolfe, J.M. (2017). Binocularity and visual search – Revisited. *Attention, Perception & Psychophysics, 79* (2), 473-483. DOI: 10.3758/s13414-016-1247-8
18. Utochkin, I.S. (2016). Visual enumeration of spatially overlapping subsets. *The Russian Journal of Cognitive Science, 3*, 4-20.

19. Utochkin, I.S., & Tiurina, N.A. (2014). Parallel size averaging is possible but range-limited: A reply to Marchant, Simons, and De Fockert. *Acta Psychologica, 146,* 7-18. DOI: 10.1016/j.actpsy.2013.11.012
20. Tiurina, N.A., & Utochkin I.S. (2014). Rol' lokal'nogo i global'nogo skhodstva priznakov v zadache zritel'nogo poiska [The roles of local and global feature similarity in visual search]. *Voprosy Psikhologii, issue 4,* 107-117.
21. Utochkin, I.S. (2013). Visual search with negative slopes: The statistical power of numerosity guides attention. *Journal of Vision, 13* (3): 18, 1-14. DOI:10.1167/13.3.18

1. Introduction

1.1. General research problem

Human visual experience is introspectively organized in an object-based fashion. That is, people normally see the world consisting of a number of various meaningful things without effort. The same can be said of visual memory: People can remember thousands of images of objects and easily recognize them among other objects, even if they have seen them only once (Brady et al., 2008; Standing, 1973).

Do we indeed have complete and stable representations of all objects we encounter in visual cognition? There are impressive demonstrations that this completeness and stability are an illusion, sometimes referred to as the Grand Illusion of our consciousness (Noë, 2002). Inattentional blindness is one such example showing a failure to spot an otherwise salient object when attention is engaged in a difficult task (Mack & Rock, 1998). Change blindness, an inability to see a large change in a visual scene when the change is masked by a global transient (such as eye movements, eyeblinks, brief occlusions), shows a failure of conscious object perception even when we intentionally look for such changes (Rensink, 2002). These and other demonstrations suggest that our capacity of conscious object perception is limited to a handful of items at one time. This processing "bottleneck" is often associated with the capacity of attention (Pylyshyn & Storm, 1988) or with working memory (Cowan, 2001). One influential idea why this bottleneck arises within the perceptual system is that correct object perception requires *binding* of independently processed *basic features* of multiple different objects (Treisman, 1996; Wolfe et al., 2011). The enormous combinatorial complexity of the visible world given the potential variety of the features (Tsotsos, 1988) makes binding hardly possible to run for all objects in parallel. If we imagine that the complete parallel object recognition would require not only correct binding of visible features but also linking these bindings to correct memory recordings, this would raise computational complexity of perception and recognition enormously (Tsotsos, 1988).

If capacity limits of full object recognition are granted, then what underlies this introspective ease of perceiving and remembering the world consisting of many different objects? Can it be accomplished with sparse representations that do not need separate

access to each object in the entirety of its details? If yes, then how is it done? How can these representations be used for various visual and memory tasks? The present work summarizes research that I and my collaborators have run to investigate these questions.

1.2. Theoretical basis of the current work includes the contemporary conceptual architectures of visual perception with "shallow" and "deep" processing (e.g., the distinction between preattentive and attentional processing – Neisser, 1967; Treisman, Gelade, 1980; selective and non-selective pathways for object and scene processing – Wolfe, Vo, Greene, Evans, 2011; feedforward and reverse hierarchies in conscious visual perception – Hochstein, Ahissar, 2002; the functional continuum of attentional states from focused to distributed attention" – Treisman, 2006, etc.); the theory of global perceptual representation of multiple objects and scenes as the computation of ensemble statistical representations (Alvarez, 2011; Chong, Treisman, 2003; Haberman, Whitney, 2012; Whitney, Yamanashi Leib, 2018; Treisman, 2006); the idea of hierarchical encoding theory in visual memory (Brady, Konkle, Alvarez, 2011; Brady, Alvarez, 2011; Corbett, 2017).

1.3. Summary of scientific novelty

1. The present dissertation suggests a general view of the format used to represent multiple objects in the visual system, from one-shot visual perception to long-term memory. The basic idea underlying this view is that, given the severe limits of deep object processing, the perception of multiple objects and their maintenance in memory can be carried out using relatively shallow and sparse feature representations of different kinds. These feature representations, in turn, can be summarized and consciously accessed as a broad spectrum of ensemble statistics.

2. We present experimental evidence that, even though complex real-world objects are subjectively experienced as holistic things (which is also reflected in some theories), the real-world objects can be stored as separate features related to various plausible independent transformations of object appearances (for example, differences between object exemplars within same basic category, or differences between states of the same object), but that the features can be misbound at retrieval.

3. We present evidence for the efficient coding of numerous properties of multiple objects in a form of ensemble representation, a generalized visual impression of multiple objects in a form of statistical summaries. We demonstrated that various types of ensemble statistics (average feature, feature variability, or the approximate number of objects) can be easily extracted from the same set of items without any loss in accuracy associated with the distribution of attention between them. In addition, we demonstrated that the visual system takes into account the rich contextual information about individual objects when summaries their ensemble statistics.

4. A new theory of ensemble-based rapid visual categorization and segmentation of multiple objects is proposed and tested in different visual tasks (e.g., visual search, texture discrimination). The theory suggests that the shape of a feature distribution in an ensemble (smooth unimodal or uniform vs. sharp, polymodal) determines whether all objects in a set are perceived as a single categorical group or parsed into several groups belonging to different categories.

5. We present new empirical evidence for the hierarchical interaction and influence of the ensemble information on individual object representations in visual working memory. In particular, we showed that the recall precision of an individual feature is not fixed and that it inherits the amount of noise from an ensemble representation this object belonged to and that the individual reports are systematically biased toward the mean feature of all objects.

1.4. Theoretical significance

The theoretical significance of the current work can be characterized by its contribution to the general cognitive theory and architecture of visual representations. Moreover, the author's work, such as the theory of rapid visual categorization and segmentation, contribute to understanding the role of various representations (e.g., ensemble representations) in the variety of visual tasks.

1.5. Applied significance is relevant for the possible use of the reported findings and conclusions for psychologically grounded principles of efficient information displays and visualization. The reported findings about ensemble summary statistics can be useful

for teaching regular statistics. The results of the reported work are partially used in undergraduate courses, “Cognitive Psychology” and “Psychology and Neurophysiology of Perception and Attention”, at the HSE University.

1.6. Statements for the defense

1. Although the visible world is introspectively perceived and remembered in an object-based manner, the information about big sets of objects can be conveyed without the need to deeply process each object as a whole. Sparse representations of various features and ensemble statistics of these features can serve as efficient proxies of complete and detailed object representations. Feature, object, and ensemble representations can be parts of a flexible hierarchical system that gives an access to a general impression of the environment filled with multiple objects and also guides deep processing of individual objects.

2. Meaningful features of complex real-world objects (the features corresponding to physically separable variations in appearance, such as exemplar and state features) are stored independently in visual long-term memory

3. Ensemble representations of large sets of objects in the form of feature summary statistics (for example, an average feature, feature variability, or the number of items) are an effective means of organizing the information about the set given the fundamental capacity limitations on deep object processing. In particular, different summary statistics are read out from the same set in parallel, that is, without the cost of dividing attention between these statistics. In addition, the computation of ensemble statistics correctly takes into account context features in which individual objects are presented (for example, distance and depth cues are taken into account when the average size is estimated), and, hence, the resulting summary statistics adequately reflect veridical properties of visible objects in the real world.

4. The shape of an ensemble feature distribution can be used as a cue for rapid visual segmentation and categorisation of multiple spatially intermixed objects. Therefore, information about the qualitative diversity of objects in the visual scene can be accessed via ensemble statistics, without the need for deep processing of all objects.

5. Ensemble statistical representations can be used to organize and optimize information about individual objects under the limited capacity of holding the information about individual objects, for example, in a working memory task with several individuals. These organizing and optimizing effects are demonstrated by findings that memory reports of individual object features inherit the statistical properties of an ensemble these objects belong to.

1.7. Data collection

Six out of seven articles selected for the defense describe sets of psychophysical experiments. Overall, twenty separate experiments are described in these papers, with over 800 observers taken part in these experiments. Observers were tested either in a laboratory (n = 340), or online via Amazon Mechanical Turk (n = 496). The laboratory experiments were run at the Cognitive Research Laboratory (HSE University, Moscow, Russia), Vision and Memory Laboratory (University of California, San Diego, USA), Visual Attention Laboratory (Brigham and Women's Hospital and Harvard Medical School, Boston, USA).

1.8. Public presentations on the topic and grant support

The results of the present work have been publicly presented since 2012 to 2020 in 46 talks and posters at 20 conferences in Russia and worldwide. These included: Annual Vision Sciences Society Meeting (2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020), European Conference on Visual Perception (2013, 2015, 2018, 2019), Conference "Cognitive Science in Moscow: New Research" (2013, 2019), IACS International Conference on Cognitive Science (2014, 2016), etc. Eight colloquium talks have been presented in the HSE Laboratory for Cognitive Research (2019), Visual Attention Laboratory at Brigham and Women's Hospital (2015, 2018, 2020), Vision and Memory Laboratory at University of California San Diego (2016, 2019), Vision Science Laboratory at Harvard University (2015), Schacter Memory Laboratory at Harvard University (2018).

The studies presented as different parts of the dissertations have been funded by the Russian Science Foundation (grant #18-18-00334, 2018-2020), by the Russian

Foundation for Basic Research (grants #12-06-31223, 2012; #15-06-07514, 2015-2016), and by the Basic Research Programme at the HSE University (2012-2019).

2. Features vs. objects as representational units of perception and memory

2.1. The concept of feature and the problem of feature diagnostic criteria

Distinguishing between objects and features is critically important for many theories of information processing and storage in vision. The discussion around these concepts and attempts to find clear boundaries between feature representations and object representations are related to the search for fundamental elements of visual experience and for the architecture of the system that builds conscious images on the basis of these elements. In one of our recent reviews (Wolfe & Utochkin, 2019), we give a detailed analysis of the concept of "preattentive feature" and theoretical problems with the use of this concept. The term "preattentive" refers to an idea suggested in frameworks such as the Feature Integration Theory (FIT, Treisman & Gelade, 1980) that these basic structures are available in the visual field at early stages of processing before attention deploys and provides limited access of the fraction of the visual field to the "bottleneck" of deep object processing (binding, working memory maintenance, recognition, long-term consolidation, etc.). One of the important properties of preattentive processes is that they are not subject to processing bottlenecks and, hence, can operate over the entire visual field in parallel. As such, the features registered preattentively can be potential candidates for being important components of visual awareness beyond the limited capacity of deep object processing.

Although the idea of purely preattentive features giving birth to conscious visual representations without any processing bottleneck has been challenged at some point (Joseph, Chun, & Nakayama, 1997), the idea that some information is broadly available without the need to focus on each particular thing in the visual field is still of use. We will address the way this information is extracted, the "how"-question, below (see section 3). Here, we try to answer the "what"-question: What criteria should be met to diagnose some stimulus aspect as a preattentive feature? We review these criteria (many of them were suggested by Anne Treisman - e.g., Treisman, 2006), as well as empirically

grounded caveats for using these criteria.

The first criterion of a feature is that it should guide attention in a visual search task, which means that some distinct properties of multiple objects can be used to limit the number of objects that are likely candidates to be attended (efficient set size). For example, a single red dot among many green dots will immediately grab attention in a bottom-up manner. Another example is a search for a red letter "T" among other red and black letters will be limited to inspecting only a red subset of letters (Egeth, Garbart, & Virzi, 1984) which means redness guides search in a top-down manner (via knowledge of target features). The important caveat for this criterion is that not all properties that guide are simple features because sometimes they can be decomposed to simpler features. For example, search for a red vertical line among green vertical and red horizontal lines can be efficiently guided but in fact it is guided by redness and verticalness simultaneously (Wolfe, Cave, & Franzel, 1989).

The second criterion is the so called search asymmetry that refers to a finding that a target with a certain feature will be found faster among distractors without that feature than vice versa (e.g., finding a letter "C" among letters "O" is easier than finding a letter "O" among letters "C" - probably because the gap is a feature, Treisman & Souther, 1985). There are also a number of cautions about this criterion, including inherent methodological issues of many search asymmetry experiments (Rosenholtz, 2001a) or problems with the interpretation of subtle asymmetries as evidence of a basic feature (e.g., slightly more efficient search of angry faces among neutral faces than vice versa - Eastwood, Smilek, & Merikle, 2001).

The third criterion is that preattentive features should support global effortless grouping and segmentation (e.g., an ability to see digits in a mess of interspersed differently colored dots in a color blindness test). This criterion logically follows the supposedly parallel nature of preattentive processing. Yet, this criterion is also not absolute and there are examples showing that even most conventional basic features (e.g., orientation) fail to form easily segmented global groups under some conditions (Inverso et al., 2017).

The fourth criterion is proneness to selective adaptation that often manifests as

contrast aftereffect illusions. For example, adaptation to a grating tilted to the right will make a subsequent vertical grating being perceived slightly tilted to the left (tilt aftereffect). This is also true for many other sensory dimensions (color, motion, etc.), as well as for complex perceptual dimensions (facial expression, race, gender, etc.). But there are many exceptions. Some features that provide good guidance and strong search asymmetries can show no evidence for aftereffects. And some complex properties that are obviously reducible to simpler features do show adaptation aftereffects (for example, the McCollough effect, adaptation to a certain color-orientation association).

The fifth criterion coming from neurophysiology suggests that the basic features are stimulus properties that distinct regions (the so called retinotopic "feature maps") in the visual cortex selectively respond to, such as colors, orientations, motion, etc. (Zeki, 1978; Zhaoping, 2016). However, the main problem with this approach is that simply knowing that some aspect of a stimulus is selectively processed by a distinct brain region does not tell us what kind of conscious experience it correlates with and if there is any correlated conscious experience (Hochstein & Ahissar, 2002).

Summarizing, we end up with a complicated picture of what counts as a basic, preattentive feature. Perhaps, identifying the complete list of such elementary features, as well as the list of unambiguous featural criteria is not always possible. Therefore, we arrive at a pragmatic, operational definition of a "feature" as of a stimulus property that can guide attention (that is, it provides information about things in the visual field before they are attended and this information is used to determine where attention should likely go) and that are not decomposed to simpler features (although it can be not very easy to establish in some cases).

2.2. Features as independent storage units in visual long-term memory

Article selected for the defense: Utochkin & Brady (2020a)

Whereas the concept of the preattentive feature implies that some meaningful discrimination about objects can made early and in parallel without the need of deep, attentional processing, the paper by Utochkin and Brady (2020) focuses on a different aspect of object processing, namely encoding and storage in long-term memory. If

preattention deals with features or roughly segmented “proto-objects”, one might intuit that items getting attention are certainly represented as whole objects and that this is a way they are further represented in working and long-term memory. In fact, however, the format of memory representations - whether it is object-based or feature-based - is a debated topic. This debate is particularly strong in the visual working memory community (e.g., Fougnie, Marois, & Asplund, 2010; Luck & Vogel, 1997; Wang et al., 2017). However, this question is also relevant for long-term memory and our study addresses this question.

Brady and his colleagues (Brady et al., 2013) have suggested an original paradigm to test whether long-term memories are unitized (object-based) or independent (feature-based). They showed observers with a long string of pictures of real-world objects from a variety of categories (a mug, a car, a backpack, etc.), one object per category. The observers then had to do a recognition memory test choosing the exact picture they have seen among four items. All four items were objects of the same category and independently varied in two feature dimensions. These dimensions could be color and state of an object (e.g., *red mug filled with coffee* vs. *red mug empty* vs. *green mug filled with coffee* vs. *green mug empty*) or exemplar and state (e.g., *backpack A closed* vs. *backpack A open* vs. *backpack B closed* vs. *backpack B open*). Note that these features are not necessarily basic visual features, like those discussed in Wolfe and Utochkin (2019) (see section 1.1). Rather, these are parts of object appearance that can be carefully manipulated in an experiment preserving the meaningfulness of an image. Measuring the probability of correct recognition for each feature separately at different time delay between the study phase and test phase, Brady et al. (2013) found that these features were forgotten at different rates (e.g., loss in correctly remembered colors increased faster with time than loss in correctly remembered states). This led Brady et al. (2013) to conclude that objects are in fact not stored as unitized entities in memory. This conclusion, however, was challenged in another recent work using a similar paradigm (Balaban et al., 2020).

In our current work (Utochkin & Brady, 2020), we tested a stronger hypothesis about the representational format of long-term memory for objects. Whereas Brady et al. (2013) were interested in feature forgetting, our major prediction concerned a situation

when the features are present in memory. Even when people are reasonably good at remembering object features, they can be worse at *binding* these features together, that is, they do not remember which exact features belonged to which object. We tested this prediction in Experiments 1A (n = 20) and 1B (n = 19). The only difference between these two experiments was the use of verbal interference, a method diminishing verbal rehearsal of visual stimuli during the study phase: Verbal interference was absent in Experiment 1A and present in Experiment 1B.

In Experiments 1A and 1B, we used a stimulus set from Brady et al. (2013) with varying exemplars and states of real-world objects. The experiments consisted of two tasks presented in separate blocks, Exemplar-State task and Exemplar task. Each task consisted of a study phase and a test phase. In the study phase of the Exemplar-State task, participants were sequentially shown 240 images of real-world objects which the participants were instructed to remember in as much detail as possible. These 240 images included 120 object categories with two exemplars of each category. Critically, for half of categories the exemplars were presented in the same state (e.g., *mug A filled with coffee, mug B filled with coffee*), whereas for the other half they were presented in different states (e.g., *toolkit A open* and *toolkit B closed*). In the test phase of Exemplar-State, recognition memory was tested. In each test trial, both exemplars from the same category were presented in both possible states (four images in total). The participants had to choose a correct state for each exemplar. We were interested in whether participants correctly recalled that states had been same or different (state memory) and whether they recalled which exemplar each state went with (conjunction memory). In the study phase of the Exemplar task, participants were shown 120 new object categories, 2 exemplars in each (stimulus set from Konkle et al., 2010 was used). In the test phase, two old and two new exemplars of the same category were presented, so that participants had to recognize two old exemplars (exemplar memory). Therefore, we obtained separate measures of memory for both types of features (exemplar memory and state memory) and for their conjunctions.

We found that participants showed reasonably good memory for exemplars (80% correct) and that they discriminated whether the exemplars of a given category had been shown in same or in different states (67% choices of two same states for exemplars shown

in same states vs. 36% choices of two same states for exemplars shown in different states). Overall, participants reported 74% correct state-exemplar conjunctions for categories with exemplars shown in same states but their performance dropped to chance level (53% correct) for categories with exemplars shown in different states. This pattern was basically replicated in Experiment 1B, though with overall worse performance which is explainable given concurrent verbal interference. We interpret this pattern of results as evidence for relatively independent storage of exemplar features and state features of real-world objects. Indeed, we found that, despite relatively preserved memory for exemplars and states separately, participants were at chance reporting conjunctions, as the "different-states" test trials showed. Note that good conjunction memory in the "same-state" condition does not require binding at all, as it can be accomplished simply by remembering the correct state without any care about exemplars.

The idea of independent feature storage is two-sided. On one hand, if features are independent it is difficult to bind them when needed (which we tested in Experiments 1A and 1B). On the other hand, it should be easy to "unbind" the features when the task requires to recognize them separately. We addressed this prediction in Experiments 2A, 2B, and 2C.

In Experiment 2A (n = 20), we tested the aforementioned prediction for exemplar and state features, likewise Experiments 1A and 1B. In the study phase, participants were shown 120 images, each was one exemplar in one state from different categories (e.g., only *toolkit A* ***open***, only *coffee mug B* ***empty***). In each trial of the test phase, an old exemplar was shown paired with a new exemplar from the same category. The task was to choose the old exemplar. In different test conditions, we manipulated interference from state information. The Baseline condition included an old exemplar in an old state against a new exemplar in an old state (*toolkit A* ***open*** vs. *toolkit B* ***open***), which is a standard recognition task. In the Generalized condition, we showed an old exemplar in a new state and a new exemplar in a new state (*toolkit A* ***closed*** vs. *toolkit B* ***closed***), which was to test how observers recognize an exemplar when state is changed. In the Misleading condition, we showed an old exemplar in a new state and a new exemplar in a old state (*toolkit A* ***closed*** vs. *toolkit B* ***open***), that is, not only the correct exemplar changed its

state but the incorrect exemplar acquired a familiar state.

We found that observers were good at exemplar recognition in the baseline condition (80% correct) and that the generalized and the misleading conditions yielded just slightly lower performance (78% and 74% respectively) that were found non-significantly different from the baseline. Suspecting that some small effects could be missed due to insufficient statistical power, we ran Experiment 2B, an online replication of Experiment 2A with a larger sample (n = 100). Average percentages of correct answers were very close to Experiment 2A, yet they were found significant due to greater statistical power. Therefore, there was a slight interfering effect of state manipulations on exemplar recognition. However, the absolute differences between the conditions are noteworthy. They were small (loss of -4 to -7% in percent correct). This is a substantially different scale of effect than in Experiments 1A and 1B with near-chance performance in exemplar-state reports. Therefore, recognizing exemplar features unbound from state change is easier than recognizing two features bound together.

As we obtained evidence for some small interference of state information with exemplar information in Experiment 2B, we wanted to show how performance would look if participants had to report features that are hard to unbind. These are so called integral feature dimensions, properties that can change independently physically but affect one another in perception. One example of integral dimensions are hue and luminance, two components of perceived color (Garner & Felfoldy, 1970). In Experiment 2C (n = 100, online) we used the same paradigm as in Experiments 2A-B but manipulated hue and luminance instead of exemplars and states. In the study phase, participants were shown 30 silhouette pictograms of real-world objects, each being “bright” or “dark” in luminance and having a random hue from the CIE Lab color space. In the test phase, participants were shown two copies of the same silhouette in two different luminances (one old, one new) and had to recognize the old luminances - thus, luminance was treated in analogy with exemplars in Experiments 2A-B. Hues were manipulated in analogy with states in Experiments 2A-B (Baseline, Generalized, and Misleading conditions). Overall, we found quite high performance in the baseline condition comparable with that in Experiments 2A-B (78% correct in Experiment 2C vs. 80% in Experiment 2B) but

interference from manipulated hues in the generalized and misleading conditions was much greater than in Experiment 2B (respectively, -13% and -19% correct in Experiment 2C vs. -3% and -7% correct in Experiment 2B). Therefore, even despite a much shorter list of learned stimuli, observers in Experiment 2C showed far stronger interference caused by an irrelevant integral dimension than did observers in Experiments 2A-B. This suggests that exemplar and state features are indeed less integral than hue and luminance and that one can be recognized relatively independently from another.

Overall, our experiments show that semantically meaningful features of real-world objects (which are not necessarily the same thing as preattentive features reviewed by Wolfe & Utochkin, 2019) can be stored independently from one another. An important novel addition to the previous data (Brady et al., 2013) is that we demonstrate that people can successfully remember the features but still fail to bind them. Our most recent data (Markov, Utochkin, & Brady, 2021) suggest that this works not only for long-term memory but for working memory as well. Real-world objects are often thought of as natural representational units of our everyday perception and memory (Bastin et al., 2019; Scholl, 2001). It is probably impossible to introspect otherwise: How can one imagine such property as "state" separately (pure "openness" or "coffee-fullness")? Yet, our data show that an introspectively unified object representation can be in fact reconstructed from its features or from features of different objects. We also discuss the benefits of this relative feature independence: (1) it reduces the computational complexity of the storage, (2) provides a reasonable degree of memory invariance (an object can be recognized when some of its features naturally change), and (3) even when some features are forgotten, we still can retrieve object information using other features.

3. The richness of ensemble perception

Speaking of features (especially preattentive features) as of sparse representations that can provide a lot of useful information about objects before these objects get access to deeper processing, we can ask a question: How can these features be computationally processed simultaneously without the need to individuate each particular feature at each location? In the past 1.5-2 decades, the idea of *ensemble perception* as an ability to extract

summary statistics from large collections of objects has become popular in vision science. Ensemble perception is often thought of as a powerful tool that our visual system exploits to overcome the severely limited bottleneck of attention and working memory in the representation of individual objects (Cowan, 2001; Pylyshyn & Storm, 1998) and to create a rich representation of the world that we see far beyond just a handful of objects (Cohen, Dennett, & Kanwisher, 2016). This coarse ensemble summary is sometimes more useful to give us a "big picture" than the precise representation of a few objects. For example, if you are in a farmer market and you have two baskets filled with tomatoes, which basket would you choose to pick the tomatoes from? If you want ripest tomatoes you need to pay attention to their redness. You can focus on each tomato you see in each of the baskets, but it will probably take a plenty of time, as you need to attend to each tomato in turn. Beside that, you should update the intermediate results of your search in memory and store them to be able to eventually count the number of ripe tomatoes in each basket. This looks like a very time and effort consuming strategy. Instead, you can roughly estimate the mean redness of each basket and compare them. Research in ensemble perception shows that people are indeed good in rapidly judging the mean feature along numerous sensory and perceptual dimensions (such as size, orientation, color, speed, even facial expression, and animacy – see Haberman & Whitney, 2012; Whitney & Leib, 2018 for reviews). Interestingly, the mean feature of a set can be extracted very precisely for a very short viewing time, whereas individual members of exactly the same set are often discriminated at near chance (Ariely, 2001). Therefore, ensemble perception can be superior over individual object perception under short stimulus presentation.

For the long time, the literature on ensemble perception was focused on an ability to summarize multiple objects in the form of their average feature. The average is a first-moment descriptive statistic that can be useful as a compressed single approximate characteristic of all objects that helps deal with the bottleneck issue. Indeed, it is intuitively clear that capturing a single mean feature of many objects would impose less load on the limited-capacity systems of attention and working memory than trying to capture all individual features. This is what researchers mean when they call ensemble

statistics a tool to “compress” the visual input. However, the final goal of data compression is to store more information with less capacity and being able to restore the original data. From this point of view, encoding only the mean feature definitely would make the information about the whole scene “over-compressed”, probably turning the impression of multiple variable objects into the impression of a single “average” item. This is certainly not the case. For example, we do not see an average car moving with an average speed when we look at road traffic; we see a number of cars, we can say that they are different in sizes, colores, and speeds. In order to do that, observers should be able to encode a rich set of ensemble properties not limited to only the average. Indeed, studies show that human observers are capable of reasonably good estimates of some other summary statistics of multiple objects, such as variability (or range or variance - e.g., Dakin & Watt, 1997; Norman, Heywood, & Kentridge, 2015; Solomon, Morgan, & Chubb, 2011) and an approximate number, or numerosity (Burr & Ross, 2008; Chong & Evans, 2011; Halberda, Sires, & Feigenson, 2006). This provides potentially an effective basis for the experienced richness of the world consisting of objects without representing each object (Cohen et al., 2016).

This section summarizes our work that advances the idea of rich ensemble representations that allow observers to know a lot about the multi-object visual environment in a short time. The paper by Khvostov and Utochkin (2019) focuses on the computational capacity and the architecture of the system that supports access to various ensemble properties at the same time. Tiurina and Utochkin (2019) present evidence that the visual system can take into account contextual information for individual objects when summarizes their features in ensemble. Finally, a set of papers (Utochkin, 2015; Utochkin & Yurevich, 2016; Utochkin, Khvostov, & Stakina, 2018) introduces and tests a new theory that ensemble perception goes beyond a limited number of summary statistics and gets access to the whole feature distribution that is used for rapid categorization and segmentation of multiple objects.

3.1. Independent and parallel representation of different ensemble statistics

Article selected for the defense: Khvostov & Utochkin, 2019

While abilities to perceive mean features, feature variance (range), or numerosity

of multiple items have been documented in numerous studies, only few of them have addressed links between these abilities. The basic question asked in these studies is a question about the functional architecture of ensemble perception. Looking at correlations between various ensemble tasks, researchers try to figure out whether there is a common source of variance for all these tasks that can indicate a "general statistician" that does all summary computations. And if there is such a single statistical processor, does it follow rules of regular statistics (e.g., computing mean as a sum of all individual features divided by numerosity)? For example, comparing correlations between accuracies in judging the total size (sum), numerosity, and mean size of multiple circles, Lee, Baek, & Chong (2016) found that the accuracy of averaging is not well predicted by the accuracy of sum or number estimation, which suggests that ensemble perception does not directly follows regular statistical rules (a similar conclusion can be found in Raidvee et al, 2020 who used a different method). Other examples demonstrate the lack of correlation between mean and variance judgments (Yang, Tokita, & Ishiguchi, 2018), mean and numerosity judgments (Utochkin & Vostrikov, 2017) and even limited correlations between averaging in various feature dimensions (Haberman, Brady, & Alvarez, 2015). These demonstrations consistently suggest *independence* between different kinds of ensemble statistical representations and probably no common mechanism to read them out from the same image.

If various ensemble representations show strong independence, how are they accessed together to provide the overall rich impression of the environment? Can several independent statistics of the same set of objects be accessed in parallel and without interference? Or, is each independently computed statistic accessed only in turn? For example, if an observer is focused on the mean color of leaves on a maple in autumn, will he or she be able to estimate the variety of shades of these leaves, and vice versa? This important question, that we term the *parallelism* question, has not gained much attention before, unlike the independence question. The study by Khvostov and Utochkin (2019) systematically addresses both these questions in the same set of tasks. This study follows another study from our group (Utochkin & Vostrikov, 2017) but presents improved methodology and extends the list of tested ensemble properties.

In Experiment 1 (n = 24), we tested the independence and parallelism of reporting the mean size and numerosity of the same set of objects. We employed a *dual-task paradigm* that aims to track the cost of dividing attention between two unrelated tasks simultaneously, which is quite a straightforward measure of parallelism. In each trial, participants were shown a sample set of 7 to 36 differently sized circles for 500 ms. Then they had to report either the mean size of circles, or their number, or both in turn (in a random order) by adjusting the size of a single test circle or a numeric value. Single tasks (report only mean or only number) were used as baseline conditions for performance requiring no division of attention, whereas the dual task (report both statistics in turn) were used to estimate the cost of divided attention compared to the single tasks. The two types of single tasks and the dual task were presented in separate blocks. To test our pattern for independence between the two summaries, we estimated the correlations across participants. Of principal interest were the cross-correlations between the precision of mean reports and the precision of numerosity reports. An important addition to this across-observer correlation analysis was the trial-by-trial correlation analysis. It became possible due to the fact that we asked participants to report both mean size and the number of items within each trial of the dual task. This analysis allowed us to clarify the exact nature of the correlation patterns on the across-observer level. For example, it could distinguish whether the lack of across-observer correlation is caused (1) by the genuinely uncorrelated changes in each single judgment or (2) by swinging reallocation of attention to one statistic at cost of another leading to a negative trial-by-trial correlation but yielding no correlation between data points after they are averaged within each observer.

Our results showed that there was no substantial dual-task cost for either mean reports or number reports compared to the corresponding baseline single tasks, when the order of report was matched to the single task - that is, when we took reports given first in the dual task. We also found no evidence for correlations between these two tasks across observers. Noteworthy, auto-correlations between the same statistics under different conditions (dual vs. single task) were highly correlated showing good consistency of our measurements. The trial-by-trial correlations between mean and numerosity report precision were also negligible in the vast majority of participants

corroborating the idea of their genuine uncorrelatedness. Therefore, our results suggest that mean size and numerosity of the same set of objects are accessible in parallel and that they are likely provided by different mechanisms.

Experiment 2 addressed the link between mean and range judgments. It consisted of two parts, Experiment 2A (n = 16) and 2B (n = 19). Experiment 2A was a prerequisite for Experiment 2B. It tested whether observers can transfer their impression of range from a sample set with one mean size to a test set with another mean size. It was necessary in order to attest that mean and range can be manipulated independently in a dual task, that is, a test display used for the range report can have a mean other than a sample but observers are still able to adjust the range. In each trial, we presented a sample set of 16 circles with a standard mean size and broadly varied size range. It was followed by a test set of 16 circles whose mean size was drawn from the interval -60% to 60% of the standard mean (step 10%) and a randomly preset size range. Using a mouse wheel, observers had to increase or decrease the range to match the range of the sample (note that the mean size of the test was intact during range adjustment). We found in the result that the difference between the mean sizes of the sample and the test sets did not have an effect on the precision of range adjustment. This suggests that observers indeed are able to transfer size range across various mean sizes.

Armed with this core result of Experiment 2A, we have run Experiment 2B that was an exact replication of Experiment 1 but for mean and range, rather than mean and numerosity. Accordingly, instead of using variable numbers of circles we used sets with the fixed number of 16 circles but variable in size range. Participants adjusted the mean on a single test circle (as in Experiment 1) and they adjusted the range on another set of 16 items with the mean randomly differing from the sample (as in Experiment 2A). The results strongly replicated the pattern from Experiment 1: No substantial decline in the dual task compared to the single-task baselines and no evidence of correlation between the precision of mean reports and the precision of range reports both across observers and across trials. Therefore, we conclude that mean and range are likely processed independently and in parallel.

Overall, our study demonstrates that a spectrum of at least basic ensemble statistics

such as mean, range, and numerosity are not accounted for by a common source of variance which indicates the distributed nature of their representation by independent mechanisms. This conclusion is in line with the previous literature. In addition, we show that these statistics are also available in parallel, with no cost for each other. That is, the whole spectrum of ensemble information is extracted from the same set of objects at one time. Both independence and parallelism of these computations provide a strong basis for rather rich and elaborated representation of multiple objects in a scene without the need to deeply process each individual object.

3.2. Contextual features are taken into account in the computation of ensemble statistics

Article selected for the defense: Tiurina & Utochkin (2019)

Our previous analysis of rich ensemble representations concerned experimental situations where sets of multiple objects were isolated from any other variable cues that naturally persist in real perception. These cues are very important because they inform an observer about the environmental context which a current stimulus is presented in. One obvious example of strong context dependence is perceptual constancy, an ability to rescale a retinal image of an object in accordance with contextual variables which allows perceive the physical properties of an object relatively invariant of changing observation conditions. For example, an object at different distances from the observer will have different retinal sizes (visual angles) but distance and depth cues can compensate for the variations of visual angle, so the perceived size will not be strongly affected (Holway & Boring, 1941).

In this study, we asked whether context information can be efficiently taken into account when the features of multiple objects are summarized as an ensemble. Since multiple objects are normally presented under conditions varying from item to item, can the visual system build an ensemble representation rescaled for these variable conditions? For example, cars on a road can be at different distances from an observer, so a correct estimate of their mean size or mean speed would require rescaling angular sizes and speeds in accordance with different distances. Moreover, simply knowing how far the

cars are in general (e.g. in the form of an average distance) is not enough to correctly compute their mean sizes and speeds. The visual system should know which distance each angular size or speed goes with, which implies feature binding. The combinatorial complexity of a stimulus with multiple angular sizes (or speeds) and multiple distances imposes an extremely high computational load (Tsotsos, 1988). To remind, prominent theories such as FIT suggest that such complexity can be dealt with by focusing attention on individual items in turn (Treisman & Gelade, 1980). We tested if this can be accomplished for a large set of objects simultaneously, when an ensemble summary is extracted.

In Experiment 1 (n = 30), we tested whether observers can rescale the mean size of multiple objects in accordance with an apparent distance. We varied the apparent distance using mirror stereoscopes that made things look farther or closer by manipulating binocular disparity, i.e. horizontal displacement relative to the bifixation point, between the left-eye and the right-eye images. In each trial, observers first fixated a cross in the center of the visual field, which set a reference point with zero disparity. Then the observers were shown a sample set of differently sized circles for 1000 ms. This set was presented in one of three apparent planes: Foreground (a plane in front of fixation), Middle (the plane of fixation), or Background (a plane behind fixation). The sample set was followed by a test circle always presented in the center of the middle plane. Observers had to adjust the size of the test to match the mean size of the sample set. We found in the result that the observers tended to systematically overestimate mean sizes of sets presented on the background compared to the middle and foreground sets. To remind, the background plane simulated a far distance where the stimulus with a given angular size is perceived larger than the stimulus with the same angular size presented closer to an observer. The observed mean size overestimation is in this direction. We did not find a symmetrical underestimation of mean size in the foreground, which we explain by some asymmetries in the disparity-distance space and by the noisiness of the mean representation itself (compared to representations of individual, fully attended sizes). Therefore, we take our results as rough evidence that mean size is rescaled.

In Experiment 2 (n = 26), we tested whether the rescaled mean size takes into

account distances to individual items. Alternatively, the observers could estimate the mean angular size of all items and the mean distance to all items and multiply the mean size by the mean distance. The difference between these two hypothetical mechanisms is that the former one requires "knowing" how individual sizes are combined with individual distances before averaging the sizes, whereas the latter one does not require such knowledge and combines two average features after computing each of them separately. In order to distinguish between these two possibilities, we designed our stimuli so that we could manipulate size-distance conjunctions keeping both the mean size and the mean depth unchanged. This was provided by different size-distance correlations. We distributed eight differently circles in four depth planes having gradually increasing apparent distances from the observer. Two circles were presented in each plane. In the Positive correlation condition, the greater were the angular sizes, the farther they were located in apparent distance. This made small circles look even smaller and large circles even larger if they could be rescaled correctly. In the Negative correlation condition, the greater were the angular sizes, the closer they were located in apparent distance. This made small circles look larger and large circles look smaller if they could be rescaled correctly. Therefore, if observers were able to correctly rescale individual sizes, then the whole set of circles would look more diverse in the positive size-distance correlation and less diverse in the negative correlation, although angular sizes and apparent distances are the same. We also used a control condition in which all circles were shown in the plane of fixation. As in Experiment 1, observers had to adjust the mean size of a sample set on a single test circle shown in the center of the fixation plane.

Since angular sizes were generated from the same distributions across the conditions and depth planes were also the same, the apparent mean sizes did not differ between the conditions. If so, how could averaging performance tell anything about rescaling individual sizes? We used an interesting property of visual averaging, namely the *range effect*. The range effect is the growth of error (imprecision) in reporting the mean of a set if the diversity (range) of individual features also grows. As different size-distance correlations in our experiment induced differences in the apparent range of sizes, we expected that the absolute error in adjusting the mean size would also change as a

function of these correlations. For example, it can be predicted that the positive correlation should cause a greater error because it exaggerates the range of apparent sizes. In contrast, the negative correlation reduces the apparent range, so the error should tend to diminish. This is basically what we have found in the result. Indeed, the absolute error in reporting the mean size in the negative correlation condition was smaller than that in the positive correlation condition. Although the intermediate, control condition did not show any significant difference from either the positive or negative correlation conditions (probably due to relatively small increments between apparent ranges), the direction of the effect fit the predictions from the direction of the apparent range changes. Hence, it turns out that our observers did see different apparent ranges in different size-distance correlations. This suggests that the observers were able to rescale individual items correctly based on corresponding depth cues.

Overall, this study (Experiment 2 in particular) provides another powerful demonstration of rich ensemble information that people can extract from a scene. Even when observers do not need to know anything about individual objects and have to report only their average feature, they still take a lot of information about the distribution of objects (e.g. the range) and, critically, about contexts which the objects go with. Therefore, quite elaborated visual representations of multiple objects can exist beyond the well documented severe capacity limits for processing individual objects (Cowan, 2001; Scholl, 2004).

3.3. Ensemble representations as a basis for rapid categorization and segmentation

Articles selected for the defense: Utochkin (2015), Utochkin & Yurevich (2016), Utochkin, Khvostov, & Stakina (2018)

So far, we discussed ensemble representations as a tool to capture the summary information about the whole set of visible objects. However, summarizing everything that falls into one's field of view with single set of summaries (e.g., single mean, single variance) is not always useful. Imagine berry picking. A person is looking at several bushes and has to decide which bush has the ripest berries. This can be accomplished quickly by comparing the average "redness" of the berries on each bush and choosing the

bush with the biggest mean redness. However, before the picker can estimate the average redness of the berries, how does he or she know which objects are exactly berries? In fact, they are typically intermixed with leaves that definitely will not help determine the average redness: Leaves are usually green and more numerous than berries, so they will interfere with averaging of berry colors. The question is again, can we discriminate all berries from all leaves any way other than via the limited-capacity bottleneck of deep processing to recognize objects (e.g., Wolfe et al., 2011). Can we judge ensemble properties of different object kinds independently at all, if their number definitely exceeds the capacity of the hypothetical bottleneck?

Rapid ensemble-based categorization framework. In Utochkin (2015), I put forward a theory of *rapid visual categorization (or segmentation) of multiple objects*. The core idea of this theory is that rich ensemble representations can be used for rapid categorization without any need in the recognition of each individual object. In other words, the visual system reads out the distributional properties of a visual ensemble to make a decision whether the observer sees a bunch of objects belonging to one category or belonging to several different categories. To illustrate the difference between these two cases, I suggest looking at two real-world images of objects on Figure 1a). Both images have approximately the same range of various colors distributed among objects. In one case (Figure 1a, upper panel), even if the image is shown only briefly, a reader will see only one category of fall leaves. In another case (Figure 1a, bottom panel), the reader will clearly see two categories: green leaves and yellow lemons. If we look at how the physical distributions of hues look in these two cases (Figure 1b), we will see a clear difference in shapes. For one category (Figure 1b, upper panel), it looks like a single-peak distribution (something resembling a standard Gaussian or a uniform distribution); for different categories (Figure 1b, bottom panel), the distribution is rather two-peak. If we half-split these physical distributions, each half-set of hues will contain a random subset of leaves for the upper picture; and there will be a perfect categorical split for the bottom picture (either the leaves, or the lemons). From other behavioral and physiological evidence, we know that the visual system is somehow capable of reproducing the shape of the physical distribution of ensemble attributes (Chetverikov, Campana, & Kristjansson, 2017a, 2017b;

Treue, Hol, & Rauber, 2001) that can be subsequently used to decide whether all objects are drawn from a same or from different categories. If features are clustered around several local averages and there is a large gap between the clusters, then these clusters are likely to represent categorically separable sets of features (Figure 1c, right panel). Elsewhere, I term such distributions "segmentable" (Utochkin, Khvostov, Stakina, 2018; Utochkin, Yurevich, 2016). If the values are distributed more evenly within the same range, they are likely to be represented as a single Gaussian corresponding to one category (Figure 1c, left panel), although this category can be sparse, that is, including quite different exemplars (like yellow and green fall leaves). This kind of distribution and representation is called "non-segmentable". In general, the use of the distributional shape allows efficient determining whether all objects belong to the same or different categories (which I termed *primary categorization*) without processing each item separately.

Another important function of rapid categorization is splitting the scene into subsets (even if they are strongly intermixed in space) and break down the subsequent analysis into separate streams. This subsequent analysis termed *secondary categorization* can run in at least two directions: in-depth and in-breadth. *In-depth categorization* implies that, once objects are broken down into several categories by one feature dimension (e.g. by color), one of these categories is selected and another act of categorization is applied to a new feature dimension. For example, if a display consists of multiple red-vertical, red-horizontal, green-vertical, and green-horizontal bars, an observer can firs select all red bars (because red is categorically different from green) and look whether there are more vertical or horizontal bars among this red subset (vertical and horizontal are also highly distinct categories). *In-breadth categorization* implies that the observer compares whether several primary categories are same or different in terms of the other dimension. For example, red and green bars can have lots of different lengths. Primary categorization by distinct color will allow us to estimate whether red bars are longer on average than green one. Thus, even if sizes of all bars taken together are distributed smoothly over a broad range they can be estimated as distinguishing categorical features (the red category is also longer on average) when marked by the primary categorical features, i.e. by colors.

Summarizing, the theory of ensemble-based rapid categorization and segmentation

suggests another example of how elaborated visual impression about the whole scene can be acquired from the rich ensemble representation (here, the shape of a feature distribution) of a single feature dimension (primary categorization) or of a combination of several dimensions (secondary categorization).

Our further work was dedicated to empirical tests of the rapid ensemble-based categorization theory.

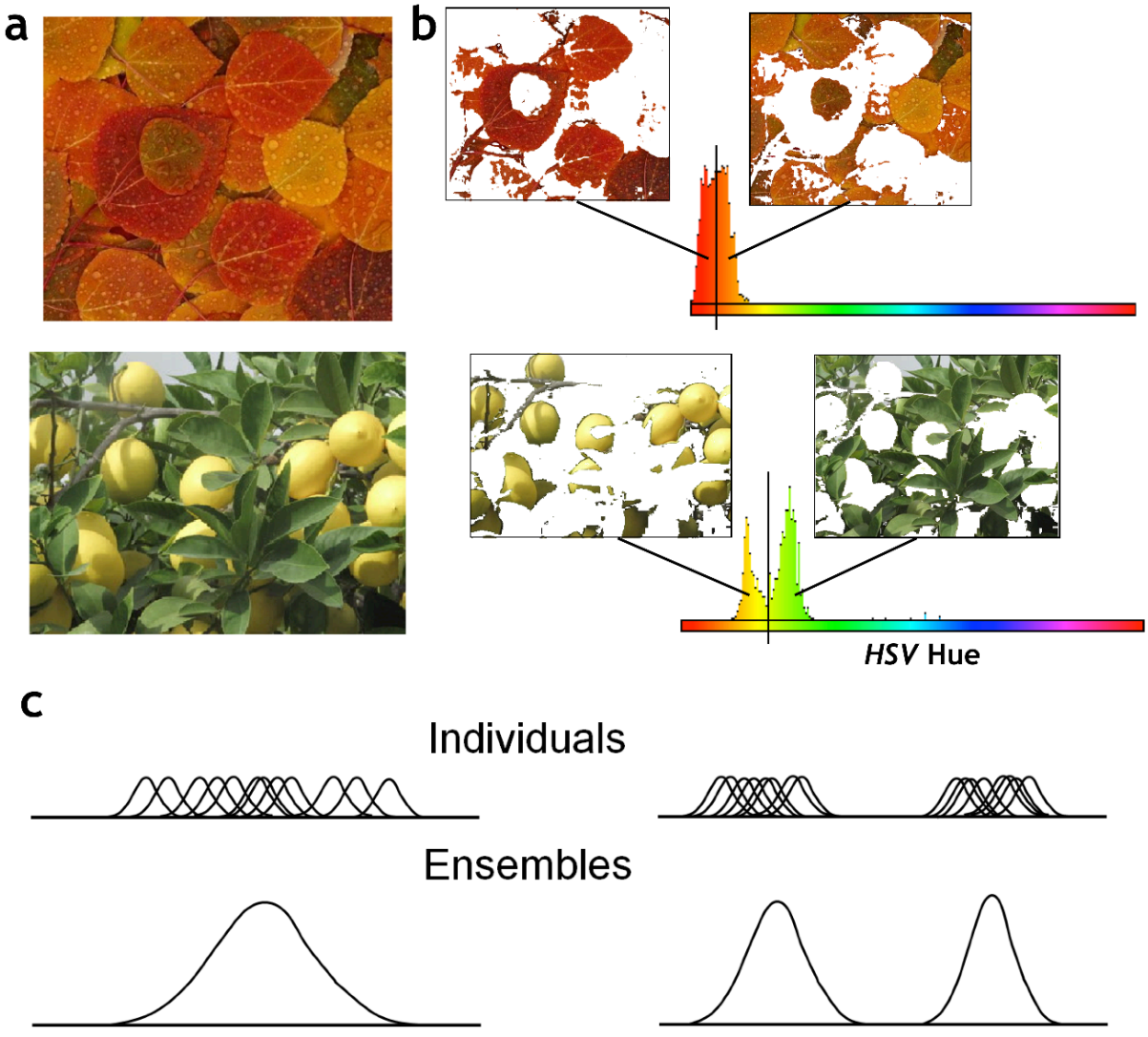

Figure 1. Real-world examples of feature distributions that can be used for rapid categorization in the color domain: (a) original images of multiple objects belonging to one (top panel) or two (bottom panel) categories; (b) the physical distributions of hues (in the HSV color space) with the images half-split along these hue distributions; (c) proposed noisy visual representations of individual items and pooled ensemble

representations for the single-category (left panel) and the two-category (right panel) examples. Adapted from: Utochkin (2015, Figure 1).

Evidence from visual search. To test whether the "segmentable" or "non-segmentable" shape of the feature distribution supports categorical separation or the formation of a single category, we have run three experiments using the visual search paradigm (Utochkin & Yurevich, 2016). In a typical visual search task, observers have to find as quickly as possible a predefined target among a number of other irrelevant items, called distractors or non-targets, or to answer that the target is absent. It is a well established finding that the ease of visual search for a target item strongly depends on distractor similarity, or homogeneity (Duncan & Humphreys, 1989). For example, if one looks for a target (e.g., a line tilted to the right), it is much easier to find it among distractors that are all vertical or all tilted to the left than among vertical and left-tilt distractors. A standard theory behind this pattern is perceptual grouping by similarity that makes all homogeneous distractors easily grouped, attended, and rejected altogether if they do not look like a target (Duncan & Humphreys, 1989). If the distractors are heterogeneous they tend to form a lot of separate groups and are attended serially which makes search more difficult. In terms of ensemble statistics, heterogeneous distractors form a broad-range distribution that makes the whole display looking noisier and the target harder to detect (Rosenholtz, 2001b).

Our hypothesis was that the shape of distractor distribution plays a mediating role in how heterogeneous distractors form groups to be inspected in visual search. More specifically, we hypothesized that, when the range of distractor heterogeneity is carefully controlled, the search will be easier among smoothly (uniformly) distributed than among sharply (bimodally) distributed distractors. This is because the smooth distribution is supposed to form a single continuous ("non-segmentable") ensemble representation corresponding to one category, whereas the sharp distribution is supposed to form several categories that are harder to treat as a single group ("segmentable" distributions).

Experiments 1 (n = 20) and 2 (n = 24) had similar designs but different feature spaces were tested in each. In Experiment 1, participants looked for a size singleton

(either the smallest or the largest circle) among distractors with middle sizes. In Experiment 2, participants looked for a singleton bar tilted either 45º, or 135º among vertical bars and/or bars tilted away from target. distractors could form one of four feature distributions. Two of them were critical to test our hypothesis. They were heterogeneous, that is, they contained distractors of different sizes or orientations. The range of distractor variation was carefully controlled across these two conditions. For example, in Experiment 1, if the target was small (0.7º of visual angle) then distractors covered the range from medium to largest possible (1.1º to 1.8º). In Experiment 2, if the target was tilted by 45º then distractors covered the range from vertical to exactly counter-tilted orientations (90º to 135º). Critically, within these strictly fixed ranges, distractor features could be distributed either sharply or smoothly. If the distribution was sharp (bimodal), then only its extreme values were present (e.g., only 90º and 135º bars) in equal proportions. If the distribution was smooth (uniform), then multiple intermediate feature values were present along with these extremes (e.g., 90º, 95º, 100º, …, 135º), also in equal proportions. The rest two types of the distributions were homogeneous, when all distractors had exactly the same size or orientation. These two conditions served as baselines to establish how well the target feature is distinguished from both boundaries of the distractor range (that is, how well a 45º-target is found when all distractors are 90º or when they all are 135º).

In Experiments 1 and 2, participants were shown displays consisting of 13, 25, or 37 items (often termed *set sizes*). They had to respond as quickly as possible whether the target is present or absent. The target was either a very small, or a very large circle in Experiment 1 (target identity changed unpredictably from trial to trial which also shifted the distractor distribution in a mirror direction). Similarly, in Experiment 2, participants had to find a target tilted away from the distractors (either 45º or 135º). We analyzed the absolute reaction time (RT), which reflected the overall task difficulty, and the slope of the RT function (RT increment as a function of set size), which reflects how many items or groups of items are inspected before attention arrives at the target or before search is terminated.

As a major result, we found in both experiments that the sharp distributions yielded the overall slower RT than the smooth distributions, suggesting that the former task was more difficult. At the same time, we found no effect of set size on the RT, that is, search did not substantially depend on the number of to be inspected items. We interpret this as evidence for global distractor grouping that allowed simultaneous rejection of each group regardless of its size (Duncan & Humphreys, 1989). Then, if the groups are attended and rejected as wholes then the increment in the absolute RT can reflect the number of serially attended and rejected groups. Therefore, the longer RT in the sharp condition can reflect a series of operations with separate ("segmentable") groups.

One argument against interpreting the findings from Experiments 1 and 2 in terms of categorical segmentability might be the fact that overall feature variance could confound with the distribution shape. Indeed, even when the range is fixed, a bimodal distribution has a greater variance than a uniform one. Therefore, the greater variance could explain the greater search difficulty among sharply distributed distractors. We addressed this issue in Experiment 3 (n = 25). It was basically the replication of Experiment 2 with orientation search. We had exactly the same set of distractor distributions but added one more, that we called the sharp transition distribution. It consisted of extreme values within the range (e.g., 90º and 135º) and also included one intermediate orientation half-way between the extremes (112.5º). That is it was an intermediate type between the bimodal and the uniform distributions. Importantly, the step between the orientations that we used to create the intermediate orientation (22.5º) was shown to be effectively discriminated preattentively (Foster & Ward, 1991), so it could support global segmentation. It is easy to see that variance increases monotonously across the distributions in the following sequence: smooth - sharp transition - sharp (bimodal). Therefore, if the variance account is correct, we would also observe the monotonous growth of the RT in accordance with this sequence.

However, the results of Experiment 3 violated this monotonous prediction. In fact, the search among smoothly distributed distractors was the fastest one among the heterogeneous conditions; the search among "extremely" distributed distractors (sharp, bimodal distribution) was slower; and the sharp transition yielded the slowest search.

This non-monotonic pattern could not be predicted by changes in the overall variance. But it can be explained by the internally represented shape of the distractor distribution. Presumably, the sharp and the smooth shapes of orientation distributions corresponded to ensemble representations that were called "segmentable" and "non-segmentable" above. Sharp transitions between the feature values provide the internal distribution with peaks corresponding to each presented value and large gaps between these peaks, which should lead to segmentation of the set into categorically different subsets. So, in extremely sharp distributions we had two peaks leading to the perception of two separate groups; while in sharp distributions with three values we had three peaks leading to the perception of three separate groups. In contrast, smooth transition provided a single-peak broadband internal distribution without large gaps, which leads to the representation of all the items as one sparse group. Each subset is analyzed as a separate chunk and rejected serially, making search among sharply distributed distractors slower (Duncan & Humphreys, 1989). This "segmentability" account was quite good at predicting the observed RT ranking across the conditions.

Evidence for rapid ensemble-based segmentation along two sensory dimensions at the same time. Utochkin and Yurevich (2016) provided evidence for the role of the feature distribution in categorization based on a single dimension (e.g., only size or only orientation). However, in real perception we rarely see ensembles varying along only one dimension. Often, several features are variable in objects at the same time. Here, we tested whether observers can discriminate between ensembles when none of their basic dimensions alone is informative for discrimination but only statistics of their combinations are informative. Critically, we were interested in the role of "segmentability" in our ability to discriminate between such ensembles.

We have run three experiments on texture discrimination, a task that requires to locate the boundary between two surfaces consisting of elements assigned by two different stimulus rules (for example, one surface made of vertical lines and another made of horizontal lines). An ability to correctly locate the boundary suggests that the observers can globally discriminate stimulus properties which underlie this rule. In all our experiments, the rule was always set as a correlation between the length and orientation

of textural elements. We had a square 8x8 grid filled with 64 lines. This square field could be divided by two halves, either horizontal or vertical. Each half of the field consisted of lines whose lengths and orientations have been drawn from exactly the same distributions. That is, neither lengths nor orientations alone could be used to discriminate between the textures placed in these two halves. The difference was defined as the direction (sign) of length-orientation correlation: In one texture, the rule was “the longer - the steeper”, whereas in another texture the rule was “the longer - the shorter”. Therefore, successful texture discrimination could be warranted only if observers could perceive the global variation in both dimensions simultaneously. Like in the study by Utochkin and Yurevich (2016), our key manipulation was the shape of the feature distributions within fixed ranges. We used sharp, bimodal distributions (only longest and shortest lengths, only flattest and steepest orientation) to create “segmentable” features. We used smooth, uniform distributions to create “non-segmentable” features. Segmentability could be manipulated orthogonally across the feature dimensions (that is, length and orientations could be both segmentable within one display, both non-segmentable, or one could be segmentable and another non-segmentable). Figure 2 depicts examples of all four segmentability combinations.

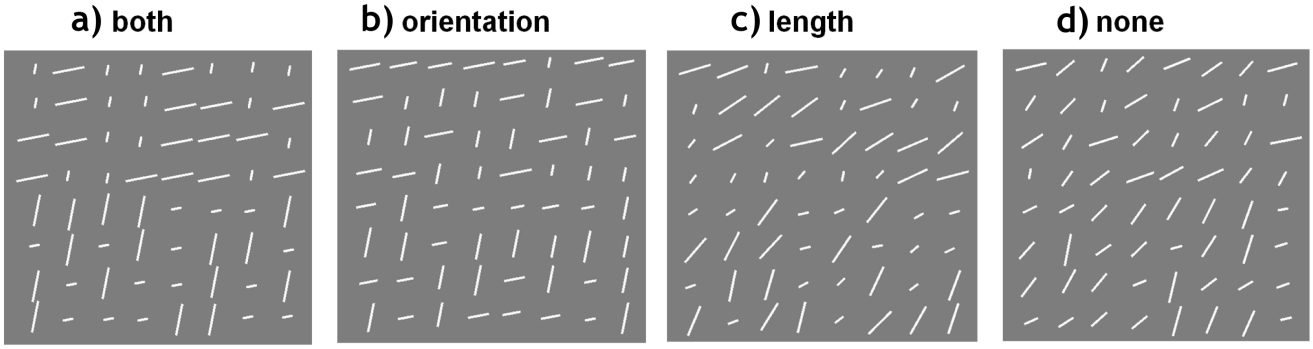

Figure 2. Example stimuli used in the experiments by Utochkin et al. (2018). Here, each square texture is divided by two halves horizontally. The top halves contain lines whose lengths and orientations are correlated with r = -1 (the longer, the flatter), whereas the bottom halves contain lines whose lengths and orientations are correlated with r = +1 (the longer, the steeper). In (a) both length and orientations are segmentable (have sharp bimodal dustributions), in (b) only orientations are segmentable, whereas lengths are not (have smooth uniform distributions), in (c) only lengths are segmentable, and in (d)

neither lengths, nor orientations are segmentable. Adapted from: Utochkin et al. (2018, Figure 4).

In Experiment 1 (n = 5: 4 experienced and 1 naive observer), we tested the texture discrimination as a function of segmentability using the psychophysical method of constant stimuli. Observers were shown the textures, as described above. The textures were presented for 200 ms and then masked. Their task was to answer which patch - top or bottom from the central fixation point - had "longer-steeper" lines. The correlation contrast defined the difference between the patches. For example, if length-orientation correlation in the top patch was +1.0 then it was -1.0 in the bottom patch; if the correlation in the top patch was -.75 then it was +.75 in the bottom patch, etc. Feature distributions in this experiment were either both segmentable, or both non-segmentable (as in Figures 2a and 2d, respectively). We measured the frequency of "top" responses as a function of the signed correlation contrast in each segmentability condition. We then built psychometric functions for each observer and each segmentability condition using logistic regression. The critical parameter we estimated was the σ of the psychometric function that characterizes discrimination (the smaller the σ, the better discrimination). We found that all observers showed substantially smaller σ in the segmentable (σ = 0.35-0.55) compared to the non-segmentable (σ = 0.81-2.5) condition. Therefore, our observers were more sensitive to length-orientation differences between two textures when these textures were segmentable.

In Experiment 2 (n = 21), we also tested texture discrimination as a function of segmentability but we used a different paradigm. Observers were shown the differently correlated texture patches divided by either vertical or horizontal boundary. The correlation contrast between the patches was always +1 vs. -1. The textures were presented for 100, 200, 350, or 500 ms and then masked. These variable durations were used to estimate the speed of texture discrimination. All possible combinations of length and orientation segmentability were tested, as shown in Figure 2. The observer's task was to answer whether a boundary between the textures was vertical or horizontal. Based on these answers we calculate a standard psychophysical discriminability index, $d`$. We

found in the result that in all segmentabilty conditions but the "both segmentable" condition, *d`* were very close to 0 suggesting almost null discriminability of the textures. In contrast, when both distributions were segmentable, they provided reasonable discriminability (*d`* = 0.7-0.8). Interestingly, this advantage seemed to be early: It took up to 200 ms to appear and it was not improved by longer durations. So, it appears that the segmentability of both dimensions supports categorical discrimination between textures and that this discrimination does not benefit from longer sampling.

In Experiment 3 (*n* = 23), that followed-up Experiment 2, we tested whether the results of Experiment 2 were due to global sampling of the whole texture or by local (attentional) sampling of only a few elements near potential boundaries. We removed elements from the corners of our textures and kept elements only near potential boundaries (the configuration looking like a cross). We compared this condition with complete textures, as in Experiment 2. Stimulus duration was fixed at 200 ms. Again, we found that two segmentable distributions supported reasonably good texture discrimination (*d`* = 0.7-0.8) and that discrimination was overall better for complete textures than for those containing only near-boundary elements. This suggests that we need some rather global statistics to perform this discrimination better.

Summarizing, we observed the advantage of displays with both segmentable length and orientation consistently, across three experiments. The explanation is based on the concept of segmentability supporting rapid subset categorization within overall textures. Observers are bad at capturing the whole length-orientation correlation. Instead, they can parse the whole display into subsets using one feature dimension (above, it was termed *primary categorization*). Then, observers can select one of these primary categories and look for differences in a second dimension (*secondary in-depth categorization*). For example, an observer can selectively attend to only long lines and compare different patches of long lines by their average orientation. Good segmentability certainly helps both stages of this strategy. First, it diminishes confusion between the primary subsets (it is much easier to see *very long* lines separately from *very short* lines than to see *longer* lines separately from *shorter* lines). Second, when a primary category is isolated and selected for further processing, the shape of another feature distribution affects the ease

of mean comparison. Obviously, the mean contrast between half-distributions is greater if the original distribution is sharp than when it is smooth. This was corroborated by our last experiment (Experiment 4) where we artificially isolated this hypothetical second stage of texture discrimination.

In Experiment 4 (n = 16) where we tested how people discriminate between the same textures when a half of the elements are removed (for example, a half of longest elements or a half of flattest elements). This artificially simulates ideal primary categorization and subset selection. Here, the only way to discriminate between two patches is comparing mean features along a remaining dimension (mean orientation if half-split is based on length, and vice versa). We found that the shape of this remaining distribution influenced discrimination. Again, sharp distributions raised sensitivity substantially. Additionally, we found that $d`$ in Experiment 4 were overall larger than in Experiments 2 and 3, which suggests that subset splitting of full sets and subset selection is in fact an imperfect, noisy process.

To conclude, rapid categorization and segmentation of multiple objects seems to be a prospective subject where the theory of ensemble perception can be applied to make some working predictions and suggest some idea of potential mechanisms. They show that ensemble representations are functional not only as a compressed summary of many same-type objects (e.g. in the form of the average feature) but also as appropriate grounds for establishing the difference between objects of different types.

Overall, our work collected in this section advocates the idea of rich ensemble representations that convey a lot of information about multiple objects in a scene without any need to deeply process individual objects one by one. Indeed, we showed that various statistical properties of ensembles can be extracted independently from each other and with no substantial attentional cost, that contextual features such as depth cues can be correctly taken into account, and that multiple objects can be efficiently categorized and segmented even if they are spatially intermixed. Of course, ensemble perception is not a substitute for deep attentional processing of individual objects. Ensemble perception is a tool to get the coarse “gist” of a scene but often inappropriate for exact localization,

perception, and recognition of objects. Moreover, ensemble representations are also not free of limitations of different kinds (e.g., Halberda et al., 2006; Im & Chong, 2014; Khvostov et al., 2020). Nevertheless, ensemble representations appear to strongly contribute to the general impression of completeness and stability of perception and to play an important role in various visual tasks.

4. Ensemble statistics shape representations of individual objects in working memory

Article selected for the defense: Utochkin & Brady (2020b).

In the previous section, we showed that ensemble representation can be an efficient means of perceiving large sets of objects when the number of objects obviously exceeds any established limits of focused attention and working memory. In the present study, we move further and demonstrate that ensemble information can be used to organize the limited-capacity processes requiring conscious access to individual object representation. Specifically, we tested the role of ensembles in visual working memory for individual objects.

Working memory (as well as short-term memory) is commonly defined as a limited-capacity system holding a small amount of information most important for an ongoing task (Baddeley, 1986). Working memory capacity is usually measured in structural units, or chunks whose number is severely limited (Cowan, 2001). For visual working memory, an individual object is considered to be a natural chunk (although it is a highly debated question whether objects can be further subdivided into smaller independent entities – features, see Section 1.2 for details of this debate). This intuitively appealing idea of objects as chunks is reflected in experimental approaches to study visual working memory and in theories. For example, in a typical visual working memory experiment, researchers often manipulate memory load showing different set sizes of objects and varying their features. Researchers believe that, when they simply manipulate the number of individual items as experimental conditions, they can estimate some important parameters such as the probability of recall and recall precision (e.g., Zhang & Luck, 2008). Importantly, by doing so, they assume that these parameters reflect internal

representations of items in working memory (the number of objects stored in memory and representational fidelity). In other words, these items are assumed to be encoded, stored, or forgotten as separate units independent from other items. However, we know from the literature that individual items are probably not that independent. For example, they can be affected by the ensemble information about the remembered items altogether, which makes observers recall individual features shifted to the average feature of all remembered items (e.g., Brady & Alvarez, 2011).

In our study, we present new evidence that critical parameters of visual working memory representations (in particular, precision that is so important for many models) in fact inherit those of ensembles. This study is a follow-up of the study by Brady and Alvarez (2011) that demonstrates that not only mean feature but also range affect observers' reports of individual items.

In Experiment 1 (n = 16), we had participants perform visual working memory tasks with orientations of simple objects, triangles. In different blocks, they had to remember one orientation, four orientations, or the average (or the ensemble) orientation of four triangles. These were three tasks of interest in this experiment. In a typical trial, the participants were first shown a precue indicating the location of one particular item (Remember 1 task), or all four items (Remember 4 task), or the average item (Remember Mean task). The precue was followed by a sample stimulus with four differently oriented isosceles triangles for 300 ms. Observers had to maintain the triangles in working memory for 1000-ms (blank screen) and then they had to adjust the orientation of a test triangle to match the orientation of either an individual item, or the average. We had three ranges of orientations of the sample stimulus: 30º, 60º, and 120º, which provided their "goodness" of organization into ensemble.

We analysed the error distributions in the 360º circular orientation space. The space was centered on the correct answer (0º error) and errors were plotted as signed (clockwise or counterclockwise) angular differences from 0º (that is, Error = Observer's answer - Correct answer). Note that for individual items (that is, Remember One and Remember Four tasks), we plotted the error in a way such that a positive error was always towards an ensemble average and a negative error was always away from the mean, which was

done in order to estimate whether there is a systematic bias of the error distribution towards or away the mean orientation. In the Remember One task, the error distribution was always narrow (small error) and centered at 0º (unbiased) regardless of the overall range. But in the other two tasks, the distributions clearly became wider as the stimulus range increased. Importantly, the circular deviation of the distributions (which is inverse precision) grew as a function of range in the ensemble task (typical *range effect*, as described above in section 3.2), and so did the deviation in the "Remember Four" task. Moreover, in the "Remember Four" task, the distributions were positively biased, that is, they were biased toward mean. So, there was a strong resemblance between what people did in the ensemble task and in the "Remember Four" task.

One might argue that this resemblance might be nothing special but just doing the averaging task instead of the "Remember Four", because remembering the average is easier than trying to remember four insdividual orientations. To test this, we simulated how the "Remember Four" distributions would look if observers just averaged. The simulations were based on what we know about the distributions in the averaging task. The simulated distributions were progressively more biased than the data as a function of range. In fact, in the smallest range (30º) the simulated and actual distributions were almost the same, but they were substantially more different in the other two ranges. Interestingly, although the absolute bias away from the correct answer grew as a function of range, it decreased as a proportion of the distance between the correct answer and the mean. That is, when the range increases individual responses are relatively less biased, which suggests that the observers did not rely simply on averaging, especially in large ranges.

Could the resemblance between the ensemble task and the "Remember Four" task found in Experiment 1 be due to the fact that our observers were exposed to both these tasks? Specifically, could experience in the "Remember Mean" task cause the explicit transfer of the strategy to the "Remember Four" task? We tested this possibility in Experiment 2 (n = 16), where participants had to complete only the "Remember Four" task. We had the same three ranges of orientations as in Experiment 1. In order to further encourage remembering individual objects we added filler trials with a range of 360°

where no "averageable" ensemble information is available. Basically, the results replicated those of Experiment 1 regarding the "Remember Four" task. This suggests that the observers reproduced some of ensemble properties even when they did not have an idea of averaging as an explicit strategy.

So far, we tested the influence of ensemble on memory for orientations that were intentionally organized to produce an ensemble with a certain range. But in typical working memory experiments, individual features are usually generated randomly and, thus, more independently. Would the ensemble effect be preserved in this case? We tested it in Experiment 3 (n = 296, online via Amazon Mechanical Turk). We pregenerated 48 displays of three randomly oriented triangles and showed them 296 MTurk workers, so each display yielded a good error distribution with 296 estimates. The displays covered a very broad variety of ranges from 24º to 180º. We find a high correlation between the physical range of these randomly generated orientations and the error measure, circular deviation (r = .72, p < .001). We also found the positive bias in error distribution (bias toward mean) and that this bias in proportion to the physical range decreased as a function of that range (r = -.61, p < .001). It suggests that the observers tended to less rely on reporting just the average when the range was larger. Both the error (angular deviation) and bias in Experiment replicated the patterns of Experiments 1 and 2.

To conclude, we see that recall precision of individual items strongly correlates with the precision of ensemble representation in visual working memory, and that individual reports are biased toward the ensemble mean. The more similar the items in an ensemble, the more they are biased. These findings are consistent with the *hierarchical encoding* framework earlier introduced by Brady and Alvarez (2011; Brady, Konkle, & Alvarez, 2011) suggesting that information in working memory is stored and retrieved at different levels of abstraction (e.g. as a combination of individual and ensemble representations). The information stored at different levels and in different formats can be efficiently combined for the optimization of an ongoing task.

5. Conclusion

The studies collected in this dissertation concerned various aspects of one basic question: What kind of representations and how are used to provide us with the

impression of easily seeing and remembering tens and hundreds of objects? The discussion of this question was built around three representational formats: features, objects, and ensembles.

First, we investigated the relationship between features and objects. Here, we reviewed the concept of "preattentive" feature and found that, although it is not always easy to empirically diagnose such features, they are real in a sense that they guide the selection of potentially relevant information for further deep attentional object processing and, thus support an impression of quickly available objects. Moreover, we then showed that even fully attended and encoded real-world objects are in fact stored in a set of relatively independent functionally meaningful features (such as exemplar and state features) in long-term memory. Therefore, we showed that relatively shallow feature representation sometimes can be efficient proxies of fine and accurate object perception and memory.

Second, we demonstrated that ensemble representations of multiple objects and their features in a form of various statistical properties can convey rich information about the whole set of objects without any need of knowing everything about each individual object. We showed that various ensemble summary statistics representing different aspects of the set (mean, range, numerosity) can be extracted independently from each other and in parallel. Moreover, when summarizing ensemble information along one sensory dimension (e.g. size), the visual system takes into account contextual information about variations in other dimensions (e.g., distance) to rescale the resultant summary accordingly. Finally, we demonstrated that ensemble representations can be used not only to summarize all visible items but for parsing them into categorically distinct subsets when these items are very different. Third, we showed that, apart from being an efficient tool for representing all objects or features together, ensemble statistics can be used to organize representations of individual objects when deeper encoding and independent storage of each is required within the limited-capacity system of working memory. Overall, features, objects, and ensembles can be co-existing representational formats reflecting the hierarchical organization of the visual system. These representational formats flexibly interact to form the visual image of the world in perception and memory.

References

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12 (2), 157–162.

Baddeley, A. D. (1986). *Working memory.* Oxford, UK: Clarendon Press.

Balaban, H., Assaf, D., Arad Meir, M., & Luria, R. (2020). Different features of real-world objects are represented in a dependent manner in long-term memory. *Journal of Experimental Psychology: General, 149*(7), 1275-1293.

Bastin, C., Besson, G., Simon, J., Delhaye, E., Geurten, M., Willems, S., & Salmon, E. (2019) An integrative memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences, 42,* e281: 1–60.

Brady, T. F. and 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. & 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., Konkle, T, & Alvarez, G.A. (2011). A review of visual memory capacity: Beyond individual items and towards structured representations. *Journal of Vision*, *11*(5):4, 1-34.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA*, 105 (38), 14325-14329.

Brady, T. F., Konkle, T., Alvarez, G.A., & Oliva, A. (2013). Real-world objects are not represented as bound units: Independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, 142(3), 791-808.

Chetverikov, A., Campana, G., & Kristjánsson, A. (2017a). Representing color ensembles. *Psychological Science*, *28*(10), 1510 – 1517.

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2017b). Learning features in a complex and changing environment: A distribution-based framework for visual attention and vision in general. *Progress in Brain Research*, 1–24.

Chong, S. C., & Evans, K. K. (2011). Distributed versus focused attention (count vs estimate): Distributed versus focused attention. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2 (6), 634–638.

Cohen, M. A., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences, 20*, 324-335.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Research*, 37 (22), 3181–3192.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433-458.

Eastwood, J.D., Smilek, D., & Merikle, P.M. (2001). Differential attentional guidance by unattended faces expressing positive and negative emotion. *Perception and Psychophysics, 63*, 1004-1013.

Egeth, H. E., Virzi, R. A., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 32–39.

Foster, D. H., & Ward, P. A. (1991). Asymmetries in oriented-line detection indicate two orthogonal filters in early vision. *Proceedings of the Royal Society London: Series B, 243,* 75–81.

Fougnie, D., Asplund, C. L., & Marois, R. (2010). What are the units of storage in visual working memory? *Journal of Vision*, *10*(12), 27–27.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1,* 225-241.

Haberman, J., & Whitney, D. (2012). Ensemble perception: Summarizing the scene and broadening the limits of visual processing. In J. Wolfe and L. Robertson (Eds.),

From Perception to Consciousness: Searching with Anne Treisman. Oxford University Press, 339-349.

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.

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science, 17,* 572-576.

Hochstein, S. & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron, 36,* 791-804.

Holway, A. H., & Boring, E. G. (1941). Determinants of apparent visual size with distance variant. *American Journal of Psychology, 54*(1), 21–37.

Im, H.Y., & Chong, S.C. (2014). Mean size as a unit of visual working memory. *Perception*, 43, 663-676.

Inverso, M., Sun, P., Chubb, C., Wright, C.E., & Sperling, G. (2016). Evidence against global attention filters selective for absolute bar-orientation in human vision. *Attention, Perception, & Psychophysics, 78,* 293.

Joseph, J.S., Chun, M.M., & Nakayama, K. (1997). Attentional requirements in a "preattentive" feature search task. *Nature, 387,* 805-807.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3), 558-78.

Khvostov, V.A., Markov, Y.A., Brady, T.F., & Utochkin, I.S. (2020). Limitations on animacy categorization in ensemble perception. *PsyArXiv.* https://psyarxiv.com/d4za6/

Lee, H., Baek, J., & Chong, S. C. (2016). Perceived magnitude of visual displays: Area, numerosity, and mean size. *Journal of Vision*, 16 (3): 12, 1–11.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279-281.

Mack, A. & Rock, I. (1998). *Inattentional Blindness.* Cambridge, MA: MIT Press.

Markov, Y.A., Utochkin, I. S., & Brady T. F. (2021). Real-world objects are not stored in holistic representations in visual working memory. *Journal of Vision, 21*(3): 18, 1–24.

Neisser, U. (1967). *Cognitive Psychology.* Appleton-Century-Crofts.

Noë, A. (2002). Is the visual world a grand illusion? *Journal of Consciousness Studies, 9*(5-6), 1–12.

Norman, L. J., Heywood, C. A., & Kentridge, R. W. (2015). Direct encoding of orientation variance in the visual system. *Journal of Vision*, 15 (4): 3, 1–14.

Pylyshyn, Z. W. & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3,* 179 – 197.

Raidvee, A., Toom, M., Averin, K., & Allik, J. (2020). Perception of means, sums, and areas. *Attention, Perception, and Psychophysics, 82,* 865–876.

Rensink, R. (2002). Change detection. *Annual Review of Psychology, 53,* 245-277.

Rosenholtz, R. (2001a). Search asymmetries? What search asymmetries? *Perception and Psychophysics, 63,* 476–489.

Rosenholtz, R. (2001b). Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search. *Journal of Experimental Psychology: Human Perception and Performance, 27*(4), 985-999.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1/2), 1-46.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, 11 (12): 13, 1–11.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25,* 207–222.

Treisman, A. & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General, 114,* 285-310.

Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology, 6,* 171–178.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition, 14*(4-8), 411-443.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* 97-136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*(1), 15–48.

Treue, S., Hol, K., & Rauber, H. J. (2000). Seeing multiple directions of motion – physiology and psychophysics. *Nature Neuroscience, 3,* 270–276.

Tsotsos, J. K. (1988). A `complexity level' analysis of immediate vision. *International Journal of Computer Vision, 2,* 303 – 320.

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.

Wang, B., Cao, X., Theeuwes, J., Olivers, C. N. L., & Wang, Z. (2017). Separate capacities for storing different features in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(2), 226–236.

Whitney, D. & Leib, A. (2018). Ensemble perception. *Annual Review of Psychology, 69,* 105-129.

Wolfe, J. M., Vo, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in Cognitive Sciences*, *15*(2), 77-84.

Wolfe, J.M., Cave, K. R., & Franzel, S.L. (1989). Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance,* 15(3), 419-433.

Wolfe J.M., & Utochkin I.S. (2019). What is a preattentive feature? *Current Opinion in Psychology,* 29, 19-26.

Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is there a common summary statistical process for representing the mean and variance? A study using illustrations of familiar items. i-Perception, 9 (1), 204166951774729.

Zeki, S.M. (1978). Functional specialisation in the visual cortex of the rhesus monkey. *Nature, 274,* 423-428.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Zhaoping, L. (2016). From the optic tectum to the primary visual cortex: Migration through evolution of the saliency map for exogenous attentional guidance. *Current Opinion in Neurobiology, 40,* 94-102.