Complex and simple cells form a structured representation of identities and attributes

Pietro Berkes,* Richard Turner, and Maneesh Sahani
Gatsby Computational Neuroscience Unit
Alexandra House, 17 Queen Square, London WC1N 3AR, United Kingdom

Abstract

Many computational models have offered functional accounts of the organization of the sensory cortex. However, most have lacked the structure needed to extract the high-order causes of the sensory input. Here we present a generative model of visual input based on the duality between the identity of image features and their attributes. The presence of a feature is encoded by a binary identity variable, while its appearance is modeled by a multidimensional manifold, parametrized by a set of attribute variables. When applied to natural image sequences, the model finds attribute manifolds spanned by localized Gabor wavelets with similar positions, orientations, and frequencies, but different phases. Thus the inferred activity of attribute variables after learning resembles that of simple cells in the primary visual cortex. Identity variables indicate the presence of a feature irrespective of its position on the underlying manifold, making them phase-insensitive, like complex cells. The dimensionality of the learnt manifolds and the relationships between the wavelets correspond closely to anatomical and functional observations regarding simple and complex cells. Thus, this generative model makes explicit an interpretation of complex and simple cells as elements in the segmentation of a visual scene into independent features, with a parametrization of their episodic appearance. It also suggest a possible role for them in a hierarchical system that extracts progressively higher-level entities, starting from simpler, low-level features.

1 Introduction

The properties of cells in the cerebral cortex are known to be linked to the structure of the sensory environment. One (Helmholtzian) view for why this might be, is that the goal of a perceptual system — to infer from sensation the environmental causes most likely to be responsible — compels it to reflect the generative causal structure of the environment. Recent theoretical work that links receptive fields in the visual cortex to the statistics of natural images may be viewed in this light. An assumed model specifies the properties of causes and how they combine to generate images; the parameters of the model are fit to an ensemble of natural images; and then inference within the learnt model is compared to the response of cortical cells. However, the generative models assumed tend to be elementary: the effects of hidden causes superimpose linearly in the image; the causes are homogeneous *a priori*; and their distributions are either *independent* and *sparse* [1, 2], or (in video sequences) *independent*, but *temporally stable* or *predictable* [3, 4, 5]. Despite their simplicity, such models have been notably successful in mirroring response properties of visual cortical neurons.

The true causal structure of images is more complex. One departure from the simplistic model can be seen in the failures of an algorithm that has largely been successful in extracting high-level properties from a simple 1D environment. The Slow Feature Analysis (SFA) algorithm uses a statistical model in which data are generated by a number of slowly-varying sources [6, 7]. When exposed to a simple

* http://www.gatsby.ucl.ac.uk/~berkes/, berkes@gatsby.ucl.ac.uk

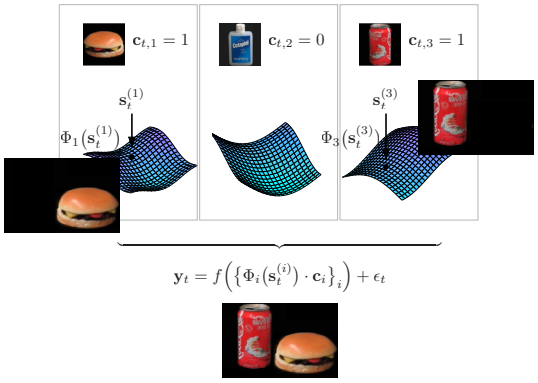

Figure 1: Illustration of the basic structure of the model. Each object or feature is represented by a binary variable $c_{t,i}$ that indicates its presence or absence and is modeled by a manifold formed by the set of its episodic poses, defined by a mapping Φ_i and parametrized by variables $s_{t,ij}$ that are interpreted as attributes of the object or feature. The episodic poses are multiplied by the state of the identity variables, so that absent objects give no contribution, and then combined through a function f to generate the observations $\mathbf{y}_t$.

environment formed by translations of random objects, SFA learns two sets of variables: one set whose response relates to the form of the object, independent of its position (*what* information), and one which gives the position of the object, irrespective of its form (*where* information) [6]. Although these results are encouraging, there are important limitations. First, although the two kinds of signal differ semantically, the model gives all variables the same *a priori* meaning. A readout system that needed to access just one of the signals would face the difficult problem of distinguishing between them. Second, the input sequences used for these experiments contained only individual objects. When multiple objects are present, a what/where division still emerges, but different objects are typically mixed into single features, as their signals have similar temporal scales (Anonymous 2005, unpublished results). Further, if every object needed to be characterized by more than one attribute (for example, if it varied in position *and* scale), attributes that belonged to the same object would not be bound. These problems are not due to the particular prior over latent variables assumed by SFA (on the contrary, our results suggest that the variables in our model are best described by a slowly-varying dynamics), but instead come from the structural mismatch between the SFA model and the environment, and should be expected in any model that assumed homogeneous variables.

We therefore propose a different class of models, in which the duality of object or feature identity on the one hand and the ensemble of its attributes on the other, is represented explicitly. One possible class of models is *bilinear*. While such models have been studied before [8, 9], this earlier work was based on explicitly labeled objects or features in training data (that is, different views of the same item were labeled as such). Here, we show that a simple bilinear model trained in an entirely unsupervised way from natural image sequences, naturally learns biologically plausible features, with low dimensional manifolds of attributes. Many aspects of the learnt representation correspond closely to anatomical and functional observations regarding simple and complex cells in the primary visual cortex (V1). This offers a functional interpretation for the presence of two main classes of cells in V1. Complex cells represent the probability of presence of an oriented feature, while simple cells parameterize the precise appearance of the feature in the visual input.

2 The model

We implemented the basic distinction between identity and attributes using a generative model with two coupled sets of variables with distinct semantics. The *identity* of external causes is represented by binary variables $c_{t,i}$ that indicate the presence or absence of cause i at time t. The appearance of each cause in the input is modeled by a manifold formed by the set of its episodic poses, i.e. every point on the manifold is a possible configuration of the object or feature in the input space. The manifold is defined by a mapping Φ_i and parametrized by variables $s_{t,ij}$, that are interpreted as *attributes* (Fig. 1). To make this concrete within a cartoon example, consider the rightmost panel of Figure 1, which contains the model for a beverage can. The arrow indicates the point on the manifold where the can has a particular position and viewpoint in the input visual space. If one of the attribute variables corresponds to the orientation of the can, changing its value would trace a trajectory on the manifold, which would result in a rotation of the object in the image space.

As shown in Figure 1, these two sets of variables interact to form the input data. To generate the observations $\mathbf{y}_t$, the episodic poses $\Phi_i(\mathbf{s}_{t,i})$ are multiplied by the state of the identity variables $c_{t,i}$,

so that absent causes give no contribution, and then combined through a function f:

$$\mathbf{y}_t = f\Big(\{\Phi_i(\mathbf{s}_{t,i}) \cdot c_{t,i}\}_i\Big) + \epsilon_{\mathbf{t}}\,, \tag{1}$$

where $\epsilon_{\mathbf{t}}$ is an additive, independent noise term.

Here, we follow [8, 9], and define the mappings $\Phi_i(\mathbf{s}_{t,i})$ to be linear (equivalently, we define the attribute manifolds to be hyperplanes) and f to sum its arguments. This gives a bilinear mapping

$$\mathbf{y}_t = \sum_{i=1}^{d_c}\sum_{j=1}^{d_s} \mathbf{w}_{ij}\, s_{t,ij} c_{t,i} + \epsilon_{\mathbf{t}}\,. \tag{2}$$

Assuming that the noise term is Gaussian with variance $\sigma^2_{y,d}$ along dimension d, we can write the probability of observing an input sequence conditioned on a setting of the latent variables:

$$P(Y|C,S) = \prod_{t=1}^{T} P(\mathbf{y}_t|\{c_{t,i},\mathbf{s}_{t,i}\}_{i=1,\ldots,d_c}) = \prod_{t=1}^{T} \mathcal{N}_{\mathbf{y}_t}\left(\sum_{i,j} \mathbf{w}_{ij}\, s_{t,ij}\, c_{t,i},\, \mathrm{diag}\big(\sigma^2_{y,d}\big)\right)\,, \tag{3}$$

where $\mathcal{N}_{\mathbf{x}}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is a Gaussian distribution over $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Here and in the following capital letters stand for the set of all variables with the corresponding lowercase letter (e.g., $C = \{c_{t,i}\}$ for $t = 1,\ldots,T$ and $i = 1,\ldots,d_c$).

A complete probabilistic model also requires a prior distribution on the latent variables. In this case, we might expect objects or features to appear in a visual scene independently of one another and for extended periods of time, and their appearance to vary in a continuous way. This translates into a prior distribution over identity and attribute variables as follows. Identity variables are modeled as independent, binary Markov chains:

$$P(C) = \prod_i \left(P(c_{1,i}) \prod_{t>1} P(c_{t,i}|c_{t-1,i})\right) \tag{4}$$

$$P(c_{1,i}=1) = \pi_0\,, \quad P(c_{t,i}=a|c_{t-1,i}=b) = T_{ba}\,, \quad a,b \in \{0,1\}\,. \tag{5}$$

Our intuition that objects are persistent in time is respected when the probability of remaining in the current state is larger than that of switching, i.e. when the transition probabilities T_{00} and T_{11} are larger than $1/2$. Attribute variables are modeled with a State Space Model (SSM):

$$P(S) = \prod_i \left(P(\mathbf{s}_{1,i}) \prod_{t>1} P(\mathbf{s}_{t,i}|\mathbf{s}_{t-1,i})\right) \tag{6}$$

$$P(s_{1,ij}) = \mathcal{N}_{s_{1,ij}}\big(0,\sigma_s^2\big)\,, \qquad P(\mathbf{s}_{t,i}|\mathbf{s}_{t-1,i}) = \mathcal{N}_{\mathbf{s}_{t,i}}(\boldsymbol{\Lambda}_i \mathbf{s}_{t-1,i}, \boldsymbol{\Sigma}_i)\,. \tag{7}$$

The matrices $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{ij})$ and $\boldsymbol{\Sigma}_i$ are defined to be diagonal, meaning that attributes are uncorrelated, and are related by the equation $\boldsymbol{\Sigma}_i = \mathbf{1} - \boldsymbol{\Lambda}_i^2$, so that the variance of the attribute variables is 1 in the prior [7]. This imposes an absolute scale, eliminating rescaling degeneracy. Slowly-varying variables have a positive autocorrelation, and would thus have parameters λ_{ij} between 0 and 1, with larger values corresponding to slower variables.

Ideally, the dimensionalities of the model — the numbers of objects and associated attribute variables — are also learnt from data. We use a Bayesian approach to determine these values, in which we assume an *Automatic Relevance Determination* (ARD) prior over the weights $\mathbf{W}$ [10, 11]:

$$P(W) = \prod_{ij} P(\mathbf{w}_{ij}) = \prod_{ij} \mathcal{N}\Big(\mathbf{0}, \mathrm{diag}(\gamma_{ij})^{-1}\Big)\,. \tag{8}$$

These zero-centered Gaussian priors discourage large weights. The widths of the distributions are set by the precision hyperparameters γ_{ij} which are learnt alongside the other parameters. Since the weights of redundant attribute dimensions are free to match the prior, and as this is centred on the origin, they are driven to zero. The precision hyperparameter can then diverge to infinity, effectively pruning the weight from the model. As a result, only the dimensions of the attribute manifold that are required to describe the data without overfitting remain active after learning [10, 11].

For the rest of the parameters we choose conjugate priors (see the caption of Fig. 2)[1]. The complete directed graphical model showing the dependencies between variables is depicted in Figure 2.

[1]Conjugacy means that the posterior distribution has the same functional form as the prior, resulting in

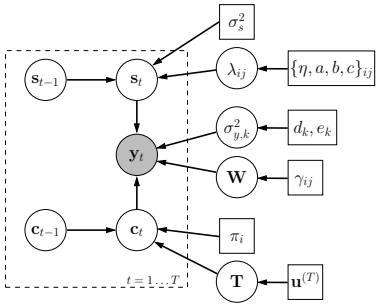

Figure 2: Directed graphical model. Circles represent random variables, and rectangles represent hyperparameters. Gray shaded elements are observed variables. The dashed plate indicates that its content is replicated T times (the length of an input sequence) in the complete model. The prior over the input noise precision $1/\sigma_{y,k}^2$ is a gamma distribution with parameters d_k, e_k, the prior over the transition matrix T_{ba} is Dirichlet with parameters $\mathbf{u}^{(T)}$, and the prior over λ_{ij} is a nonstandard distribution (due to the coupling between mean and variance of $s_{t,ij}$) in the exponential family that requires 4 hyperparameters to be specified (η, a, b, and c).

3 Learning

In the Bayesian formulation the parameters of the model are formally equivalent to latent variables, differing only in that their number does not increase with the number of data points. The goal of learning is then to infer the posterior joint distribution over variables and parameters given the data:

$$P(C, S, \Theta|Y, \Xi)\,, \tag{9}$$

where Θ indicates the ensemble of all parameters and Ξ all hyperparameters (in the following for simplicity we will omit the dependency on Ξ). Although this distribution is intractable (as in most non-trivial models), it is possible to use a *structured variational approximation* to obtain a tractable system. The idea is to introduce a new factored distribution $Q(C, S, \Theta)$ in which some dependencies between the variables are neglected, while keeping the rest of the distribution intact. Learning proceeds by functional minimization of the Kullback-Leibler divergence between the factorized and the real posterior $KL(Q(C, S, \Theta)||P(C, S, \Theta|Y))$. It can be shown that this minimization maximizes a lower bound of the marginal likelihood $P(Y)$ [11].

The key factorization underlying the Variational Bayes Expectation Maximization algorithm (VBEM) [11] is the one between latent variables and parameters

$$P(C, S, \Theta|Y) = Q(C, S)Q(\Theta)\,. \tag{10}$$

Given this basic factorization, the algorithm proceeds by iteratively inferring the latent variable distribution $Q(C, S)$ given the observations and averaging over the parameters (*E-Step*), and the parameter distribution $Q(\Theta)$ given the observations and averaging over the latents (*M-Step*). We need two further factorizations to achieve a tractable algorithm: one between the distribution of weights and input noise, and one between different identity variables at different times. Note that these approximations do not completely eliminate dependencies between the factorized variables, which still influence each other through their sufficient statistics (for example their means). In particular, the method is much less constraining than the commonly used approach of Maximum A Posteriori (MAP) estimation, where the entire posterior distribution is collapsed to a single point by taking the values of latents and parameters at the mode. Although the derivation of the learning equations requires long algebraic computations, they are derived from the VBEM setting without any noteworthy deviation, and are thus omitted here due to space limitations.

4 Results

In the following, we present the representation learned by the model when applied to natural videos sequences, and compare it to the representation found in V1. The sources of our input data are the CatCam videos [12], which consist of several minutes of recording taken from a camera mounted on the head of a cat freely exploring a novel natural environment. Since some sections of the video contain recording defects (block artifacts or pixel saturation) we selected a subset that showed minimal distortion (labeled `b0811lux` in the dataset). Observations consist of the time-series of pixel intensities in fixed windows of size 20×20 pixels. The windows were placed to cover (without overlaps) the central 200×200 region of the video. In this way we obtained a total of about 300,000

tractable integrals. Conjugate priors are intuitively equivalent to having previously observed a number of imaginary *pseudo-observations* under the model. By choosing the number of pseudo-observations of the prior we can regulate how informative the prior becomes.

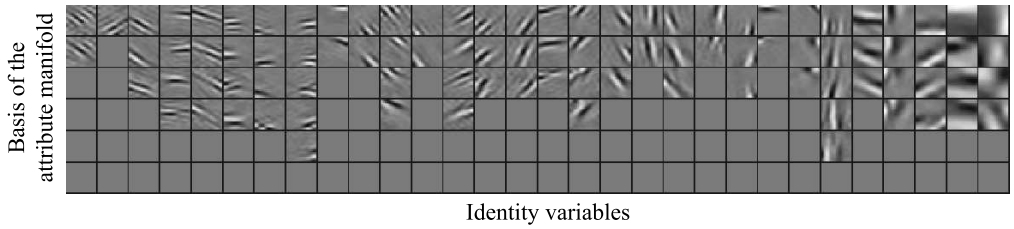

Figure 3: Basis vectors learned from natural videos. The basis vectors $\mathbf{w}_{ij}$ spanning the attribute manifold of identity i are shown in the ith column. Each weight vector is normalized to improve visibility. Gray, empty boxes indicate weights that were pruned by the algorithm. Identity variables are sorted by decreasing frequency and the basis vectors are sorted by increasing precision γ_{ij}.

frames. The input data were preprocessed by removing the mean of each frame to eliminate global changes in luminance and to compensate for the camera's global gain control mechanism. The data were then reduced in dimensionality from 400 to 81 dimensions with equalized variances, using principal component analysis.

We initialized the model with 30 identity variables and attribute manifolds of 6 dimensions and let the algorithm learn the model size by reducing the number of active attribute dimensions by ARD hyperparameter optimization. The mean of the weights $\mathbf{w}_{ij}$ was initialized at random on the unit sphere, and the priors over the parameters were chosen to be non-informative for the input noise (1 pseudo-observation, $\sigma^2_{y,k} = (0.3)^2$) and more informative for the dynamic parameters (2000 pseudo-observations), favoring persistent identity variables and slowly-varying attributes ($\langle T_{00} \rangle = 0.9, \langle T_{11} \rangle = 0.8$, $\langle \lambda_{i,1:d_s} \rangle = (0.3, \ldots, 0.1)$). We perform 500 VBEM iterations, using at each iteration a new batch of 60 sequences of 50 frames taken at random from the entire dataset. After 300 iterations we start learning the precision parameters γ_{ij}, updating their values every 20 iterations.

When presented with a new set of observations, the model infers a distribution over the values of the latent identity and attribute variables. To make comparisons with neurons in the visual cortex, we identified the mean of the distributions with the neural firing rate. This choice is necessarily arbitrary, since we lack an established theory of how to map probabilistic models to neural hardware. In particular, the brain is quite likely to represent more than a single value, carrying information about uncertainty in order to be able to weight alternative interpretations of the data. Fortunately, however, the model learns to infer the values of the latent variables with high confidence for stimuli at high contrast. Thus, the probability distributions tended to concentrate around the mean, and many different choices of neural correlates would give similar results.

Figure 3 shows the learned basis vectors. Each column displays the basis vectors of the attribute manifold corresponding to one identity variable. Since the manifold is a hyperplane, each feature is modeled by all linear combinations of the basis vectors (Fig. 4d). For every manifold, the basis vectors are shaped like Gabor wavelets with similar position, orientation, and frequency, but different phase (Fig. 5a–c). Thus every point on the manifold has a similar shape, orientation, and frequency but varies in phase (and possibly amplitude). When presented with a drifting sine grating of orientation and frequency similar to the one of the basis vectors, the probability of the feature being present $P(c_{t,i} = 1|\mathbf{y}_t)$ rapidly approaches 1 and remains constant, while the attribute variables oscillate to track the position of the sine grating on the manifold, as illustrated in Figure 4. Attribute variables thus behave similarly to simple cells in V1, in that they respond optimally to a grating-like stimulus and oscillate when its phase changes; while identity variables respond like complex cells, being insensitive to the phase of their optimal stimulus.

To explore this connection further we compared properties of simple cells RFs in V1 as reported in the physiological literature with the RFs of the attribute variables. Because the model (due to the multiplicative interaction of identity and attribute variables) and the inference process (notably because of *explaining away* effects) are nonlinear, we computed the best linear approximation to the input-output function by linear regression using colored noise input. The resulting filters were visually indistinguishable from the basis vectors in Figure 3 and are thus not shown. We then computed the parameters for the resulting RFs by fitting a Gabor function to the filters.

Figure 5 (a–c) shows the distribution of orientation, frequency, and phase for each pair of RFs belonging to the same identity variable (for instance, a variable with a 4D attribute manifold would

contribute 6 points to each graph). Thinking that attributes corresponding to the same feature might cluster in the visual cortex, we compared these plots to the data reported in [13] for pairs of simple cells recorded from the same electrode in area 17 of the cat visual cortex (Fig. 5d–f). In both cases we observed clustering of the pairs primarily in orientation and somewhat less in frequency, while no relation was apparent in phase[2]. The distribution of preferred frequencies and orientations in the attributes RFs are shown in Figure 6 (a,c). The distribution of frequencies is quite broad compared to that found in models based on sparse coding or ICA [14, 15], where frequencies tends to cluster around the highest representable value, and compares well with the width of the distribution in simple cells (Fig. 6b) [16]. The joint distribution of orientation and frequency (Fig. 6d) covers the parameter space relatively homogeneously. Note that the CatCam input data show less high-frequency power at horizontal orientations, which is reflected in the results. Figure 6e shows the joint distribution of RF width and length in normalized units (number of cycles) in our model and for simple cell RFs as reported by Ringach [17] for area V1 in the macaque. The aspect ratios are similar in both cases (again, unlike ICA results), although the model results tend to have larger RFs, possibly again due to the particular content of the video.

Initially, the algorithm learns a representation with attribute manifolds of full dimensionality. Many attribute dimensions, however, are later found to be redundant or unnecessary, and are thus eliminated by the ARD prior. At the end of learning the representation is slightly overcomplete, with 96 basis vectors representing an 81-dimensional input space, and the dimensionality of each feature manifold is typically between 2 and 4 (Fig. 7). This can be compared with the number of input dimensions that influence the response of a complex cell, as estimated by the number of statistically significant non-zero eigenvalues in the Spike-Triggered Covariance matrix. Touryan et al. [18] report a distribution of significant dimensions highly peaked at 2, with only a few complex cells influenced by 1, 3, or 4 dimensions. However, they consider as significant eigenvalues that are both larger than expected by chance *and* whose difference from the preceding eigenvalue is sufficiently small. This latter criterion is arbitrary, and so we take their results to lower bound the actual distribution. Rust et al. [19] perform a similar analysis using spatio-temporal stimuli and report 2 to 8 significant dimensions for complex cells. Since our weights are instantaneous, and to represent temporal changes would require additional dimensions, we take this to be an upper bound. Moreover, their distribution of significant dimensions is quite broad, which is consistent with our results[3].

The posterior distributions over the dynamical parameters T_{ba} and λ_{ij} confirm that the learned causes are indeed stable in time. Another set of simulations that does not make use of the temporal prior (not shown due to space constraints) results in a model that requires more basis vectors to describe the data but is in general inferior as measured by its free energy (the lower bound on the marginal likelihood) and by its match to physiological data. Temporal stability seems thus to be an important cue to recover external causes [cf. 4, 5].

[2]Phase difference is estimated here by fixing the global orientation and frequency of an identity to the one of the best fitted RF, and re-fitting only the phase parameter to the RFs of the other attribute variables.

[3]The stimuli used in [19] were random bars fixed to the preferred orientation and size of the cells. The learned RFs are thus fundamentally 2D (one spatial and one temporal dimension). Additional basis vectors that would be needed to model changes in the RF in the direction of the optimal orientation could thus be missing.

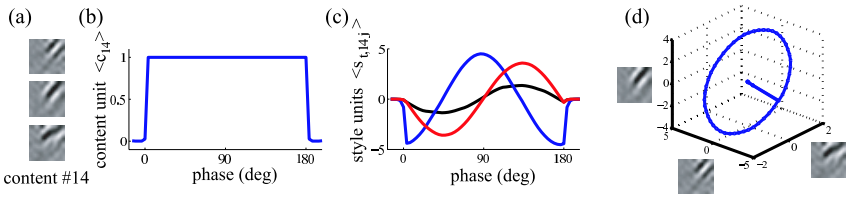

Figure 4: Interpretation as complex and simple cells. (a) Basis vectors corresponding to one of the identity variables in the learnt model (no. 14 in Fig. 3). (b–d) Response to a drifting sine grating at the preferred orientation and frequency. (b) Response of the identity variable, $\langle c_{t,14} \rangle$. (c) Response of the attribute variables, $\langle s_{t,14\,j} \rangle$. (d) Response of the attribute variables as in (c), displayed as a trajectory over the 3D attribute manifold.

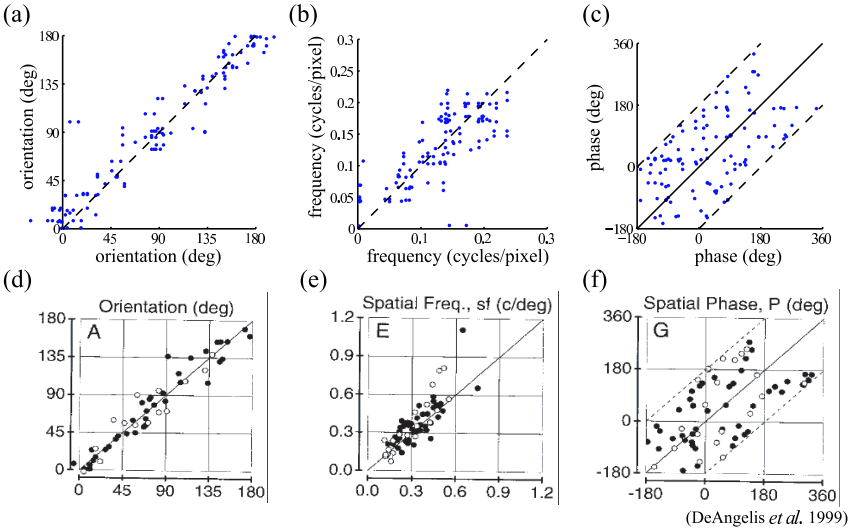

Figure 5: (a–c) Distribution of orientation, frequency, and phase for pairs of attributes belonging to the same identity variable. (d–f) Similar plots for pairs of simple cells recorded from the same electrode in area 17 of the cat visual cortex [13].

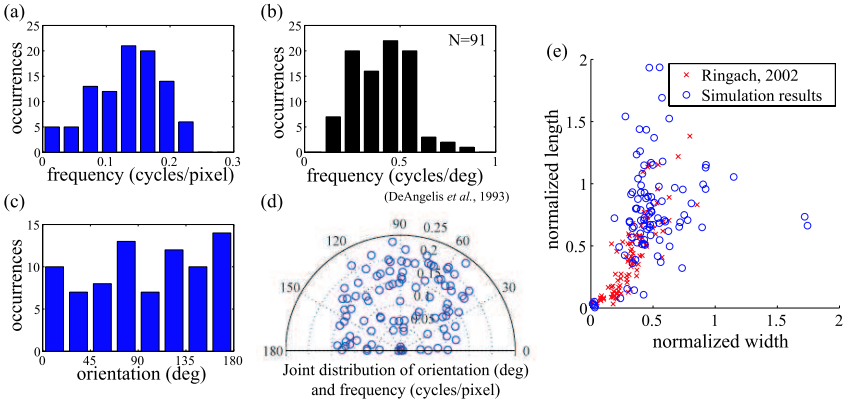

Figure 6: RF statistics. (a,c) Distribution of preferred frequency and orientation of the RFs of attribute variables in our model. (b) Distribution of preferred frequency in simple cells area 17 of the cat visual cortex [16]. (d) Joint distribution of preferred orientation and frequency in the model. (e) Comparison between the joint distribution of RF width and length (in number-of-cycles units) in our model and as reported by Ringach [17] for cells in area V1 in the macaque.

5 Conclusions

We have proposed a generative model for images based on the fundamental duality between the identity of an object or feature and its attributes. By explicitly considering the coupling between these two aspects, it is possible to extract and bind together attributes that belong to the same object, and at the same time construct an invariant representation of the object itself. We modeled identity with a set of binary variables, each coding for the presence or absence of different objects or features. Their attributes were described by a manifold parametrized by a set of attribute variables. Identity variables were assumed to be stable, and their attributes to vary smoothly in time. The interaction

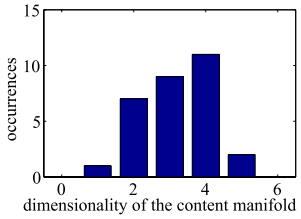

Figure 7: Distribution of the dimensionality of the attribute manifold. We considered active attribute directions with precision parameter $\langle \gamma_{ij} \rangle < 500$.

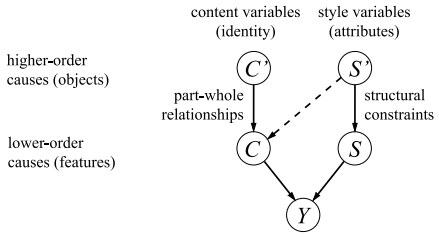

Figure 8: Schematic illustration of a two-layer identity/attributes hierarchy. The dotted line represents cases where the attributes influence the presence of objects parts. For example, in the case a face seen from behind, node, mouth, and eyes would not be visible and do not need to be generated.

between these two aspects was captured using a product nonlinearity that combines the two sets of variables to generate the input. We were also interested in determining the size of the model, i.e., the number of attribute and identity variables required to optimally describe the input data. This was achieved by performing a Bayesian analysis of the model and by defining appropriate priors over the generating weights. As a result, after convergence, only the weights needed to effectively match the data remained active and all redundant attribute directions were pruned out, avoiding overfitting the input data. The algorithm was applied to natural image sequences, in order to learn a low-level representation of visual scenes. The filters associated with the individual attribute variables were shown to have characteristics similar to those of simple cells in V1. The RF of attributes associated with the same identity variable had similar positions, orientations, and frequencies, but different phases. As a consequence, the corresponding identity variable became invariant to phase change and behaved like a complex cell.

In the standard energy model of complex cells and in several previous computational models, complex and simple cells form a hierarchy. Simple cells have the role of subunits and are considered as an intermediate step to build complex cells. Their phase-dependent information is then discarded as a first step toward the construction of an invariant representation. Here complex and simple cells do not form a hierarchy, but rather two parallel population of cells with two different functional roles: the first coding for the presence or absence of oriented features in its RF, the latter parameterizing some local parameters of the features (mainly their phase). This interpretation is reminiscent of a what/where stream segregation at the level of the primary visual cortex.

The key motivation behind the proposal of a structured model for sensory input was the potential to extract high-level causes from natural data. Figure 8 illustrates how the model might be extended in a hierarchical way to achieve this goal. In the schematic, high-level identity variables representing, for instance, entire objects generate lower-order entities, like parts of an object or image features. For example, the activity of an identity variable corresponding to a face would activate with high probability at the lower level variables coding for the presence of eyes, nose, and mouth. Similarly, high-level attributes like the size and viewpoint of the face would influence low-level attributes like the position of its individual parts. The hierarchy would be repeated down to individual image features. Such a structure would allow the visual system to benefit from the advantages of a Recognition-by-Components architecture, including the ability to reuse known parts to form novel objects, and to express the wide range of possible configurations of articulate objects [20, 21]. The implementation of such a hierarchical system to learn a representation of multiple, composite objects will be the object of future work.

References

[1] B. Olshausen and D. Field. *Nature*, 381(6583):607–609, 1996.
[2] A. Bell and T. Sejnowski. *Vision Res*, 37(23):3327–3338, 1997.
[3] R. Rao and D. Ballard. *Nat Neurosci*, 2(1):79–87, 1999.
[4] K. Körding, C. Kayser, W. Einhäuser, and P. König. *J Neurophysiol*, 91(1):206–212, 2004.
[5] P. Berkes and L. Wiskott. *J Vis*, 5(6):579–602, 2005.
[6] L. Wiskott and T. Sejnowski. *Neural Comput*, 14(4):715–770, 2002.
[7] R. Turner and M. Sahani. *Neural Comput*, 19(4):1022–1038, 2007.
[8] J. Tenenbaum and W. Freeman. *Neural Comput*, 12(6):1247–1283, 2000.
[9] D. Grimes and R. Rao. *Neural Comput*, 17(1):47–73, 2005.
[10] C. Bishop. In *ICANN 1999 Proceedings*, pp. 509–514, 1999.
[11] M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
[12] B. Betsch, W. Einhäuser, K. Körding, and P. König. *Biol Cybern*, 90:41–50, 2004.
[13] G. DeAngelis, G. Ghose, I. Ohzawa, and R. Freeman. *J Neurosci*, 19(10):4046–4064, 1999.
[14] J. van Hateren and A. van der Schaaf. 265:359–366, 1998.
[15] Y. Karklin and M. S. Lewicki. In *Adv Neural Info Processing Sys*. MIT Press, 2006.
[16] G. DeAngelis, I. Ohzawa, and R. Freeman. *J Neurophysiol*, 69(5):1091–1117, 1993.
[17] D. Ringach. *J Neurophysiol*, 88(1):455–463, 2002.

[18] J. Touryan, G. Felsen, and Y. Dan. *Neuron*, 45:781–791, 2005.
[19] N. Rust, O. Schwartz, J. Movshon, and E. Simoncelli. *Neuron*, 46:945–956, 2005.
[20] I. Biederman. *Psychol Rev*, 94(2):115–147, 1987.
[21] D. Ross and R. Zemel. *J Mach Learn Res*, 7(11):2369–2397, 2006.