Human-Computer Cooperation for Fine-Grained Visual Categorization

Brett D. Roads
Department of Computer Science and Institute of Cognitive Science
University of Colorado
Boulder, CO 80309-0430
brett.roads@colorado.edu

September 17, 2015

Abstract

The ability to perform fine-grained visual categorizations is critical in many professions (e.g., medical diagnosis, forensic analysis, botany, and zoology). Automatic classification systems cannot yet match the performance level of human experts in many domains, and humans typically require many years of training to become experts. This article discusses three strategies that have been used to allow computers and untrained individuals to cooperate and, through a synergy, achieve levels of performance higher than either could do on their own. First, *decision support* systems help human agents reach a categorization decision by reshaping the demands of the task so that everyday visual abilities can be exploited. Second, *human-in-the-loop* applications help computer agents reach a categorization decision by injecting human capabilities into an automated classification system. Third, *efficient training* applications leverage predictive models to optimally teach individuals how to make unaided categorizations. Even though the time scale and distribution of responsibility varies across each strategy, four recurring techniques provide a common thread that runs through all the applications. Each technique is discussed in turn, showing the commonality across the three strategies, revealing gaps in the current literature and suggesting new ways to synthesize the existing approaches.

Visual categorization requires an agent (machine or human) to classify visual stimuli into conceptual categories. This article is primarily interested in the ability to make *fine-grained* category distinctions of *naturalistic* stimuli (i.e., stimuli that represent the real world). Fine-grained categorization involves stimuli that are both visually and semantically similar (e.g., great gray owl vs. great horned owl). Fine-grained categorization may involve two categories (melanoma vs. not melanoma) or many categories. Expertise of this form is pervasive in day-to-day life and ubiquitous across a wide range of professions (e.g., dermatology, mammography, botany and ornithology). Yet becoming an expert is typically difficult and time-consuming.

In part, visual categorization is difficult because successful discrimination between categories requires attending to details that are often subtle. In fine-grained categorization, categories exhibit a high degree of feature overlap such that differences can be non-obvious. Furthermore, not all differences are meaningful. Agents must be able to distinguish between feature variation that signals a different category versus variation that does not. The difficulty of identifying relevant variant is magnified by the fact that within-class variation can be relatively large compared to between-class variation. Honing in on the right features, in a vast space of potential features, poses an appreciable challenge to a novice agent.

Three human-computer cooperation strategies have emerged for achieving expert levels of visual categorization. Each strategy merges the respective strengths of humans and computers, achieving performance levels that are difficult outside of the partnership. *Decision support* applications help human agents reach a categorization decision by reshaping the demands of the task so that everyday visual abilities can be exploited. *Human-in-the-loop* applications help computer agents reach a categorization decision by injecting human capabilities into a computer categorization system. *Efficient training* applications leverage predictive models to intelligently teach human agents the visual expertise necessary to make unassisted categorizations.

All of these strategies are connected by the use of four general techniques. Each technique is a unique conceptual method for promoting expert-levels of performance. *Highlighting* directs an agent's attention to specific regions or parts of an image. *Task transformations* modify a categorization task so that human agents require less-domain specific knowledge. *Structure sensitive selection* employs knowledge of the underlying category structure to intelligently select exemplars shown to the user. *Response-adaptive selection* uses previous responses to vary the the selection of exemplars at each point in time. Table 1 sorts the selected works by the strategy and technique that they employ. Each technique is reviewed in turn, detailing the common thread that runs across strategies. Taken together, the presented work reveals gaps in the current literature and suggests new ways to synthesize existing strategies and techniques

0.1 Highlighting

Highlighting can be used to help an agent focus on the features that are task-relevant. To assist a human agent, attention can be directed towards diagnostic regions using various image manipulations. Image manipulations can vary from being direct (e.g., circling a diagnostic feature) to subtle (e.g., manipulating the relative contrast of a diagnostic region). To assist a computer agent, pixels in diagnostic regions can be given much more relative weight that pixels in non-diagnostic regions. Guiding an agent to the relevant features diminishes the burden of knowing the diagnostic features and helps agents perform a categorization task.

0.1.1 Decision support applications

In decision support applications, a human agent is provided with highlighting-based assistance. In order to highlight an image, some information must be known about the the location of the diagnostic regions. Decision support applications have primarily utilized three different sources of information for highlighting. An image can be highlighted by exploiting expert-provided annotations, recorded eye movements and computer vision features.

	Decision Support	**Human-in-the-Loop**	**Efficient Training**
Highlighting	Krupinski, Nodine, & Kundel, 1993 (p. 1) Litchfield, Ball, Donovan, Manning, & Crawford, 2010 (p. 1) Berg & Belhumeur, 2013 (p. 2)	Deng, Krause, & Fei-Fei, 2013 (p. 4) Duan, Parikh, Crandall, & Grauman, 2012 (p. 3)	Dror, Stevenage, & Ashworth, 2008 (p. 4) Roads, Mozer, & Busey, submitted (p. 4)
Task Transformation	Aldridge, Glodzik, Ballerini, Fisher, & Rees, 2011 (p. 6) Robertson, McIntosh, Bradley-Scott, MacFarlane, & Rees, 2014 (p. 5)	Jia, Abbott, Austerweil, Griffiths, & Darrell, 2013 (p. 7) Branson et al., 2010 (p. 7) Swanson et al., 2015 (p. 8)	Wahlheim, Dunlosky, & Jacoby, 2011 (p. 8) Kirchoff, Delaney, Horton, & Dellinger-Johnston, 2014 (p. 8)
Structure Sensitive Selection	Brodley et al., 1999 (p. 10) Kumar et al., 2012 (p. 10) Roads & Mozer, submitted (p. 10)		Hornsby & Love, 2014 (p. 10)
Response-Adaptive Selection	Ferecatu & Geman, 2009 (p. 11)	Wah et al., 2014 (p. 12) Wah, Maji, & Belongie, 2015 (p. 12)	Birnbaum, Kornell, Bjork, & Bjork, 2013 (p. 14) Mettler & Kellman, 2014 (p. 13)

Table 1: The selected works arranged based on the human-computer cooperation strategy they employ. Work that only partially implements a technique is shown in gray font. The (clickable) page numbers indicate the location where the work is introduced.

A rich research tradition has developed around decision support applications for medical images. Categorizing medical images, such as x-rays, can be challenging for novices and experts alike. Various methods have explored how to effectively cue individuals to diagnostic regions of an image, such as circling the location of a tumor (e.g. Krupinski et al., 1993). Expert-provided annotations can be used as the basis of highlighting, if the annotations are available. In practice, expert annotations are often unavailable. An alternative is to highlight an image on the basis of another person's eye movements. By overlaying a visualization of where other individuals have looked, a new viewer can be alerted to regions worth further investigation. Highlighting images on the basis of other's eye movements has been shown to improve categorization performance on chest x-rays (Litchfield et al., 2010; Donovan, Manning, & Crawford, 2008). Interestingly, these works have shown that both novice and experts benefit from viewing the search behavior of another novice or expert, although the benefit is sometimes larger for novices.

Another approach to highlighting is to develop computer vision systems that locate diagnostic regions. Berg and colleagues (Berg & Belhumeur, 2013; Berg et al., 2014) developed a system that uses computer vision features to highlight differences between similar bird species.[1] Starting with the part-annotated CUB-200-2011 database (Wah, Branson, Welinder, Perona, & Belongie, 2011), Berg and Belhumeur created a vocabulary of *part-based one-vs-one features* (POOFs). Each POOF attempts to discriminate between two classes based on a given bird part and set of computer vision

[1] The visual guide is publicly available at birdsnap.com.

features (i.e., color or spatial frequency).[2] For example, a POOF may be designed to discriminate between a red-winged blackbird and a rusty blackbird by looking at the color of the wing. A POOF that does a good job distinguishing between categories can be assumed to represent a diagnostic feature. For example, a well-performing POOF may indicate that the color of the beak is highly diagnostic. To aid users in seeing the diagnostic features, a screen is shown with two *reference exemplars* belonging to different but visually similar categories. The POOFs are used to circle the diagnostic locations on the two images. The reference exemplars are selected such that the diagnostic features are slightly exaggerated and the variability on non-diagnostic features (including pose) is minimized.

0.1.2 Human-in-the-loop applications

The difficulty of fine-grained categorization means that computer systems can benefit from highlighting as well. In contrast to decision support applications, the direction of assistance is flipped. Human agents provide constraining information that helps computer vision systems focus on critical regions. Utilizing different types of human input, both Duan et al. (2012) and Deng et al. (2013) demonstrate that providing highlighting to computer vision systems can boost categorization accuracy.

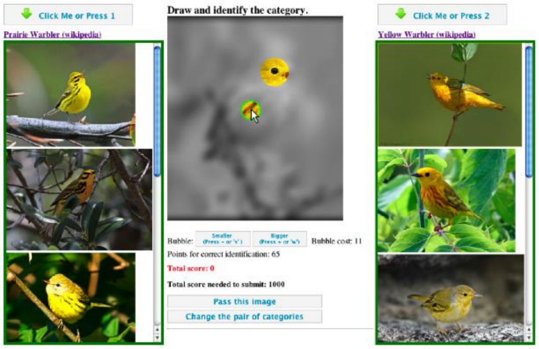
Duan et al. (2012) used human judgments to reduce a set of machine detectable candidate features down to a subset of semantically meaningful attributes. Uncovering diagnostic and semantically meaningful attributes is valuable for automatically annotating an image dataset and aligning computer vision features with human vision features. Starting with a domain of images and corresponding category labels, the first stage runs the images through a hierarchical segmentation algorithm (Arbelaez, Maire, Fowlkes, & Malik, 2011) that produces a set of regions at different scales. In the second stage, an iterative procedure is used to filter the regions down to a subset of diagnostic and semantically meaningful attributes. In the first step of the iterative procedure, a latent conditional random field (L-CRF) (Quattoni, Wang, Morency, Collins, & Darrell, 2007) automatically finds K discriminative candidate attributes between two categories (a *category split*). The objective function is defined so that the recovered candidates are common within a category (in terms of appearance, scale and location), are different between categories, and have minimal region overlap. A new category splits is determined by a greedy procedure that finds the two categories that are most similar in terms of the presence and absence of the attributes discovered so far.

In the second step of the iterative procedure, the most discriminative of the K candidates are presented to 10 Amazon Mechanical Turk (AMT) participants.[3] Each user views a screen showing an attribute visualization containing a handful of reference exemplars each superimposed with a 2-D Gaussian contrast spotlight on the location of the attribute. The user is asked to provide a name for the highlighted region, provide a descriptive word, and rate the confidence of their judgment (e.g., head, red, very certain). If the users find the candidate meaningful (i.e., nameable with high confidence), then the candidate is added to the attribute vocabulary bank. If not, the next most discriminative candidate is presented. If none of the K candidates are considered meaningful, the procedure proceeds to the candidates of the next category split. When the original set of hierarchical regions is filtered down to a set semantically meaningful attributes, these attributes can be leveraged by a computer agent to perform the categorization task.

Duan et al. compared their proposed approach to three alternatives. The *upper bound* alternative establishes a theoretical upper bound where at each iteration the most discriminative candidate is added to the vocabulary bank. The *discriminative only* alternative is similar to the theoretical upper bound, but drops any candidates that are not semantically meaningful. The difference between performance on the upper bound and the discriminative only condition captures the cost of using only semantically meaningful candidates. The *hand-listed* alternative uses expert-generated attributes. In two categorization test using 25 categories from the CUB-200-2011 dataset (Wah et al., 2011) and 10 categories from the Leeds Butterflies dataset (Wang, Markert, & Everingham, 2009), the features

[2] For each POOF, a second bird part is also used in order to rotate and scale the images into alignment.

[3] www.mturk.com



Click Me or Press 1
Prairie Warbler (wikipedia)
Draw and identify the category.
Click Me or Press 2
Yellow Warbler (wikipedia)
Bubble:
Smaller (Press - or 'x')
Bigger (Press + or 'w')
Pass this image
Change the pair of categories

Bubble cost: 11

Points for correct identification: 65

Total score: 0

Total score needed to submit: 1000

Figure 1: The web-based game developed by Deng et al. (2013) uses human input to determine discriminative regions for bird categorization. When training a computer vision system, the human-selected regions are given special emphasis. (Taken from Figure 3 of Deng et al., 2013)

proposed by Duan et al. perform significantly better than the discriminative only approach. In addition, the proposed approach does better than the hand-listed approach for the 25 bird categories.[4].

Instead of requiring individuals to verbalize labels for candidate regions, Deng et al. (2013) requested the discriminative regions more directly. Inspired by Bubbles (Gosselin & Schyns, 2001), Deng et al. developed a web-based game that used human input to determine discriminative regions for bird categorization (Figure 1). In the game, a user earns points by correctly categorizing a test image into one of two categories. Correct categorization of the test image results in earned points, while an incorrect categorization results in steep penalty of negative points. To aid in the task. the user is presented with two scrollable lists of reference exemplars, one list for each category. When a trial begins, the test image is blurred and de-saturated, revealing only coarse spatial frequency and gray-scale information. A user can reveal a small circle of the original test image by clicking on a region. Once a user reveals about 30% of the object's bounding box, the amount of earnable points drops to zero. Users are incentivized to reveal only was is necessary to successfully categorize the test image. When training a computer vision system to classify birds from the CUB-200-2011 dataset, emphasis can be given to the human-selected regions, resulting in a 6% improvement in accuracy (32.8%) over the previous best result (26.7%).

0.1.3 Efficient training applications

In training, the long-term objective is transitioning the individual towards correctly recalling the label of an unknown test image. In a sense, the high cost in time, money, and effort to train novices has motivated decision support and human-in-the-loop strategies. Substantial research has focused on how to make training more efficient and thus less costly. A major hurdle is learning the relevant features for a categorization task. Highlighting can be used to make the features more salient to a human learner and potentially accelerate the learning process. Relevant features can be made more salient using a number of different methods.

To help novices learn where to look on fingerprint categorization task, Roads et al. (submitted) trained novices to direct their attention in a more expert-like fashion. Participants were tasked with categorizing a series of fingerprint pairs as coming from the same individual or different individuals. In an *expert gaze* condition, participants were cued to the fixation locations of an expert using a flashing red Gaussian bump. In a *novice gaze* condition, participants were cued to the fixation locations of a novice. In a *incongruent expert gaze* condition, participants were cued to expert fixation locations of a different pair of fingerprints than the pair being viewed. Using a pre- and post-test comparison without highlighting, participants in the expert gaze condition showed a larger shift towards expert-like gaze behavior than participants in the novice gaze condition. Performance between the expert gaze and incongruent expert gaze was not significantly different, suggesting that participants were learning about broad-scale features, but not fine-scale features.

Diagnostic features can also be highlighted by exaggerating them. In fine-grained categorization, within-class variation can be of a comparable magnitude to between-class variation, making it

[4] A Hand-listed comparison was not possible for the Leeds Butterflies dataset because there were no expert-generated labels associated with the dataset

difficult for a novice to know what variation is diagnostic. To make the relevant feature variation more salient, Dror et al. (2008) created caricature versions of a set of aircraft stimuli. Using results from a cluster analysis of similarity ratings (see Ashworth & Dror, 2000), eight aircraft stimuli were evenly divided into two groups. The *heterogeneous* aircraft all exhibited low inter-similarity (i.e., from different clusters) and the *homogeneous* aircraft all exhibited high inter-similarity (i.e., from the same cluster). For each group, the four exemplars were morphed together to create an average exemplar. Caricature exemplars were created by exaggerating the differences between the original exemplar and the corresponding average.

In an experiment that tested the usefulness of caricatures, participants were randomly assigned to one of four conditions. Each condition varied the set of images (original or caricature) used for training and test. Heterogeneous and homogeneous aircraft were trained and tested in separate halves of the experiment. In the training phase, participants saw 160 study (40 per category) and then 800 recall trials (200 per category). Study trials showed a labeled exemplar for 5 s. Multiple choice trials showed an unlabeled image and required users to make a 4-way multiple choice. A test was given halfway through the procedure and at the end of training. Accuracy was close to ceiling and was not analyzed. Response times decreased across training, with categorizations of heterogeneous aircraft being faster than the homogeneous aircraft and the caricature aircraft being faster than the original aircraft. The benefit of the caricature stimuli was larger for the homogeneous group. The pattern of response times during test shows a large disadvantage to switching from the caricature to the original homogeneous stimuli at the first test, but this disadvantage mostly dissipates by the second test. Together the results support the idea that caricatures provide the most gain when the task requires discriminating between similar stimuli.

0.2 Task transformation

Task transformations are used to help human agents perform difficult categorization tasks. In the most demanding situation, an individual is presented with a test image and is responsible for providing the correct label. Successful completion requires that the individual *recall* both the relevant features and the corresponding label. This scenario mirrors the demands faced by an expert. By the very nature of the task, novices will be unlikely to give a correct response. Instead of asking individuals to perform a direct categorization by recall, difficulty can be reduced by re-framing the task in a variety of ways.

A powerful approach is to transform a categorization task into a *visual comparison* task. In a visual comparison task, individuals compare a test image to a set of candidate reference exemplars (Figure 2). By choosing reference exemplars, a user can indicate similarity between the test image and the reference exemplars. Depending on the application, high similarity can indicate that the test image belongs to same category as the chosen reference exemplar or the presence of a particular attribute. Visual comparison tasks have two important consequences. First, visual comparisons decouple the task of knowing the relevant features and knowing the correct category label, allowing a user focus solely on visual information. The second important consequence is that co-presented images permit individuals to surmise relevant features, enabling non-experts to have a reasonable chance at successfully completing the task. Although notions of similarity will vary by individual (Murphy & Medin, 1985), individuals are likely to have shared beliefs regarding informative features (Brooks, Norman, & Allen, 1991). For example, when categorizing images such as skin lesions, everyone is likely to consider occlusion, lighting changes and orientation irrelevant.

0.2.1 Decision support applications

Building on the exploratory work of Brown, Robertson, Bisset, and Rees (2009), Robertson et al. (2014) evaluated the benefit of using visual comparisons to categorize skin lesions as either melanoma or not melanoma (i.e., benign lesions). Novice participants were assigned to one of three conditions. In the *multiple choice* condition, participants saw only the test image and two category labels from which to choose from. In the *feature exemplars* condition, participants saw the test

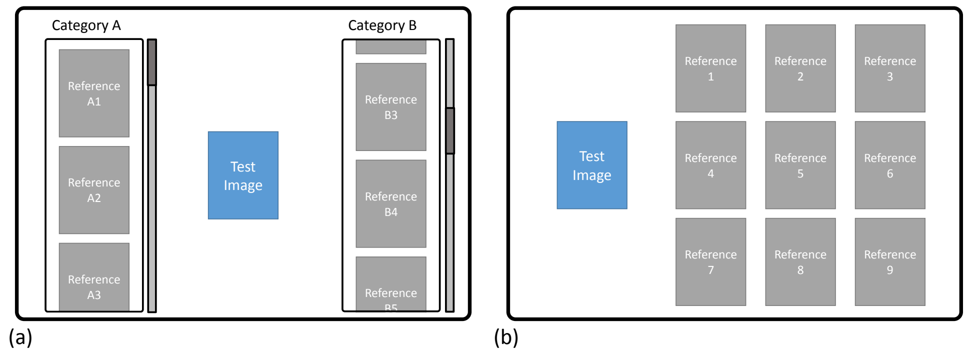


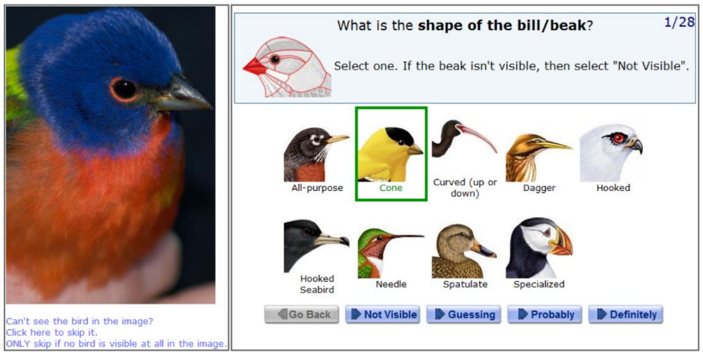
Figure 2: Cartoon illustrating two prototypical screen arrangements that utilize visual comparisons. (a) The category membership of the reference images is made clear to the user. (b) The category membership of the reference exemplars is hidden from the user.

image along with three reference exemplars that each illustrated a diagnostic feature of melanoma: asymmetry in shape, blurred borders, and non-uniform color. Each feature exemplar was accompanied by a short textual explanation of the informative feature. In the *category exemplars* condition, participants saw a test image in the center of the screen flanked by a scrollable view of 21 melanoma exemplars and a scrollable view of 21 benign exemplars.

On a test of 48 images (12 melanoma, 36 benign), average participant accuracy with the feature exemplars (.61) and category exemplars (.62) was significantly better than chance, while the multiple choice condition was not. The lack of a statistical difference between the two exemplar conditions suggests references exemplars are helpful but leaves many open questions regarding other differences between the two conditions: the number of references, the use of text, and the reference selection procedure. The exemplars for the feature condition were hand-picked while the exemplars for the category condition were randomly selected from a database. While randomized selection minimizes assumptions about the typicality of particular lesions–as Robertson et al. point out–randomized selection suffers from the fact that it will not appropriately sample from subcategories if any exist.

In another skin lesion application, Aldridge et al. (2011) selected reference exemplars more systematically by exploiting domain knowledge of exemplar typicality. Sixty-eight reference exemplars were hand-selected from five categories based on technical quality and because the authors considered the exemplars representative of a particular class. The reference exemplars were situated in a multi-stage decision support system in which users made a sequence of three visual comparisons about a test image. The final visual comparison implicitly categorized the test image into one of the five categories. Aldridge et al. make clear that the connection between the different screens was based on the experimenters' opinion of visual similarity and, to a lesser degree, the lesions' underlying pathological diagnosis (p. 280)."

Participants categorized 12 images of lesions into one of five categories by recall or using the decision support system. Using the decision support system, average accuracy for dermatology students (.99) and lay members of the public (.96) was near ceiling. Using recall, accuracy for dermatology students (.16) was much lower. Even after the dermatology students completed a 10-day clinical dermatology attachment, the decision support system fostered higher accuracy (.99) than the recall group (.51). The approach demonstrates that it is not always necessary to provide users with a clear indication of category labels during the decision process. Aldridge et al. point out that a limitation is that the reference exemplars and the screen groupings were selected and arranged by hand. A more scalable solution must specify a procedure for automatically selecting and arranging reference exemplars.



What is the **shape of the bill/beak**?

Select one. If the beak isn't visible, then select "Not Visible".

Can't see the bird in the image?
Click here to skip it.
ONLY skip if no bird is visible at all in the image.

Figure 3: The user interface used by Welinder et al. (2010) and Wah et al. (2011) to collect attribute labels using attribute-based visual comparisons. (Taken from Figure 5 of Wah et al., 2011)

0.2.2 Human-in-the-loop applications

Task transformations partner everyday human abilities with the computational power of computer systems to create accurate human-in-the-loop applications. It is challenging for computers to identify semantically meaningful parts (e.g., a beak) and define semantically meaningful similarity relationships. But what is hard for computer vision systems is relatively easy for human agents. Human agents can quickly locate parts and can rapidly assess similarity between stimuli. Human judgments will tend to be semantically meaningful because individuals can exploit general knowledge about the world to rule out many uninformative features. By using task transformations, image datasets are enriched with human judgments and computer vision systems receive a boost in categorization performance.

Jia et al. (2013) used in-group/out-group judgments in order to construct a database of images organized around coherent hierarchical categories (e.g., dalmatian, domestic dog, animal, organism). AMT participants were shown five reference exemplars belonging to a single category and asked to judge whether a test image belonged to the category implied by the reference exemplars. Using the same reference set, participants categorized a sequence of twenty test images as ’yes,’ the test image belongs to the category or ’no,’ it does not. In all, 20,000 sequences were categorized. On average, participants only needed about three seconds per test image. As a result, Jia et al. were able to embed a database of images within a hierarchical structure of similarity relations. The enriched dataset was used to evaluate different visual concept learning models, demonstrating a system that appropriately generalizes labeled exemplars to other category levels.

Instead of making visual comparisons based on the entire image, individuals can also make restricted comparisons about particular regions or features of an image. Restricted comparisons may be desirable if the complexity of the images leaves substantial flexibility in how to perform visual comparisons. Restricted comparisons also provide one method for producing image annotations for large image datasets. For example, the Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset (Welinder et al., 2010; Wah et al., 2011) contains over 11,000 annotated images for 200 different bird species. Annotations were obtained by having AMT participants make visual comparisons along 25-28 specific attribute dimensions and select the most similar attribute exemplar (Figure 3). Users were shown a test image and a guiding question about a particular attribute dimension, (e.g., what is the shape of the bill/beak?). Users could then select from a set of attribute values that all displayed a corresponding prototypical exemplar image (e.g., cone, hooked, needle, spatulate). In addition to selecting the attribute exemplar they considered most similar to the query image, users indicated their level of confidence. The end result is that every image is annotated with attribute values.

Given a database with supplementary information, interesting computer categorization systems can be built. By leveraging the location and attribute annotated CUB-200 database, Branson and colleagues (Branson et al., 2010; Branson, Van Horn, Wah, Perona, & Belongie, 2014) built a human-in-the-loop categorization system. The system begin by extracting computer vision features from an unknown test image. Using the computer vision features as a starting point, the system determines the most intelligent questions to pose to a user. For example a user may be asked to click on a specific part (e.g., the beak) or indicate the attribute value of a specific dimension (e.g., a hooked bill).

Users are aided in answering these questions through the use of attribute exemplars. Such a system makes it possible to categorize an unknown image with improved accuracy.

If a human-computer system can correctly categorize unknown images, it becomes possible to study a number of higher-order questions. For example, if images of animals can be correctly categorized it may be possible to develop population estimates and migration models. Swanson et al. (2015) deployed 225 camera traps across a 1,125 km^2 area in the Serengeti National Park, Tanzania from 2010-2013. The motion-activated cameras produced 1.2 million image-sets (each image-set contains 1-3 photographs taken in a single burst). Using the citizen science website Snapshot Serengeti (www.snapshotserengeti.org), volunteers categorized the collected images into one of 48 categories (e.g., cheetah, Grant's gazelle, Thomson's gazelle). Snapshot Serengeti helps novice users categorize an image-set by providing users with the ability to perform holistic visual comparisons and attribute-based visual comparisons. Users can select a category label from a drop-down list to see reference exemplars, a brief textual description and a list of similar categories. Users can also narrow the list of categories by specifying the attribute values of five attribute dimensions (i.e., pattern, color, horns, tail, build). Collecting multiple categorizations for each image-set results in a consensus classification. As a validation procedure, volunteer responses were compared to a subset of expert-labeled image-sets. The consensus classification was 96.6% accurate on the expert-labeled image-sets. Swanson et al. envision the classified images being combined with time and place information in order to conduct various ecological analyses. Citizen science projects demonstrate how visual comparison tasks can be used to achieve high levels of categorization accuracy and massive throughput.

0.2.3 Efficient training applications

When undergoing training to categorize a domain of images, trainees are faced with the difficult task of learning the relevant features. Although the end goal is to be able to recall the correct category label without assistance, task transformations can provide developmental stepping stones. In the same way that visual comparisons help novices in decision support and human-in-the-loop applications, visual comparisons have the potential to help novices by facilitating the feature discovery process and mitigating initial learning demands. Some research has examined task transformations in the service of training, but a systematic exploration is missing.

A common task transformation used in training is a *same/different* classification task. Instead of providing a category label, participants sequentially view two images, one at a time, and then report whether the two images belong to the same category or different categories. A major benefit of using the same/different task is that it eliminates the need to learn an association between visual features and a label. Instead, learners can focus on learning the relevant visual features. The same/different task has become a common method of evaluating novice and expert participants in perceptual expertise research precisely because it eliminates the possibility of errors where participants know the relevant features but do not know the correct label (e.g., Gauthier, Williams, Tarr, & Tanaka, 1998; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999; Scott, Tanaka, Sheinberg, & Curran, 2008; Tanaka, Curran, & Sheinberg, 2005; Wong, Palmeri, & Gauthier, 2009). Variants of the same/different task have been used in other applications. In two visual learning software programs, Kirchoff and colleagues (Kirchoff et al., 2014; Burrows, Krebs, & Kirchoff, 2014) used a number of different task transformations to help novices learn to categorize plants. Inspired by the sequential same/different task, participants performed a simultaneous same/different task in which two images were co-presented and participants indicated if the exemplars came from the same category or different categories. Categories learned using the software resulted in higher categorization performance than categories learned outside the software.

Simultaneously presenting exemplars provides the learner with an opportunity to notice similarities and differences Wahlheim et al. (2011) examined the influence of co-presented exemplars on learning. Participants were trained to categorize images of birds into one of twelve visually similar bird families (i.e., families from the same taxonomic order). During the training phase of the first experiment, participants studied 72 different labeled exemplars (six per category), one time each. Subjects were randomly assigned to one of two presentation conditions. In both presenta-

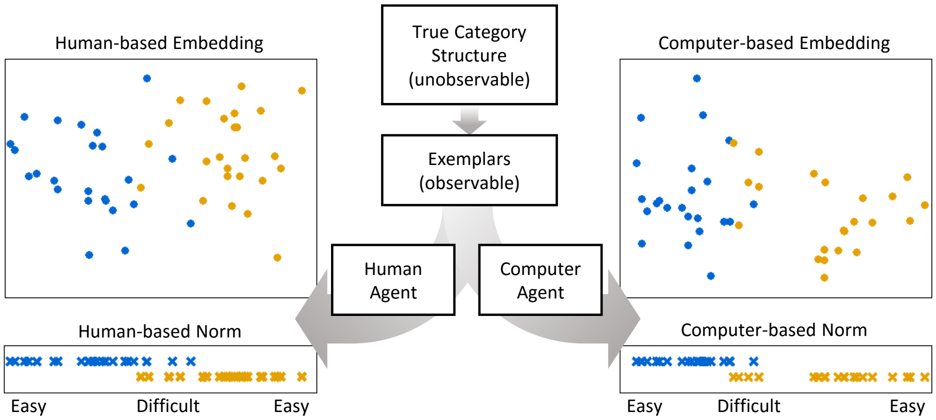


Figure 4: The category structure of a given domain describes how different exemplars are related to on another. Information obtained by querying human or computer agents can be used to infer a model of category structure. Depending on the collected information, a model of category structure may infer an multidimensional embedding or information about the relative difficulty of an exemplar (i.e., a norm). In general, the human-based and computer-based category structure models will be different.

tion conditions, each trial block contained six exemplars. In the *singles* condition, each study trial contained one labeled exemplar, yielding six trials per block. In the *pairs* condition, each study trial contained two labeled exemplars, yielding three trials per block. Study time was held constant across conditions, allowing 8 s per exemplar. In addition to a presentation condition, each category was evenly assigned to one of two *sequence* conditions. In the *blocked* condition, a given block showed exemplars from only one category. In the *interleaved* condition, each block showed one exemplar from each of the six categories. The conditions interacted such that in the *pairs-interleaved* case, co-presented exemplars were from different categories; while in the *pairs-blocked* condition, co-presented exemplars were from the same category.

During a test phase, participants performed a 12-way multiple choice categorization for 72 seen exemplars (six per category) and 48 novel exemplars (four per category). Categorization accuracy was better for interleaved categories than for blocked categories. Furthermore, accuracy for co-presented exemplars from different categories (pairs-interleaved) was better than co-presented exemplars from the same category (pairs-blocked). There was no corresponding difference for singles-interleaved and singles-blocked categories. These results are consistent with the hypothesis that co-presenting exemplars from different categories helps learners identify discriminative features and enhances learning.

0.3 Structure sensitive selection

Structure sensitive selection employs knowledge of the underlying category structure to intelligently select exemplars shown to the user. The category structure of a given domain describes the perceived relationship between different exemplars. Although not directly observable, category structure can be modeled as a multidimensional space where the dimensions specify relevant feature variation (see Figure 4). Exemplars are modeled as points such that nearby points are similar. A model of category structure is an embedding that identifies a subset of perceptually relevant dimensions among an infinite number of existing dimensions. In the simplest case, a model of category structure captures the distance of an exemplar from a category boundary.

Human-based and computer-based models of category structure are likely to emphasize different feature dimensions. Human-based category structure can be obtained by collecting human judgments and using multidimensional scaling algorithms to recover an embedding (e.g., Shepard, 1962; Carroll & Chang, 1970; van der Maaten & Weinberger, 2012). Computer-based category structure is defined using extracted computer vision features. When using computer vision features to model category structure, there is a larger risk of a semantic gap between predicted similarity and human-perceived similarity.

A model of category structure permits exemplars to be selected in various ways. Given some exemplar, it is possible to select the most similar exemplar. Furthermore, a category structure allows you to select prototypical exemplars that represent the average features of a set of exemplars. In addition, exemplars near a category boundary can be identified and avoided if they are difficult to categorize.

0.3.1 Decision support applications

By modeling exemplars using computer vision features, a decision support application can select reference exemplars that are visually similar to a test image. Brodley et al. (1999) developed a system that helps users diagnosis high resolution computed tomography images. The user provides an undiagnosed test scan, which is compared to a labeled database using low-level computer vision features. The system returns the top four matching reference exemplars. Each exemplar is composed of 20-50 cross-sectional slices (with the key slices identified) accompanied by additional textual information (e.g., the scan date and doctor) and the corresponding diagnoses. One limitation of using low-level computer vision features is that there may be a semantic gap between the computer-vision features and the features that individuals use.

Although computer vision features risk a semantic gap, some domains are well described by computer vision features. Leafsnap (Belhumeur et al., 2008; Kumar et al., 2012) leverages the fact the shape profile of a leaf–which is highly diagnostic–can be well captured by segmentation algorithms. A user takes a photograph of an leaf they wish to categorize and Leafsnap extracts useful computer vision features. Similarity of different leaf species is computed in a feature space and reference exemplars of the top few matches are returned to the user. The burden of categorizing a leaf into one of hundreds of possibilities is reduced to visually comparing the unknown leaf to a small number of likely candidates.

In many applications it is important to select reference exemplars while taking into account within-category variability. If a given category has distinct sub-categories or large within-category variability, randomly selected reference exemplars may mislead novices when they make visual comparisons. Roads and Mozer (submitted) created a decision support system that uses an exemplar-level psychological embedding in order to select an optimized set of references. The system is designed to operate with minimal information about the test image. Using a cognitive model of the user, R reference exemplars are selected such that the expected categorization accuracy–across the entire category structure–is maximized. When tested on AMT participants making visual comparisons of faces, optimized reference sets produced better performance that randomly selected reference sets.

0.3.2 Efficient training applications

Category structure can also be exploited to make training more efficient. By drawing on category structure information, Hornsby and Love (2014) created a set of idealized training exemplars from which to teach individuals to categorize mammograms as normal or tumorous. An idealized set of exemplars was assembled by removing exemplars that were relatively difficult. Relative difficulty was determined by norming a set of mammography images. In the norming study, AMT participants categorized mammograms as normal or tumorous. Categorization of the test image was facilitated by a labeled normal reference exemplar and a labeled tumorous reference exemplar. Based on accuracy, exemplars were split into three groups of difficulty (i.e., easy, medium and hard) based on average accuracy.

The idealized set of exemplars were evaluated by training a second group of AMT participants to categorize mammograms using trial-and-error learning. Participants were randomly assigned to one of two training conditions. In the *idealized* condition, the stimuli were drawn solely from group of easy exemplars. In the *actual* condition, the stimuli were drawn equally from easy, medium, and hard exemplars. On each training trial, participants categorized a single test image as normal or tumorous and received corrective feedback. Following the training phase, subjects completed 18 novel test trials (3 easy, medium, hard images from each category). Participants were more accurate in classifying novel test items in the idealized condition than the actual condition. Examining performance by exemplar difficulty, accuracy was better in the idealized condition for easy and medium exemplars, but worse for the hard exemplars. While idealized training results in greater categorization performance overall, training on idealized items hurts generalization performance to hard exemplars.

0.4 Response-adaptive selection

Response-adaptive selection allows the selection of exemplars to vary over time as a function of a user's previous responses. A user's response reveals information about the similarity of a test image to category exemplars. Depending on the application, this information can be used to update a system's model of the category structure, predicted label of the test image, or model of a learner's knowledge. Once updated, the system can select new exemplars, prioritizing exemplars that will provide the biggest gain to the system's objective.

0.4.1 Decision support applications

By integrating user similarity responses over a sequence of similarity judgments, Fang and Geman (2005) demonstrated how the selection of reference exemplars can be dynamically prioritized. The approach also showed how to gradually bridge the semantic gap between computer vision features and psychological similarity. As an initial test of their approach, participants were asked to hold a test image in mind (i.e., the facial identity of specific individual) and make similarity comparisons between the mental test image and images of individuals. The category structure of the faces was modeled using computer vision features derived from kernel Fisher's discriminant analysis. Users were first presented with a set of R reference exemplars that spanned the category structure. After making a selection, the user's response was combined with the modeled category structure in order to select a new set of R reference exemplars. By making a sequence of similarity judgments, users successively refined the set of reference exemplars until a match occurred.

Given a user selection, the next set of eight reference exemplars were selected by attempting to maximize the potential information of the user's next selection. First, all the reference exemplars in the database were given a weight based on how similar they were to the selections made so far. All the exemplars in the database were then clustered into R groups based on proximity in the KFDA embedding space. Group assignments are manipulated such that the sum of the total weight for each group was approximately equal. Thus, not all groups had the same number of exemplars, but the sum weight of each group was approximately equal. Once this was achieved, the highest weighted member from each group was used as a reference exemplar in the next reference set (Figure 5). The benefit of using this approach, is that it helps ensure that useful information will be learned at each stage. In contrast, displaying the R exemplars with the highest weights may tell you very little if all of exemplars are bad choices.

Ferecatu and Geman (2009) extended the approach of Fang and Geman to handle larger databases and more complex semantic classes (art images, architecture, and history). In addition to providing a more rigorous theoretical justification of the main algorithms, Ferecatu and Geman explored the difference between holding a test image in memory versus having a test image available to examine. They found that displaying examples of the target class did not significantly improve results. This is potentially due to the fact that the category distinctions occur at a relatively ordinate level (e.g. lions vs. leaves). It is possible that displaying a target exemplar would have produced a significant difference with finer-grained category distinctions. In both the approach used by Fang and Geman and Ferecatu and Geman, the semantic gap between computer-based similarity and human-based similarity still provides a major challenge to performing efficient image retrieval. It

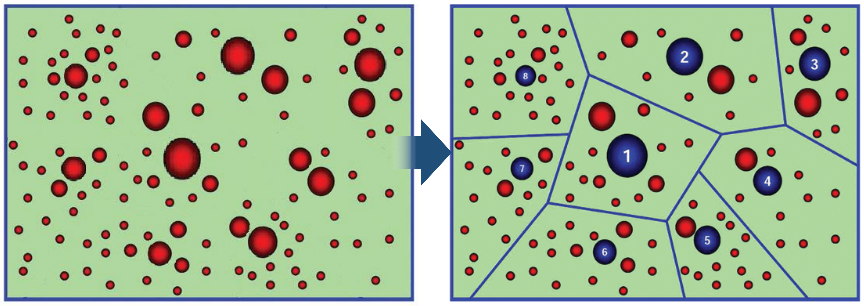


Figure 5: Illustration of a reference selection procedure based on information maximization. This conceptual approach is used by Fang and Geman (2005), Ferecatu and Geman (2009), Wah et al. (2014), and Wah et al. (2015). Figure modified from Figure 2 in Ferecatu and Geman (2009).

would be valuable to know ahead of time how images in a database relate to one another based on human similarity.

0.4.2 Human-in-the-loop applications

Using response-adaptive selection, human-in-the-loop approaches can continuously update beliefs regarding the test image and select reference exemplars that will maximize the information gain of the next user response. Using AMT participants, Wah et al. (2014) first collected human similarity judgments on a database of bird images (CUB-200-2011). Participants saw a test image along with a 3×3 grid of reference exemplars. Participants were instructed to select the reference exemplars that were clearly dissimilar from the test image. Each screen yielded a set of images marked as similar (similar set) and a set of images marked as dissimilar (dissimilar set). By assuming that every item in the similar set is more similar to the test image than every item in the dissimilar set, 8 to 20 similarity triplets can be inferred. Each similarity triplet states that the test image is more similar to reference image i than reference image j. After collecting a large number of similarity triplets, t-distributed stochastic triplet embedding (t-STE) (van der Maaten & Weinberger, 2012) was used to infer a psychological embedding.

Simulated participants were used in experiments to assess the added value of human similarity judgments. Simulated participants were modeled using both a deterministic choice function and a stochastic choice function with the free parameters trained on real user responses. The deterministic simulated users were specified by a choice function that always selected the reference exemplar most similar to the test image. The noisy simulated users were specified by a choice function that probabilistically selected a reference exemplar in proportion to its similarity to the test image (see Appendix A for modeling details). Following the selection procedure developed by Geman and colleagues (Fang & Geman, 2005; Ferecatu & Geman, 2009), computer vision features were integrated with (simulated) user similarity judgments and reference sets were dynamically selected in an information-maximizing manner. Simulated users were shown a test image and a 3×3 grid of reference exemplars and made a most-similar judgment based on the specified choice function.

Without computer vision, deterministic users needed to judge an average of 4.32 screens. In combination with computer vision, users only needed to judge an average of 2.72 screens. Thus computer vision reduces the number of screens a user must judge in order for the system to classify an unknown test image. If computer vision is used, but reference sets are selected randomly, an average of 3.28 screens needed to be judged. Response-adaptive selection reduces the number of screens that need to be shown to a user.

If it possible to specify the locations of discriminative regions, it is also possible to constrain user similarity judgments to specific regions of an image. Wah et al. (2015) learned specific regions from a bird image dataset (CUB-200-2011) through a combination of automatic region discovery and manual selection. While the regions are not required to have a pure semantic interpretation, they roughly correspond to regions like the head, chest, and back. Similar to to Wah et al. (2014), Wah et al. (2015) presented users with a test image along with a 3×3 grid of reference exemplars. However, the test image and reference images were cropped to one particular region. Users were told to select the cropped reference exemplar that was most similar to the cropped test image. User similarity judgments were collected for all five regions. After collecting a sufficient number of similarity judgments, t-STE was used to infer a psychological embedding for each region.

Similar to Wah et al. (2014), Wah et al. (2015) explored the impact of using guided similarity judgments. A small experiment found that on average, it takes a human user less time to carry out a similarity judgment on a localized image than a nonlocalized image (11.35 ± 10.17 s and 16.36 ± 14.31 s respectively), although statistical significance is not reported. Interestingly, consistency between users on the localized and non-localized task are very similar. One might have supposed that restricting the amount of available information would increase the extent to which user judgments agree. It seems that the primary advantage with localized similarity judgments is that users can perform it more quickly. As before, the embeddings can be used in combination with computer vision to classify unknown test images. A noisy user model was trained on real human data to serve as a surrogate participant in a categorization experiment. Noisy simulated user responses indicate that the fewest screens are needed (9.85 screens) when localized and nonlocalized similarity judgments are used in combination with computer vision. Localized only (9.99 screens) and nonlocalized (11.53 screens) require incrementally more screens on average. These results suggest that localized similarity judgments are more informative than non-localized, but that using both types of similarity judgments is useful. Wah and colleagues (Wah et al., 2014, 2015) demonstrate an interesting approach to collecting similarity judgments to enrich an existing database and show how response-adaptive selection can be used to intelligently request help from non-expert human users.

0.4.3 Efficient training applications

During training, a learner will see many trials but is constrained to see one trial at time. A training system must prioritize which exemplars to use on a trial according to what is most beneficial for the learner. Response-adaptive selection allows the training system to adaptively select exemplars based on the users performance on past trials.

The most direct performance measure is categorization accuracy. While it is immensely beneficial to leverage accuracy information, other behavioral measures are available as well. Kellman and colleagues (Kellman, 2013; Mettler & Kellman, 2014; Rimoin, Altieri, Craft, Krasne, & Kellman, 2015) prioritize items for study using the Adaptive Response-Time-based Sequencing (ARTS) system that takes into account accuracy, response times, and the number of trials since the last presentation. Categories that are incorrectly categorized are given high priority while categories that are correctly categorized are given priority proportional to the response time on the last category presentation. The longer a category has not been presented, the higher priority it gets. The relative contribution of each component is set by a number of free parameters.

Mettler and Kellman (2014) tested their performance-based scheduling by training participants to categorize images of butterflies into 12 categories. Participants were randomly assigned to one of three scheduling conditions. Each scheduling condition continued until performance on all categories was as criterion (5 out of 6 correct with RT less than 3 s). In some scheduling conditions, when performance for a category reaches criterion, no additional trials use the category, i.e., the category is retired. In the *random* condition, scheduling was purely random and had no category retirement. In the *adaptive* condition, scheduling was done with adaptive category sequencing with category retirement. In the *mini-blocks adaptive* condition, scheduling was first done on the basis of blocks and then used adaptive category with retirement.

In the first experiment, participants were trained to criterion in classifying 12 categories of butterflies and saw eight different exemplars per category. Each category corresponded to one taxonomic genus and three species. In the pre-test phase participants were shown 12 trials (one per category). Each trial required the subject to read a label and select one of four images (each from a different category). Participants were given no feedback. In the training phase, participants saw a variable number of trials. Each training trial showed a category label and two images from different categories. Participants selected the image corresponding to the label and were given corrective feedback. In the immediate post-test phase, subjects were tested on the same 12 exemplars from the pre-test and 12 novel exemplars (one per category). Each screen showed one label and four images. One week later, participants underwent a delayed post-test that was identical to the immediate post-test.

Since training to criterion results in a different number of trials per participant, Mettler and Kellman (2014) compared scheduling conditions using a measure of learning efficiency, defined as accuracy on the post-test items divided by the number of trials invested during training. Learning efficiency for the adaptive condition was higher than the mini-blocks adaptive condition and the random condition, although the differences were only marginally reliable. Learning efficiency between the random and adaptive conditions was significant when averaging across post-tests, differed marginally on the immediate post-test and was significant for the delayed post-test. These results suggest that adaptive scheduling is advantageous for learning efficiency.

In addition to using past performance, test exemplars can be dynamically prioritized based on category similarity. Substantial work has investigated whether category items should be sequenced so that within-category members occur consecutively (*massed*) or whether it is better to *interleave* exemplars from different categories (e.g., Carvalho & Goldstone, 2014a, 2014b; Kang & Pashler, 2012; Kornell, Castel, Eich, & Bjork, 2010). Should trials be sequenced such that a learner is able to see the commonalities between highly similar stimuli of the same category? Or should the learner be able to see the differences between less similar stimuli from different categories? The initial findings in more simplistic domains have recently been extended to learning naturalistic fine-grained visual categories.

Birnbaum et al. (2013) performed three experiments to better understand the benefits of interleaving and blocking while controlling for spacing effects. In Experiment 1, AMT participants learned eight species of birds. All participants were shown 32 study trials (four per category), each for 4 s. Each study trial consisted of labeled exemplar. Participants were randomly assigned to three training conditions, all training conditions used interleaving schedules. In the *contiguous* condition, participants saw study trials with no interjected trivia. In the *alternating-trivia* condition, every study trial was preceded by a single trivia question. In the *grouped-trivia* condition, eight trivia questions after every eight trials. After the study phase, participants performed a three min distractor task and then a post-test. The post-test required participants to categorize eight novel exemplars (one per category) in a 8-way multiple choice. Accuracy in the contiguous and grouped-trivia conditions were not significantly different. However, accuracy in the alternating-trivia condition was significantly lower. These results are consistent with the discriminative-contrast hypothesis. Suggesting that interleaving better supports the discovering of discriminative features.

In Experiment 2, participants learned to classify 16 butterfly species. During the training phase, all participants saw 64 study trials (4 per category), each study trial showing a labeled exemplar. Each species was assigned to either an *interleaved* or *blocked* training condition. Half the participants were assigned to a contiguous presentation condition and were half assigned to an alternating-trivia presentation condition with 10-s for each trivia question. The trials were arranged in blocks of four and followed a organizational pattern of I I B B I I B B I I B B I I B B. An I I sequence of blocks randomly interleaved one exemplar from each of the eight categories assigned to the interleaved condition. A B block presented all four exemplars of one of the categories assigned to the blocked condition. In an immediate post-test, participants categories 16 novel butterfly images (one per category) in a 16-way multiple choice. Performance in the interleaving condition was superior to that in the blocking condition when exemplars were presented contiguously. This result is consistent with discriminative-contrast hypothesis and suggests that the benefits of spacing and interleaving are subadditive.

In Experiment 3, participants learned to classify 16 butterfly species. During the training phase, all participants saw 64 study trials (4 per category), each study trial showing a labeled exemplar. All participants experienced interleaved study trials. Participants randomly assigned to one of two spacing conditions. In the *small-spacing* condition, the average number of trials between repetitions of the same category was three trials. In the *large-spacing* condition, the average number of trials between repetitions of the same category was 15 trials. In an immediate post-test, participants categories 16 novel butterfly images (one per category) in a 16-way multiple choice. Accuracy in in the large-spacing condition was significantly better than accuracy in the small-spacing condition. This result suggest that there are gains from the spacing effect in addition to the gains from interleaving.

0.5 Gaps in the literature

The presented work reveals three general strategies that have emerged for performing fine-grained categorization. Across these strategies, four different techniques have been employed. Jointly reviewing the relevant literature reveals a number of gaps. It should be noted that many of the presented works span multiple techniques but were presented within the technique that they most exemplify. For example, the approach of Wah and colleagues (Wah et al., 2014, 2015) is a strong demonstration of response-adaptive selection, but also utilizes task transformations and structure sensitive selection. In addition, some work only partially explores a relevant technique. In order to obtain a more systematic evaluation, certain gaps will have to be filled. In particular, decision support and human-in-the-loop applications are well positioned to tackle open questions regarding structure-sensitive selection. Efficient training applications have many avenues to explore and would benefit by drawing on ideas from decision support and human-in-the-loop applications.

0.5.1 Gaps in decision support and human-in-the-loop applications

Visual comparisons provide a foundational platform from which to aid users in making and learning categorization judgments. However, there are many considerations that are poorly understood. The presentation of multiple reference exemplars provides the user with a sense of context and suggests features that may be important for performing categorization. However, the elicited context can be challenging to model explicitly. Gomes, Welinder, Krause, and Perona (2011) ran into context issues while crowdsourcing categorization judgments for various image datasets. AMT participants were presented with a interface showing a 6×6 grid of exemplars and asked to group them into as many different categories as they liked. Gomes et al. found that depending on the set of shown exemplars, individuals grouped images differently. If a user was presented with mostly outdoor scenes with a couple of indoor scenes, users typically grouped all the indoor scenes into one category. However, if a user is presented with only indoor scenes, they group images into finer-grained indoor categories such as kitchen, living room, and office. The set of reference exemplars has the potential to influence perceived similarity, and it is not clear how to best model these effects so that they may be incorporated into a reference selection procedure.

Related to the issue of context is determining the number of reference exemplars to show at one time. Geman and colleagues (Fang & Geman, 2005; Ferecatu & Geman, 2009) found that eight references was optimal because, using many fewer or many more has adverse consequences with real users (Fang & Geman, 2005, p. 643). It would be advantageous to have a better understanding of how the number of reference exemplars influences categorization performance. It is possible the optimal number of reference examples primarily reflects a limit on attentional resources. With a small number of reference exemplars, adding one more is advantageous because it creates a better context. However at a certain point, the high number of reference exemplars could result in perceptual crowding, potentially hurting performance. Alternatively, it is possible that the optimal number of reference exemplars primarily depends on the similarity relations of the categories as well as the amount of variability within categories. In other words, categories with higher variability should be permitted to have more reference exemplars.

In a number of applications, participants made a series of visual comparisons on one test image. Geman and colleagues provide one strategy for intelligently sequencing screens in an online fashion. Another approach would be to learn beforehand an optimal tree of stages in the same spirit as Aldridge et al. (2011). Both context and the number of reference exemplars would presumably play a large role in the optimal staging structure.

0.5.2 Gaps in efficient training applications

Guiding attention with highlighting

Highlighting has a strong toe-hold in decision support and human-in-the-loop applications, but a comparable research agenda is absent from efficient training applications. The forms of highlighting used in decision support and human-in-the-loop applications may be useful in efficient training applications. In decision support applications, substantial evidence suggests that some forms of highlighting can be used in a beneficial manner(e.g. Krupinski et al., 1993; Litchfield et al., 2010; Donovan et al., 2008). The decision support highlighting used by Berg and colleagues (Berg & Belhumeur, 2013; Berg et al., 2014) also seems promising and warrants a controlled evaluation with users. If the general approach of Berg and colleagues is successful in improving user performance, other approaches could be recruited in a similar spirit. For example, both of the human-in-the-loop applications outlined by Duan et al. (2012) and Deng et al. (2013) yield information regarding diagnostic regions and could be used as the basis of highlighting.

Existing work using highlighting in efficient training has much to gain by borrowing from decision support and human-in-the-loop applications. Roads et al. (submitted) demonstrate one parallel by using cueing techniques like those used in medical decision support. The caricature training used by Dror et al. (2008) is similar in spirit to the approach of Berg and colleagues, in that the relative saliency of informative features is enhanced. However, the existing training applications could be greatly expanded by importing the computer system machinery that excels at finding diagnostic regions. Using highlighting may help novices discover the relevant features more quickly and make training more efficient.

Expanding the use of task transformations to calibrate difficulty

Decision support and human-in-the-loop applications have taken great advantage of task transformations in order to boost categorization performance. Efficient training applications have underutilized this technique. Some efficient training applications (e.g., Wahlheim et al., 2011; Kirchoff et al., 2014; Burrows et al., 2014) have begun to investigate task transformations, but a more comprehensive research agenda is needed. Task transformations have the potential to help learners by softening the initial learning curve. Instead of having study trials with labeled exemplars (Wahlheim et al., 2011) or match/mismatch trials, learners could be presented with an unlabeled test image and labeled reference exemplars. Learners must then select the exemplar that is in the same category as the test image. Such an approach forces the learner to take an active role but also permits learners to observe similarities and differences between the reference exemplars. Other potentially useful task transformations include attribute-based visual comparisons (e.g., Robertson et al., 2014; Branson et al., 2010). Having learners complete trials that teach diagnostic attributes may accelerate the learning process. Lastly, including textual descriptions may improve learning efficiency (e.g., ?, ?; Swanson et al., 2015). By borrowing from decision support and human in the loop applications, the difficulty of a trial can be calibrated to the ability of the learner, potentially increasing the efficiency of training.

Exploiting response-adaptive selection to efficiently schedule trails

Decision support and human-in-the-loop applications have utilized response-adaptive selection to great effect, but efficient training applications have only begun to explore this technique. The prioritization scheme used by Mettler and colleagues (Mettler & Kellman, 2014; Kellman, 2013) embodies the computer system's belief about what the learner currently knows and what trials are difficult. A more comprehensive model could also incorporate knowledge about the category structure to more accurately calibrate the difficultly of trials and promote discriminative contrast. training more efficient.

A comprehensive prioritization scheme could take into account category structure in order to more accurately predict difficulty. When learning names of novel faces, Pantelis, van Vugt, Sekuler, Wilson, and Kahana (2008) showed that as the number of faces in the embedding vicinity of the test face increased, recall performance decreased. Furthermore, incorrect recalls were more likely to come from nearby faces. Taken together, these results suggest that more difficult categories will reside in dense regions of the category structure and should be given a priority that reflects this source of difficulty.

A prioritization scheme could also consider category structure in order to promote discriminative contrast. In training, discriminative contrast has primarily been studied using blocked and interleaved paradigms. Scheduling of this sort captures a coarse-level similarity distinction of being in the same category or in different categories. Research suggests that the effectiveness of interleaving and blocking is contingent on the similarity of the to-be-learned categories (see Richler & Palmeri, 2014 for a review). Interleaving exemplars from different categories is advantageous when categories are very similar. When categories are highly discriminable, it is more effective to block trials. Although this distinction is a useful post hoc explanation, it remains unclear how to predict a priori whether interleaving or blocking will be more effective for a set of categories. For example, what category level (e.g., animal, bird, perching bird, swallow, cave swallow) should determine a switch from blocking to interleaving? It seems a broader view of similarity-based scheduling is necessary.

Instead of making scheduling distinctions at a particular category level (which correlates with the degree of similarity), one could attempt to model the between-category similarity directly. Given between-category similarities, trials could be scheduled along a gradient of blocked and interleaved, as well as structured in a hierarchical fashion. For example, if categories A1, A2, and A3 form a similarity cluster and B1, B2, and B3 from another similarity cluster, then trials could be scheduled as A1 A2 A3 A1 A2 A3 B1 B2 B3 B1 B2 B3. Even though all categories are at the same level of distinction, more similar categories are interleaved and less similar categories are blocked. Such an approach could be further modified to be stochastic rather than deterministic. In a stochastic scheme, the prioritization of the next test exemplar would reflect previously seen exemplars. Using a common currency of priority, both exemplars could be dynamically prioritized to promote discriminative contrast of highly similar categories.

When dynamically prioritizing exemplars, the protocol for assigning priority could shift over time. Mettler and colleagues use a prioritization function that effectively places the strongest priority on the most difficult exemplars. Alternatively, the prioritization function could change over time. Initially relatively easy exemplars could be given priority, but as the learner becomes proficient, priority could be shifted to more difficult exemplars (e.g., Pashler & Mozer, 2013). This same idea could be extended to the selection of the task. Initially learners could perform trials involving visual comparisons, but as they become more proficient, priority could shift to recall trials. Response-adaptive selection provides the opportunity to migrate to more efficient and practical training systems.

Appendix A

Models of Visual Categorization

There are many ways to model human visual categorization. A cognitive model typically distinguishes between the knowledge representation and the mental processes that operate on the representation. In categorization, the cornerstone of the knowledge representation is the *category structure*. A formal model of category structure specifies the relationship between exemplars. In categorization, there are two major approaches to formally modeling category structure (e.g., Sattath & Tversky, 1977). *Geometric* models assume that stimuli can be modeled as points in a D-dimensional geometric space. The relationship between stimuli is a function of the distance between their corresponding points. *Network* models, such as trees, represent each stimulus as a node in a connected graph. Relations between stimuli are a function of the path length between their corresponding nodes.

Each approach has strengths and weaknesses. For example, geometric models impose an upper bound on the number of points that can share the same nearest neighbor (Tversky & Hutchinson, 1986). If a given domain contains hierarchical categories (e.g., animal, bird, sparrow), a network model may be more appropriate (Sattath & Tversky, 1977). Despite the limitations of geometric models, they have been successfully applied to many aspects of visual categorization (e.g., Nosofsky, 1986, 1987, 1992). When modeling category structure, the the majority of the presented works utilize geometric models. The remainder of this section covers geometric models in more depth.

A.1 Geometric models

Geometric models are one common class of cognitively-plausible categorization models. Geometric models assume that stimuli (i.e., category instances) can be represented as a point in a D-dimensional geometric space. Stimulus i is therefore represented as the point $\boldsymbol{z}_i$ in this space. Typically, the similarity between two stimuli is the primary relationship of interest. The similarity between two stimuli is determined by the *distance function* ($d\colon \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$) and the *generalization function* ($g\colon \mathbb{R} \to \mathbb{R}$),

$$s_{ij} = g(d(\boldsymbol{z}_i, \boldsymbol{z}_j)), \tag{A.1}$$

where s_{ij} is the similarity between stimulus i and j. Given a distance function and a generalization function, the probability a user makes a specific choice (e.g., selecting a label, selecting an exemplar) can be modeled as

$$P(r|q) \propto g(d(\boldsymbol{z}_q, \boldsymbol{z}_r)). \tag{A.2}$$

This equation might represent the probability of selecting reference exemplar r, given some test image q. This formulation provides a powerful way to model and predict user behavior.

A.1.1 Distance functions

Given an D-dimensional space, a number of distance functions can be used. Depending on the application, a particular distance function may be most appropriate (e.g., Lesot, Rifqi, & Benhadda, 2009). A commonly used distance function is the Minkowski distance of order ρ,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} |x_i - y_i| \right)^{\frac{1}{\rho}}. \tag{A.3}$$

The Minkowski distance is a valid metric for any value of ρ such that $\rho \geq 1$. For most applications, ρ is set to 1 (L_1/Manhattan distance) or 2 (L_2/Euclidean distance). The decision to use L_1 and L_2 distance is typically based on the psychological nature of the stimulus dimensions (e.g., Ashby & Townsend, 1986). The L_1 distance is typically used if the the stimulus dimensions are believed to be psychologically separable (e.g., width and height of a rectangle). The L_2 distance is used for domains where the stimulus dimensions are psychologically integral (e.g., hue and saturation).

If stimulus dimensions are separable, judgments of similarity can depend on the current allocation of attention to different dimensions. Category learning can be characterized in terms of the adaptation of weights to emphasize discriminative features (e.g., Jones, Maddox, & Love, 2005). A weighted Minkowski distance can be used to capture the flexible nature of attention,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} w_i |x_i - y_i| \right)^{\frac{1}{p}}, \tag{A.4}$$

where w_i indicates the weight given to the ith dimension and the weights are constrained such that $\sum_i w_i = 1$.

In some applications, there is no appreciable difference between using L_1 and L_2 distance (e.g., Fang & Geman, 2005; Wah et al., 2014). Under these circumstances, L_2 is typically used for its ease in interpretation and convenient mathematical properties.

A.1.2 Generalization functions

The generalization function g defines the degree of generalization from one stimulus to another. The generalization function is typically assumed to be isotropic and a monotonically decreasing function of distance (Shepard, 1987). Integrating the models of Jones, Love, and Maddox (2006); Jones, Maddox, and Love (2006); Nosofsky (1986); Shepard (1987) into their most general form, we obtain the generalization function:

$$g(d) = \gamma + \exp(-\beta d^{\tau}), \tag{A.5}$$

where γ, β, and τ modulate the gradient of generalization.

A variety of alternative generalization functions are possible (e.g., Lesot et al., 2009; Kruschke, 2008). For example, the generalization function used by Wah and colleagues (Wah et al., 2014, 2015) replaces the gradual influence of the γ parameter with more abrupt behavior:

$$g(d) = \max(\theta, (1 - \theta) \exp(-\beta d^{\tau})), \tag{A.6}$$

where θ, β and τ are free parameters that control the gradient of generalization. Demonstrating another alternative, Fang and Geman use a generalization function where $g(d) = \frac{1}{d}$.

A.1.3 Embeddings

So far, we have assumed that a set of points $\boldsymbol{z}_i$ is known for every stimulus. A number of approaches are available for deriving the set of points $\boldsymbol{z}_i$. Computer vision features or human judgments can be used to define an embedding. When using computer vision features, the feature values are the basis of the points. Alternatively, an embedding can be derived using human judgments.

Multidimensional scaling (MDS) is a class of techniques for inferring a psychological embedding $\mathbf{z}$ given proximity data p_{ij}. Proximity data may be derived from many types of data, such as response times, confusion matrices, similarity judgments and dissimilarity judgments. The *representation function* $f(p_{ij})$ specifies how the proximities should be related to the distances d_{ij}

$$f\colon p_{ij} \to d_{ij}(\mathbf{z}), \tag{A.7}$$

where the particular choice of f specifies the *MDS model*.

References

Aldridge, R. B., Glodzik, D., Ballerini, L., Fisher, R. B., & Rees, J. L. (2011, May). Utility of non-rule-based visual matching as a strategy to allow novices to achieve skin lesion diagnosis. *Acta Dermato-Venereologica*, *91*(3), 279-283. Retrieved from `http://europepmc.org/articles/PMC3160473` doi: 10.2340/00015555-1049

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011, May). Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(5), 898-916. doi: 10.1109/TPAMI.2010.161

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological review*, *93*(2), 154-179. doi: 10.1037/0033-295X.93.2.154

Ashworth, I. I., Alan R.S., & Dror, I. E. (2000). Object identification as a function of discriminability and learning presentations: The effect of stimulus similarity and canonical frame alignment on aircraft identification. *Journal of Experimental Psychology: Applied*, *6*(2), 148-157. Retrieved from `http://0-search.proquest.com.libraries.colorado.edu/docview/614337495?accountid=14503`

Belhumeur, P. N., Chen, D., Feiner, S., Jacobs, D. W., Kress, W., Ling, H., ... Zhang, L. (2008). Searching the world's herbaria: A system for visual identification of plant species. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer vision - eccv 2008* (Vol. 5305, p. 116-129). Springer Berlin Heidelberg. Retrieved from `http://dx.doi.org/10.1007/978-3-540-88693-8_9` doi: 10.1007/978-3-540-88693-8_9

Berg, T., & Belhumeur, P. (2013, Dec). How do you tell a blackbird from a crow? In *Computer vision (iccv), 2013 ieee international conference on* (p. 9-16). doi: 10.1109/ICCV.2013.9

Berg, T., Liu, J., Lee, S. W., Alexander, M., Jacobs, D., & Belhumeur, P. (2014, June). Birdsnap: Large-scale fine-grained visual categorization of birds. In *Computer vision and pattern recognition (cvpr), 2014 ieee conference on* (p. 2019-2026). doi: 10.1109/CVPR.2014.259

Birnbaum, M., Kornell, N., Bjork, E., & Bjork, R. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory I& Cognition*, *41*(3), 392-402. Retrieved from `http://dx.doi.org/10.3758/s13421-012-0272-7` doi: 10.3758/s13421-012-0272-7

Branson, S., Van Horn, G., Wah, C., Perona, P., & Belongie, S. (2014). The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization. *International Journal of Computer Vision*, *108*(1-2), 3-29. Retrieved from `http://dx.doi.org/10.1007/s11263-014-0698-4` doi: 10.1007/s11263-014-0698-4

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., & Belongie, S. (2010). Visual recognition with humans in the loop. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer vision - eccv 2010* (Vol. 6314, p. 438-451). Springer Berlin Heidelberg. Retrieved from `http://dx.doi.org/10.1007/978-3-642-15561-1_32` doi: 10.1007/978-3-642-15561-1_32

Brodley, C., Kak, A., Shyu, C., Dy, J., Broderick, L., & Aisen, A. M. (1999). Content-based retrieval from medical image databases: A synergy of human interaction, machine learning and computer vision. In *Proceedings of the sixteenth national conference on artificial intelligence and the eleventh innovative applications of artificial intelligence conference innovative applications of artificial intelligence* (pp. 760–767). Menlo Park, CA, USA: American Association for Artificial Intelligence. Retrieved from http://dl.acm.org/citation.cfm?id=315149.315448

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*(3), 278-287. Retrieved from http://dx.doi.org/10.1037/0096-3445.120.3.278

Brown, N. H., Robertson, K. M., Bisset, Y. C., & Rees, J. L. (2009). Using a structured image database, how well can novices assign skin lesion images to the correct diagnostic grouping? *Journal of Investigative Dermatology*, *129*(10), 2509-2512. Retrieved from http://dx.doi.org/10.1038/jid.2009.75 doi: 10.1038/jid.2009.75

Burrows, G. E., Krebs, G. L., & Kirchoff, B. K. (2014). Visual learning agricultural plants of the riverina a new application for helping veterinary students recognise poisonous plants. *Bioscience Education*. Retrieved from http://journals.heacademy.ac.uk/doi/pdf/10.11120/beej.2014.00028 doi: http://dx.doi.org/10.11120/beej.2014.00028

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, *35*(3), 283-319.

Carvalho, P., & Goldstone, R. (2014a). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, 1-8. Retrieved from http://dx.doi.org/10.3758/s13423-014-0676-4 doi: 10.3758/s13423-014-0676-4

Carvalho, P., & Goldstone, R. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481-495. Retrieved from http://dx.doi.org/10.3758/s13421-013-0371-0 doi: 10.3758/s13421-013-0371-0

Deng, J., Krause, J., & Fei-Fei, L. (2013, June). Fine-grained crowdsourcing for fine-grained recognition. In *Computer vision and pattern recognition (cvpr), 2013 ieee conference on* (p. 580-587). doi: 10.1109/CVPR.2013.81

Donovan, T., Manning, D. J., & Crawford, T. (2008). Performance changes in lung nodule detection following perceptual feedback of eye movements. In *Medical imaging 2008: Image perception, observer performance, and technology assessment.* International Society for Optics and Photonics.

Dror, I. E., Stevenage, S. V., & Ashworth, A. R. S. (2008). Helping the cognitive system learn: exaggerating distinctiveness and uniqueness. *Applied Cognitive Psychology*, *22*(4), 573–584. Retrieved from http://dx.doi.org/10.1002/acp.1383 doi: 10.1002/acp.1383

Duan, K., Parikh, D., Crandall, D., & Grauman, K. (2012, June). Discovering localized attributes for fine-grained recognition. In *Computer vision and pattern recognition (cvpr), 2012 ieee conference on* (p. 3474-3481). doi: 10.1109/CVPR.2012.6248089

Fang, Y., & Geman, D. (2005). Experiments in mental face retrieval. In T. Kanade, A. Jain, & N. K. Ratha (Eds.), *Audio-and video-based biometric person authentication* (pp. 637–646). Springer-Verlag.

Ferecatu, M., & Geman, D. (2009, June). A statistical framework for image category search from a mental picture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(6), 1087-1101. doi: 10.1109/TPAMI.2008.259

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, *2*, 568–573.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training greeble experts: a framework for studying expert object recognition processes. *Vision Research*, *38*(1516), 2401 - 2428. Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698997004422 doi: http://dx.doi.org/10.1016/S0042-6989(97)00442-2

Gomes, R. G., Welinder, P., Krause, A., & Perona, P. (2011). Crowdclustering. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 558–566). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4187-crowdclustering.pdf

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261 - 2271. Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698901000979 doi: http://dx.doi.org/10.1016/S0042-6989(01)00097-9

Hornsby, A. N., & Love, B. C. (2014). Improved classification of mammograms following idealized training. *Journal of Applied Research in Memory and Cognition*, *3*(2), 72-76. Retrieved from http://www.sciencedirect.com/science/article/pii/S2211368114000321 doi: http://dx.doi.org/10.1016/j.jarmac.2014.04.009

Jia, Y., Abbott, J. T., Austerweil, J., Griffiths, T., & Darrell, T. (2013). Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 1842–1850). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5205-visual-concept-learning-combining-machine-vision-and-bayesian-generalization-on-concept-hierarchies.pdf

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 316–332.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. In *Proceedings of the 27th annual meeting of the cognitive science society* (pp. 1066–1071).

Jones, M., Maddox, W. T., & Love, B. C. (2006). The role of similarity in generalization. In *Proceedings of the 28th annual meeting of the cognitive science society* (pp. 405–410).

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. Retrieved from http://dx.doi.org/10.1002/acp.1801 doi: 10.1002/acp.1801

Kellman, P. J. (2013). Adaptive and perceptual learning technologies in medical education and training. *Military Medicine*, *178*(10S), 98-106. Retrieved from http://dx.doi.org/10.7205/MILMED-D-13-00218 doi: 10.7205/MILMED-D-13-00218

Kirchoff, B. K., Delaney, P. F., Horton, M., & Dellinger-Johnston, R. (2014). Optimizing learning of scientific category knowledge in the classroom: The case of plant identification. *CBE-Life Sciences Education*, *13*(3), 425-436. Retrieved from http://www.lifescied.org/content/13/3/425.abstract doi: 10.1187/cbe.13-11-0224

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498-503.

Krupinski, E., Nodine, C., & Kundel, H. (1993). Perceptual enhancement of tumor targets in chest x-ray images. *Perception I& Psychophysics*, *53*(5), 519-526. Retrieved from http://dx.doi.org/10.3758/BF03205200 doi: 10.3758/BF03205200

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (pp. 267–301). New York, NY: Cambridge University Press.

Kumar, N., Belhumeur, P., Biswas, A., Jacobs, D., Kress, W., Lopez, I., & Soares, J. (2012). Leafsnap: A computer vision system for automatic plant species identification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer vision - eccv 2012* (p. 502-516). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-33709-3_36 doi: 10.1007/978-3-642-33709-3_36

Lesot, M.-J., Rifqi, M., & Benhadda, H. (2009). Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *1*(1), 63-84. Retrieved from http://www.inderscienceonline.com/doi/abs/10.1504/IJKESDP.2009.021985 doi: 10.1504/IJKESDP.2009.021985

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010, Sep). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *Journal of Experimental Psychology: Applied*, *16*(3), 251-262. doi: 10.1037/a0020082

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*(0), 111 - 123. Retrieved from http://www.sciencedirect.com/science/article/pii/S0042698913003015 (Perceptual Learning - Recent advances) doi: http://dx.doi.org/10.1016/j.visres.2013.12.009

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289-316.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87-108. doi: http://dx.doi.org/10.1037/0278-7393.13.1.87

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (p. 363-393). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pantelis, P., van Vugt, M., Sekuler, R., Wilson, H., & Kahana, M. (2008). Why are some peoples names easier to learn than others? the effects of face similarity on memory for face-name associations. *Memory & Cognition*, *36*(6), 1182-1195. Retrieved from `http://dx.doi.org/10.3758/MC.36.6.1182` doi: 10.3758/MC.36.6.1182

Pashler, H., & Mozer, M. C. (2013, Jul). When does fading enhance perceptual category learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1162-1173. doi: 10.1037/a0031679

Quattoni, A., Wang, S., Morency, L.-P., Collins, M., & Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *29*(10), 1848-1852. doi: doi:10.1109/TPAMI.2007.1124

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 75-94. Retrieved from `http://dx.doi.org/10.1002/wcs.1268` doi: 10.1002/wcs.1268

Rimoin, L., Altieri, L., Craft, N., Krasne, S., & Kellman, P. J. (2015). Training pattern recognition of skin lesion morphology, configuration, and distribution. *Journal of the American Academy of Dermatology*, *72*(3), 489-495.

Roads, B. D., & Mozer, M. C. (submitted). Improving human-machine cooperative classification via cognitive theories of similarity.

Roads, B. D., Mozer, M. C., & Busey, T. (submitted). Using highlighting to train attentional expertise.

Robertson, K., McIntosh, R. D., Bradley-Scott, C., MacFarlane, S., & Rees, J. L. (2014). Image training, using random images of melanoma, performs as well as the abc (d) criteria in enabling novices to distinguish between melanoma and mimics of melanoma. *Acta Dermato-Venereologica*, *94*(3), 265-270.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319-345. Retrieved from `http://dx.doi.org/10.1007/BF02293654` doi: 10.1007/BF02293654

Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2008). The role of category learning in the acquisition and retention of perceptual expertise: A behavioral and neurophysiological study. *Brain Research*, *1210*, 204 - 215. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0006899308004113` doi: http://dx.doi.org/10.1016/j.brainres.2008.02.054

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125-140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015). Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, *2*.

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science*, *16*(2), 145-151. Retrieved from `http://pss.sagepub.com/content/16/2/145.abstract` doi: 10.1111/j.0956-7976.2005.00795.x

Tversky, A., & Hutchinson, J. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*(1), 3-22.

van der Maaten, L., & Weinberger, K. (2012, Sept). Stochastic triplet embedding. In *Machine learning for signal processing (mlsp), 2012 ieee international workshop on* (p. 1-6). doi: 10.1109/MLSP.2012.6349720

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset* (Tech. Rep. No. CNS-TR-2011-001). California Institute of Technology.

Wah, C., Horn, G. V., Branson, S., Maji, S., Perona, P., & Belongie, S. (2014, June). Similarity comparisons for interactive fine-grained categorization. In *Computer vision and pattern recognition (cvpr).* Columbus, OH.

Wah, C., Maji, S., & Belongie, S. (2015). Learning localized perceptual similarity metrics for interactive categorization.

Wahlheim, C., Dunlosky, J., & Jacoby, L. (2011). Spacing enhances the learning of natural concepts: an investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750-763. Retrieved from `http://dx.doi.org/10.3758/s13421-010-0063-y` doi: 10.3758/s13421-010-0063-y

Wang, J., Markert, K., & Everingham, M. (2009, September). Learning models for object recognition from natural language descriptions. In *British machine vision conference (mbvc).* London, UK.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). *Caltech-UCSD Birds 200* (Tech. Rep. No. CNS-TR-2010-001). California Institute of Technology.

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects: Becoming a ziggerin expertbut which type? *Psychological Science*, *20*(9), 1108-1117. Retrieved from `http://pss.sagepub.com/content/20/9/1108.abstract` doi: 10.1111/j.1467-9280.2009.02430.x