Integrating Episodic and Semantic Information in Memory for Natural Scenes

Pernille Hemmer (phemmer@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA, 92697-5100

Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA, 92697-5100

Abstract

Recall of objects in natural scenes can be influenced not only by episodic but also by semantic memory. To model the statistical regularities that might be encoded in semantic memory, we applied a topic model to a large database of labeled images. We then incorporated the learned topics in a dual route topic model for recall that explains how and why episodic memories are combined with semantic memories. The dual route model was applied to an empirical study in which people recall objects from scenes under varying amounts of study time. The dual route model explains how the trade-off between episodic and semantic memory is affected by study time, output position, and also congruity of the object with the scene context.

Keywords: Episodic Memory; Semantic Memory; Natural Scenes; Bayesian models; Reconstructive memory

Introduction

Semantic knowledge can exert strong influences on episodic recall. In the verbal domain, the use of highly related words on a study list can lead to intrusions of related words in free recall (Roediger & McDermott, 1995). Similarly, expectations about objects in scenes can lead to recall of objects that were not present in the scene. For example, people can recall seeing books in an office where there were no books present (Brewer & Treyens, 1981). These intrusions demonstrate the influence of semantic knowledge on recall. Some researchers have viewed such intrusions as demonstrations of shortcomings of the memory system. However, semantic knowledge can also serve as an aid to episodic memory and lead to improvements in recall performance (e.g. Hemmer & Steyvers, 2009; Konkle & Oliva, 2007; Huttenlocher et al. 1991).

Dual retrieval accounts of memory propose that reconstruction from memory requires accessing either the verbatim memory trace or semantic information relevant to the event (Brainerd et al., 2002). The verbatim – or episodic memory – trace is a representation close to the original event, while the semantic information is an abstraction of the event, often referred to as 'gist' or 'schema'. Previous dual route models have not explained in detail how the semantic information is represented (or extracted from the environment) and have not fully described the detailed mechanisms for the interaction between episodic and semantic information.

In this research, we build on the framework of rational memory models that assume that the memory system is exploiting environmental regularities when recalling information about past events (Anderson, 1990; Steyvers & Griffiths, 2008). We develop a dual route memory model and apply it to the problem of recalling objects from natural scenes. We assume that an observer is presented with a scene during study and is instructed to retrieve from memory objects that occurred in the scene. The goal for the observer is to reconstruct the objects from the scene optimally combining the available information. We assume that the available information is based on noisy episodic memories and also on encoding based on the semantic context. Previous research has shown that people are sensitive to the contextual information in scenes and can quickly extract a high-level semantic representation of a scene (Potter et al., 2002).

In this paper, we will first present an empirical study on scene recall and investigate how recall accuracy varies as a function of study time and what the accuracy is if there is no episodic information at all and recall is based on semantic information only. The experimental data allow us to assess how people trade off between episodic and semantic memory. We then present a topic modeling analysis (Griffiths & Steyvers, 2004; Griffiths, Steyvers & Tenenbaum, 2007) for a large database of labeled images. The extracted topics serve as approximations to the kinds of statistical regularities that people might have encoded in semantic memory. Lastly, we will show how a dual route topic model (Steyvers & Griffiths, 2008) that mixes episodic and semantic information during encoding can account for the empirical findings. We also show how the model can explain the Von Restorff effect, where people have better memory for objects that are incongruous with the scene context.

Empirical Study on Scene Recall

We conducted a series of behavioral experiments using natural scenes such as kitchens and offices to quantify the relative contribution of semantic knowledge on recall. In a memory experiment, we showed images of natural scenes for varying amount of study time. We expected that by decreasing the amount of study time, recall would be based more on semantic memory and would lead to a larger number of errors. To assess the prior knowledge people

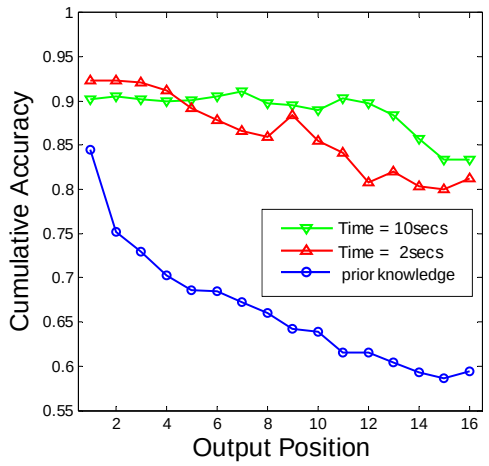

Figure 1. Cumulative accuracy as a function of study time and output position. The figure also shows the simulated performance when one treats the responses from the prior knowledge experiment as responses in the memory experiment

have about certain types of scenes, we also conducted a norming study where we asked participants to name the objects they expected to appear in certain types of natural scenes, without actually showing them any image. Finally, we ran a perception experiment, using the same images as used in the memory experiment, where participants were asked to name all the objects that they perceived in the image. This perception experiment allowed us to assess the ground truth of which objects were perceived to be present in each image, which can be used to score the accuracy of responses in the memory experiment.

Methods

Participants were undergraduate students at the University of California, Irvine. There were 22 participants in the prior knowledge experiment, 25 participants in the perception experiment, and 49 participants in the memory experiment.

Materials. We sampled 10 images from the LabelMe database (Russel & Torralba, 2008) where we chose 2 images each of 5 different scene types. The scene types correspond to kitchen, dining, office, hotel room, and urban scenes.

Prior Knowledge Experiment. To assess prior (semantic) knowledge about specific scenes, we asked participants to list objects that they would expect to occur in a given scene type (which was described by the verbal label). Participants entered their responses on a computer screen and were required to make responses for a minimum of 60 seconds before continuing to the next question.

Perception Experiment. In this experiment, we assessed the ground truth for the occurrence of objects in each of the 10 images. Materials were presented on two computer screens. The image was presented on the left screen while response instructions and a response box were presented on the right screen. Participants were asked to list the objects present in each image and were required to make responses for a minimum of 6o seconds. They received feedback based on matching their responses to those of previous participants. Images were presented in random order

Memory Experiment. For the memory experiment, participants studied an image for either 2 or 10 seconds. After completing a short distracter task, participants were asked to list all the objects they recalled seeing in the presented image. Study images were presented in random order. Each participant only saw 5 images, one from each scene type, to avoid carryover effects where the memory from one scene type affects recall of another image of the same type.

Response Normalization. Responses for all experiments were corrected for spelling, plurals, and qualifiers (e.g., numbers, color, size and location). For example, "chair" and "chairs" were mapped to the single entry "chair", and "silver car" was mapped as "car".

Results and Discussion

To measure performance in the memory experiment, we checked whether a given recalled object was part of any of the responses that were given by participants in the perception experiment. If it was, it was scored as a correct response. If it was not, we manually checked whether the recalled object could still be considered as a description of an object that was part of the image. Only if it was not, the response was scored as incorrect. We calculated cumulative accuracy in the memory experiment as a function of the output position. In other words, we calculated the mean accuracy for the first item recalled, first two items recalled, etc. Figure 1 shows the cumulative accuracy as a function of output position and study time. Overall, cumulative accuracy decreases as a function of output position. Therefore, more intrusions are made later in recall, a finding compatible with results from the verbal memory domain (Roediger & McDermott, 1995). Cumulative accuracy was highest for the short study time condition for the first five output positions. After the sixth output position, the cumulative accuracy was best for the long study time conditions. Therefore, the somewhat counterintuitive finding here is that shorter study times do not necessarily lead to worse performance – the first few items remembered are *more* likely to be correct compared to a condition with longer study times (however, the *total* number of correct responses is greater with longer study times; for 2 and 10 second conditions, there were an average of 7 and 9 correct responses respectively per subjects per image).

We can explain this finding as an effect of the trade-off between episodic memory and semantic knowledge. For short study times, only a few objects might have been observed. Some of these objects can be encoded episodically without running into interference or capacity constraints. These few objects can subsequently be output with fairly high accuracy. On the other hand, if a scene is

studied for a longer period, more objects overall are noticed and will need to be encoded. This longer list might not be encoded entirely by episodic means and part of the encoding might be based on generalized semantic knowledge. This will lead to lower accuracy for the first few items recalled but to higher accuracy at later output positions because of the enhanced semantic encoding.

Figure 1 also shows the performance one can expect from prior knowledge in the absence of any episodic information. This is the case where the image was not studied at all (corresponding to zero second study time). Even though we did not actually run this in the memory experiment, we can consider the responses from the prior knowledge experiment as reasonable guesses to the objects of an image in a particular scene. We ran an analysis where we treated the prior knowledge responses for each scene type as memory responses for the image (for the same type), preserving the order of the responses. Figure 1 shows that the performance of this condition is fairly high. The first item guessed in the prior knowledge experiment leads to 85% accuracy in the memory experiment, even though the response is not associated with any episodic knowledge of the task. For later responses, accuracy does decrease but cumulative accuracy is still higher than 55% even after guessing 16 items. The difference between the performance from prior knowledge and actual recall reveals the contribution of episodic memory, which might be smaller than one might expect. These results demonstrate that general knowledge of scenes can greatly contribute to the accuracy of recalling objects from natural scenes.

A Model for Object Recall in Natural Scenes

One conclusion from our empirical study is that semantic knowledge can lead to good baseline performance in scene memory. When recalling objects from a kitchen that has never been seen before, recall can be reasonably good if the guesses are based on general knowledge of kitchen scenes (e.g., guesses such as "refrigerator", and "sink"). Of course, performance improves when actual episodic memories of the particular image can be retrieved. This raises the question of how the interaction between episodic and semantic memory can be modeled. We will first discuss a topic model for scenes that approximates the semantic knowledge people might have about objects in scenes and then develop a dual route topic model that integrates both episodic as well as semantic memory information.

A Topic Model for Scenes

Probabilistic topic models have been developed as a method to automatically learn semantic representations for documents by analyzing the statistical relationships between words and the documents they occur in (e.g. Griffiths & Steyvers, 2004; Griffiths, Steyvers & Tenenbaum, 2007). In the topic model, each document is expressed as a mixture of topics that can be thought of as the gist of a document, and each topic represents a probability distribution over words. Here, we apply the topic model to a subset of 13,572 images

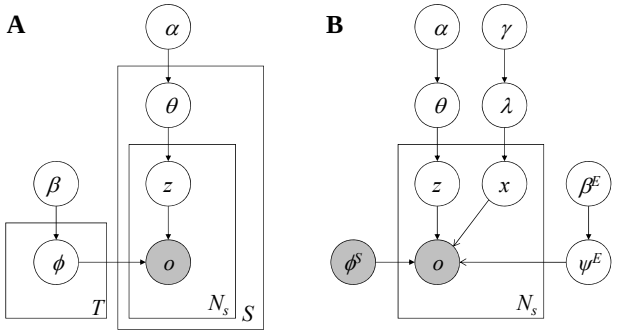

Figure 2. The graphical model representation for A) the standard topic model and B) the dual route topic model.

of the LabelMe database (Russel & Torralba, 2008). These images were annotated by volunteers resulting in a total of 87,152 labels and 3782 unique types. The subset contains images of natural scenes, such as urban street scenes and indoor scenes of kitchens and offices.

We treat each scene from the database as a mixture of topics and each topic as a distribution over image objects. This specifies a generative model in which objects in a scene are selected by first sampling a topic from the topic distribution associated with the scene and then sampling an object from the topic. Specifically, the conditional distribution of an object o in a scene s is given by,

$$P(o \mid s) = \sum_{t=1}^{T} P(o \mid z = t) P(z = t \mid s) \qquad (1)$$

where $p(o| z=t)$ is the multinomial distribution over objects given topic t and indicates which objects are important to a topic, and $p(z=t| s)$ is the multinomial distribution over topics given scene s and indicates which topics are important to a particular scene.

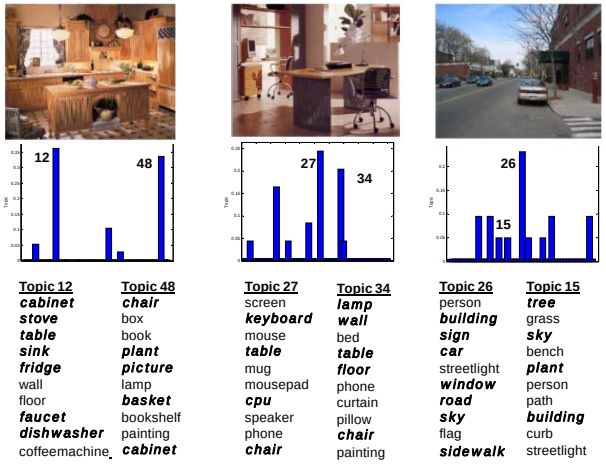

Figure 3. Model predictions for three scene types: kitchen, office and urban. The bar graphs show the distribution over 50 topics for a scene with topic indices for the two most likely topics. The rank-ordered object distributions corresponding to these topics are shown below. Objects labeled in bold were part of the original image annotations.

Figure 2, panel A shows a graphical model representation of the topic model. Shaded nodes represent observed variables while nodes without shading represent unobserved variables. The arrows indicate the conditional dependencies between the variables, and the plates show the replications of sampling steps. There are S scenes and each scene has N_s objects. The variable θ is the scene-topic multinomial and ϕ is the topic-object multinomial. The priors on the multinomials are Dirichlet distributed with hyperparameters α and β. We treat α and β as constants in the model (we set $\alpha = 0.1$ and $\beta = 0.01$).

We applied the topic model with T=50 topics to the LabelMe image database and used Gibbs sampling to infer both $p(o|\ z=t)$ and $p(z=t|\ s)$. Several examples of topic distributions are illustrated in figure 3. The figure shows images from three different scene types: kitchen, office, and urban with the inferred topic distribution for that image. For example, topic 12 is the most likely topic for the particular kitchen image and topic 27 is the most likely topic for the particular office image. Some of the likely topics are illustrated in at the bottom of figure 3. This shows the list of most likely objects associated with each topic. Overall, the model shows that the topics for each image qualitatively capture the semantic context of the image. The likely objects in the topics associated with scenes are objects that can reasonably be found in the respective scenes, and seem to describe the 'gist' of the scene.

A Dual Route Model for Object Recall

The topic model itself cannot be a complete model for reconstructive memory. The topic distribution for a scene provides a generalized representation for the occurrence of objects in scenes (e.g., offices), which is useful to characterize the "gist" of a scene. However, the distribution over topics is insufficient to represent the exact set of objects present in an image. In human memory, recall can be quite accurate, if given enough study time. Therefore, to give a more complete account of human memory, we need to expand the topic model with an additional component that allows the model to reconstruct the specific objects present in a scene.

We will now describe an extension to the standard topic model called the dual route topic model introduced by Steyvers and Griffith (2008). We will apply the model to the problem of scene recall. In the model, recall of objects in a scene is a result of two processes: episodic recall and recall based on the semantic context. The semantic information is an abstraction based on the statistical regularities of the collection of scenes. For each image, the semantic context is encoded by a probability distribution over topics. The episodic information is based on a noisy encoding of the actual list of objects present in the image to be remembered. In the model, we will implement episodic noise by a simple sampling process. We assume that the episodic sampling process is based on a multinomial distribution over objects with a symmetric Dirichlet prior,

$$o \mid \psi \sim \text{Mult}(\psi) \quad (2)$$

$$\psi \sim \text{Dirichlet}(\beta^E) \quad (3)$$

where β^E is the hyperparameter that controls the amount of smoothing. Note that this process is not just defined over the observed objects in the image, but over all object types (i.e., the object vocabulary). In this process, it is possible to give high probability to a variety of objects, making them likely to be retrieved from the episodic route. However, with the Dirichlet prior, a capacity constraint can be built in. With small values of β^E, it is unlikely, a priori, that the probability over objects is distributed over a large number of objects, therefore encouraging a sparse representation of objects. Therefore, the smoothing parameter determines how much of the retrieval process focuses on the observed objects versus other objects in the vocabulary.

If recall is based strictly on this episodic component, performance should be accurate, at least for a subset of items on the list, but it could potentially fail to fully retrieve the whole list. If recall is based strictly on semantic information performance might not be as accurate but the topic distribution allows retrieval of a larger number of items. The dual route topic model allows recall to be a mixture of these two extremes. The weighting is such that recall is neither too specific nor too general. In the model a mixing process determines if an object is generated using the episodic route or using semantic information. An indicator variable x, acts as a switch such that if x=1, the object is sampled from the semantic route, and if x=0, the object is sampled from the episodic route. We assume that the probability of a route assignment is distributed Bernoulli with a symmetric Beta prior:

$$x \mid \lambda \sim \text{Bernoulli}(\lambda) \quad (4)$$

$$\lambda \sim \text{Beta}(\gamma) \quad (5)$$

Therefore, the conditional distribution of an object o given a scene s, is given by:

$$p(o \mid s) = p(x=1 \mid s)\sum_{t=1}^{T} p(o \mid z=t)p(z=t \mid s) + p(x=0 \mid s)p'(o \mid s) \quad (6)$$

where the first term is the distribution over objects predicted by the topic model weighted by the probability of a route assignment in favor of a semantic encoding. The second term is the object distribution $p'(o|\ s)$, predicted by the episodic route weighted by the probability of a route assignment in favor of an episodic encoding.

Note that this model specifies a generative procedure for producing objects in a given scene. Figure 2B shows a graphical representation of the complete model. Note that we assume that the distribution over objects in each topic, ϕ^S, is observed and estimated by the topic model in a prior learning phase.

The main use of the model is as an encoding model where the goal is to infer the encoding parameters conditional on the observed set of objects in an image. In other words, the goal is to find an encoding such that during retrieval, the model is likely to reconstruct the observed set of objects in an image, taking into account the probabilistic constraints of the model – the built in capacity constraint for the episodic

route and the overgeneralization of the semantic route. Because the model assumes that each object originates from a single memory route, the goal of encoding is to infer which objects can be encoded via the episodic route and which objects can be reconstructed by a probability distribution over topics (specific for the image studied).

The latent variables z and x can be inferred using Gibbs sampling (the remaining latent variables can be integrated out). The topic and route assignment for the i-th object can be jointly determined conditional on all other assignments:

$$p\left(x_i = r, z_i = j \mid z_{-i}, x_{-i}, \mathbf{o}, \varphi^S\right) \propto \begin{cases} \phi^S_{j,o_i}\left(n_{j,-i}+\alpha\right)\left(n_{1,-i}+\gamma\right) & r = \\ \dfrac{n_{o_i,-i}+\beta^E}{n^V_{-i}+M\beta^E}\left(n_{j,-i}+\alpha\right)\left(n_{0,-i}+\gamma\right) & r = 0 \end{cases} \quad (7)$$

where M is the number of unique objects labels in the LabelMe database, $n_{0,\text{-i}}$ is the number of time the episodic route is assigned, $n_{1,-i}$ is the number of times the semantic route is assigned, and $n_{o,-i}$ is the number of times a specific object o is assigned to the episodic route. The subscript $-i$ indicates that the assignment for the i-th object is not included in the counts. We treated the hyperparameters α, β^E and γ as constants in the model (we set α= 0.1, β^E= 0.000001, and γ=0.3).

We applied the dual route topic model to a small number of images from the LabelMe image set. We selected a set of 10 images to correspond with the 10 images used in the memory experiment. The images used in the simulation were selected based on having a relatively large number of annotations (30-60).

Up to this point, the model specifies a retrieval probability $p_i^{retrieve} = p(o_i \mid s)$ for each object i. Ideally, one would recall objects from this distribution strictly in order of decreasing probability. However, we assume that people cannot determine the strict order of probabilities. Therefore, we incorporate noise in the recall sampling process by letting the actual recall probability be based on a soft-max sampling process:

$$p_i^{recall} = \exp\left(\frac{1}{\tau} p_i^{retrieve}\right) \Big/ \sum_j \exp\left(\frac{1}{\tau} p_j^{retrieve}\right) \quad (8)$$

where τ is the parameter that controls the sampling noise. We set τ =0.008. In the experiment, participants were not allowed to repeat previous answers. To simulate this with the model, we sampled objects without replacement from the recall distribution.

To simulate the effect of study time we selected two subsets of the annotation word list for the images: a set of 80% of the annotations and a set of 20 % of the annotations. This corresponds to the idea that when studying an image for a restricted period of time not all the objects in the image are noticed. Subsets were created by drawing a random sample of objects from the full object set. Figure 4 shows the model predictions plotted in the same way as the results of our empirical study. The results show a qualitative fit to the experimental data. Objects from the smaller subset, corresponding to short study times, have initial higher

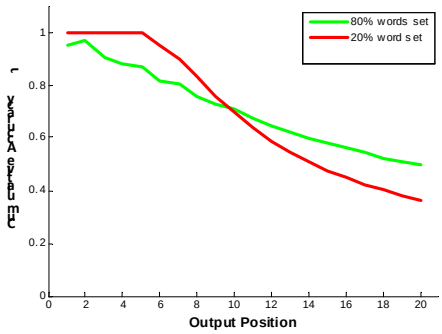

Figure 4: Model predictions: cumulative accuracy by output position when 80% and 20% of the objects have been perceptually encoded. The two conditions simulate the effect of long and short study times respectively.

accuracy, while a larger subset has initial lower accuracy followed by a cross-over. This models the somewhat counter-intuitive finding of our empirical study that the first few objects recalled for short study times are more likely to be correct than for longer study times. The model explains this finding because of different weightings of the two encoding routes, episodic and semantic. If a scene is studied for a longer period of time more objects are noticed and encoded, but it is more difficult to accurately store the longer list of object in memory because of the sparsity constraint in the episodic memory route. This leads to a greater number of objects encoded by the semantic route. While this route cannot fully reconstruct the objects present in the image, it is able to "guess" a larger number of objects, leading to relatively higher cumulative accuracy for later output positions. In contrast, seeing a scene for a shorter period of time, leads one to notice fewer objects but these objects can be encoded more effectively by the episodic route. However, the semantic context is not as well encoded in this case, leading to poorer performance for later output positions. Figure 5 show the probabilities of route assignments for three conditions: full set of objects, and the 80% and 20% subset conditions corresponding to long and short study times. Smaller word sets lead to greater episodic contributions, while larger word sets lead to almost equal contributions of episodic and semantic encoding routes.

The relative contribution of episodic and semantic information in recall can also account for other standard

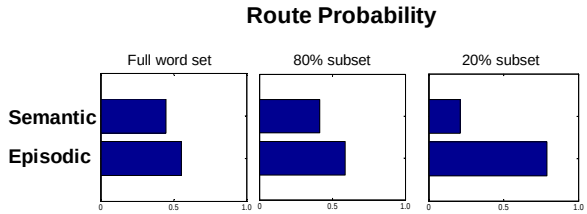

Figure 5: Model predictions for the full response set and for two sub sets of 80% and 20% of responses respectively.

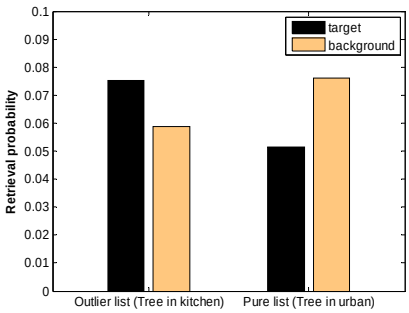

Figure 6: Model predictions an object that is either incongruent (a tree in a kitchen) or congruent (a tree in an urban scene).

memory phenomena, such as the semantic isolation effect (von Restorff, 1933). An object is more likely to be recalled when it is part of a list where it violates the semantic context than when it is presented in a list where it is congruent with the semantic context. This finding can be explained by the dual route model because the route assignment to episodic and semantic memory routes is dependent on the context. Objects consistent with a scene (e.g., typical kitchen objects in a kitchen) can be explained by the semantic route, whereas an object that is not part of the semantic context of the scene (e.g., a 'tree' in a kitchen) can be explained by the episodic route assignment.

To simulate the semantic isolation effect we created an artificial image where we manually determined the presence of objects. We selected an object that had an average recall probability within its semantic context – a tree in an urban scene (these are the same 10 scenes used in the previous two simulations). Figure 6 shows that 'a tree in an urban scene' is recalled with a slightly lower probability than the mean recall probability of all other objects in the scene. We then placed the 'tree' into a semantic context where it did not fit (e.g., a kitchen) by randomly removing an annotation in each of 10 kitchen scenes and replacing that annotation with 'tree'. The urban and kitchen scenes were equated for the number of annotations. We set α= 0.1, β^V= 0.01, and γ=0.3

Figure 6 shows the recall probability for the target object 'tree' and mean recall for all other objects on the list. Recall was higher for the target word than for the other objects. This is consistent with the finding for semantic isolation effects, as well as the idea that objects incongruent with the semantic context of a scene are recalled using episodic information.

Conclusion

We have given an account of reconstructive memory, where reconstruction of objects in a scene is based on a mix of episodic memory traces and semantic context. Short study times lead to recall guided by episodic memory, whereas recall after longer study times is more influenced by semantic information. This counter-intuitive notion that longer study times lead to less reliance on episodic memory, is consistent with our empirical data showing that longer study times lead to an initially lower performance followed by a cross-over in accuracy. Given a dual route topic model account of reconstructive memory, where recall probability is given by the ability of an encoding route – episodic or semantic - to explain the occurrence of an object in a scene, this is to be expected. The model can also account for semantic isolation effects by favoring episodic encoding for objects that are not consistent with the semantic context of a scene.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Brainerd, C. J., Wright, R., & Reyna, V. F. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language, 46*, 120–152.

Brewer, W. F., & Treyens, J. C. (1981) Role of schemata in memory for places. *Cognitive Psychology, 13,* 207-230.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114,* 211-244.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101,* 5228-5235.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review, 16, 80-87.*

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in establishing spatial location. *Psychological Review, 98,* 352-376.

Konkle, T., & Oliva, A. (2007). Normative representation of objects: Evidence for an ecological bias in perception and memory. In D. S. McNamara & J. G. Trafton (Eds.), *Proc.s of the 29th Annual Cognitive Science Society, (pp. 407-413), Austin, TX: Cognitive Science Society.*

Potter, M. C., Staub, A., Rado, J., & O'Connor, D. H. (2002). Recognition memory for briefly presented pictures: The time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 1163-1175.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not present in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21,* 803-814.

Russel, B. C., & Torralba, A. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision, 77,* 157-173.

Steyvers, M., & Griffiths, T. (2008). Rational analysis as a link between human and information retrieval. In: N. carter and M. Oaksford (eds.) *The Probabilistic Mind: Prospects from Rational Models of Cognition.* Oxford University Press.

von Restorff, H. (1933). Uber die Wirkung von Bereichsbildungen im Spurenfeld [the effects of field formation in the trace field], *Psychologische Forschung, 18*, 299–342.