Coding Strategies in Memory for 3D Objects: The Influence of Task Uncertainty

Christopher J. Bates (cjbates@g.harvard.edu)

Samuel J. Gershman (gershman@fas.harvard.edu)
Department of Psychology, Harvard University

Abstract

Memory is limited in capacity, which means that we must choose what information to prioritize for storage. Part of knowing what to prioritize is predicting future needs. For example, if you view a 3D object, later on you may wish to recall exactly how it was oriented. Alternatively, you might need to remember its shape, independent of viewpoint. Given this kind of uncertainty, a good strategy would be to store multiple kinds of information about the objects we observe, and then decode in a task-dependent manner. We tested whether people apply these strategies in the specific domain of short-term memory for novel faces. To test whether people store various kinds of information about a face, and then decode in a task-dependent manner, we modeled their responses in a memory task using features (extracted from deep neural networks) that varied in how much 3D information they carried. We found strong evidence for a mixed-storage strategy, which did not vary in response to task demands. Our results suggest that in order to fully understand resource allocation and retrieval strategies in human memory, it may be critical to consider not just the distribution over tasks in people's natural environments, but also task uncertainty at the time of encoding.

Keywords: Visual working memory; rate-distortion theory; deep neural networks; face perception

Introduction

Normative frameworks in cognitive science seek to understand aspects of behavior by comparison to the ideal strategy in the task of interest. For systems with limited informational capacity, such as memory, the optimal strategy is given by a branch of information theory called rate-distortion theory (Berger, 1971; Sims, 2016), which defines a constrained optimization problem that balances capacity constraints against performance objectives. Given the limited capacity of human memory, the brain must select what information to store.

The rate-distortion framework conceptualizes memory encoding and retrieval processes using a communication channel. A channel consists of an encoder and a decoder, where the encoder is a function that maps from the source (or stimulus) to an abstract "code" vector, z, and the decoder is a function that reverses that process. If there were no capacity constraint, there would be no advantage to this remapping. However, when the channel has limited capacity, the encoder can be carefully designed so that z carries only the most crucial information, and thus requires fewer bits to transmit. In this work, we will consider z to be the memory trace corresponding to a stimulus. We will also refer to the process of retrieving information from z as "decoding" from memory.

While studies have applied the rate-distortion framework to predict how people adapt their visual encoding strategies in response to changing demands in their environment (Sims, Jacobs, & Knill, 2012; Bates, Lerch, Sims, & Jacobs, 2019; Bates & Jacobs, 2020, 2021), they have not yet addressed one critical component of the problem facing people outside the lab: There are many tasks we might need to perform in the future, and most tasks only require a subset of stimulus features. If we knew which tasks were going to be performed, we could save the critical subset of features with high fidelity and forget the rest. For example, if you view a 3D object, later on you may wish to recall exactly how it was oriented. Alternatively, you might need to remember aspects of its shape, independent of viewpoint. If there is uncertainty about future needs, then a smart strategy would be to store a set of features that could subserve either task, as needed (note that this strategy also requires context-dependent retrieval).

Critically, this strategy may need to be learned over extended periods of time. In lab settings, there has been intense interest in understanding people's ability to flexibly reallocate on the fly between different objects or feature dimensions in response to task demands or cues (Ye, Hu, Ristaniemi, Gendron, & Liu, 2016; Maxcey-Richard & Hollingworth, 2013). Outside of the lab, however, contextual cues are not usually as explicit, which means it makes sense to learn a default strategy over time that minimizes errors in expectation.

Do people implement such a strategy? To our knowledge, this question has not been studied before, but a related problem has been studied within Anderson's rational analysis of memory, which uses "need probabilities" to predict the availability of information in memory (Anderson & Milson, 1989). Anderson considers the problem of efficient information retrieval, where the goal is to minimize the amount of search required to retrieve a piece of information. For example, the probability of needing to retrieve an item tends to decrease with the amount of time it has been stored. Thus, search costs will generally decrease if more recent items are prioritized over older items. By contrast, our work is not concerned with search costs but rather with storage costs.

In this work, we conduct an investigation into encoding and decoding strategies in the face of future task uncertainty. First, we test the hypothesis that people i) store multiple, distinct feature sets that are useful in distinct tasks that they are likely to encounter, and ii) decode these features from mem-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

ory in a context- or task-dependent manner. Second, we test the inflexibility of people's encoding policies. To do this, we employ a simple manipulation within the standard change-detection paradigm, aimed at distinguishing between 2D versus 3D features. Our reasoning was that these two categories are likely to be useful in various natural tasks, and therefore are likely candidates to be stored together. In one type of trial, participants are tasked with detecting a change to the study object when there is a change in viewpoint between study and test. In the other trial type, the viewpoint does not change. The key intuition is that if people store both 2D and 3D features, they should be able to combine those two sources of information in our change-detection task when viewpoint does not change very much between study and test. However, 3D features are likely easier to use than 2D features when viewpoint changes a lot (e.g. a linear readout may be sufficient for 3D but not 2D features). Therefore, the goal of our modeling will be to measure to what extent participants are relying on 2D versus 3D features in each type of trial.

Our experimental design makes the critical (and arguably reasonable) assumption that some mixture of 2D and 3D features is optimal to store in the context of people's natural environments. In reality, we cannot know precisely what is optimal without careful additional study. Thus, if we do not find evidence that people store a mixture of features, it could be that this assumption is wrong.

For our stimuli, we use computer generated faces. Since our stimuli are complex and naturalistic, it is difficult to take the common approach of hand-crafting features. Instead, we take layers from deep neural networks trained to interpret faces as the set of candidate features. Based on previous work (Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020; Schrimpf et al., 2020), we presupposed that layers would run the gamut between more 2D and more 3D in nature. We also compare these features to a separate set, derived from subjective ratings of high-level facial attributes.

Experimental methods

We conducted two memory experiments, plus a third experiment in which we collected additional subjective ratings about the stimuli used in the memory experiments. The first and second experiment differed only in whether two trials types (viewpoint-change and no-viewpoint-change) were mixed within participant. In the first experiment, each participant saw only one trial type, while in the second experiment, they saw both. We conducted the second experiment in order to test the flexibility of encoding strategies in response to task demands. Specifically, while our hypothesis is that people have a relatively fixed allocation strategy on each trial, learned over many hours of experience outside the lab, another possibility is that people decide on each trial how to allocate resources across feature sets. If each participant only sees one kind of trial, they may quickly learn a fixed strategy that is tailored to that trial type. In order to distinguish between these strategies, we train participants on one trial type and test how they generalize to the other. If encoding strategies are relatively fixed across conditions, then the generalization trials should be statistically similar to their counterparts in Exp. 1. By contrast, if encoding strategies are tailored to whichever trial type is most prevalent, then the generalization trials should statistically resemble trials of the other type in Exp. 1. Finally, our third experiment reexamined the data from Exp. 1 using new models that included subjective ratings from a separate pool of participants.

Stimuli and Procedure. Stimuli in the memory experiments were cropped face images (512×512 pixels) generated using the Basel Face Model (BFM) (Gerig et al., 2018), placed on a white background. The BFM we used (2019 version) consists of 199 shape and 199 color dimensions. These dimensions are the result of applying principal component analysis to highly detailed, physical scans of 200 real faces. Faces ("identities") are sampled from the model by sampling values for these dimensions. The model also includes dimensions for facial expression, but we fixed all of these values to zero, resulting in neutral expressions.

To produce a target-probe pair, we first sampled a random target identity using the BFM. Then we sampled a nearby identity such that they were separated by a cosine distance equal to δ. Each identity that is sampled can be rendered from any viewpoint. In no-viewpoint-change trials, the viewpoint was always frontal (yaw, pitch, and roll all set to zero). In viewpoint-change trials, the target stimulus was frontal, but the probe stimulus differed in that it was always rotated $+15$ degrees in yaw (whether or not the identity changed). For viewpoint-change trials, we used $\delta = 0.75$, while for no-viewpoint-change trials, we used $\delta = 0.35$. These values were chosen based on pilot data in order to target an 80% average correct response rate.

Participants in the memory experiments performed a change-detection task (Figure 1). On each trial, the target stimulus was presented for 2 s, the retention interval was 1.5 s, and the probe stimulus stayed on screen until response. The inter-stimulus screen did not include masking (i.e., it was blank). Each participant in Exp. 1 and 2 completed 200 trials. Probe stimuli were randomly sampled per participant such that half of the trials were "change" trials (the probe identity was different than the target identity) and the other half were "same" trials (the identities were the same). They completed 4 practice trials prior to the testing phase, and received feedback (correct or incorrect) on every trial (both practice and test). Data was collected using Cloud Research, which is a service built on top of Amazon's Mechanical Turk, and includes filters to improve data quality.

In Exp. 1, each participant was randomly assigned to either see only viewpoint-change trials (N=15) or only no-viewpoint-change trials (N=15). All participants saw the same 200 target stimuli, but in a different random order. Exp. 2 was identical in methodology, except that each participant (N=36) saw 80% viewpoint-change trials and 20% no-viewpoint-change trials. (Note that Exp. 2 had more par-

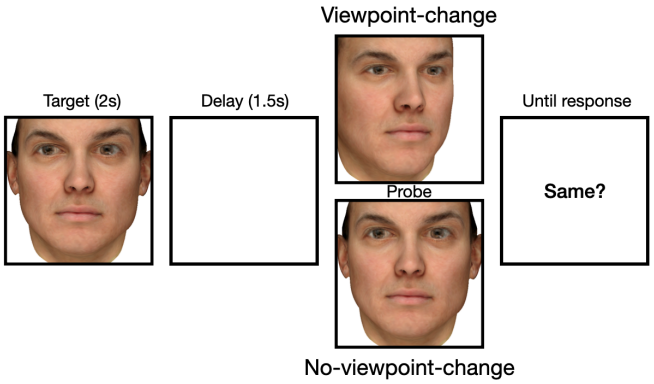

Figure 1: Change-detection procedure. On each trial, the probe either did or did not change in viewpoint relative to the target, and the target was always in the frontal view.

ticipants in order to get enough samples for the generalization trials.) Trials for each type were sampled randomly without replacement from the larger pool of 200 trials from Exp. 1, per participant, and the order was random. Thus, participants no longer saw all the same target stimuli. Of course, the strongest test of whether people are adapting their decoding strategies would be to only show one no-viewpoint-change trial to each participant, on the very last trial. However, we chose the 80/20 split to strike a balance between hypothesis testing needs and the amount of data that would need to be collected.

Modeling

The aim of our modeling was to compare human responses to a variety of visual features which varied in their 3D face-shape specificities. Toward this end, we extracted features from two different deep convolutional neural networks (DCNN), trained on faces according to different objectives. Previous work (Yildirim et al., 2020) showed that these networks acquire a hierarchy of features ranging from view-specific (more 2D) in early layers to more view-invariant (more 3D) in later layers, thus meeting our requirement.

One of the pre-trained networks we used was the Efficient Inverse Graphics engine (EIG) (Yildirim et al., 2020). This network is trained to invert a generative face model, under the assumption that the faces we encounter in the world are generated according to the BFM. Specifically, the EIG was trained to map face images (generated based on the BFM) to identities in the BFM (the $\sim$400 shape and color dimensions used to produce a face). The network demonstrated good generalization to real photographs.

The other pre-trained network we used was the VGG face network, as presented in (Parkhi, Vedaldi, & Zisserman, 2015). We used the "VGG-raw" pretrained version as in Yildirim et al. (2020). This network was trained to map a face image to one of several thousand celebrity identities. The training set was created from images freely available on the internet, and the architecture was based on VGG-16 (Simonyan & Zisserman, 2014).

Subjective Ratings. While the DCNN layers we use here may capture some important abstractions about faces, it is possible that they do not adequately capture certain high-level, behaviorally-relevant features that are salient to people. For example, people might be sensitive to changes in the masculinity or femininity of a face (Freeman, Rule, Adams Jr, & Ambady, 2010), in a way that the DCNNs are not. It is possible that including or omitting these kinds of features may alter our conclusions.

To investigate this possibility, we collected subjective ratings (N=19 per stimulus) of our face stimuli from participants along several high-level dimensions. Participants gave ratings on a one-to-five scale along masculinity/femininity, strangeness (some generated faces looked particularly unusual or striking), age, weight, and emotional valence (some faces deviated slightly from neutral, despite setting expression to neutral in the BFM). We collected 19 ratings per stimulus image, which we z-scored per question within each participant and then averaged. The result was a single score for each question for each target or probe image. This vector was then treated as another feature, just like a DCNN layer.

Logistic regression models. In order to predict same/different responses, we trained logistic regression models based on the (flattened) DCNN features. Specifically, we first produced a "psychological" distance for a particular layer by computing the cosine distance between target and probe in feature-space. That is, we compute the activations from layer i for the target, and then for the probe, and compute the cosine distance between these two vectors. The result is a number between 0 and 2, where 0 means the images are highly similar according to layer i. This procedure gives one number for each of the 200 trials, for each layer. Then we fit a standard logistic regression to map from the distance predicted by layer i on each trial and whether the participant responded "different". If layer i is a good model of the data, then its distance should be larger in trials where people reported a change more frequently. We restricted our analysis to the half of trials in which there was actually a change between target and probe, since in the other half the model distance was always zero.

Finally, we note that the procedure was slightly different for the survey data compared to DCNN layers. We assumed that the magnitude of the feature vectors in this case mattered, so instead of cosine distance, we used absolute difference. That is, for each target-probe pair, for each question k, we took the absolute difference between the average rating for target and probe. Then, we summed across questions to produce a single distance value for each target-probe pair, just like the DCNN layers.

Feature invariances. While we hypothesize that layer depth may be used as a proxy for face-shape invariance, we can also measure this property more directly. We created an invariance index by measuring how much similarity drops off as viewpoint changes. Intuitively, if a layer is perfectly invariant, there should be no drop-off with viewpoint change.

More specifically, for each layer we measured average cosine distance between the target and probe identities from the viewpoint-change "same" trials (i.e., when the target and probe identity was the same but viewpoint changed). Thus, we averaged 200 distances to get a single value per layer. Finally, we subtracted those values from 1 to ensure that they increase with face-shape invariance.

Results

Experiment 1. In this experiment, we tested the hypothesis that people i) store multiple, distinct feature sets that are useful in distinct tasks, and ii) decode these features from memory in a context or task-dependent manner. In particular, we tested whether people store face features that are relatively 3D-shape-invariant in addition to features that are less shape-invariant. To do so, we fit a standard logistic regression model to aggregated participant data for each layer in each of the two DCNNs. For each layer, we then calculated the model's log likelihood.

The result for each DCNN and each condition is shown in Figure 2. In the viewpoint-change condition, the clear trend is that the later layers of each network produce higher likelihoods. In the no-viewpoint-change condition, later layers do not provide the best fits. The function of likelihood versus layer depth appears more smooth and monotonic in the no-viewpoint-change condition in the VGG face network, compared to EIG. We suspect this is a result of having more layers, since it has been found that deeper networks acquire higher representational similarity between adjacent layers (Kornblith, Norouzi, Lee, & Hinton, 2019).

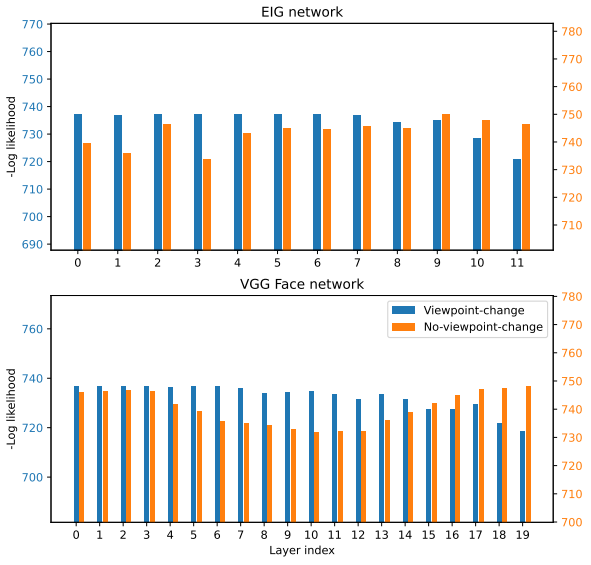

Figure 2: Negative log likelihoods of the logistic regression model based on each DCNN layer (Experiment 1).

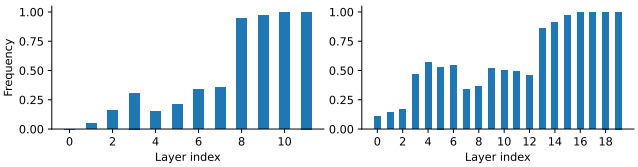

Figure 3: Summary of bootstrap analysis for Exp. 1. Each bar shows the proportion of resamples of the data for which the layer "preferred" the viewpoint-change data.

We conducted a bootstrap analysis to estimate confidence intervals by resampling responses with replacement. However, we found that this analysis was not meaningful, because errors were correlated between layers. That is, all layers tended to go up or down in likelihood together, depending on which trials were sampled. Thus, we instead measured how well the rank ordering of likelihoods was preserved across resamplings. For each resample, we recorded the rank index of layer i after sorting from lowest to highest likelihood. Then we counted how often layer i "preferred" the viewpoint-change versus no-viewpoint-change condition. For example, we say that layer i prefers the viewpoint-change condition for a particular resample of the data if it gets ranked higher for that condition than for the no-viewpoint-change condition. We conducted this procedure separately for each DCNN. Figure 3 shows the results of this analysis. The deeper layers strongly prefer the viewpoint-change condition, and preference for the no-viewpoint-change condition tends to increase toward earlier layers, in a manner consistent with Figure 2.

The pattern of results above suggests that layer depth may be a reasonable proxy for face-shape invariance. However, it is also important to test this more directly. Thus, we measured the face-shape invariance for each layer, as described in the Experimental Methods, and plot the result in Figure 4. Interestingly, our invariance measure provides even stronger support for the hypothesis that participants relied on more view-invariant features in the viewpoint-change condition, and less view-invariant features in the no-viewpoint-change condition. In particular, our measure tracks the log likelihoods in the no-viewpoint-change condition rather precisely (compare to orange bars in Figure 2). Thus, layer depth is not a perfect proxy for invariance, according to our measure, but the ways in which it differs directly support the primary hypothesis.

We used Spearman correlation to quantify the degree of agreement between DCNN features and responses. In the viewpoint-change condition, we found a correlation of $\rho = 0.37$ ($p < 0.0001$) between stimulus-averaged responses and target-probe distances derived from the highest-likelihood layer of the EIG. For VGG, the correlation was $\rho = 0.34$ ($p < 0.0001$). For the no-viewpoint-change condition, the corresponding correlations were $\rho = 0.38$ ($p < 0.0001$) and $\rho = 0.37$ ($p < 0.0001$).

Experiment 2. For this experiment, we conducted the same analyses as the first experiment (grouping by trial type). As discussed above, if we find a similar result to Experiment 1,

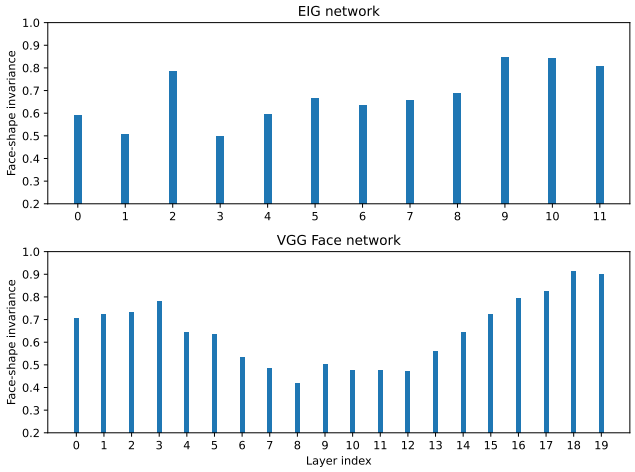

Figure 4: Face-shape invariance for each DCNN layer.

it suggests that people were storing a mixture of features, as hypothesized, rather than setting their allocation strategy on a trial-by-trial basis. Our results confirmed this prediction, as can be seen in Figure 5. We also conducted the bootstrap analysis, which looked very similar to Exp. 1.

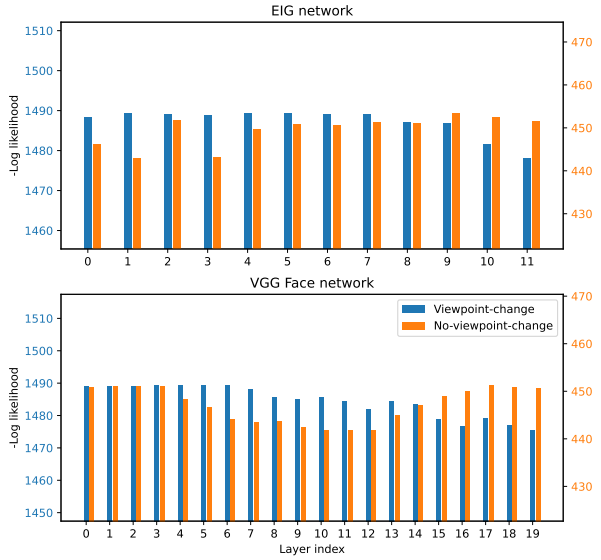

Figure 5: Negative log likelihoods of the logistic regression model based on each DCNN layer (Exp. 2). Note that the magnitudes of the likelihood values depend on the number of trials, which were uneven (80% viewpoint-change, 20% no-viewpoint-change).

Experiment 3. This experiment explored additional features that participants encode which are not captured by the DCNN features. As described above, we collected subjective ratings along several behaviorally relevant dimensions. We then estimated target-probe distances along these dimensions, in a similar manner to the DCNN layers. We asked i) how well this ratings-based model explains responses on its own, and ii) whether it explains additional variance when combined with the most explanatory DCNN layers.

We found that on its own, the ratings-based model performed as well in the viewpoint-change condition but poorly in the no-viewpoint-change, compared to the DCNN layers. Applying the same logistic regression analysis as above to Exp. 1 data, we found that the ratings-based model had a log-likelihood of -718.4 in the viewpoint-change condition (Spearman $\rho = 0.33, p < 0.0001$) and -740.8 (Spearman $\rho = 0.20, p < 0.01$) in the no-viewpoint-change condition, compared to -719.2 and -732.6 for the best DCNN layers, respectively. Thus, in the viewpoint-change condition, participants relied relatively more on features like those in our survey questions. One possible explanation for the discrepancy between conditions comes from the fact that the target-probe delta in the viewpoint-change condition needed to be larger because trials were more challenging. As a result, one might be more likely to find noticeable differences between the target and probe along high-level dimensions like masculinity. This could be true, for example, if some dimensions are coded in a more categorical manner (e.g., male vs. female) (Beale & Keil, 1995; Goldstone & Hendrickson, 2010). If so, a change would be hard to notice unless it passes a certain threshold.

Next, we repeated this analysis for a composite model that combined the best-fitting DCNN layers in each condition (based on results from Exp. 1) with the ratings-based model. We defined this composite model as $d_{comp} = \alpha d_{ratings} + (1 - \alpha) d_{DCNN}$, where d is cosine distance (for some target-probe pair) and $0 \leq \alpha \leq 1$ is a free parameter. We conducted a grid search over α values, training a logistic model for each one. In the viewpoint-change condition, we found maximum log likelihoods of -708.8 ($\alpha = 0.34$) and -707.2 ($\alpha = 0.39$) for EIG and VGG, respectively, as compared to -721.6 and -719.2 for the DCNN layers alone. In the no-viewpoint-change condition, we found a maximum log likelihood of -729.9 ($\alpha = 0.24$) and -727.9 ($\alpha = 0.27$) for EIG and VGG, respectively, as compared to -734.1 and -732.6 for each DCNN layer alone. Thus, with both networks, we found the composite model was an improvement over DCNN layers alone, though the improvement was greater in the viewpoint-change condition. In both networks and both conditions, the DCNN layers were weighted more heavily, suggesting that DCNN features were at least as important in explaining responses.

It is possible that the increase in likelihood after adding in the ratings-based model could simply reflect the greater diversity of features, rather than the particular high-level attributes we measured. For example, perhaps similar gains would be found if we repeated the same analysis but instead combined the best-fitting DCNN layer with other layers from the same network or a different network. To test this, we repeated the composite model analysis but replaced the ratings-based model with every DCNN layer in turn. We found the resulting improvements were more modest, even when mix-

ing and matching between EIG and VGG (see Figure 6).

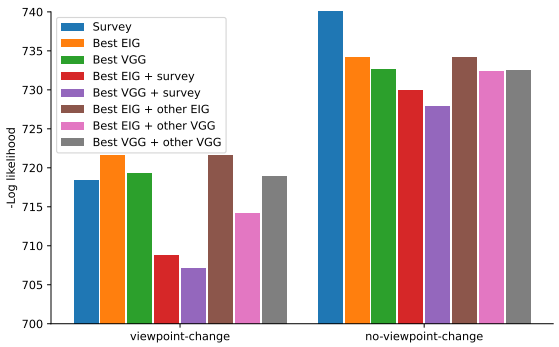

Figure 6: Comparison of models that do and do not include subjective ratings. "Best" refers to the layer that produced the highest likelihood on its own in Exp. 1, and "X + Y" denotes a linear combination of two models, X and Y (with the optimal weighting between them, see text). To produce "Best X + other X" we paired the best layer in X with each remaining layer in turn, and reported the pairing with the best result. To produce "Best X + other Y", we did the same, except we searched over layers in Y, not X. All values computed using data from Exp. 1.

Importantly, the composite model did not overturn the conclusions from our earlier analysis. We repeated the analysis from Experiment 1, except that we fit the composite model for every layer. We found that the relative differences in likelihood between layers were almost identically preserved, except that all the likelihood values were shifted higher as a result of adding the ratings-based model.

Taken together, these results support the conclusion that participants in our experiments relied on additional features that were not adequately captured in the similarity structures of the DCNN models. Moreover, they add support for our hypothesis that people store multiple distinct feature sets. Specifically, they suggest that participants were always storing these additional, high-level features alongside more purely perceptual features. The neurobiological evidence also supports this conclusion, at least in the case of sex: separate brain regions have been found to subserve categorical versus continuous representations of gender (Freeman et al., 2010).

Discussion

In this work, we considered the problem of task uncertainty in visual memory. There are many kinds of features that we could store after looking at the objects in our natural environment, and different features may be more useful in some memory tasks than others. For example, here we specifically considered a dichotomy between 2D and 3D features: 2D features are needed for recalling viewpoint-related information, while 3D features are needed for recalling shape-related information. If either kind of task is possible in the future, a good strategy would be to store both 2D and 3D features.

Data from our change-detection task suggest that people adopt some version of this strategy. However, our experiments do not directly assess the optimality of behavior. While it seems intuitively reasonable that people should store some mixture of 2D and 3D face features, addressing this question requires rigorous study of the set of memory-related tasks in people's natural environments, as well as uncertainty over future needs. For example, if people do not usually need to recall viewpoint information, then devoting too much of their storage to 2D features would be suboptimal.

Our work raises important questions regarding the design of encoding and decoding operations in memory systems and how these are coordinated. For instance, in the viewpoint-change condition from our experiment, participants needed to decode selectively from the 3D features stored in memory, largely ignoring the 2D features. While previous research has raised the point that memory should store invariant features in order to recognize the same object under new viewing conditions (Riesenhuber & Poggio, 2000), to our knowledge the problem of storing and decoding mixtures of features has not been considered. For instance, is it better to store 2D and 3D features as separate memory traces, or as part of a single, integrated trace?

Finally, while our work applies most clearly to memory for complex objects and scenes, it may also have implications for understanding performance limitations in the kinds of simplified displays more commonly used in the literature. For example, it is possible that people store features at multiple levels of abstraction, similar to our finding with faces, even in simple displays like colored squares on a blank background. If so, configural features may have a predictable impact on memory performance. In fact, there is robust evidence that the specific configuration of items in a display matters substantially (Brady, Konkle, & Alvarez, 2011; Brady & Tenenbaum, 2013; Martin & Becker, 2021; Brady & Alvarez, 2015; Orhan & Jacobs, 2013), and thus memory for one item in a display cannot be considered fully independent of memory for the other items. The framework we present here suggests a principled modeling approach for predicting the role of item configuration in memory and how that relates to capacity limits. In particular, our work suggests that people allocate storage for "extra" features that may not end up being used in a particular task, as a rational response to task uncertainty. Thus, to estimate capacity and predict performance, we need to consider the full set of features being stored. We should try to understand the features that people store when studying natural images, because they may be similar to the ones they store when studying more simplified displays.

Acknowledgments

We thank Laila Johnston, Bernhard Egger, and Ilker Yildirim for their valuable inputs and contributions. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, and the Multi-University Research Initiative Grant (ONR/DoD N00014-17-1-2961).

References

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological review*, *127*(5), 891.

Bates, C. J., & Jacobs, R. A. (2021). Optimal attentional allocation in the presence of capacity constraints in uncued and cued visual search. *Journal of Vision*, *21*(5), 3–3.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of vision*, *19*(2), 11–11.

Beale, J. M., & Keil, F. C. (1995). Categorical effects in the perception of faces. *Cognition*, *57*(3), 217–239.

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. NJ: Prentice-Hall.

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of vision*, *15*(15), 6–6.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, *11*(5), 4–4.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85.

Freeman, J. B., Rule, N. O., Adams Jr, R. B., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*(6), 1314–1322.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., & Vetter, T. (2018). Morphable face models-an open framework. In *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 75–82).

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

Martin, A., & Becker, S. I. (2021). A relational account of visual short-term memory (vstm). *Cortex*, *144*, 151–167.

Maxcey-Richard, A. M., & Hollingworth, A. (2013). The strategic retention of task-relevant objects in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 760.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological review*, *120*(2), 297.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015, September). Deep face recognition. In X. Xie, M. W. Jones, & G. K. L. Tam (Eds.), *Proceedings of the british machine vision conference (bmvc)* (p. 41.1-41.12). BMVA Press. Retrieved from `https://dx.doi.org/10.5244/C.29.41` doi: 10.5244/C.29.41

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature neuroscience*, *3*(11), 1199–1204.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, *119*(4), 807.

Ye, C., Hu, Z., Ristaniemi, T., Gendron, M., & Liu, Q. (2016). Retro-dimension-cue benefit in visual working memory. *Scientific Reports*, *6*(1), 1–13.

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science advances*, *6*(10), eaax5979.