A Dependency Capturing Code for Robust Object Representation

Rishabh Raj[1], Dar Dahlen[1], Kyle Duyck[1] and C. Ron Yu[1, 2, †]

[1]Stowers Institute for Medical Research, 1000 East 50th Street, Kansas City, MO 64110
[2]Department of Anatomy and Cell Biology, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66160

[†]Correspondence: 816-926-4334 (office)
cry@stowers.org

Abstract

The brain has a remarkable ability to recognize objects from noisy or corrupted sensory inputs. How this cognitive robustness is achieved computationally remains unknown. We present a coding paradigm, which encodes structural dependence among features of the input and transforms various forms of the same input into the same representation. The paradigm, through dimensionally expanded representation and sparsity constraint, allows redundant feature coding to enhance robustness and is efficient in representing objects. We demonstrate consistent representations of visual and olfactory objects under conditions of occlusion, high noise or with corrupted coding units. Robust face recognition is achievable without deep layers or large training sets. The paradigm produces both complex and simple receptive fields depending on learning experience, thereby offers a unifying framework of sensory processing.

One line abstract

We present a framework of efficient coding of objects as a combination of structurally dependent feature groups that is robust against noise and corruption.

Introduction

A key insight in our understanding of brain function is that neurons serve as feature detectors whose activity patterns form the basis of internal representations of the external world. Central to this idea is the concept of receptive fields of individual neurons, which correspond to the sensory features that elicit optimal response in the neuron. Classic models of sensory processing are based on the hierarchical organization of sensory features and the efficient coding hypothesis (*1-3*). In these models, sensory features that define an object are faithfully and efficiently represented in the brain, through receptive fields with minimal redundancy (*1, 3*). However, sensory stimuli are often ambiguous and incomplete in the natural environment. Noise, occlusion and other forms of corruption lead to altered representations (**Figure 1A-C**). Additional mechanisms are required to infer the identity of the stimuli. These mechanisms require the storage of reference representations in the brain, an unsolved problem on its own.

An alternative idea, espoused by Marr and Nishihara, is that the brain generates a consistent representation of the object presented in different forms (*4*). This perceptual invariance requires object representation to be stable under various conditions. Meanwhile the representation must maintain sensitivity to distinguishing features among different objects. These two diametrically opposed requirements are difficult to reconcile; no current computational model allows the formation of a common representation from different forms of the same inputs without extensively learning these forms. How the receptive field properties develop and are utilized to form robust internal representation remain unsolved. Here we present a computational framework that provides a solution to the problem. It is conceptualized from a novel interpretation of efficient coding, with a

focus on efficient coding of objects instead of features. We show that an algorithm developed from this new formulation allows redundant feature coding, thereby improve both robustness and sensitivity of object representations as suggested (*5*).

Theoretical Basis

We first provide a reasoning of the framework based on Information Theory. To simplify the problem, we define an object as a collection of coherent, structurally related features corresponding to an entity. For example, a 2-dimensional face image is defined by lines and shades that have a specific spatial arrangement; an odor is defined by the odorant receptors it activates. In a given system, we consider a finite set of ***N*** *objects* (***O***) to be encoded, represented, and recognized. We also consider a simplified model with two levels of sensory coding, performed by a set of primary encoders (e.g., retinal cells) that encodes sensory input at the individual feature level, and a set of representational encoders (e.g., cortical neurons) that encodes the objects. In the classic interpretation of efficient coding, efficiency is achieved when the capacity of the coding system conforms to the entropy $H(X)$ associated with occurrence probability of stimulus *features* (*X*) (*6*). Classic studies have demonstrated that in early sensory processing stages, efficient code of sensory features can be obtained by adapting to the statistics of the natural stimuli (*7-11*).

For the representational encoders, according to the classic interpretation, an efficient coding system ought to match its capacity to the entropy of *objects*, $H(O)$, associated with the probability of occurrence of the *objects* $\mathbb{P}(O_i)$(*1*). However, just as the prior appearances of head or tail in multiple tosses of a coin have no bearing on the

outcome of the next toss, the prior occurrences of the objects do not predict their future appearances. They are not particularly useful in their identification. According to the principle of indifference, when there is no reason to believe that one outcome should occur more frequently than any others, every event should be assigned equal probability (*12*). Therefore, we propose that, for recognition purposes, the essential goal of sensory processing is to seek maximal efficiency in representing *objects* by adapting the code to a uniform distribution of *objects* (i.e., $\mathbb{P}(O_i) = 1/\boldsymbol{N}$). With this assumption, the entropy of the object ensemble is maximal, $H_{max}(O) = \ log\ \boldsymbol{N}$. Incidentally, with redundancy expressed as:

$$R = 1 - \frac{H(O)}{C} \qquad (1)$$

where C represents the capacity of the coding system, minimal redundancy is reached with $H_{max}(O)$ for any value of capacity C. Thus, with this new interpretation of efficient coding, maximal efficiency can be achieved when the coding capacity of the representational encoders matches $H_{max}(O)$.

Efficient coding of a uniform distribution is achievable through maximizing distinction among the representations. As object identities are embedded in complex statistical dependencies among its features, a code set that captures the structural relationships among the object features and utilizes minimal combinations of units should produce maximally distinct representations. Accordingly, an algorithm that preserves information about the object while separating their representations should result in independent encoding of the objects, make this framework *object*-centered. This characteristic distinguishes our framework from that aiming for efficient coding of *features* (*13-15*). Coding local features may allow faithful "reconstruction" of the sensory

inputs (*4, 16, 17*), but consistent representations of objects in cases of corruption, occlusion, and neuronal loss cannot be easily attained without esoteric mechanisms and learning from large examples (**Fig. 1A-C**). Missing or confounding features may generate a representation that resembles something different (**Fig. 1C**). In our formulation, on the other hand, individual features can be redundantly encoded, and their existence can be inferred from the structural dependencies among features such that partial signals may provide sufficient information to produce the same representation as the full-featured stimulus (**Fig. 1D, E**). We refer to this formulation as dependency-capturing (D-Capturing) framework.

What would be an algorithm that can allow the coding elements to capture structural dependencies among sensory features? A ubiquitous observation in the sensory system is an expansion of encoders from primary to high order centers. We therefore adopt a dimensionally expanded representation in our framework. However, dimension expansion results in many possible representations consistent with the input. To obtain a unique representation of the objects, we constrain the system to settle on the sparsest representation. We devise a learning process that maximizes mutual information between the primary and the representational encoders while imposing sparsity constraints. In addition to sparsity, we impose non-negativity constraint on the representations as well as the dictionary. Specifically, we use an objective function to minimize both the *Frobenius* and L_0 norms (implemented as minimizing L_1 norm) under non-negative constraint (see methods). This approach of seeking a sparse code is similar to but distinct from those used before (*14, 15, 18, 19*). A key difference is that, in our model, non-negativity constraint is combined with dimension expansion and sparsity to effectively capture the

morphological components of the sensory inputs (*20*). As the representation of an object can be considered as a vector of coefficients describing the linear combination of dictionary elements (receptive fields), this constraint, while serving as a proxy of neuronal responses carried by spikes, prevents the use of negative coefficients to subtract features from arbitrarily complex dictionary elements. Sparsity constraint can be interpreted as to force the grouping of distinguishing morphological features of stimuli into receptive fields such that the fewest elements can capture the information about the source. Together, these constraints force the dictionary elements to extract structural relationship among features that are naturally present in the stimuli and maximize distinction in representations. Utilization of the coding elements is non-linear; when combined, the representation may not precisely recapitulate the original signal. However, it allows individual features to be inferred from its relationships with others when occlusion occurs.

Robustness of D-Capturing Framework

We illustrate this framework using a simplistic two-layer encoding system with a set of 256 (16x16) primary encoders and a set of 500 representational encoders (**Figs. 2A, S1A**). We train the two-layer network to represent black and white symbols from world languages as objects, because the information of the symbol set can be readily quantified. Dictionaries and representations can be obtained through sequential learning or through non-negative blind source separation (*21*). We estimate the efficiency of the D-Capturing paradigm by comparing redundancies (see methods) present in primary and representational encoders. For the primary encoders, redundancy remains constant

regardless of the number of symbols (**Fig. 2B**). In contrast, when the number of symbols matches that of the representational encoders (***N*** =500), we observe much lower redundancy in the representational than the primary encoders (**Fig. 2B**). This observation indicates a high efficiency of the network in representing the symbols. Additionally, the correlation matrix of representational encoders is very close to identity, indicating the distinctiveness of representations (**Fig. S1C**). To examine whether the dictionary elements capture the structural relationships among pixels in the symbols, we calculate *KL* divergence between the distributions of each primary encoder in symbol and dictionary spaces. The low divergence at low ***N*** indicates that the dictionary structures match the symbols, and the representational encoders are capturing dependencies among symbol features (**Fig. 2C**). This could be further verified by the similarity of correlation matrices of primary encoders in the dictionary and the symbol spaces (**Fig. S1B**).

As ***N*** becomes larger, both redundancy in the representational encoders and *KL* divergence increase, indicating that the structures of the dictionary elements diverge from those of the symbols, and the code becomes less efficient (**Fig. 2B and 2C**). Nevertheless, representational efficiency is still higher than that of the primary encoders.

We next assess representation robustness by adding Gaussian noise to the input signal (**Fig. 2D, ii**), corrupting through randomly silencing the primary encoders (**Fig. 2D, iii**), or occluding part of the symbols (**Fig. 2D, iv**). In each case, representations of the corrupted input patterns are nearly identical to those of uncorrupted signals (**Fig. 2D, i**). Remarkably, when we inversely reconstruct images from the representations using the dictionary elements, the recovered images are the original symbols (**Fig. 2D**). Note that the high level of performance does not require learning from corrupted symbols. To

quantify the specificity of representations, we calculate the pairwise cosine distances between a corrupted input and all learned symbols in the representation space and Z-scored the distances. We observe high Z-scores for the correct input - symbol pairs, indicating that the network generates representations that are highly specific in that they match well with the original uncorrupted input but not others. In Monte Carlo simulations with randomly missing primary encoders, high representational specificity is achieved with as few as 60 (23.4% of the 256) encoders (**Fig. 2E**). We also find that the decoding algorithm is sensitive to small differences in the input patterns. For example, two highly similar inputs are represented differently and robustly (**Fig. 2F**). Thus, D-Capturing can achieve robust and sensitive representations of independent objects.

We next apply the model to complex signals by training it to recognize human faces (**Fig. 3A**). The learned dictionary elements are neither parts of faces, nor do they resemble whole faces. Instead, they are composed of complex assemblage of facial features (**Fig. 3B**). This observation suggests that the algorithm captures structural relationships among the features present in the training set. Representations of the faces are stable and robust against common alterations such as headwear, facial hair or eyewear **(Fig. 3C)**. The same person is represented nearly identically when a mustache, a pair of sunglasses, or both were added. Even when half of a face was blocked in different positions, the model produces the same representation (**Fig. 3D**). Inversely reconstructed images from the representations are similar to those of unadulterated faces even when the faces are half blocked (**Fig. 3C and D**). Notably, the robustness is achieved from learning 2000 "clean" examples, in contrast to many approaches using large of numbers of variegated examples as training sets.

We compared our "face code" against a recently proposed code based on principal components analysis (PCA)(*22*). Using dictionaries obtained through PCA, the same face with different parts occluded generated different representations. Image recovery produces occluded, but not uncorrupted images (**Fig. 3E**). Quantification of specificity using Z-scores show that only the D-Capturing network generates representations that are highly specific in matching the original face (**Fig. 3F)**. Recovered images from representation of corrupted inputs are highly similar to the originals (**Fig. 3G**). PCA-based decoding does not exhibit such selectivity or similarity (**Fig. 3F and 3G**). Thus, our framework suggests a robust combinatorial face code distinct from the one proposed before (*22, 23*)

Near Optimal Sparse Recovery Enables Robustness of D-Capturing

The key to representational stability in the D-Capturing framework is that different forms of the same stimulus converge onto the same representation. We find that stability in our coding framework relies on L_1 minimization (*24*), which can find the near-optimal solutions (i.e., closest to L_0 solutions) for sparse signals (*25, 26*). Classic linear models such as non-negative factorization, or non-linear input transformations are not sufficient to generate representational stability (*20*). We compared L_1 minimization against Locally Competitive Algorithm (LCA) (*27*) and the classic Linear Non-Linear (LNL) approach, in generating representations of occluded faces. L_1 minimization produces representations with the highest specificity (**Fig. 4A**). Although LCA performs better than the LNL, neither produces representations with the specificity of L_1 minimization algorithm. When we compare the sparsity of the solutions, it is evident that LNL does not produce sparse solutions (**Fig. 4B**). LCA produces a solution that only utilizes a subset of components

derived from L_1 minimization. Currently, we do not have a network implementation of the L_1 minimization procedure. It will be interesting to investigate this further.

In the D-Capturing framework, a large set of representational encoders could maintain optimal sparsity and hence efficiency in encoding growing numbers of objects. In an ideal example, if the number of representational encoders $\boldsymbol{K} > \boldsymbol{N}$, one may use $\boldsymbol{N}$ units, with each representing an object uniquely. The system entropy is thus maintained at $\log N$ and redundancy is minimal. As $\boldsymbol{N}$ grows larger, the distinctiveness among representations must be compromised to ensure no loss in information. A relaxation in distinctiveness results in an increased redundancy because the coding capacity grows beyond $H_{max}(O)$ (**Fig. 2**). This corollary suggests that $\boldsymbol{K}$ should be as large as possible. Indeed, a ubiquitous observation of sensory systems is that high brain areas contain much larger numbers of neurons than that of the receptor sheets. From the perspective of theory proposed here, one could argue that a large number of cortical neurons is needed to ensure sparsity of the representation through dimension expansion, and to accommodate a large number of objects to be encoded while maintaining sparsity and coding efficiency.

D-Capturing Permits Generation of Simple and Complex Projective Fields

This rationale raises a question as to what type of dictionary structures emerge if the number of objects needs to be encoded far exceeds the number of neurons. To answer this question, we trained the network with varying numbers of image patches taken from natural images as individual stimuli (*28*) (**Fig. S2A**). To simulate the On and Off channels in the mammalian visual system, the images are parsed into On and Off channels during training. Increasing the size of the training set gradually produces dictionary elements with

localized and orientation-selective projective fields that are similar to the receptive fields of V1 simple cells (*17*) (**Fig 5A**). Despite high correlations among the images, representations are highly decorrelated (**Fig. S2B and 2C**). Interestingly, in all training conditions we also find complex projective fields similar to the V1 complex cells (*17*). With low number of training images, the dictionary elements are relatively complex (**Fig. 5B**). The percentage of simple projective fields in the dictionary increases with the size of the training set. With a fixed set of training images, an increase in the representation dimension reduces correlation among the representations but increases the correlations among dictionary elements (**Fig S2C**). Thus, in this D-Capturing framework, localized tuning features naturally emerge when large numbers of independent objects are encoded. It can be argued that for areas such as V1, the requirement to encode extraordinarily large number of independent stimuli in the natural environment forces it to produce localized tuning features even though the drive of the algorithm is to maximally distinguish stimulus representations. Importantly, complex tuning is always present in our model without requiring it to be synthesized from simple cells as the classic model predicts (*17*).

D-Capturing Permits Robust Odor Identification

The examples using images show that our framework fulfills the criteria of uniqueness, stability and sensitivity proposed by Marr and Nishihara in assessing representational models for object recognition (*4*). Can this model be applied in other sensory systems? Some sensory modalities do not detect external stimuli with pixel-like spatial segregation of the input patterns. In the mammalian olfactory system, for example, an odor activates a disparate set of glomeruli in the olfactory bulb (*29, 30*). The pattern is

transformed into sparse activities in the piriform cortex, where odor identities presumably are represented (*31, 32*). This two-stage system, without explicitly decoding elemental features of the chemicals, has a remarkable ability in identifying individual odorants (*30*). It is also resilient against neuronal loss; rodents can recognize trained odors even when large portions of the olfactory bulb are removed (*33*). We, therefore, test whether our framework can decipher physiological responses recorded from the olfactory system. In a previous study, we have collected the responses of 94 glomeruli to 40 odorants recorded from the dorsal surface of the mouse olfactory bulb (*30*). Training the model to form sparse and independent representations of the odors in a 150-dimensional space, we find nearly identical representation of odors can be generated from the responses of small subsets of glomeruli (**Fig. 6A**). For example, odor representations generated from a random set of 16 glomeruli are nearly identical to those from the full set, suggesting that odor recognition can be achieved with far fewer glomeruli (**Fig. 6A**). Moreover, nearly identical representation of the same odor is achieved using different, arbitrary sets of glomeruli (**Figure 6B**). We have performed Monte Carlo analyses using the responses of different numbers of randomly selected glomeruli and find a rapidly decrease of error rate in odor identification when the number of glomeruli increases (**Fig 6C**). 100% of odorants are correctly identified with an average of 15 or more glomeruli randomly selected from the set (**Fig. 6C**). Decoding glomerular patterns is also robust against noise. We have added Gaussian noise to the glomerular responses and calculated the performance of odor identification from Monte Carlo analyses. Increasing noise level reduces performance (**Fig. 6D**). Nevertheless, odor identification is resilient against noise. At 10% noise level, nearly perfect identification is achieved with 20 glomeruli.

Even when the noise level reached 40% of the signal, 60% of odorants can be identified using the responses from 30 glomeruli.

Discussion

Taken together, by seeking efficiency in object representation, arguably a more pertinent goal of the sensory systems (*34*), the coding paradigm presented here allows relationships arising from structural dependencies among the features to be directly captured, creating representations that are robust against different forms of signal corruption. It also allows a system to learn from small number of examples, requiring neither multiple layers of processing nor learning from large number of variants to enable representational robustness. D-Capturing provides a unifying framework to understand robust object representation across different sensory modalities and across animal species with various levels of complexity of the nervous system. It can explain the appearance of both simple and complex cell like tunings simply based on the tasks and the number of objects that need to be encoded. Importantly, in this framework, object recognition can be achieved solely through forming the representations, without having to solve the inverse problem or requiring additional steps to infer from the responses (**Fig. 1D**). This framework is in contrast with most current theoretical paradigms, which treats the decoding of sensory input as the inverse of sensory-triggered response to obtain stimulus features (**Fig 1A**). Without the need to compare objects in the feature space, generating a representation can be considered as encoding and decoding at the same time; encoding and decoding is one and the same.

Author Contributions

C.R.Y conceived the idea and supervised the research. C.R.Y and R.R. developed key concepts and co-wrote the paper. R.R and D.D. performed analyses and modeling. K.D. generated face database.

Acknowledgments

We thank Drs. K. Si, J. Unruh, P. Kulesa, M. Klee and members of the Yu laboratory for insightful discussions. The work is supported by funding from Stowers Institute and the NIH R01DC 014701. This work fulfills, in part, requirements for RR's and KD's Ph.D. thesis with the Open University, United Kingdom.

Competing interests

The work presented is the subject of a patent application by Stowers Institute, including U.S. Provisional Application No. 62/693,136. The inventors include the CRY, RR and DD. If the invention is commercialized, the inventors are entitled to a share of the income.

Data and materials availability

Data and codes are available at https://github.com/Yulab2019/D-Capturing

Reference:

1. H. B. Barlow, Possible principles underlying the transformations of sensory messages. (1961).
2. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nature neuroscience* **2**, 1019 (1999).
3. J. J. Atick, Could information theory provide an ecological theory of sensory processing? *Network* **3**, 213-251 (1992); published online Epub1992/07/01 (10.3109/0954898X.2011.638888).
4. D. Marr, H. K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B* **200**, 269-294 (1978).
5. H. Barlow, Redundancy reduction revisited. *Network: computation in neural systems* **12**, 241-253 (2001).
6. V. L. Ming, L. L. Holt, Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America* **126**, 1312-1320 (2009).
7. J. J. Atick, A. N. Redlich, Towards a theory of early visual processing. *Neural Computation* **2**, 308-320 (1990).
8. Y. Dan, J. J. Atick, R. C. Reid, Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of Neuroscience* **16**, 3351-3362 (1996).
9. S. Laughlin, A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c* **36**, 910-912 (1981).
10. M. S. Lewicki, Efficient coding of natural sounds. *Nature neuroscience* **5**, 356 (2002).
11. E. C. Smith, M. S. Lewicki, Efficient auditory coding. *Nature* **439**, 978 (2006).
12. J. M. Keynes, Chapter iv: The principle of indifference. *A treatise on probability* **4**, 41-64 (1921).
13. A. J. Bell, T. J. Sejnowski, The "independent components" of natural scenes are edge filters. *Vision Res* **37**, 3327-3338 (1997); published online EpubDec (
14. B. A. Olshausen, D. J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607-609 (1996); published online EpubJun 13 (10.1038/381607a0).
15. B. A. Olshausen, D. J. Field, Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res* **37**, 3311-3325 (1997); published online EpubDec (
16. D. Marr, A theory of cerebellar cortex. *The Journal of physiology* **202**, 437-470 (1969).
17. D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* **160**, 106-154 (1962); published online EpubJan (
18. P. O. Hoyer, in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*. (IEEE, 2002), pp. 557-565.
19. P. O. Hoyer, Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research* **5**, 1457-1469 (2004).

20. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999); published online EpubOct 21 (10.1038/44565).
21. J. Rapin, J. Bobin, A. Larue, J.-L. Starck, Sparse and non-negative BSS for noisy data. *Signal Processing, IEEE Transactions on* **61**, 5620–5632 (2013); published online Epub2013 (
22. L. Chang, D. Y. Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013-1028. e1014 (2017).
23. C. F. Stevens, Conserved features of the primate face code. *Proceedings of the National Academy of Sciences* **115**, 584-588 (2018).
24. E. Candes, J. Romberg, l1-magic: Recovery of sparse signals via convex programming. *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf* **4**, 14 (2005).
25. E. J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on* **52**, 489–509 (2006); published online Epub2006 (
26. E. J. Candes, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory* **52**, 5406-5425 (2006).
27. C. J. Rozell, D. H. Johnson, R. G. Baraniuk, B. A. Olshausen, Sparse coding via thresholding and local competition in neural circuits. *Neural Comput* **20**, 2526-2563 (2008); published online EpubOct (10.1162/neco.2008.03-07-486).
28. J. H. Van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences* **265**, 359-366 (1998).
29. A. L. Fantana, E. R. Soucy, M. Meister, Rat olfactory bulb mitral cells receive sparse glomerular inputs. *Neuron* **59**, 802-814 (2008).
30. L. Ma, Q. Qiu, S. Gradwohl, A. Scott, Q. Y. Elden, R. Alexander, W. Wiegraebe, C. R. Yu, Distributed representation of chemical features and tunotopic organization of glomeruli in the mouse olfactory bulb. *Proceedings of the National Academy of Sciences* **109**, 5481-5486 (2012).
31. C. Poo, J. S. Isaacson, Odor representations in olfactory cortex:"sparse" coding, global inhibition, and oscillations. *Neuron* **62**, 850-861 (2009).
32. D. D. Stettler, R. Axel, Representations of odor in the piriform cortex. *Neuron* **63**, 854-864 (2009); published online EpubSep 24 (10.1016/j.neuron.2009.09.005).
33. X.-C. Lu, B. Slotnick, Olfaction in rats with extensive lesions of the olfactory bulbs: implications for odor coding. *Neuroscience* **84**, 849-866 (1998).
34. D. J. Field, What is the goal of sensory coding? *Neural computation* **6**, 559-601 (1994).
35. P. Comon, Independent component analysis, a new concept? *Signal processing* **36**, 287-314 (1994).
36. P. Comon, C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*. (Academic press, 2010).
37. G. I. Allen, L. Grosenick, J. Taylor, A generalized least-square matrix decomposition. *Journal of the American Statistical Association* **109**, 145-159 (2014).

38. J. Rapin, J. Bobin, A. Larue, J.-L. Starck, Sparse Regularizations and Non-negativity in BSS. *Proceedings of SPARS, Lausanne, Switzerland*, 83 (2013).
39. D. L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization. *Proceedings of the National Academy of Sciences* **100**, 2197-2202 (2003).
40. D. L. Donoho, Compressed sensing. *Information Theory, IEEE Transactions on* **52**, 1289-1306 (2006).
41. D. L. Donoho, For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics* **59**, 797–829 (2006); published online Epub2006 (
42. E. J. Candes, T. Tao, Decoding by linear programming. *IEEE transactions on information theory* **51**, 4203-4215 (2005).
43. S. S. Chen, D. L. Donoho, M. A. Saunders, Atomic decomposition by basis pursuit. *SIAM review* **43**, 129-159 (2001).
44. D. L. Donoho, Y. Tsaig, I. Drori, J.-L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE Transactions on Information Theory* **58**, 1094-1121 (2012).
45. S. Boyd, L. Vandenberghe, *Convex optimization*. (Cambridge university press, 2004).

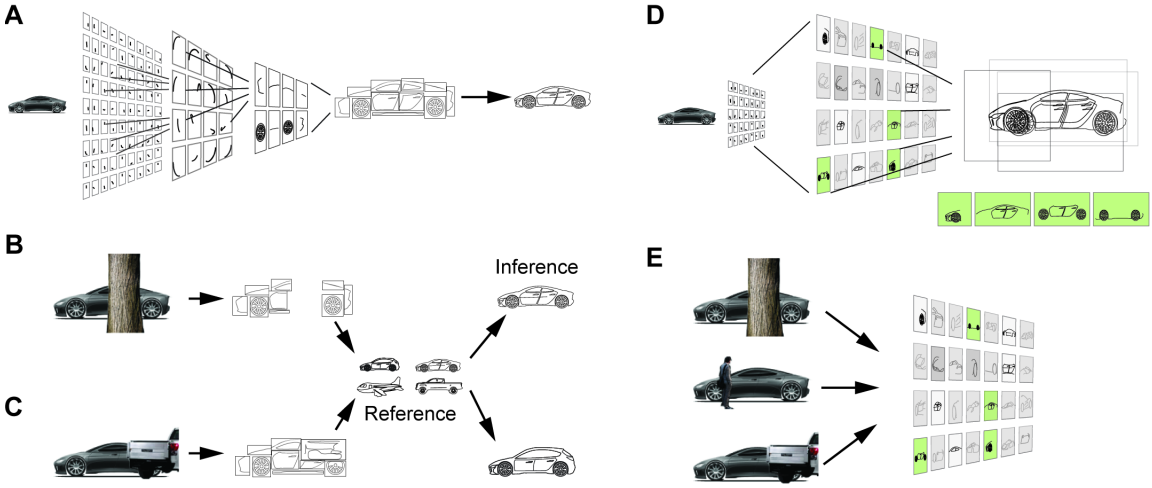

Figure 1. Illustration of the classic and D – Capturing frameworks for object representation: A. In the classic framework, representation is compositional. Simple features detected by the sensory organs are integrated in hierarchically organized circuits to generate increasingly complex receptive fields (RFs). The set of RFs that are activated by an object forms a representation of the object. **B – C.** Occlusion presents a problem for the classic framework. With only partial features detected, representations of the objects differ from the unobstructed one. Additional mechanisms such as inference from a stored reference are required to identify the object (B). Ambiguity in confounding representation may also lead to misidentification of the object (C). Where and how the references are stored is not known. **D – E.** D-Capturing framework generates independent cortical models of objects without explicitly recovering stimulus features. **(D)** In D-Capturing framework, encoders capture features as well as the arrangement of these features with respect to each other. This allows embedding of complex structural relationships among features in the receptive fields of the encoders without requiring (nor exclusive of) a hierarchical assembly. Only a sparse set of units are consistently active (highlighted and displayed in green) to represent an object even though more units have RFs consistent with the input. Because individual features are redundantly encoded, missing features can be inferred from their relationship with those stored in the RFs. **(E)** D-Capturing allows the same objects, occluded or not, to generate the same representation.

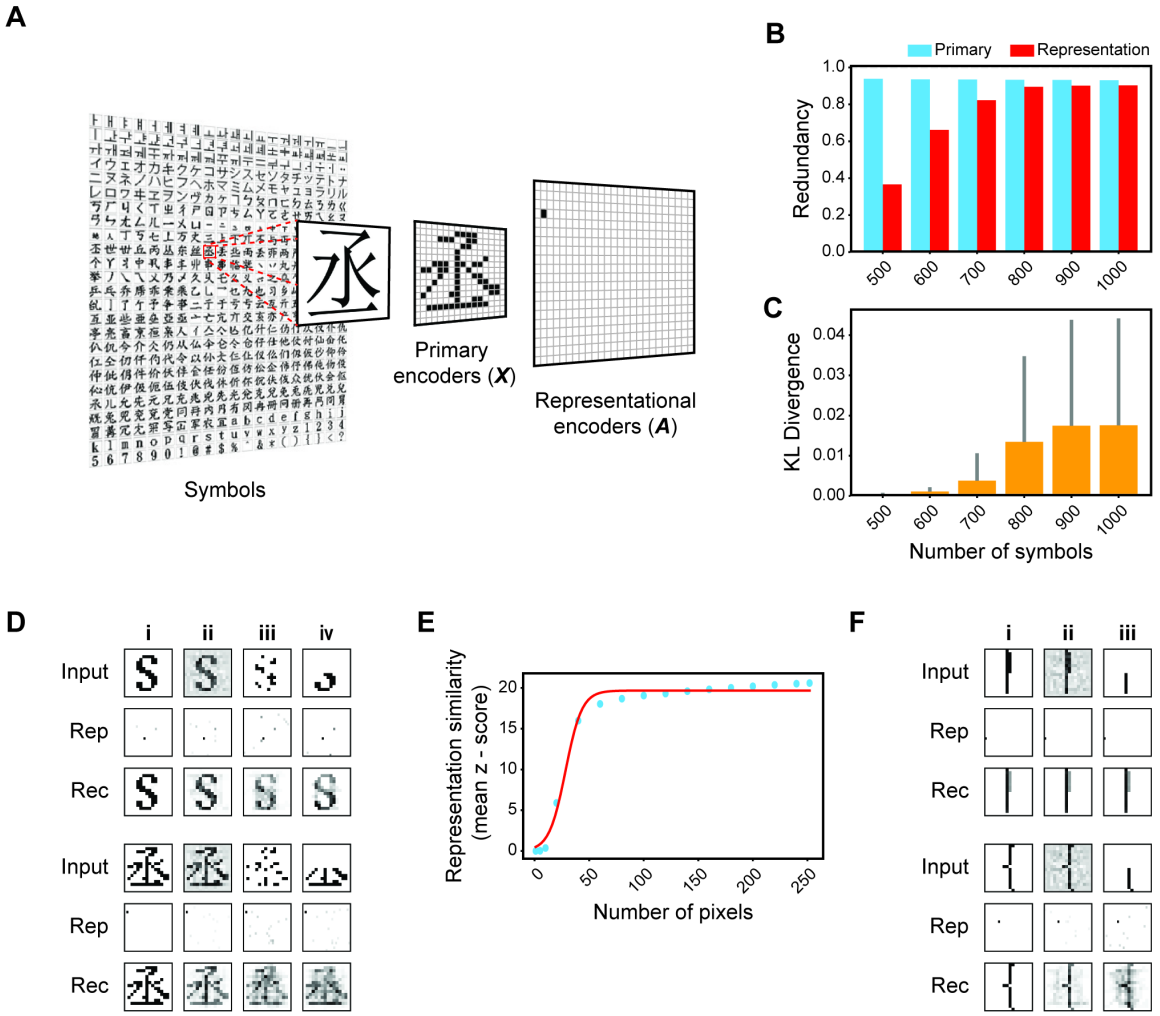

Figure 2. Efficiency and robustness of the D-Capturing framework: A. Illustration of a system with two levels of encodings in the D-Capturing framework. A set of 256 (16x16) primary encoders (***X***) encodes local features (pixels) of the symbols from world languages and a set of 500 (20x25) representational encoders (***A***) encodes symbol identities. An example of the process is demonstrated using a Chinese character. This character happens to be encoded by a single representational encoder which indicates the ability of representational encoders to encode complex structural features. **B.** Analysis of redundancies in primary and representational encoders while encoding increasing numbers of symbols. Redundancies are calculated as $R_X = 1 - \frac{H_{max}(O)}{C_X}$ and $R_A = 1 - \frac{H_{max}(O)}{C_A}$ respectively, where C_X and C_A are corresponding channel capacities and $H_{max}(O) = \log N$. **C.** Analysis of *KL* divergence of pixels between dictionary and symbol spaces increases with increasing number of symbols. **D.** Examples of decoding

corrupted input under noisy **(ii)**, pixel loss **(iii),** and occlusion **(iv**) conditions. Representations (Rep) nearly identical to the originals (**i**) were generated. Reconstructed images (Rec) from the representations resemble the original symbols. **E.** Z- scored similarity between representations of corrupted and original signals as a function of total pixels in the input layer (randomly selected). Selectivity scores at different input sizes (blue dots) is fit to a sigmoidal curve (red line). **F.** Example of two highly similar symbols being distinctly and robustly represented.

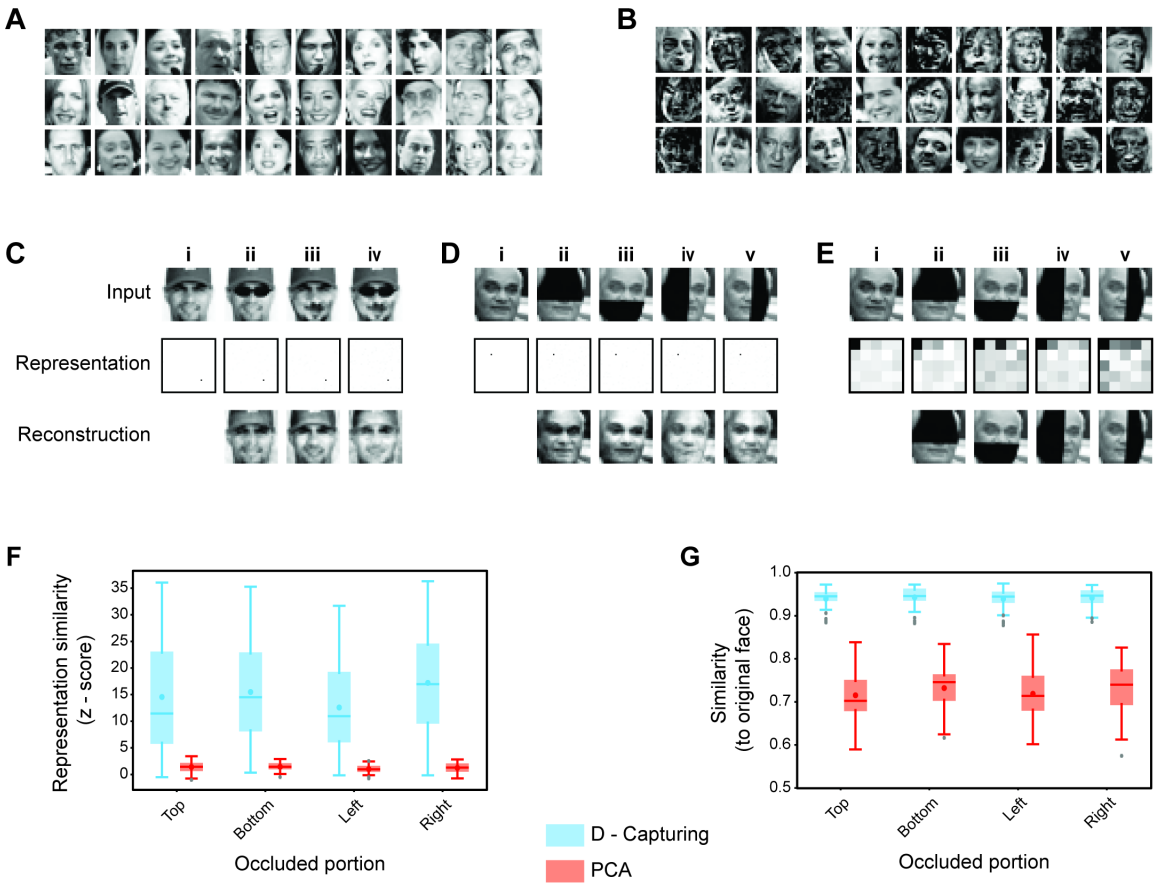

Figure 3. Representation of human faces using D-Capturing: A. Examples of face images in a library of 2000 (1000 male and 1000 female) used to train a two-layer network. **B.** Examples of dictionary elements learned from the face library. Note that they incorporate complex combinations of facial features but are not necessarily part or whole of any specific face. **C.** Representation and recovery of faces with different alterations. A face (**i**) was altered to wear a mustache (**ii**), a pair of sunglasses (**iii**) or both (**iv**). Representations of the altered faces were nearly identical to the original even though these examples were not in the training set. Images reconstructed from the representations based on the dictionary were similar to the original images. **D.** Representation and recovery of occluded faces. Four different occlusions of a face - top (**ii**), bottom (**iii**), left (**iv**), and right (**v**) were generated. Representations of the occluded faces were highly similar to the original one. Reconstructions also matched the original face. **E.** Face identity was not preserved in representations based on PCA performed on the 2000 training faces. Representations of original (**i**) and occluded faces (**ii-v**) were obtained in the principal space (the first 25 components are shown to highlight the differences). Representations of occluded faces are different from original. Reconstructed images match the corrupted rather than original faces. **F – G.** Specificity (Z-score) of face

representations **(F)** and similarity (cosine distance) between reconstructed and original images **(G)** were calculated for D-capturing (cyan) and PCA (pink). 50 faces were chosen randomly from the training set to create the four occluded versions.

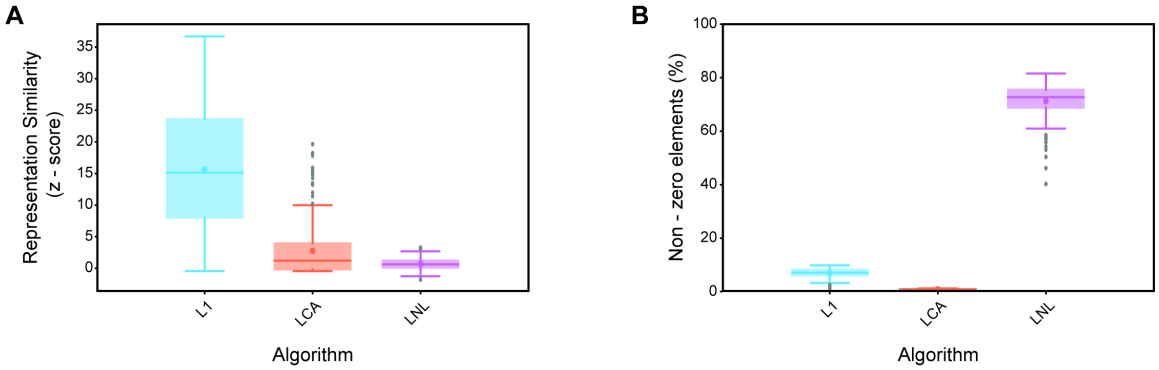

Figure 4. Near-optimal sparse solutions are necessary for high specificity decoding: A. Comparison of specificity among three algorithms in obtaining sparse representations of 50 occluded facial images as in Fig. 3F and 3G. L_1 minimization method produced representations with the highest specificity. **B.** Comparison of sparsity of the representations (measured as the number of non-zero components in the representations) obtain through the different methods. Representations from L_1 minimization are optimal and sparse. LCA recovers more sparse representations because it only recovers specific components of the optimal solution. LNL does not produce sparse representations.

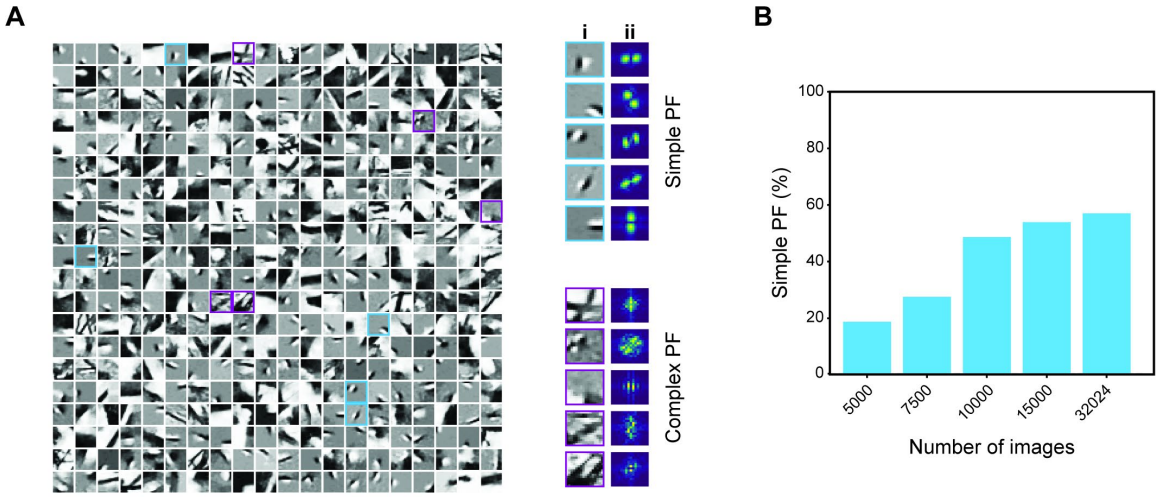

Figure 5. Emergence of simple and complex projective fields from training with natural images: A. Examples of projective fields (PFs) of individual representation elements. Both simple (cyan) and complex (magenta) projective fields are observed (i) and can be determined using Fourier transforms of the projective fields (ii). **B.** As the number of training images increases, the tuning properties of encoders become more localized and the percentage of simple PFs increases.

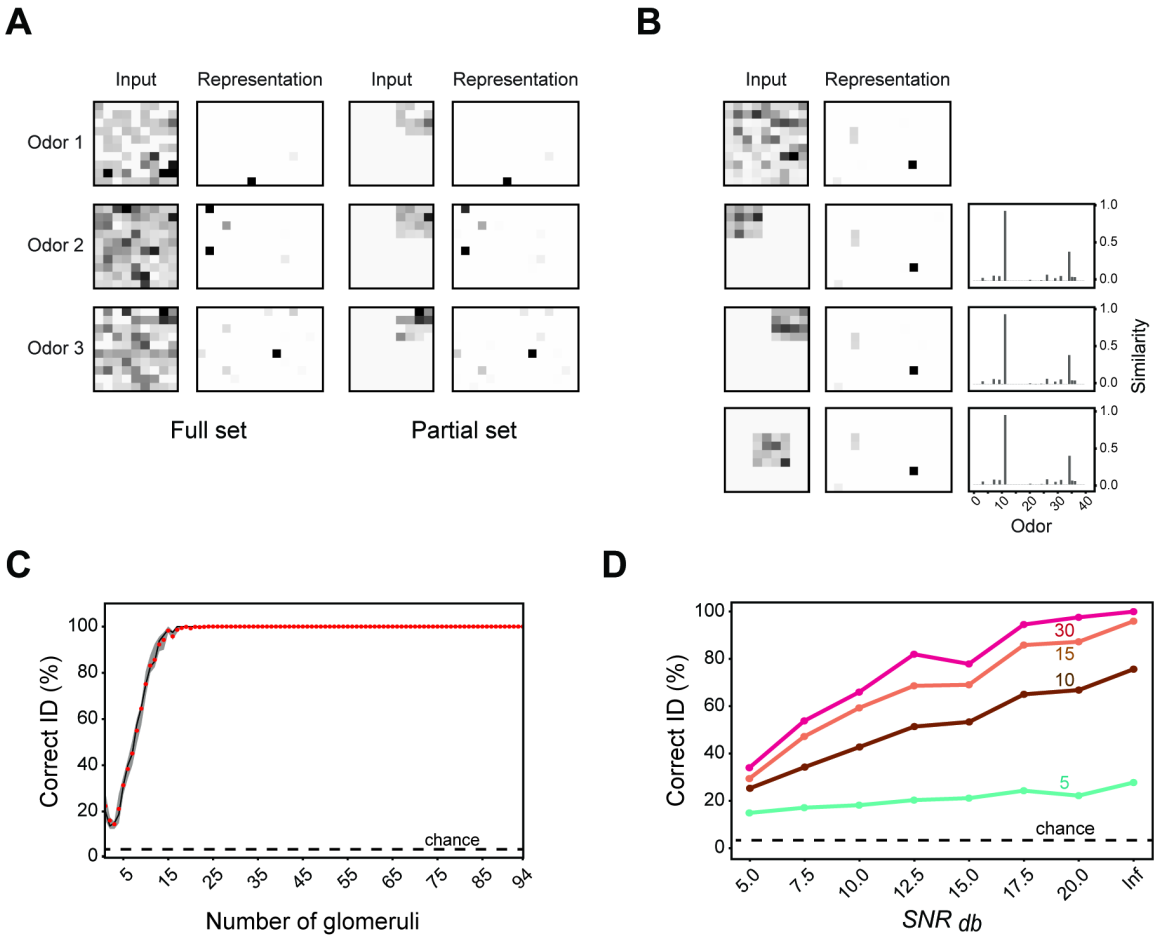

Figure 6. Odor identification using D-Capturing: A. A model with 150 representational encoders was trained to represent 40 odors recorded from 94 glomeruli (90 of which are shown in a 9x10 array). Three different odors and their learned representations are shown in the left columns (Full set). The right columns show the representation of the same 3 odors generated using the response of a subset of 16 glomeruli (Partial set). Note the similarity between representations. **B.** Odor identification based on different subsets of glomeruli. Responses of the full set and three different subsets of glomeruli (left column) are used to generate representation of the same odor (middle column). Note the similarity among the representations. For all three subsets, the representation was identified to have the highest similarity coefficient with the original odor. **C.** Percentage of odors correctly identified based on glomerular response in Monte Carlo analysis. Increasing the number of glomeruli increases accuracy. Mean (red dots) and median (black line) indicate percentage of correctly identified odors from randomly selected sets of glomeruli. Gray band: interquartile range of the distribution. **D.** Performance in odor identification under noisy conditions. Different levels of Gaussian

noise are added to the glomerular responses and at each level of noise a Monte Carlo analysis was performed. The percentage of correctly identified odors are plotted as a function of SNR. Increasing the number of glomeruli (number marked on the lines) improves accuracy.

SUPPLEMENTAL MATERIALS
Methods:

Learning algorithm

Dictionary learning is treated as a blind source separation (BSS) problem (*35, 36*). An input signal is modeled as the response of $\boldsymbol{M}$ primary encoders. In the case of images, $\boldsymbol{M} = \boldsymbol{m_1} \cdot \boldsymbol{m_2}$, where $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are the horizontal and vertical dimensions of the images. For olfactory data analyses, $\boldsymbol{M}$ equals the number of glomeruli. For a set of $\boldsymbol{N}$ signals used in a training set, the training set is presented as an input matrix $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{M} \times \boldsymbol{N}}$, representing the response of $\boldsymbol{M}$ pixel (glomeruli) to $\boldsymbol{N}$ patterns (odors). The matrix $\boldsymbol{X}$ is then factorized into two matrices $\boldsymbol{A}$ and $\boldsymbol{\Phi}$, so that $\boldsymbol{X} = \boldsymbol{\Phi A}$. Here, $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{K} \times \boldsymbol{N}}$ is the matrix representation of $\boldsymbol{N}$ patterns (odors) in a $\boldsymbol{K}$ dimensional basis set defined by $\boldsymbol{\Phi} \in \mathbb{R}^{\boldsymbol{M} \times \boldsymbol{K}}$.

To get the factor matrices through BBS, we imposed restriction on $\boldsymbol{A}$ to be sparse. The measure of sparsity was chosen to be L_0 norm, but the solution is achieved through minimizing L_1 norm. In addition to this sparsity constrain, we demanded both $\boldsymbol{A}$ and $\boldsymbol{\Phi}$ to be non-negative.

Several possible BSS algorithms could result in an appropriate matrix decomposition under the given constraints (*19, 21, 37*). In particular, we used non-negative blind source separation algorithm nGMCA (*21, 38, 39*). When a L_1 measure of sparseness is used, then the sum of the absolute values of coefficients of $\boldsymbol{A}$ is minimized. The minimization problem takes the form of:

$$argmin\ \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Phi A}\|_2^2 + \lambda\|\boldsymbol{A}\|_1, \qquad subject\ to\ \boldsymbol{A} \geq 0;\ \boldsymbol{\Phi} \geq 0$$

Thus, the process to solve this problem requires the minimization of the *Frobinius* norm difference (i.e., the Euclidean Distance) between the two sides of the equation and the minimization of the L_1 norm.

Each time BSS is performed, the $\boldsymbol{\Phi}$ matrix was seeded with random numbers. Optimization was performed until convergence or when predefined number of iterations was reached.

Sparse coding:

Once the dictionary $\boldsymbol{\Phi}$ is learned, any input pattern can be transformed into its corresponding representation. Transformation of input patterns is the process of finding the representation $\boldsymbol{a}$ that satisfies the equation: $\boldsymbol{x} = \boldsymbol{\Phi a}$. In our case, the dimension of the representational layer is chosen to be higher than that of the input layer, i.e., $\boldsymbol{K} > \boldsymbol{M}$. Here, decoding becomes an under-determined problem. Theories developed independently by Donoho (*39-41*), and by Candes and Tao (*25, 26, 42*) show that a unique solution can be obtained by imposing a sparseness constraint to the equation when solving the optimization problem. The most common use of sparsity definition includes L_0 and L_1. In our approach we perform L_1 minimization to solve:

$$\min\|\boldsymbol{a}\|_1\ subject\ to\ \|\boldsymbol{x} - \boldsymbol{\Phi a}\|_2 \leq \epsilon$$

The L_1-minimization problem can be implemented by a standard convex optimization procedure, which can be found in several publications (*41-45*).

Redundancy measurement:

To measure redundancy in encoding objects, we treated the objects as following a uniform distribution, i.e., $\mathbb{P}(O_i) = 1/\boldsymbol{N}$, where $\boldsymbol{N}$ is the total number of objects. The entropy of the ensemble of the objects is therefore $H_{max}(\boldsymbol{O}) = log\, \boldsymbol{N}$. We then calculated the capacity of the primary encoder set (C) using the probabilities of occurrence of each encoder, $\mathbb{P}(x_i = 1) = p_i$:

$$C = \sum_{i=1}^{M} p_i \, log \frac{1}{p_i} + (1 - p_i) \, log \frac{1}{(1 - p_i)}$$

Redundancy was calculated as

$$R = 1 - \frac{H_{max}(\boldsymbol{O})}{C} = 1 - \frac{log\, \boldsymbol{N}}{C}$$

The redundancy for representational encoders was calculated in a similar way, the only difference being that the representations were converted to binary forms using a Heaviside step function so that their L_0 norms could be considered while calculating probability of occurrence of individual encoders.

KL Divergence between dictionary and images

We used the Kullback–Leibler divergence (KL Divergence) to quantify the structural differences between symbols and dictionary elements. KL divergence ($D_{KL}(\mathbb{P}||\mathbb{Q})$) is a measure of information gained when a posterior probability distribution $\mathbb{P}$ is used to calculate the entropy instead of the prior distribution $\mathbb{Q}$. Denoting $\mathbb{Q}$ to be distribution over the states of a single pixel in symbol space and $\mathbb{P}$ to be the distribution over states of the same pixel in dictionary space, $D_{KL}(\mathbb{P}||\mathbb{Q})$ measures the information gained in considering the pixel to be coming from a dictionary element rather than symbols. A low divergence for all the pixels indicates that there is no gain in information if we consider any pixel to be coming from dictionary, indicating that the structure of the dictionary elements is same as structure of the symbols.

To calculate the distribution over the states of pixels in the dictionary space, all dictionary elements were binarized using a Heaviside step function. Probability of occurrence of individual pixels was calculated based on the number of dictionary elements in which the pixel is active. For instance, if a particular pixel x_i was active in n out of $\boldsymbol{K}$ dictionary elements, then the probability of occurrence of pixel x_i was calculated as

$$\mathbb{P}(x_i = 1) = \frac{n}{\boldsymbol{K}}$$

Probability of occurrence of the same pixel in symbol space is calculated based on the number of symbols m in which it is active i.e.

$$\mathbb{Q}(x_i = 1) = \frac{m}{\boldsymbol{N}}$$

Here $\boldsymbol{N}$ is the number of symbols being encoded. Finally, the KL Divergence between the two distributions is calculated as

$$D_{KL}(x_i) = \mathbb{P}(x_i = 1)\, log \frac{\mathbb{P}(x_i = 1)}{\mathbb{Q}(x_i = 1)} + \big(1 - \mathbb{P}(x_i = 1)\big)\, log \frac{\big(1 - \mathbb{P}(x_i = 1)\big)}{\big(1 - \mathbb{Q}(x_i = 1)\big)}$$

Specificity calculations:

To quantify the specificity of a representational vector in representing the original object, we computed Z-scored similarity. Cosine similarity score between the representation of the test object ($\boldsymbol{a_{test}}$) and all objects in the training set ($\boldsymbol{A_{training}}$) were calculated and Z-scored. A high Z-score indicated high similarity between the representations of the test object and a particular object in $\boldsymbol{A_{training}}$. In the figures, we plot the Z-scores for altered images with their unadulterated counterpart, which show high specificity in representing the original object.

Simulating corrupted signals:

To test the robustness of object representation by the D-Capturing framework, signals from the training set were selected and corrupted. The corrupted signals were subject to sparse decoding to generate their representations, which were then compared with those of the signals in the training set. We performed the following three types of corruption:

<u>Noise-added corruption:</u> we introduce noise by adding a Gaussian i.i.d. matrix $\boldsymbol{\mathcal{N}}$ of varying standard deviation to the input matrix $\boldsymbol{X}$. i.e. $\boldsymbol{X_{\mathcal{N}}} = \boldsymbol{X} + \boldsymbol{\mathcal{N}}$, where $\boldsymbol{X_{\mathcal{N}}} \in \mathbb{R}^{M \times N}$, is a matrix representation of noisy input. For Monte Carlo analysis, as described below, each time a simulation was performed, a different noise matrix $\boldsymbol{\mathcal{N}}$ was introduced.

<u>Pixel corruption</u>: For a given signals, a fraction of the $\boldsymbol{M}$ pixels (glomeruli) were selected from the input. Their values are maintained whereas the coefficients of the rest were set to zero.

<u>Occlusion</u>: For images, a contiguous set of pixels were selected, and their coefficient values were set to zero.

Monte Carlo analysis:

We performed Monte Carlo simulations by applying pixel corruption to the input signals and varying the number of corrupted pixels. 100 random sets (numbers varied from 2 to $\boldsymbol{M}$) of pixels (glomeruli) were selected. Using each of these randomly chosen sets, we performed sparse decoding to generate representation of the input patterns.

Input Identification

To calculate the correct identification of the object, we used the representation of each input in the training set as a library. The representation of each corrupted signal was

compared with that in the library and cosine errors were computed. An input pattern was considered correctly identified if the cosine error between its representation and that of the original signal was minimum (smaller than with representation of other patterns).

Data:

Symbols: A set of 1000 symbols from word languages were obtained and digitized to 16x16 pixel arrays.

Natural images: Natural scenes from Van Hateren data base (*28*) were digitized as grayscale pictures. Image patches of 16x16 size were randomly selected from the images. A total of 3000 patches were used to form a training set.

Facial images: For face recognition, 2000 frontal faces were obtained from Google search of publicly available images, trimmed and resized to 25 x 25 pixels.

Olfactory bulb recordings: The data used was previously published (*30*) calcium imaging of the olfactory bulb, in which we have imaged the response of dorsal olfactory bulb of GCaMP2 mice to 189 chemicals. Of these chemicals, ~150 did not elicit significant responses in the glomeruli. Since non-responding stimuli provide no information for our analyses, we removed them from further analysis. To accomplish this, we calculated the Euclidean length of each response and plotted a histogram of response amplitude. The data set used in the paper contains $\Delta F/F$ responses of 94 glomeruli to various odorants. At a threshold of 0.1 (10% $\Delta F/F$), 40 chemicals are shown to elicit a measurable response. These odor-evoked responses were used as a training set.

Projective fields generation from natural images:

The mammalian visual systems process visual information in On and Off channels. On channel images were the normal images whereas Off channel images were the inverted images. To simulate parallel processing of the two channels, the On and Off images were concatenated along the rows and dictionary elements were generated by performing BSS on the concatenated matrix. The projective fields were constructed by superposing the On-channel portion of the dictionary element with the negative of Off-channel portion of the dictionary element.

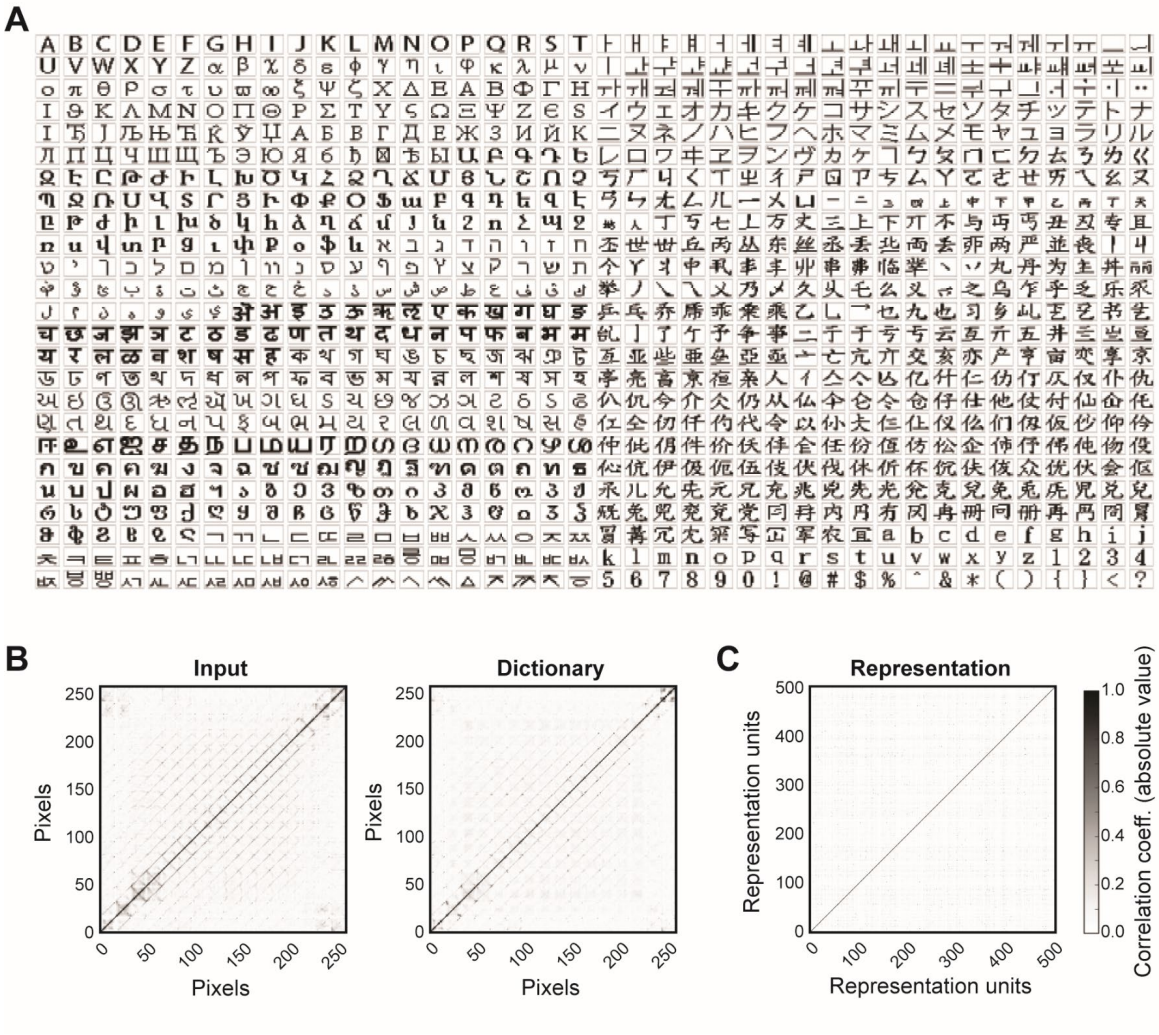

Figure S1. Symbol representations using D-Capturing: A. 1000 symbols used in the simulation. **B.** Correlation between pixels as they occur in symbols and dictionary spaces. **C.** Correlation between representational encoders.

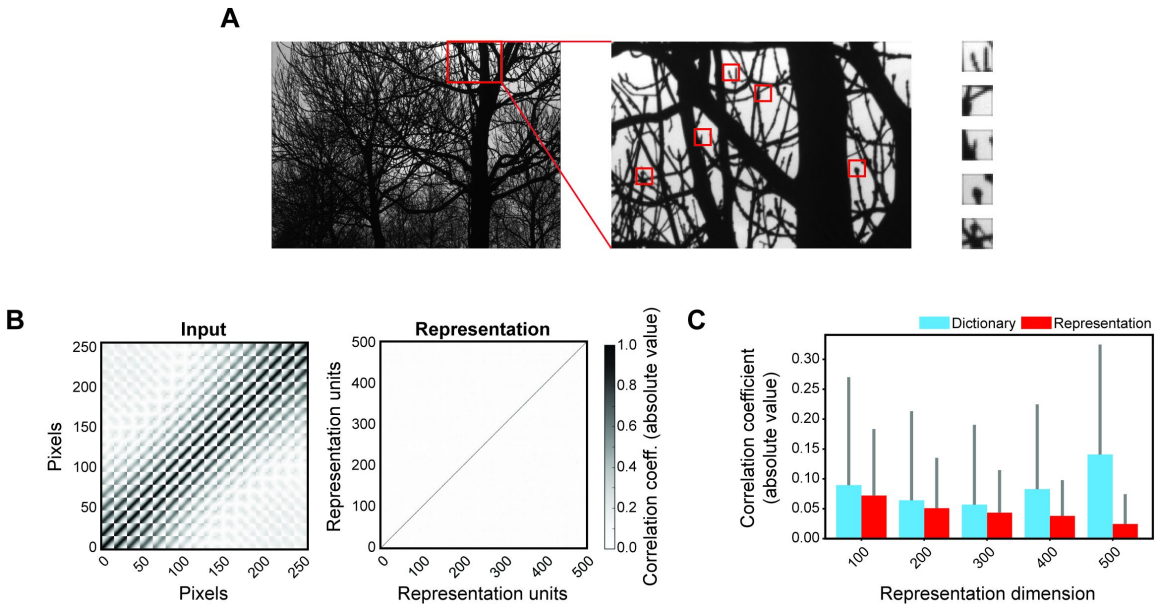

Figure S2. Natural image analysis: A. Examples of image patches from natural images used for training. **B.** Correlation among image pixels as they appear in images and among representational components (bottom panel) indicate decorrelation among representational units compared to image pixels. **C.** Correlation among representational units decrease as their numbers is increased. Bars indicate the mean of observed correlation coefficients (absolute values) and the whiskers indicate the observed standard deviation in the values.