

Preferential Encoding of Features Distinctive for Multiple Categories

Michael Fink

Interdisciplinary Center for Neural Computation

The Hebrew University of Jerusalem

Jerusalem 91904, Israel

`fink@cs.huji.ac.il`

Gershon Ben-Shakhar

Department of Psychology

The Hebrew University of Jerusalem

Jerusalem 91905, Israel

`mskpugb@pluto.mscc.huji.ac.il`

Shimon Ullman

Faculty of Mathematics and Computer Science

Weizmann Institute of Science

Rehovot 76100, Israel

`shimon.ullman@weizmann.ac.il`

Running Head:

Abstract

Understanding how features are encoded during category acquisition is a fundamental challenge in the study of human learning. The current work proposes that in order to maintain accurate generalization in large scale categorization systems, features that are useful in discriminating multiple categories must be actively preferred. This multi-class hypothesis stands in contrast with theories in which features are encoded for each categorization task individually, as well as theories that focus solely on encoding coincidences in input patterns. The current paper provides a methodology for empirically testing the proposed hypothesis under controlled conditions. It is shown that in the process of acquiring a sequence of new categories, features that are informative for recognizing *several* categories are preferably encoded. Moreover, evidence is provided that after acquiring these features, representations of categories learned in the past are actively reconstructed by the newly encoded features. Finally, it is demonstrated that encoded features play a role in facilitating future category acquisition. These results are observed using both perceptual and semantic stimuli. It is suggested that preferring features that provide maximum information on multiple categories is a general characteristic of the human categorization systems.

Introduction

The environment in which humans live requires fast and accurate recognition of numerous categories. These requirements must often be met after only few examples of a novel category have been encountered. Nevertheless, humans seem to deal with this challenge effectively, recognizing a large number of categories (Biederman, 1987) with high speed and accuracy (Thorpe, Fize, & Marlot, 1996). It remains a puzzle how new categories are acquired reliably, overcoming the difficulties associated with generalization from a small sample (Goodman, 1972). Theories of categorization are typically based on measuring the similarity of an incoming stimulus to an internal representation of the category, such as a prototype or a set of exemplars (Medin & Smith, 1984; Barsalou, 1985; Nosofsky, 1988). The tradition set by Tversky (1977) emphasized the role of having the appropriate features for measuring this similarity; however, the criteria for encoding the appropriate features, remains unclear. Thus, a major question is what features are used for evaluating similarity in an ever-increasing range of categories. Attempts to address this question have emphasized that in a compositional framework, a small vocabulary of basic features can be reused in the representation of many categories (see review and references in Goldstone, 2003). Initially, this vocabulary was characterized as including a set of fixed primitives, such as the geon-based visual alphabet (Biederman, 1987). More recent theories have proposed the use of adaptive feature sets (Schyns, Goldstone, & Thibaut, 1998), which are class-dependent, and are acquired during a learning stage.

Two theoretical frameworks have suggested computational criteria as to when a feature set should be dynamically extended. The first theoretical framework emphasizes the importance of features that encode suspicious coincidences between input patterns, e.g., jointly encoding wings and feathers (Barlow, 1989). A number of studies have supported this view by demonstrating that features sensitive to input statistics can be created in the absence of any

categorization feedback (Billman & Knutson, 1996; Rosenthal, Fusi, & Hochstein, 2001; Edelman, Hiles, Yang, & Intrator, 2001; Fiser & Aslin, 2002).

In contrast to this approach, the class-based framework suggests that the encoded features are those which are maximally discriminative in the context of a target categorization task. Discriminative power is often quantified using mutual information (Battiti, 1994; Ullman, Vidal-Naquet, & Sali, 2002; Fleuret, 2004; Vasconcelos & Vasconcelos, 2004). Several empirical findings provide indirect evidence for this theoretical framework. Goldstone (1994) has demonstrated that different categorization tasks induce selective feature sensitization. Similarly, Archambault, O'Donnell, and Schyns (1999) show that general vs. specific categorization tasks might influence the perceived properties of the same distal object. More direct evidence for feature creation induced by a categorization task, was provided by Schyns and Murphy (1994), Schyns and Rodet (1997), Goldstone (2000) and Goldstone (2003). Recently, Goldstone, Rogosky, Pevtsov, and Blair (2004) have demonstrated the process of encoding discriminative features during perceptual and semantic category acquisition while characterizing the interplay between discriminative value and feature naturalness. Evidence for selectivity to discriminative features has also been established through neuronal recordings from monkey inferior temporal (IT) cortex (Baker, Behrmann, & Olson, 2002; Sigala & Logothetis, 2002). However, these different lines of research do not directly address the fact that categorization systems must be capable of acquiring many thousands of categories.

The current study examines the hypothesis that in order to achieve efficient generalization when classifying numerous categories, the underlying organizing principle is to preferentially encode features which are informative for multiple categories. In the past, accurate generalization in novel categorization tasks has been attributed to an acquired domain theory (Ahn, Mooney, Brewer, & DeJong, 1987; Pazzani, 1991). Our research hypothesis suggests that a

significant part of domain knowledge, is manifested by encoding the features suitable for discriminating categories within the target domain. A formal description of this informativeness criterion is provided in Appendix I. The assumption supporting the proposed criterion is that categories often share common informative sub-structures. For example, tails are common to different animals, and wheels are shared by different vehicle classes. Thus, representing features that are highly informative for multiple known categories is likely to contribute to the encoding process of future categories as well. This contribution is especially significant when it is necessary to learn the manner in which complex perceptual features can change their appearance (e.g. an eye may be open or closed and a tail may bend in certain angles but not in others). When learning is done independently for each class, the complexity of such features might require observing a large sample. The use of shared features allows interclass transfer of this learned variability and thereby contributes to reliable classification of new categories (e.g. handling eye appearance under changing viewing conditions).

The perceptual-conceptual continuum view (Jones & Smith, 1993; Eimas, 1994; Quinn & Eimas, 1996, 1997; Goldstone & Barsalou, 1998; Barsalou, 1999), emphasizes the fact that categorization tasks in the perceptual domain and in the semantic domain are challenged by the similar fundamental requirement of discriminating instances originating from a large number of categories. We therefore suggest that features that are informative for multiple classes might assist generalization not only in perceptual categorization, but in semantic categorization as well. For example, once the rich cultural correlates of semantic features such as a *female* or *teenager* are encoded, they can be instrumental in characterizing many future semantic categories such as *cheerleaders* or *anorexia-nervosa*.

The hypothesis examined in this study is that features which are discriminative for several different categories are identified and encoded. If in fact selecting features that are

informative for multiple categories is a basic organizing principle, it can be hypothesized that novel features that are discriminative to multiple categories collectively, will be encoded, even in the absence of suspicious coincidences in the input statistics and even if they do not maximize the information provided for any *individual* category. Furthermore, it is hypothesized that during the dynamic process of acquiring multiple categories the principal of constantly maintaining a shared set of discriminative features can induce representational shifts of categories learned in the past and affect the representation of subsequent classes. These hypotheses are summarized in three predictions:

1. features informative for recognizing multiple categories are preferentially encoded
2. features revealed as informative for multiple categories can change former category representations
3. features encoded due to their information content for multiple categories, can later facilitate the acquisition of future categories

Four experiments were designed to test the three predictions stated above. Experiment I examined the first two predictions while Experiment II examined the third. Both experiments utilized controlled sets of semantic stimuli (job candidates descriptions). To test the three predictions in lower levels of the perceptual-conceptual continuum, Experiments III and IV replicated the first two experiments utilizing perceptual stimuli (configurations of colored cubes). The following four sections describe the four experiments, followed by a discussion on the scope and limitations of the proposed theory and research methodology.

Experiment I: Preferential Encoding of Semantic Features Informative for Multiple Categories

Experiment I was designed to examine whether features that are informative for recognizing multiple categories are preferentially encoded and whether this process might actively reconstruct the representations of categories acquired in the past. Studying the intricacies of feature creation processes in adult categorization systems is a challenging endeavor. This is especially true if the proposed hypothesis is correct and when confronted with new categorization tasks, humans do in fact employ their rich arsenal of existing features. The challenge lays in isolating the contribution of feature discriminative value for multiple classes collectively, while controlling the effects of stimuli co-occurrence and the information provided by features for each category, individually. Therefore, the experimental setup makes use of novel categories in which the experimental conditions can be designed in a controlled manner using features which do not resonate with existing representations.

Method

This section starts by presenting an abstract categorization task. The categorization task is specifically designed to test whether features that are most discriminative for multiple categories collectively, are indeed preferentially encoded. The empirical implementation of this abstract task using semantic stimuli, is described later in the materials and procedure sections of Experiment I. An implementation using perceptual stimuli is provided in Experiment III.

The abstract categorization task consisted of eight binary input elements $x_{i,i=1,\dots,8}$, jointly notated as \mathbf{x} . Four target categories $C_{n,n=1,\dots,4}$ were defined over the input vector set $\{\mathbf{x}\}$. As described in Fig. 1:Left, each category was fully characterized by four specific input elements

being in an *on* position. Thus, for example, $\mathbf{x} = [- - + + + + - +]$ is an exemplar of category, $C_2 = [* * + + + + **]$. It will now be shown why this specific category structure can be used to test the research predictions. The subset of all second order conjunctions of input elements (including $\binom{8}{2} = 28$ features) was selected as the focus of the proposed analysis ¹. Mutual information was evaluated between each of the 28 candidate features and the individual categories (using Eq. (1) in Appendix I). Then, the information was evaluated for the four categories *collectively* (summing Eq. (1) over all category assignments of C). As shown in Fig. 1:Left, categories C_1 and C_2 require two common input elements to be in an *on* state (x_3^+ and x_4^+). This common requirement was termed a *pair-feature*. In fact, each of the four categories shares a common *pair-feature* with two other categories. These *pair-features* are formally defined as:

- $v_1 \equiv x_1^+ \cap x_2^+$
- $v_2 \equiv x_3^+ \cap x_4^+$
- $v_3 \equiv x_5^+ \cap x_6^+$
- $v_4 \equiv x_7^+ \cap x_8^+$

As depicted in Fig. 1:Right, the *pair-features* are not more informative than other pairs for category C_1 individually (providing 0.264 or 0.008 bits). However, they do appear as the maximally informative features for the four categories collectively (each providing 0.581 bits) ². Thus, the first research prediction would be manifested if a salient representation of the *pair-features*: v_1, v_2, v_3 and v_4 , emerges while acquiring the four categories. As described in the introduction section the proposed setting must also control the two alternative computational criteria that induce feature creation. This goal was achieved by finding a representation that is comparable in all aspects to the *pair-feature* structure but is not as informative for the categorization tasks collectively. Such an alternative representation is generated by arbitrarily

segmenting the eight input elements into four *incongruent-pairs*, that appear each in just one category, i.e.:

- $h_1 \equiv x_1^+ \cap x_3^+$
- $h_2 \equiv x_4^+ \cap x_6^+$
- $h_3 \equiv x_5^+ \cap x_7^+$
- $h_4 \equiv x_2^+ \cap x_8^+$

Insert Figure 1 about here.

Participants Twelve undergraduate students from the Hebrew University of Jerusalem participated in Experiment I for payment or course credit. Participants, aged 20 to 30 years, were screened for normal vision.

Materials The categories defined in Fig. 1 were implemented in a job assignment task, requiring participants to sort applicants to four business firms. Eight binary characteristics encoded the input description of each candidate. For each characteristic, one value was selected to function as the *on* state and another as the *off* state e.g. Department Preference: Local + and International – (see Fig. 2). Thus, each of the four business firms required four specific characteristics to be in an *on* state. For example, Firm 3 required all candidates to be male, lawyers, living in New York and preferring the local department. Beneath the applicant sheet, an array of five target buttons was displayed (see setting in Fig. 2). Each of the four peripheral buttons was associated with one of the four firms. The central button was reserved for applicants

not qualified for any firm. In order to control feature saliency the position of the fields in the display were randomly permuted in each trial.

The intersections of the categories' relevant characteristics defined the set of *pair-features*:

- $v_1 \equiv$ Spanish-speaker *and* Married
- $v_2 \equiv$ Private-college *and* Thirties
- $v_3 \equiv$ Male *and* Lawyer
- $v_4 \equiv$ New York *and* Local-department

It was previously predicted that an internal representation of these *pair-features* should evolve if features are not selected by their information content for each categorization task individually, but rather by their information content to all categorization tasks collectively. This will be guaranteed by comparing each *pair-feature* to an *incongruent-pair* which provides an equal amount of information to any one specific category. For example, the information provided by *pair-feature* v_1 to Firm C_1 , is equal to the information provided by *incongruent-pair* h_1 (Spanish-speaker *and* Thirties) to the same category (0.264bits, as indicated in the top pane of Fig. 1).

Insert Figure 2 about here.

Training Procedure Experiment I was composed of four training stages. At each stage, participants learned the requirements of one additional firm in a trial-and-error paradigm. In every trial a random subset of the eight input elements x was set to the *on* state while the remaining features were set to the *off* state. This random activation process was carefully designed not bias

the *pairs-features* over the *incongruent-pairs* (as explained in Appendix II). The resulting eight dimensional candidate description was displayed until the participants pressed one of the category buttons. If a wrong firm was indicated an error tone was triggered³. In the first stage, dedicated to learning the requirements of Firm C_1 , only the central *default button*, and the button associated with Firm C_1 , were presented. Later, an additional button appeared with each subsequent stage, so that participants can indicate whether the candidate is compatible with the requirements of Firm C_2 , Firm C_3 , and Firm C_4 . Each training stage was concluded only when participants reached a criterion of 100 *consecutive* successes. It should be emphasized that the random activation probability over the input set $\{x\}$ was identical during all four training stages. Thus, participants were constantly viewing stimuli generated from a fixed source, while incrementally learning to recognize which exemplars (candidates) were members of categories (Firms) C_1 through C_4 . It should be noted that for each participant the various fields were randomly assigned so that four *pair-features* of one participant could have been four *incongruent-pairs* of another. Finally, when the four training stages were concluded, a test was employed to examine whether the hypothesized *pair-features* have emerged.

Testing Procedure In the testing stage, each of the category buttons was highlighted in a sequential manner while requiring the participants to verbally report which characteristics were necessary for the associated firm. The proposed test assumes that if an internal representation of the *pair-features* had undergone a process of unitization (Goldstone, 2000), the verbal reports should be composed of two *pair-features*. In other words, if certain input elements have been consolidated into a *pair-feature*, they should be reported as one. This reporting pattern should not be exhibited if a representation based on any of the *incongruent-pairs* had emerged. It should be noted that although participants were explicitly required to verbally report each category, the sequence of reported characteristics is a purely *implicit* measurement. The advantage of this

testing method is that it refrains from presenting the hypothesized *pair-features* during the testing process, thus avoiding the encoding of *pair-features* as a byproduct of the testing phase itself. It was therefore assumed that the reporting sequence was not intentionally or unintentionally biased by the participants, and that these reports essentially reflect the internal structure that had emerged in the learning process.

Results Participants completed the four training stages in 4-6 hours, requiring approximately 1000 trials. The verbal reports provided in the testing stage were then coded by examining whether the first two characteristics composed a *pair-feature* or an *incongruent-pair*. It was observed that the frequency of reporting *pair-features* was significantly higher than that of a comparable pattern of *incongruent-pairs* (binomial test, $p \leq 0.05$, $n=12$) in all of the reports of the four firms (Fig. 3). In addition to registering the report sequence, the participants of Experiment I were recorded while verbally reporting the requirements of each firm. The reports of each firm were manually annotated by marking the starting time and the ending time of the four comprising characteristics. An annotated recording of category C_3 is presented in Fig. 3. By performing this annotation it becomes possible to measure the duration of the three gaps between the four reported characteristics and to assess whether the input elements composing each *pair-feature* are indeed temporally fused, as expected by a unitization process. The three gaps were scored according to the ascending order of their duration (the shortest gap was scored as 1, the intermediate gap was scored as 2 and the longest gap was scored as 3). When analyzing these patterns in the recordings of Category C_1 , it was observed that the second gap score ($M = 2.86$, $SD = 0.38$) was significantly larger ($t(6) = 4.86$) than the first gap score ($M = 1.83$, $SD = 0.69$) and significantly larger ($t(6) = 10.49$) than the third gap score ($M = 1.43$, $SD = 0.53$)⁴. It should be noted that only recordings of participants that reported the categories as two *pair-features* were used in this analysis and that corrupted recordings and recordings where participants did not

report only the relevant characteristics (e.g. incorporating words like AND into the report) were not used⁵. Similarly, the second gap scores were significantly larger than the first and third gaps in the reports of categories C_2 through C_4 .

One might claim that this pattern of reports results from the fact that people simply tend to take longer pauses in the second gap of a four word list, or that the recording results are do a preexistent representational bias and therefore do not reflect an emergent *pair-feature* structure. Therefore, a control group was established, where participants were required to only learn and report Firm C_1 . The eight participants of the control group were selected from the same student population as the experiment group. This between subject test, compared the score difference between the second and first gaps (e.g. the report in Fig. 3 would be scored as $3 - 2 = 1$). In this comparison a significant difference ($t(13) = 2.21$) was observed between the gap scores of the experiment group ($M = 1.00, SD = 1.00$) and the control group ($M = 0.00, SD = 1.16$). Similarly, when comparing the score difference between the second and third gaps (e.g. the report in Fig. 3 would be scored as $3 - 1 = 2$). A significant difference ($t(13) = 7.17$) was observed between the gap scores of the experiment group ($M = 1.43, SD = 0.53$) and the control group ($M = -0.38, SD = 1.19$). Thus, the temporally fused reports of the *pair-features* are not observed after the acquisition of category C_1 , but rather manifest a representational shift emerging during the later acquisition of categories C_2 through C_4 .

The reported results provide evidence that in the semantic system, features are preferably encoded if they provide information for multiple categories. However, Experiment I also addresses the second prediction, stating that encoding novel features might lead to a reconstruction of former category representations. This claim is based on observing the evolved representation of category C_1 . When participants pass the first training stage, they are equipped with a representation that enables perfect recognition of all category C_1 exemplars (due to the

highly stringent training criterion of 100 consecutive successes). At this point the *pair-features* have no salient representation because they are by definition identical to the *incongruent-pairs* until learning at least one additional category. In fact the control group participants that only learned category C_1 , showed no saliency of a *pair-features* representation. However, when analyzing the reporting results of category C_1 , registered *after* concluding the learning procedure of all four categories, it is evident that the initial representation of category C_1 had been actively reconstructed. This reconstruction maintained that the representation of category C_1 complies with the *pair-features* acquired only later in the training process. Thus, the first two research predictions have been addressed while controlling for the alternative factors that might have accounted for the emergent *pair-feature* structure.

Insert Figure 3 about here.

Experiment II: Facilitated Acquisition of Novel Semantic Categories

Experiment II examined the third research prediction, namely whether the *pair-features* facilitate learning of future categories.

Method

Experiment II included an additional training stage to the four training stages of Experiment I. In this additional stage, examples of a new category, C_5 , were presented. This new category was defined as $C_5 \equiv v_2^+ \cap v_4^+$ (see Fig. 1). Unlike the training procedure of Categories C_1 through C_4 , that continued until a criterion of perfect categorization performance had been reached, the fifth stage was restricted to a prefixed number of trials. It was predicted that when

training a fifth category that is congruent with the *pair-feature* structure, a significant facilitation would be observed. Comparing the performance of participants that have already learned the four initial categories to the performance of naive participants is meaningless, due to history confounds. Therefore, Experiment II maintained a between subject design by including a control group that shared the same history as the experimental group. This control group learned a different fifth category, $\bar{C}_5 \equiv h_2^+ \cap h_4^+$, that is equally complex as C_5 yet incongruent with the *pair-feature* structure.

Participants Twelve undergraduate students from the Hebrew University of Jerusalem participated in Experiment II for payment or course credit. Participants, aged 20 to 30 years, were screened for normal vision, and then randomly allocated to equal sized experimental and control groups.

Materials The stimuli and setting of Experiment II were similar to those described in Experiment I.

Training Procedure Participants were first required to complete training stages 1 through 4 as in Experiment I. Next, both groups learned the requirements of a fifth firm using a limited set of 48 training trials. The stimulus presentation procedure of the fifth stage was identical to the first four training stages except that trials were limited to a prefixed duration of sixteen seconds. Providing 48 training trials was aimed at terminating the training process at a point where the differential learning rate of the experimental and control groups might be manifested. Rather than sampling 48 trials as in the first (unbounded) four training stages, the fifth stage displayed in a random order, a predefined set of 24 negative examples and 24 positive

examples of either the congruent category C_5 or the incongruent category \bar{C}_5 . The detailed composition of these 48 training trials is described in Appendix III.

Testing Procedure Following the 48 training trials, participants were first tested on their capability to correctly categorize the 48 training examples, by repeating the presentation procedure in the absence of any feedback signal. Only then did participants verbally report the candidate characteristics composing the new firm's requirements. If the *pair-features* have no functional influence on future category learning, it would be expected that the learning rate of the fifth category should be equal in both groups. On the other hand, it was hypothesized that the previously encoded *pair-features* might facilitate future learning of a new congruent category. It was therefore anticipated that under such constrained training conditions, the experimental group would display a significant advantage in learning the fifth category.

Results The participants of both the experimental and the control groups reported that the training procedure of the fifth firm was difficult. This effect is probably due to the limited time provided for the 48 training trials. The participants' performance on the 48 testing trials (presented without feedback), was scored by averaging the proportion of the correctly classified examples of the fifth category with the proportion of the correctly classified remaining fillers so that chance level is 0.50. It was observed that while the control group performed slightly above chance level ($M = 0.59$, $SD = 0.11$) the accuracy of the experimental group ($M = 0.81$, $SD = 0.13$) was substantially higher ($t(10) = 2.38$). Next, the verbal reporting results were scored by subtracting the number of any incorrectly reported characteristics from the number of correctly reported characteristics. Here too, a significant difference ($t(10) = 4.65$) was observed between the experimental group scores ($M = 2.33$, $SD = 1.51$) and the control group scores ($M = -0.33$,

$SD = 1.21$). It could therefore be concluded that the emerged *pair-feature* representation was a functional tool in facilitating the acquisition of a novel semantic category.

Experiment III: Preferential Encoding of Perceptual Features Informative for Multiple Categories

The goal of Experiment III was to examine whether the results of Experiment I were specific to the semantic domain or whether evidence for encoding features that are informative for recognizing many categories might also be observed in a perceptual classification task.

Method

Due to the fact that Experiment III was aimed at replicating the structure of Experiment I, the applied method was similar to that described above.

Participants Twenty undergraduate students from the Hebrew University of Jerusalem participated in Experiment III for payment or course credit. Participants, aged 20 to 30 years, were screened for normal color vision.

Materials To test the first research prediction using perceptual stimuli the eight-dimensional binary inputs \mathbf{x} , were implemented using images composed of eight *color cubes*. For each cube, one color was selected to function as the *on* state and another color as the *off* state (Fig. 4). For each individual participant, the set of 16 colors was randomly allocated to the eight cubes⁶.

Insert Figure 4 about here.

Insert Figure 5 about here.

In the *color cube* implementation, each category required four specific neighboring cubes to be in an *on* position (see Fig. 5). Categories based on neighboring cube configurations were chosen, because they are easier to acquire. Exemplars of each category were generated by using color combinations of the remaining four non-relevant cubes. Thus for example, all Category C_3 exemplars included: Black, Orange, Yellow and Green *color cubes*.

The intersections of the categories' relevant cubes defined the set of *pair-features* (see Fig. 6). Since the stimuli presentation probabilities were identical to those described in Experiment I, the mutual information measurements from Fig. 1 remain valid, leaving the *pair-features* as the maximally informative representations.

Insert Figure 6 about here.

Procedure The training procedure of Experiment III was analogous to that described in Experiment I. Similarly to the field permutation in Experiments I, at every trial of Experiment III, the entire three dimensional cube configuration was rotated, in order to control perceptual biases between the *pair-features* and the *incongruent-pairs* (see Appendix IV). When the four training stages were concluded, participant were required to verbally report the *color-cubes* relevant for each of the four categories.

Results Participants completed the four training stages in 2-3 hours. The relatively short training duration resulted from the fact that processing the perceptual stimuli required approximately half the time than in the analogues stages of Experiment I. All twenty participants succeeded in reporting the four colors relevant to each category. The frequency of reporting according to the *pair-feature* pattern was observed to be significantly higher than that of a comparable *incongruent pair* pattern in all four categories (binomial test, $p \leq 0.05$, $n=20$). Fig. 7 depicts the number of reports congruent with the *pair-feature* structure. Thus, Experiment III validates the first research predication. As in Experiment I, the fact that the report of category C_1 reflects the later acquired *pair-feature* structure, validates the second prediction regarding the reconstruction of former category representations.

Insert Figure 7 about here.

Experiment IV: Facilitated Acquisition of Novel Perceptual Categories

Experiment IV examined whether an informative feature set can facilitate learning future perceptual categories.

Method

Experiment IV replicated Experiment II, utilizing the perceptual stimuli described in Experiment III. Thus, the fifth training stage required learning a novel perceptual category $C_5 \equiv v_2^+ \cap v_4^+$ from just a few training examples. The facilitation of acquiring category C_5 was assessed in comparison to a control fifth category, defined as $\bar{C}_5 \equiv h_2^+ \cap h_4^+$.

Participants Ten undergraduate students from the Hebrew University of Jerusalem participated in Experiment IV for payment or course credit. Participants, aged 20 to 30 years, were screened for normal color vision, and then randomly allocated to equal sized experimental and control groups.

Materials The stimuli and setting of Experiment IV were similar to those described in Experiment III. Fig. 8 depicts the four *color cubes* characterizing categories C_5 and \bar{C}_5 .

Insert Figure 8 about here.

Procedure Participants were first required to complete training stages 1 through 4. Then, both the experimental and the control groups learned a fifth category using a limited set of 48 training images (detailed in Appendix III). The stimuli presentation process of this stage was similar to that described in the first four training stages except for the fact that the trial duration was fixed to six seconds. Following the 48 training trials, participants were required to verbally report the *color cubes* composing the new category.

Results As in Experiment II, the ten participants of both the experimental and the control groups reported that the training procedure of the fifth category was difficult. However, when comparing participants' performance in the experimental and control groups, it was observed that the learning rate was significantly higher in the congruent condition (Fisher Exact Probability Test, $p \leq 0.05$, $n=10$). Members of the experimental group reported on average 2.2 correct *color cubes*, i.e. participants learned most of the new category's characteristics. Members of the control group, reported on average only 0.8 out of the four *color cubes* present in category

\bar{C}_5 . Thus, the emerged *pair-feature* representation was indeed a functional tool in facilitating the acquisition of a novel perceptual category.

Summary and Discussion

The proposed experimental setting required learning four categories, each based on a conjunction of four input elements. Pairs of input elements, termed *pair-features* were suggested as the preferred internal representation due to their information content for the target categories, collectively. Although no direct feedback was provided for the *pair-feature* structure, it was experimentally found that participants' reporting patterns corresponded to an internal structure based on these pairs. The existence of the *pair-features* cannot be attributed to their perceptual salience or frequency of appearance, since these factors were carefully controlled. It was also observed that the *pair-feature* structure actively reconstructed the previously acquired representation of the first category. Finally, it was demonstrated that the emergent *pair-features* can significantly facilitate learning of future categories. The three experimental predictions have been validated using both semantic and perceptual stimuli. One might claim that while categorizing the perceptual stimuli in Experiments III and IV, participants actually translated the observed *color cubes* into semantic entities (e.g. naming each color cube). Although possible, this claim seems to be inconsistent with the substantial time difference required for trials in Experiment I (16 seconds) and in Experiment III (6 seconds). If participants were implicitly translating all perceptual stimuli into the semantic system and then performing categorization over semantic entities, it would be expected that trials in Experiment III require at least the processing duration observed in Experiment I. It remains open as to whether a single mechanism is used for consolidating features that are discriminative for multiple categories, or whether

encoding common features is a general principal characterizing many systems throughout the perceptual-conceptual continuum.

Several questions regarding the generalization of the proposed theory might be raised. The learning task stripped the category acquisition and feature creation processes to the bare minimum, two conjunctions of binary inputs deterministically defined each category. Examining the effects of more complex settings, like continuous input dimensions, complex features and other (more natural) learning schemes is the goal of future research.

Appendix I: Formal Definition of the Multiple Category Distinctiveness Criterion

The unsupervised (statistical) learning approach suggested by Barlow (1989) emphasizes the importance of features that encode suspicious coincidences between input patterns (say x_i and x_j). Formally, this theory encodes events where $\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \gg 1$. In contrast, the class-based framework stipulates that the encoded features are those which are maximally discriminative in the context of the given categorization task. Discriminative power is often quantified using mutual information (Battiti, 1994; Ullman et al., 2002; Fleuret, 2004; Vasconcelos & Vasconcelos, 2004), which is the reduction in class uncertainty, given the state of a certain feature. Formally, let \mathcal{F} be the set of candidate features, and C be a class variable, then the class uncertainty (entropy) is defined as $H(C) = -\sum_c p(c) \log p(c)$ and the uncertainty given the feature state (conditional entropy) is defined as $H(C|F) = -\sum_{c,f} p(c, f) \log p(c|f)$. Therefore, the feature to be encoded is the feature F_* that maximally reduces the class uncertainty:

$$F^* = \operatorname{argmax}_{F \in \mathcal{F}} H(C) - H(C|F) = \operatorname{argmax}_{F \in \mathcal{F}} - \sum_{F, C} p(F, C) \log \frac{p(F, C)}{p(F)p(C)} \quad (1)$$

For a single class distinction, the information provided by the presence or absence of a feature, is estimated with respect to a binary class variable, summing Eq. (1) over two states of C (class

present c and class absent \bar{c}). This research puts forth the hypothesis that in order to achieve efficient generalization in large categorization systems the underlying organizing principle must be the preferential encoding of features informative for multiple categories. Thus, the encoded features should maximize information to all the categories collectively, summing C in Eq. (1) over all possible category assignments. It should be noted that criteria suggested by the unsupervised and by the class-based frameworks, both share a ratio evaluating the deviance from statistical independence. Although biologically based systems could not be expected to accurately evaluate these probabilities and information quantities, the proposed computational criteria formally characterize the fundamentally different factors that might govern feature creation processes. By doing so, it becomes possible to characterize the empirical setting in which the differences between these theories might be manifested

Appendix II: Experiment I Trial Composition

In preliminary experiments it had been established that a very low rate of positive examples entails a non-feasible training duration. If the random process generating the state of the input elements x samples uniformly from all 256 (2^8) possible configurations, the positive example rate would be $\frac{16}{256}$, counting the sixteen different exemplars of each firm. This means that after viewing an example of a certain category, participants would observe an average of fifteen negative examples before seeing another positive one. Learning in these conditions is extremely difficult. Therefore an alternative random generating process for the input elements was favored. This process selected with probability $\frac{1}{4}$ whether four, five, six or seven input elements would be active in the current trial. Then, with equal probability, any of the candidates complying with this constraint might be selected (see Table 1). For example, the probability of seeing a specific input pattern x with six active elements (like the category C_2 exemplar: $x = [+ - + + + + - +]$), is

$\frac{1}{4} \times \frac{1}{\binom{8}{6}} = \frac{1}{112}$. Thus, the probability of observing an instance of a certain class with four, five, six, or seven active input elements is respectively $\frac{1}{280}$, $\frac{4}{224}$, $\frac{6}{112}$, and $\frac{4}{32}$. Summing these probabilities the positive frequency rate increases to $\frac{1}{5}$, thus enabling training in a reasonable time for an experimental framework. It should be emphasized that this input generation procedure maintains equal probability rates for the *pair-features* and *incongruent-pairs*. In addition it should be noted that this input generating distribution, observed by the participants, was the basis for the mutual information measurements provided in Fig. 1:Right.

Insert Table 1 about here.

Appendix III: Experiment II Trial Composition

The fifth training stage of Experiment II displayed in random order a fixed set of 24 negative examples and 24 positive examples of either the congruent category C_5 or the incongruent category \bar{C}_5 . The 24 negative examples included 4 examples of each of the categories C_1 through C_4 and 8 non-category related fillers (see Table 2). The 24 positive examples of C_5 are described in Table 3. The composition of the 48 training trials of the incongruent control group results from replacing the 24 positive examples of category C_5 with 24 positive examples of category \bar{C}_5 (see Table 4).

Insert Table 2 about here.

Insert Table 3 about here.

Insert Table 4 about here.

Appendix IV: Stimuli Rotation

In order to control the perceptual salience of the *pair-features* and the *incongruent-pairs* in the *color cube* implementation, at every trial the entire three dimensional cube configuration was rotated in 90 degrees to a random selection of either the-X, the-Y or the-Z axis. This 90 degree rotation was displayed as a two second (18-frame) animation, appearing at the end of each trial. Rotating the cube configuration was favored to permuting the locations of the *color cubes* in order to assist participants not to lose spatial orientation in the 3D perceptual scene. As a result of this perceptual salience control, each cube configuration x , could appear in 24 different orientations (see Fig. 9).

Insert Figure 9 about here.

Acknowledgments

This work was supported by ISF Grant 7-0369, EU IST Grant FP6-2005-015803, and IFT grant 032-1478.

References

- Ahn, W. K., Mooney, R., Brewer, W. F., & DeJong, G. F. (1987). Schema acquisition from one example: Psychological evidence for explanation-based learning. *In Proceedings of the Ninth Annual Conference of the Cognitive Science Society.*
- Archambault, A., O'Donnell, C., & Schyns, P. G. (1999). Blind to object changes: When learning the same object at different levels of categorization modifies its perception. *Psychological*

- Science*, 10(3), 249–255.
- Baker, C. I., Behrmann, M., & Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nature Neuroscience*, 5, 1210–1216.
- Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1, 295–311.
- Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, 629–654.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–660.
- Battiti, R. (1994). Using the mutual information for selecting features in supervised neural net learning. *Neural Networks*, 5(4), 537–550.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115–147.
- Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: a complex whole aids learning the parts. *Journal of Experimental Psychology Learning Memory and Cognition*, 22(2), 458–475.
- Edelman, S., Hiles, B. P., Yang, H., & Intrator, N. (2001). Probabilistic principles in unsupervised learning of visual structure: human data and a model. *Proceedings of the Neural Information Processing System conference (NIPS) 2001*.
- Eimas, P. D. (1994). Categorization in early infancy and the continuity of development. *Cognition*, 50, 83–93.
- Fiser, J., & Aslin, R. N. (2002). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6), 499–504.

- Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5(4), 1531–551.
- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2), 178–200.
- Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 86–112.
- Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. *Perceptual Organization in Vision: Behavioral and Neural Perspectives*. Mahwah, New Jersey., 65, 233–278.
- Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and cognition. *Cognition*, 65.
- Goldstone, R. L., Rogosky, B. J., Pevzow, R., & Blair, M. (2004). Perceptual and semantic reorganization during category learning. In H. Cohen, & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science*. Elsevier.
- Goodman, N. (1972). *Problems and projects*. Indianapolis: Bobbs Merrill.
- Jones, S. S., & Smith, L. B. (1993). The place of perception in children's concepts. *Cognitive Development*, 8, 113–139.
- Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113–138.
- Nosofsky, R. M. (1988). The dynamics of similarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(1), 54–65.
- Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17, 416–432.

- Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. *Advances in infancy research, 10*, 1–36.
- Quinn, P. C., & Eimas, P. D. (1997). A reexamination of the perceptual-to-conceptual shift in mental representations. *Review of General Psychology, 1*, 271–287.
- Rosenthal, O., Fusi, S., & Hochstein, S. (2001). Forming classes by stimulus frequency: Behavior and theory. *Proceedings of the National Academy of Science, 98*, 4265–4270.
- Schyns, P. G., Goldstone, R. L., & Thibaut, J. (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1–54.
- Schyns, P. G., & Murphy, G. L. (1994). The ontogeny of part representation in object concepts. *The Psychology of Learning and Motivation, 31*, 305–349.
- Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*(3), 681–696.
- Sigala, N., & Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature, 415*, 318–320.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520–522.
- Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–372.
- Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.
- Vasconcelos, N., & Vasconcelos, M. (2004). Scalable discriminant feature selection for image retrieval and recognition. *IEEE Conf. on Computer Vision and Pattern Recognition*.

Footnotes

¹The role of features defined over higher order conjunctions is not ruled out; however, the goal of the proposed setting is to contrast the emergent representations of features with comparable complexity.

² When evaluating the information content of an extended feature set including all first, second and third order conjunctions, the *pair-features* remained the maximally informative representations.

³In theory, a certain candidate might fall into one of sixteen (2^4) different combinations of acceptance to the four firms. In order to simplify the training procedure, participants were required to assign each candidate to one of the appropriate firms, or press on the central, *default button*, in case the candidate was not qualified for any of the firms.

⁴Throughout the paper t-test results are evaluated using one tailed tests with $p \leq 0.05$ indicating statistical significance.

⁵These criteria eliminated five of the twelve recordings.

⁶ The more dominant color was typically selected as the *on* state (e.g. Red-*on* vs. Cyan-*off*). In addition, cube edges were colored in the opponent color to emphasize the binary role of each input element.

active inputs	4	5	6	7
$p(\mathbf{x}) =$	$\frac{1}{4} \times \frac{1}{\binom{8}{4}} = \frac{1}{280}$	$\frac{1}{4} \times \frac{1}{\binom{8}{5}} = \frac{1}{224}$	$\frac{1}{4} \times \frac{1}{\binom{8}{6}} = \frac{1}{112}$	$\frac{1}{4} \times \frac{1}{\binom{8}{7}} = \frac{1}{32}$

Table 1: *Pairs-features* and *incongruent-pairs* statistics were controlled by uniformly sampling from the input patterns \mathbf{x} having four, five, six or seven input elements in an *on* state.

	appearances	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
C_1	1	+	+	+	+	-	-	-	+
	1	+	+	+	+	-	-	+	-
	1	+	+	+	+	-	+	-	-
	1	+	+	+	+	+	-	-	-
C_2	1	-	-	+	+	+	+	-	+
	1	-	-	+	+	+	+	+	-
	1	-	+	+	+	+	+	-	-
	1	+	-	+	+	+	+	-	-
C_3	1	-	-	-	+	+	+	+	+
	1	-	-	+	-	+	+	+	+
	1	-	+	-	-	+	+	+	+
	1	+	-	-	-	+	+	+	+
C_4	1	+	+	-	-	-	+	+	+
	1	+	+	-	-	+	-	+	+
	1	+	+	-	+	-	-	+	+
	1	+	+	+	-	-	-	+	+
Default	1	-	+	+	+	-	+	-	+
	1	+	-	+	+	+	-	+	-
	1	+	+	-	+	-	+	-	+
	1	+	+	+	-	+	-	+	-
	1	-	+	-	+	-	+	+	+

	appearances	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
C_5	6	-	-	+	+	-	+	+	+
	6	-	-	+	+	+	-	+	+
	6	-	+	+	+	-	-	+	+
	6	+	-	+	+	-	-	+	+

Table 3: The 24 positive examples of the congruent category C_5 defined as a conjunction of v_2 ($x_3^+ \cap x_4^+$) and v_4 ($x_7^+ \cap x_8^+$).

	appearances	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
\bar{C}_5	6	+	-	-	+	-	+	+	+
	6	+	-	-	+	+	+	+	-
	6	+	-	+	+	-	+	+	-
	6	+	+	-	+	-	+	+	-

Table 4: The 24 positive examples of the incongruent category \bar{C}_5 defined as a conjunction of two incongruent pairs $x_1^+ \cap x_7^+$ and $x_4^+ \cap x_6^+$.

Figure Captions

Figure 1. Left: Definitions of categories C_1 through C_4 learned in Experiment I, and of categories C_5 and \bar{C}_5 learned in Experiment II. Input elements indicated by + must appear in an *on* state for the category to be present, while * denotes the category's indifference to certain input elements. *Right:* Each of the 28 columns represents a second order conjunction feature by highlighting the two relevant input elements. Features are grouped by descending information content on category C_1 (top pane) and by information content on the four categories collectively (bottom pane). Although the *pair-features* (emphasized in white) do not provide the maximal information for category C_1 individually (providing 0.264 or 0.008 bits), they all appear as the maximally informative features for the four categories collectively (providing 0.581 bits).

Figure 2. The eight binary semantic characteristics and the experimental setup.

Figure 3. Left: The number of reports congruent/incongruent (black/white) with the *pair-feature* structure for categories C_1 through C_4 . *Right:* An annotated 10-second recording (amplitude vs. time) of a participant reporting category C_3 . The *pair-features* seem to be temporally fused: the shortest gap appearing between New York and Local, the second shortest gap between Male and Lawyer while the longest gap appears between Lawyer and New York.

Figure 4. An example of two stimuli x and x' composed of eight binary *color cubes*. Each *color cube* appeared in two distinct colors indicating whether the cube input element was in an *on* or *off* state.

Figure 5. The color-cubes characterizing perceptual categories C_1 through C_4 .

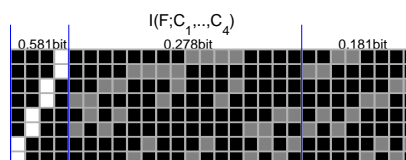
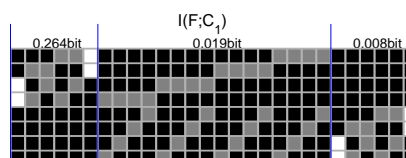
Figure 6. The color cube *pair-features*, providing the maximum information on the four target categories.

Figure 7. The number of reports congruent with the *pair-feature* structure (black) for categories C_1 through C_4 .

Figure 8. The *color cubes* defining the fifth category in the congruent condition $C_5 \equiv v_2^+ \cap v_4^+$ (left) and in the incongruent control $\bar{C}_5 \equiv h_2^+ \cap h_4^+$ (right).

Figure 9. Each cube configuration is rotated to avoid spatial biases. Depicted here are the 24 possible displays of the category C_1 exemplar [++++----].

Category:	C_1	C_2	C_3	C_4	C_5	\bar{C}_5
x_1	+	*	*	+	*	*
x_2	+	*	*	+	*	+
x_3	+	+	*	*	+	*
x_4	+	+	*	*	+	+
x_5	*	+	+	*	*	*
x_6	*	+	+	*	*	+
x_7	*	*	+	+	+	*
x_8	*	*	+	+	+	+



		<i>on</i>	<i>off</i>
x_1	Languages	Spanish	French
x_2	Marital Status	Married	Single
x_3	College	Private	Public
x_4	Age	Thirties	Twenties
x_5	Gender	Male	Female
x_6	Profession	Lawyer	Accountant
x_7	Residency	New York	New Jersey
x_8	Department	Local	International

Candidate Name: JACQUELINE LOMAX

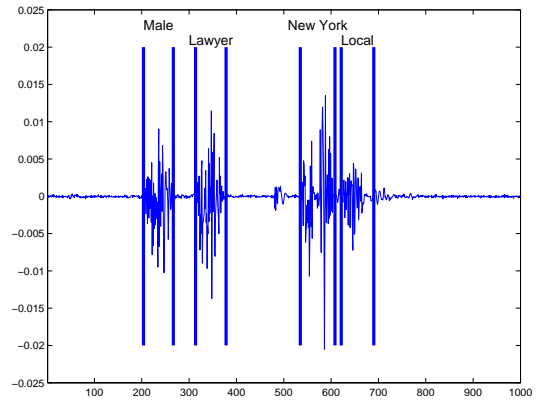
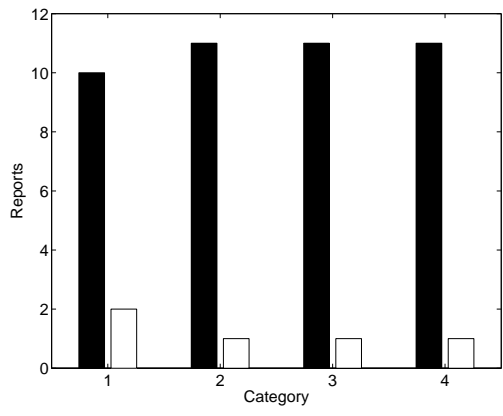
Residency: New Jersey College: Private

Department Preference: Local Languages: Spanish

Gender: Female Age: Thirties

Occupancy: Accountant Marital Status: Single

+
+ - +
+



	<i>on</i>	<i>off</i>	x	x'
x_1	Red	Cyan	-	+
x_2	Blue	Peach	+	-
x_3	Brown	White	+	-
x_4	Purple	Light green	-	+
x_5	Black	Pink	-	+
x_6	Orange	Beige	+	-
x_7	Yellow	Gray	+	-
x_8	Green	Light blue	-	+

