Sparse Coding in Early Visual Representation: From Specific Properties to General Principles

Article *in* Neurocomputing · August 2015

DOI: 10.1016/j.neucom.2015.07.070

Neil D. B. Bruce
University of Manitoba

Shafin Rahman
North South University

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Sparse coding in early visual representation: From specific properties to general principles

Neil D.B. Bruce [*], Shafin Rahman, Diana Carrier

Department of Computer Science, University of Manitoba, 66 Chancellors Cir, Winnipeg, Manitoba, Canada R3T 2N2

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form
2 July 2015
Accepted 18 July 2015
Communicated by Huaping Liu
Available online 6 August 2015

Keywords:
Sparse coding
Feature learning
Human vision
Gabor filters

ABSTRACT

In this paper, we examine the problem of learning sparse representations of visual patterns in the context of artificial and biological vision systems. There are a myriad of strategies for sparse coding that often result in similar feature properties for the learned feature set. Typically this results in a bank of Gabor-like or edge filters that are sensitive to a range of distinct angular and radial frequencies. The theory and experimentation that is presented in this paper serves to provide a better understanding of a number of specific properties related to low-level feature learning. This includes close examination of the role of phase pairing in complex cells, the role of depth information and its relationship to variation of intensity and chroma, and deriving *hybrid* features that borrow from both analytic forms and statistical methods. Together, these specific examples provide context for more general discussion of effective strategies for feature learning. In particular, we make the case that imposing additional constraints on mechanisms for feature learning inspired by biological vision systems can be useful in guiding constrained optimization towards convergence, or specific desirable computational properties for representation of visual input in artificial vision systems.

1. Introduction

Following seminal efforts linking properties of cells in visual cortex to strategies that imply efficient encoding of image statistics [16,37,40], there has been significant interest in understanding the relationship between natural image statistics and properties of cells in visual cortex. This has included deeper examination of more specific factors in the neural encoding of image content, including spatiotemporal patterns [14], color and stereo channels [22], the role of complex cells [23] and topographical arrangement of units [25].

Recently, there has also been renewed interest in representation of image content through neural networks, with sparsity [33] and autoencoders [21,20] being important driving principles behind unsupervised approaches to feature learning. Such efforts have also shown success in producing feature hierarchies [45,15] suitable for specific applications in artificial vision.

There have been significant advances in strategies for encoding natural image statistics and useful analysis of learned feature rep3resentations. However, there remains a need for targeted efforts towards understanding the nuances of representing visual input, and the impetus for observations that appear in psychophysical, brain imaging and single-cell recording studies. More precisely, there is a trade-off between the generality of features that may be derived through constrained optimization, and the specificity of characteristics expressed in learned features. For example, one might observe that a linear transformation that provides a sparse representation of image patches, results in a scheme akin to convolution with a bank of *edge filters* [35,3]. Alternatively, one might produce a very specific non-linear model of the temporal dynamics of cells that conforms to a quantitative measure of optimality in coding efficiency or information transmission [4,19]. Deriving a putative neural representation that subscribes to these more specific criteria typically requires fitting parameters associated with a more constrained definition on the space of computational operations carried out by a neuron. Efforts that generalize more specific models, or add specificity to coarse-grained models are therefore helpful in providing additional coverage intermediate to the preceding examples.

In this paper, we examine a number of specific factors in sparse coding of image statistics with an emphasis on V1-like features. This is carried out with the goal of shifting the level of specificity one step from more general sparse coding of image patches to include additional constraints or auxiliary data. Facets of early visual encoding of image statistics that are considered in this paper are as follows:

1. *Contrast polarity*: Two hallmarks of complex cells in primary visual cortex are invariance to phase [8], and to polarity [43].

[*] Corresponding author.
E-mail address: bruce@cs.umanitoba.ca (N.D.B. Bruce).

http://dx.doi.org/10.1016/j.neucom.2015.07.070
0925-2312/© 2015 Elsevier B.V. All rights reserved.

While phase invariance has received significant consideration in the context of feature learning (e.g. convolutional networks [30,28]), phase polarity and anti-phase pairing have received relatively less attention. In this work, we discuss both shift invariance, and polarity invariance in their relation to sparsity, as well as advantages for phase variation in the context of object recognition. We also explore a feature learning strategy premised on explicit phase pairing of features with sparsity as an optimization criterion. The structure of this pairing is premised on the observed properties of complex cells in early visual cortex, with initialization of features configured to promote anti-phase pairing. This includes demonstration of feature learning wherein phase signatures of features are crucial to their diversity, and a demonstration of superior recognition performance on a standard dataset for visual feature learning (STL-10) [9]. Finally, we discuss the broader implications of this line of investigation in the context of hierarchical feature learning, and general properties of neural architectures.

2. *The role of depth*: One challenging area in examining encoding of image statistics lies in the representation of depth of content in the scene in conjunction with intensity gradients and chromatic contrast. Given that image content that allows for determination of depth via stereopsis is sparse in physical space, dense depth computation presumably relies on coordination between different areas of visual cortex allowing broader representations of image content to refine the representation among units within earlier stages in visual processing [12]. This is akin to resolving ambiguity in the aperture problem by virtue of the role of higher visual areas involved in motion processing [2]. In lieu of relying on a complex representation of this form, we instead use dense depth information directly, as a surrogate for the representation that might be afforded by involvement of later stages of visual information processing and a more sophisticated characterization of depth and scene composition. Cameras that provide direct access to depth information present the possibility to examine the relationship between intensity variation, chromatic contrast and depth information directly. The goal of this exercise is to demonstrate implications of dependencies or redundancy across different sources of input, and the broader implications for visual representation.
3. *Gabor filters and image statistics*: Gabor filters are a common alternative model of cells in early visual cortex, and exist in parallel to features derived through statistical methods in computational models of human vision, machine vision and application domains. Some of the advantages of Gabor filters lie in the precise control over spectral coverage and sampling, and avoidance of the excessive dimensionality that even relatively modest sized image patches entail in feature learning. Many statistical methods for feature learning may be relatively intractable beyond a certain size of input, implying a limit on the band of frequencies that may be represented. For this reason, we consider a hybrid representation wherein the initial representation comprised of responses from a Gabor filter bank are refined according to statistical methods to examine implications for efficient coding spanning a greater spectral range than raw image patches allow. This discussion also considers the relationship between the structure of Gabor filters, log-Gabor filters and image statistics with respect to principles for information transmission and efficiency of representation.

All of these separate lines of investigation address a common thread: Given the growth in dimensionality that accompanies consideration of larger and more complex input patterns, there are inherent limits on what can be learned in an unconstrained manner. Strategies for dealing with this complexity may come in the form of imposing additional constraints on structure for learned features based on what is already known of neural information processing in visual cortex.

In the remainder of the paper, each of these points is addressed in turn, and experimental methods and supporting results are included as a subset of each section. Section 2 considers the role of contrast polarity, and introduces a strategy for feature learning that results in anti-phase pairing of filters. This also includes analysis of their properties, and efficacy in a discriminative context. Following this, Section 3 employs combined RGB and depth channels to examine properties of image coding tied to traditional factors such as intensity and chromatic variation in the presence of depth. Section 4 discusses the suitability of Gabor filters for visual representation in light of natural image statistics, and examines hybrid analytic-statistical representations and the properties of associated features. Finally, the broader implications of this analysis are discussed in Section 5 and important results summarized in Section 6.

2. Contrast polarity

Many strategies for unsupervised feature learning have been proposed, that each differ in the precise criteria for deriving optimal features. In some cases, this entails careful selection of multiple hyperparameters or number of desired features (size of basis). That said, many strategies for early feature learning, including variants of ICA [26], vector quantization [18], or deep learning [31] result in similar features for *early* feature representation comprised of Gabor-like cells reminiscent of simple cells appearing in early visual cortical areas of primates in spite of the range of parameters chosen.

In this section we outline an optimization strategy for feature learning, in which the optimization criterion is based on explicit pairing of filters, which are combined by way of the sum of squared responses. This is a configuration that is inspired by the apparent insensitivity to contrast polarity among complex cells [43]. In particular, we demonstrate that imposing this basic structure on feature learning affords a mechanism for implicitly promoting anti-phase pairing in feature learning characteristic of models of complex cells.

The precise computational details of this strategy, and demonstration of the benefits of deriving features in this manner are explored in the sections that follow. In Section 2.1.1, we discuss the precise computational details of Paired Projection Pursuit. This is followed by a qualitative examination of learned feature properties and description of the methodology for quantitative evaluation of learned features in Section 2.2.1. Comparative results for quantitative evaluation are provided in Section 2.2.2 along with some of the broader implications of this analysis.

2.1. Methods

Quadrature pairing appears prominently in the computational vision literature from energy coding through complex cells [1], to stereo vision [34] or models of motion perception [38]. Statistical approaches have also shown success in producing units with properties similar to complex cells by way of a 2-layer sparse coding strategy [24].

The typical structure of quadrature paired filters in computational models of early visual cortex, often comes in the form of combining responses of rectified Gabor filters, with a phase offset of $\pi/2$. Desirable properties of such a configuration include phase invariance, and some limited shift invariance [34]. Shift invariance may also be achieved through careful selection in sharing of weights between putative cortical layers, as is the case for convolutional networks [30]. A secondary property of complex cells comes in the form of invariance to contrast polarity [43]. It is of interest to examine implications of this invariance from an empirical standpoint to understand the impact of encouraging subtle differences in feature learning within the underlying cortical representation, and for discriminative purposes.

In this work, we combined the assumed structure of phase pairing for complex cells, with a strategy for feature learning. This is achieved in the form of constrained optimization, subject to explicit pairing of filters, sparse coding, and constraints that force both population level sparsity and dispersal. This strategy is referred to as Paired Projection Pursuit (PPP) in the section that follows. Feature learning results presented in this section are based on the 100,000 unlabeled images from the STL-10 database [9] and correspond to 1600 features. These choices were made to produce results comparable to those presented by Coates et al. [9] and Ngiam et al. [33].

2.1.1. *Paired projection pursuit*

One strategy for feature learning that is relatively devoid of the need careful tuning of parameters, or other variables associated with optimization is the proposal of sparse filtering [33] which emphasizes population sparsity, lifetime sparsity and dispersal in its constraints. In the current work, the mechanics of optimization at the level of individual cells follows the formalism presented for sparse filtering, with the additional explicit constraint of paired filters combined through a sum of squares.

Filter pairs are given by: $f_{L_j}^{(i)} = w_{L_j}^T x^{(i)}$ and $f_{R_j}^{(i)} = w_{R_j}^T x^{(i)}$ where $x^{(i)}$ corresponds to input pattern i (a vectorized image patch), and w_{L_j} and w_{R_j} correspond to the two separate linear filter pairs that share index j.

Initial weights are assigned randomly to all L cells, and R cells are assigned weights such that $w_{R_j} = \alpha\eta - (1-\alpha)w_{L_j}$ where η is an independent set of randomly selected weights of the same size as w_{L_j}. This scheme is to promote anti-phase pairing at the initialization stage while also breaking symmetry in gradients at the optimization stage between corresponding L and R pairings.

Linear filter components are subject to an L2-norm both across the cell population, and across the sample population

$$\tilde{f}_{L_j} = f_{L_j} / \| f_{L_j} \|_2 \quad \text{and} \quad \tilde{f}_{R_j} = f_{R_j} / \| f_{R_j} \|_2 \tag{1}$$

$$\hat{f}_L^{(i)} = f_L^{(i)} / \| f_L^{(i)} \|_2 \quad \text{and} \quad \hat{f}_R^{(i)} = f_R^{(i)} / \| f_R^{(i)} \|_2 \tag{2}$$

This normalization promotes both dispersal and lifetime sparsity of cells.

Finally, complex cell structure is defined according to the squared sum of normalized L and R pairs filter components

$$\begin{aligned} f_P^{(i)} &= (\hat{f}_L^{(i)})^2 + (\hat{f}_R^{(i)})^2 \\ \tilde{f}_{P_j} &= f_{P_j} / \| f_{P_j} \|_2 \\ \hat{f}_P^{(i)} &= f_P^{(i)} / \| f_P^{(i)} \|_2 \end{aligned}$$

This also includes the same stages of L2 normalization (on the sum of squared outputs) that impact the linear stage, ensuring dispersal and lifetime sparsity among complex cells.

Finally, optimization is subject to the expression

$$\text{minimize} \quad \sum_{i=1}^{N} \| \hat{f}_P^{(i)} \|_1 \tag{3}$$

which employs the L1-norm term typical of constrained optimization for sparse coding. This optimization problem is solved by way of standard L-BFGS optimization.

Dispersal and population sparsity at the level of individual L and R cells forces both dispersion of activity across cells, and across the population ensuring that both L and R cells are *active* avoiding the trivial solution where one of L or R is inactive in pairing. The constrained minimization then forces pairing that results in dispersal and population sparsity for both individual filters and paired filters.

2.2. *Results*

To understand implications of anti-phase pairing, it is important to consider both the properties of filters that arise from optimization, and the expressiveness of such filters. The first of these is examined in Section 2.2.1 in observing typical profiles for learned paired filters, including their spatial profiles, and power and phase spectra. In Section 2.2.2 we further assess the quality of the resulting features in the context of an object classification task.

2.2.1. *Properties of paired filters*

Fig. 1 demonstrates the spatial profile of learned paired filters. Note that the learned filters carry a Gabor-like profile, with an approximately inverted profile between associated filter pairs. It is also of interest that some features consist of very low frequency gradients, which are uncommon in more typical models of sparse coding (see e.g. Fig. 3).

Fig. 2 demonstrates typical filter profiles in the spectral domain, consisting of the power spectrum (top), and phase spectrum (bottom) for a number of paired filters. Note that while there is significant similarity for within pair spectral coverage, there is significant diversity in phase spectra, and this contributes significantly to the diversity of features. This also reveals some of the subtle feature properties that may not be immediately evident in examining the spatial profile of the filters alone.

2.2.2. *Quantitative evaluation*

To assess the efficacy of paired sparse filtering in producing a useful representation for discrimination tasks, we have applied paired sparse filtering to data from the STL-10 database [9] compared against alternative representations under circumstances of few labeled examples. This data set, and methodology is designed to place greater emphasis on scalability and robustness of the learned feature set. The STL-10 dataset is a subset of CIFAR-10 and is designed such that there are very many unlabeled samples, relative to a small set of labeled examples. For this reason, this is a useful evaluation set for examining the richness of the resulting feature set, and the ability to generalize. Labels correspond to 10 classes (airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck) with 500 training images per class. An additional 100,000 unlabeled images from a broader set of classes is available for feature learning. Paired filters are learned from random patches selected from the unlabeled image set. Subsequently, the 500 class specific training images are used to train a classifier based on the underlying feature set, and an additional 800 images per class used for testing. As the feature learning is based on the unlabeled feature set, this provides a good sense of the generalization or discriminative information captured by the paired filter ensemble. Performance evaluation follows the protocol established by Coates et al. [9] and performance values are based on 1600 features across all feature types.

Classification performance for various types of features including the representation appearing in this paper (Paired Filtering) demonstrates the apparent effectiveness of sparse filtering with explicit pairing of filters in a formation akin to complex cells relative to alternative linear filtering strategies (Table 1).

There are a number of interesting observations that may be made regarding the learned paired filters. First, the degree of structure and diversity in the phase spectra is of significant interest given that phase is arguably under-represented in the

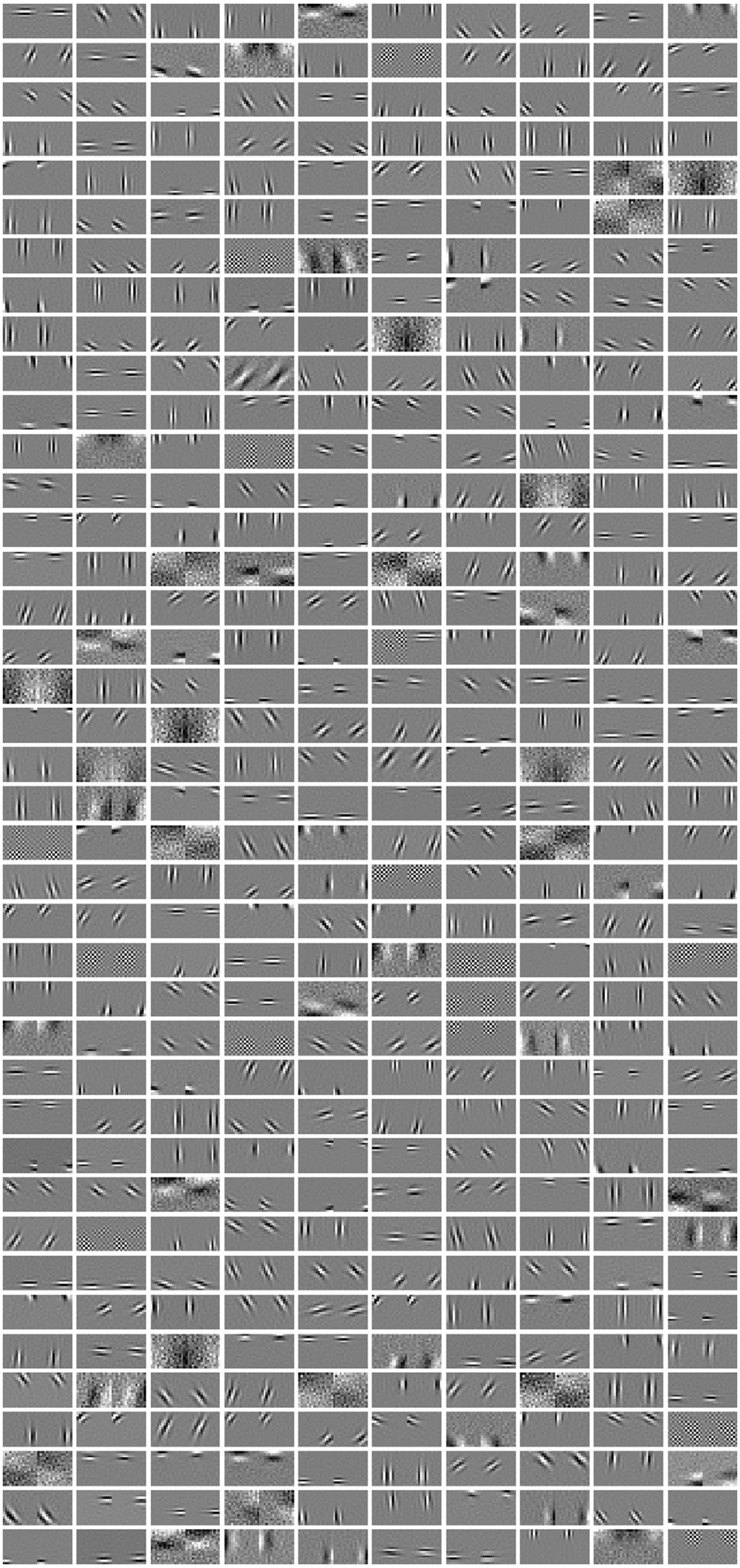

Fig. 1. Examples of the spatial profiles of coupled linear filters in paired filtering. Each set of two reveals the approximately anti-phase spatial profiles. Paired filter output is a result of the sum of squares of the two anti-phase outputs. Notice that phase pairing is preserved subject to iterative learning of edge selective filters from random and approximate anti-phase initialization.

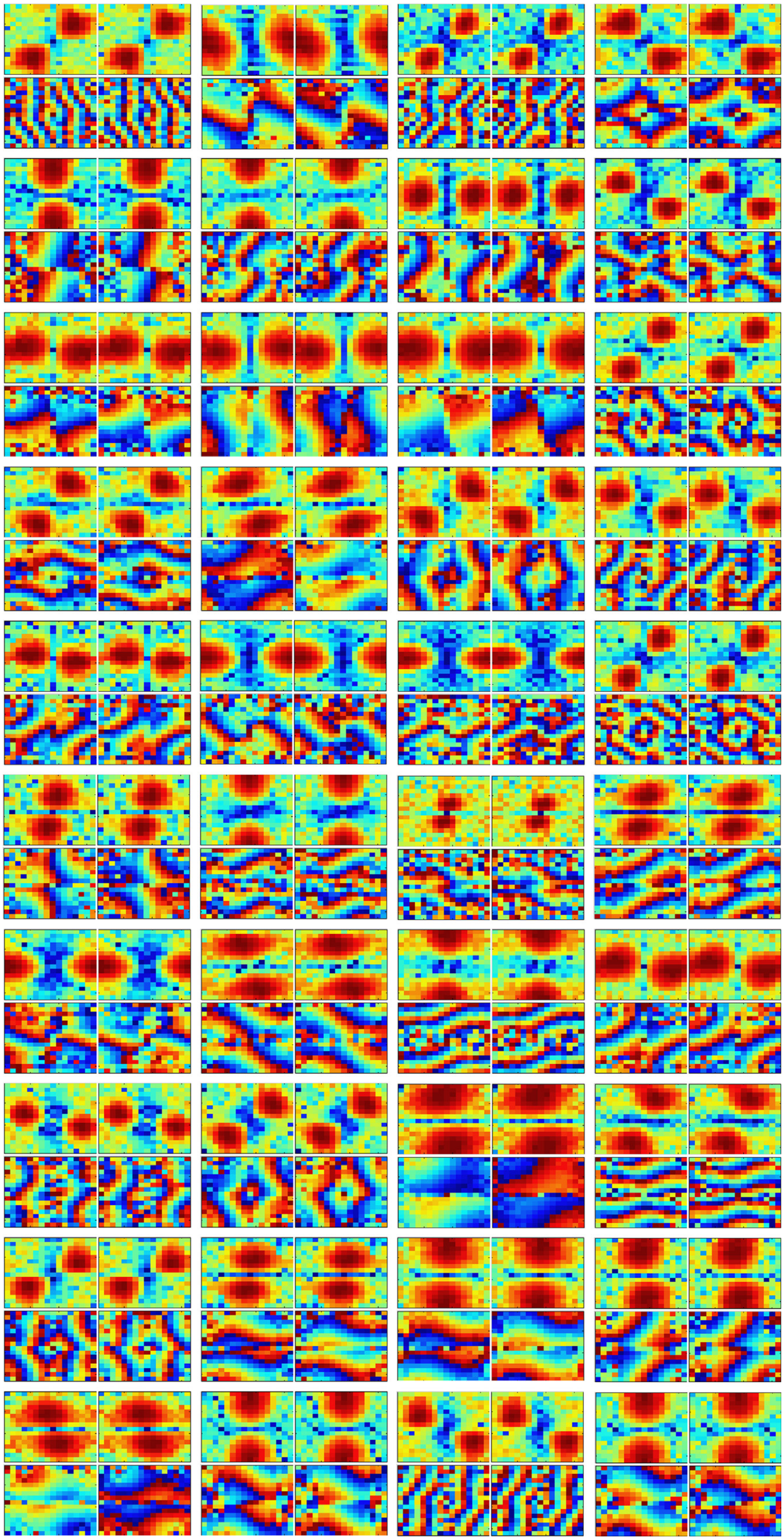

Fig. 2. Examples of filter properties in the frequency domain. For each grouping of 4: Amplitude of linear pairs (top row each group) and phase spectra (bottom pair in each group). Note that similar amplitude spectra appear, but diversity in representing image content may be attributed to the variety and complex phase spectra associated with the learned paired features. Anti-phase pairing is also evident in the coupling of pairs.

body of experiments that model early visual representation through linear models, or in assuming sparse coding. Discrimination results suggest that phase may be a relatively important consideration in visual encoding, outside of the desirable properties of phase or shift invariance. There are also additional characteristics in scene analysis and recognition tasks that may benefit from relative invariance to phase. This may include the impact of lighting, or context of an object in defining figure ground relationships and border ownership. It is also worth noting that invariance to contrast polarity presents the possibility of misclassifying instances of objects that might be successfully discriminated in the absence of such contrast polarity invariance. The results of these experiments suggest that any such deficits are outweighed by the benefits that this invariance breeds, and is a possible explanation for such computational structure featuring prominently in biological vision systems.

In addition, this formalism suggests a more general strategy for feature learning, either within a single layer of features, or as part of a hierarchy of increasingly complex features. More specifically, while imposing additional structure on a learning strategy may limit the space of representations that can be produced, it may also facilitate convergence on a set of features that would not result from optimization without a defined relationship among features, in limiting the breadth of the solution space for constrained optimization.

3. Depth

In this section we present a relatively simplistic set of experiments with largely qualitative results. These address the relationship between depth and spatial and chromatic patterns in natural images. As discussed in the introduction, the relationship between depth encoding among features (e.g. via stereopsis) and local variation of intensity and chroma presents a challenging case to study given the relative spatial sparsity of binocular depth cues, and presumed involvement of higher visual areas in this representation [12]. We therefore seek to examine such interaction at a higher level of abstraction, in the presence of absolute depth values afforded by RGB-D cameras. In this section, we describe methods associated with this analysis, and observations concerning dependencies between explicit measures of depth, and local intensity and chroma.

Table 1
STL-10 object classification performance.

Method	Accuracy
Raw Pixels [9]	31.8 ± 0.63%
ICA (Complete) [33]	48.0 ± 1.47%
K-means (Triangle) [9]	51.5 ± 1.73%
Random weight baseline [33]	50.2 ± 1.08%
Sparse filtering [33]	53.5 ± 0.53%
Paired filtering	57.9 ± 0.56%

3.1. Methods

The following analysis employs a popular method for sparse coding consisting of independent component analysis (ICA) using the extended infomax method [32]. This is a commonly used strategy for learning early visual feature representations based on large samples of input patches sampled from an ensemble of images. In this case, 600 images from the NUS-3D dataset [29] are considered, with 100 local patches sampled from each image. This dataset also includes a channel representing absolute depth from raw depth data from a Microsoft Kinect sensor.

We are interested specifically in dependencies between different input channels, and general characteristics of filters tuned to either standard RGB input, or combined RGB and depth input. A goal of this is to assess the degree of variation in the base edge filter representation in the presence of auxiliary depth information. Results corresponding to such analysis are presented in the section that follows.

3.2. Results

In this section, we show learned feature representations corresponding to either RGB input patches in isolation, or 4-D input patches consisting of RGB and depth channels. As discussed, filters are derived using extended infomax ICA. Fig. 4 reveals typical examples of spatial profiles of cells corresponding to randomly selected input patches (21 × 21) across 600 images. These carry the typical characteristics of Gabor-like filters, and color-opponency that approximately corresponds to red–green and blue–yellow opponent cells. Additional characteristics that are introduced in considering depth are revealed in the profiles presented in Fig. 5. In this case, each unit is shown as a pair corresponding to the RGB and depth components respectively. Note that the contrast of each of these is stretched to span the entire range of available RGB levels (8-bit). In terms of absolute contrast, the channels that appear to be noisier have a narrower dynamic range. There are a number of observations that can be made at an anecdotal level with respect to coupling of intensity,

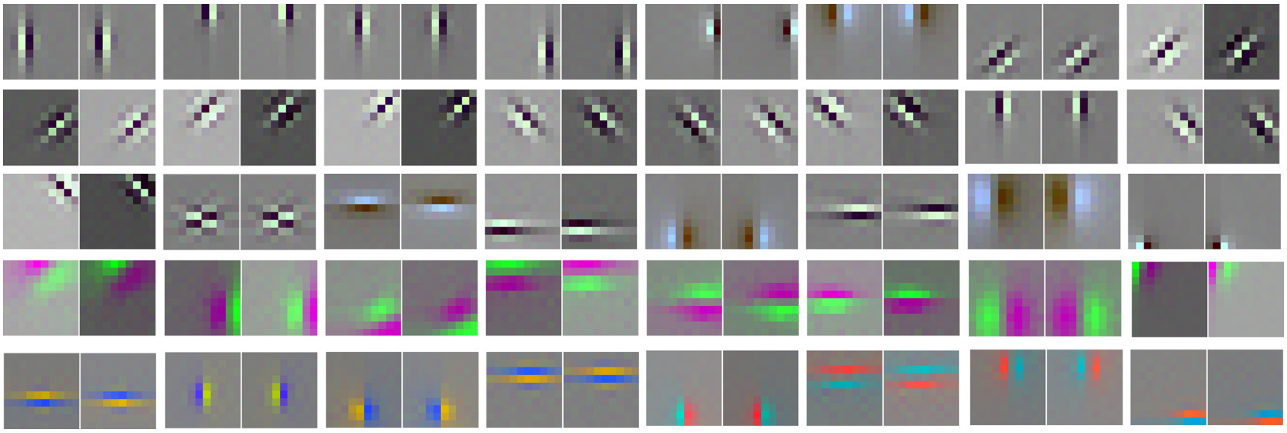

Fig. 3. Examples of paired filters derived from the STL-10 dataset. Shown are both achromatic and chromatic pairings for the small colored patches similar to output from other STL-10 evaluations, but with anti-phase characteristics in intensity and chroma. In chroma, this is reminiscent of double-opponent coupling among cells in early visual cortex. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

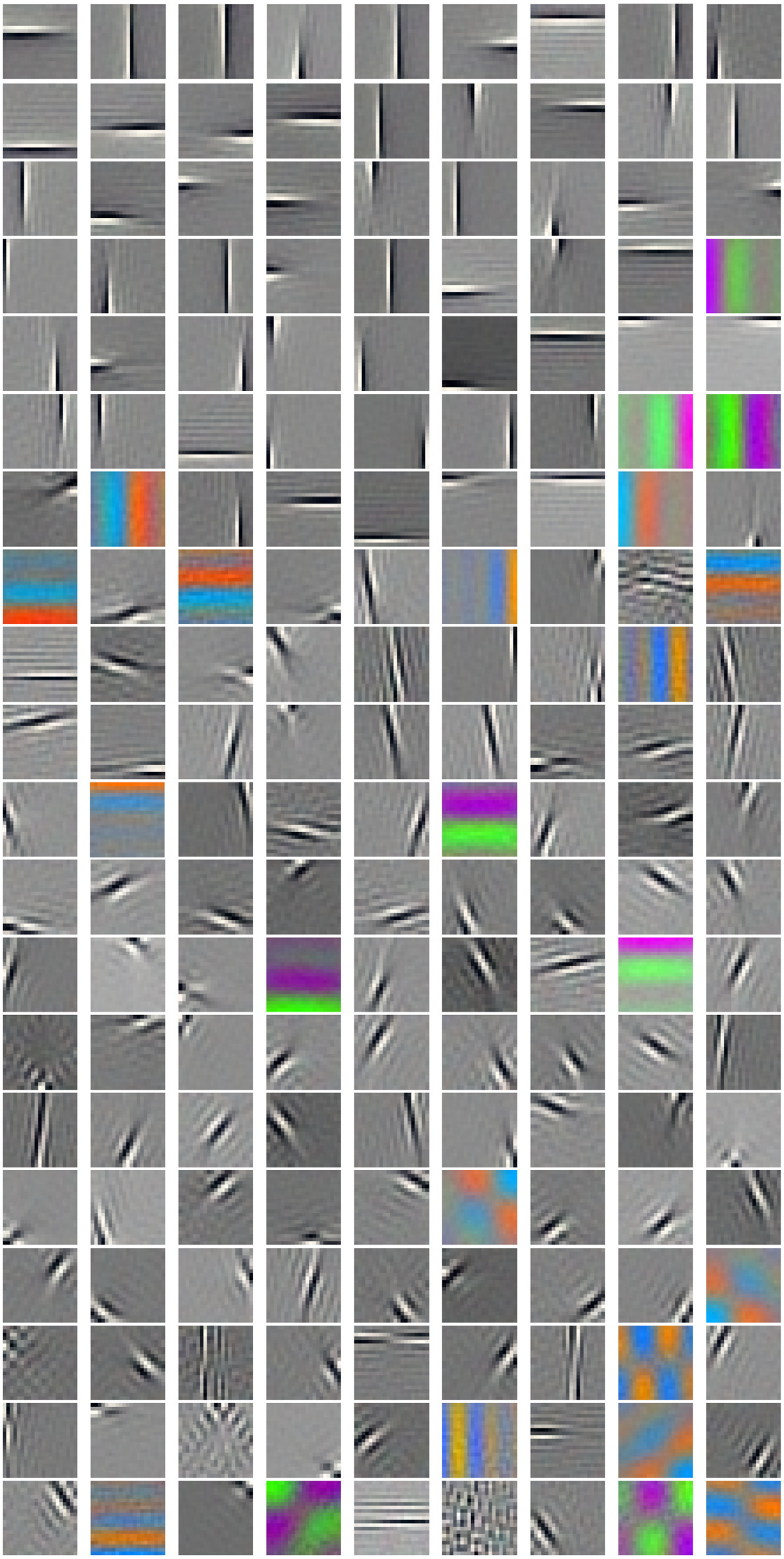

Fig. 4. Sample features derived through infomax ICA using the RGB channel only as a basis for comparing with combined RGB and depth information. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

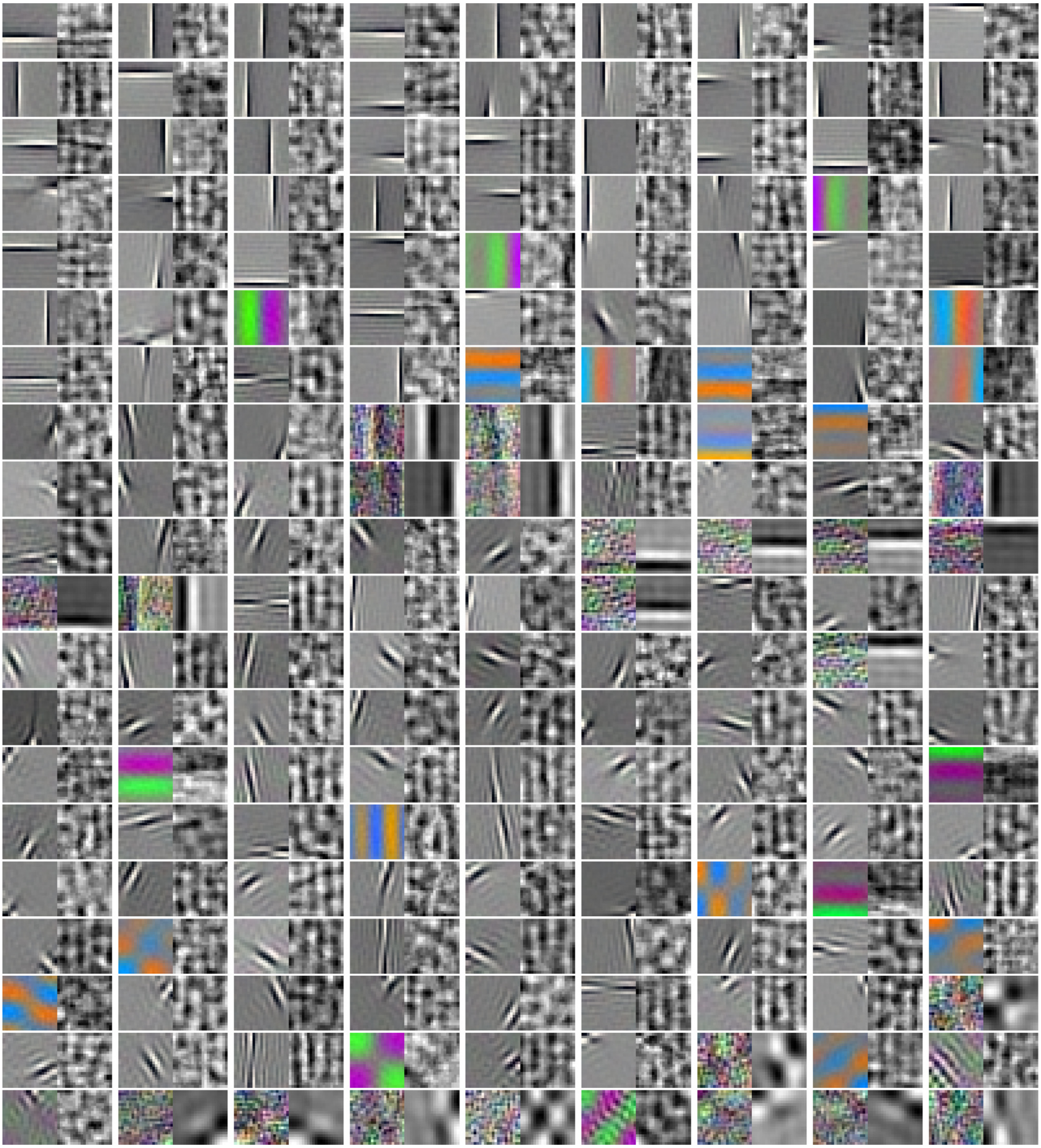

Fig. 5. Features derived through infomax ICA when both RGB and depth channels are considered together. Note that the absolute range of values is scaled to the full intensity range for greater visibility. A noisy appearance in one of the channels (color or depth) corresponds to filters with a relatively lower magnitude of weights. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

chroma and depth in considering properties of these units. First, encoding of depth information is primarily in the form of low frequency gradients. Given that this is a product of a representation that has only a weak connection to cortical representation, this does not necessarily carry any significant implications for cortical depth representation. That said, some of the additional properties of this representation are perhaps of greater interest. For example, there is relatively little coupling between intensity gradients, and depth gradients suggesting that these sources of information are relatively independent in their statistics. Secondly, there *does* appear to be some coupling between chromatic, and depth input. This is especially true for red–green channels, and for lower frequency chromatic gradients. This observation is consistent with the surprisingly large degree of pairing of chroma and disparity tuning for cells in V2 and more specifically, the existence of color coding, combined color and disparity, and disparity coding [44,41]. As a whole, this is also in line with claims of potentially distinct representations of more continuous depth gradients for surfaces and depth gradients at physical boundaries [44] respectively.

An additional sub-population of units is also depicted in Fig. 6. These units are reminiscent of cells with end-stopping characteristics [48] or capturing local curvature [13], responding to localized

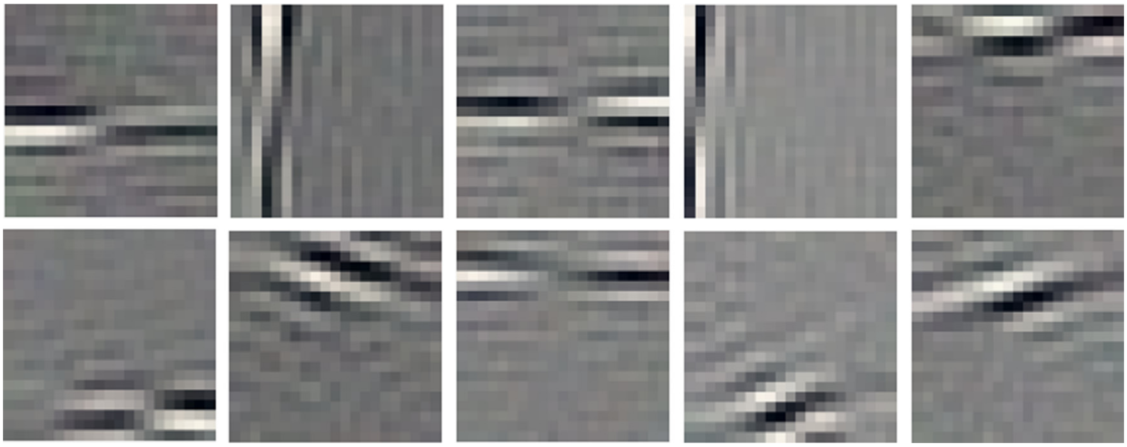

Fig. 6. A depiction of a subpopulation of units from the RGB-D condition. There is some anecdotal evidence in examining the output of a greater incidence of units of this variety. Response profiles of this type carry properties of end-stopping and encoding curvature.

gradients with a specific spatial extent. For the population associated with depth information, this type of profile appears in association with approximately 5% of units, with a figure closer to 2% for the pure RGB input. While this again is largely anecdotal, this does bring to light an additional consideration: Given input patterns with a specific scope (e.g. local RGB patches only), does a sparse encoding of such patterns change appreciably in the presence of auxiliary information? For example, additional visual encoding of input stimuli by units among higher visual areas, presence of stereo input, temporally varying input or any other sources of variation may have some impact on units tied to a representation that includes only a more limited subset of the complete set of input features. This calls for some caution in using a relatively narrow set of input patterns, purely linear filters, or limited receptive field sizes to compare hypothesized principles underlying visual encoding of input patterns. In particular, this serves as an example that feature correspondence to human data may be altered by the presence of additional input channels, or encompassing computation typical of additional functional regions in visual cortex. From a feature learning perspective, the above consideration is also important. Concatenated sets of features derived from different input channels and learned independently may differ from those that arise when all input channels are treated in concert. This is evidently due in part to correlation or redundancies expressed across different input channels, but also in a more obfuscated fashion, on the impact on path to convergence for constrained optimization in the presence or absence of different subsets of input channels.

4. Gabor filters

Gabor filters feature heavily both in computational models of vision, and in models of human vision. Features used in implemented computational vision systems range from those that are heavily driven by statistical or empirical factors (as in the prior sections), versus those that assume components with an analytic form such as Gabor filters. The latter of these categories carries some very useful properties including a greater ability to characterize behavior of individual cells, and also control over their properties as an ensemble including elements such as spectral coverage and effective range of frequencies.

In light of this, within this section we examine Gabor filters with specific emphasis on considerations that are typically more pervasive in statistically driven feature learning. This includes relating properties of Gabor filters to observed bias in natural image statistics. Moreover, we also propose a hybrid approach in which a basis of Gabor filters is modified according to statistical criteria for optimality, resulting in a feature set that draws advantages from the two distinct (analytic or statistical) strategies. These filters are referred to as Hybrid analytic-statistical filters in what follows.

4.1. Methods

Many principles that motivate putative models for the encoding of sensory signals draw their inspiration from information theoretic considerations. This includes principles such as maximizing entropy [3], sparsity [46,36], or information transmission [39]. It is therefore sensible to also consider these properties within Gabor filters, at the level of individual units, and among an ensemble of cells.

In considering individual units, Gabor filters have a profile that consists of an oriented sinusoidal pattern modulated by a local Gaussian envelope limiting the spatial extent. From an information theoretic standpoint, there is one issue that might be raised immediately concerning this structure. Natural images are often characterized as exhibiting a $1/f^{\alpha}$ dropoff as a function of frequency [16]. In employing a profile that consists of a transfer function with a Gaussian profile on a linear scale, this implies that any Gabor filter will tend to be driven more by lower frequency patterns given the higher prevalence of low frequency intensity variation in natural images. A corollary of this observation, is that the mutual information between input patterns and cell output is sub-optimal, as is the entropy of the output distribution produced by a cell. Log-Gabor filters present an alternative to Gabor filters with a Gaussian transfer function on a log-frequency scale. For log-Gabor filters, it is evident that the shape of the transfer function will help to produce a relatively even response to input patterns given the prevalence of frequencies expressed across different frequency bands in natural images. The transfer function (frequency domain) of a log-Gabor filter is given by

$$G(w)=e^{-(\log(w/w_0)^2)/2\,\log(k/w_0)^2}$$

where w_0 is the center frequency, and k/w_0 determines the filter bandwidth.

In moving from individual cells to an ensemble of cells, an additional consideration is selecting parameters that result in even spectral coverage by the complete filter set. This problem has been addressed for both standard Gabor filters [10,11], and also for log-Gabor filters [27]. With that said, the distribution of frequencies for natural image statistics is also anisotropic with respect to both angular and radial frequency (orientation and scale). For this reason, at the level of an ensemble of cells the optimality called for by information theoretic considerations also dictates uneven weighting across cells with respect to angular and radial frequencies.

One significant challenge for purely statistical methods for visual feature learning, or examining possible ties to cortical computation lies in the combinatorics associated with the dimensionality of input (e.g. patch size). Many strategies for statistical feature learning are limited in the dimensionality of input that can be considered either in complexity as a function of dimensionality, or convergence on meaningful features. Analytic models on the other hand allow for precise specification of the nature of cells and their properties including control over spectral coverage and bandwidth. In what follows, we therefore consider a statistically derived basis that assumes the response of Gabor filters for input. The goal of this is to produce an alternative means to derive a feature representation from image statistics that carries advantages of both statistical and analytic approaches to defining early visual features. This also presents an alternative vantage point for considering the relation to computation in early visual cortex in humans.

4.1.1. *Hybrid analytic-statistical filters*

There are numerous different algorithms suitable for learning sparse codes including independent component analysis (ICA) or those for which sparsity is promoted through regularization (e.g. minimizing the L1 norm). In considering algorithms for ICA, there are significant differences in the extent to which algorithms scale as a function of the dimensionality of input data. In building on filter outputs rather than raw pixels one brings the possibility of reducing the initial dimensionality of the input while allowing for lower spatial frequency patterns with a larger spatial extent to be considered.

In this line of experimentation, images from the UPenn Natural Image Database [42] were used. 400 individual images were selected across the dataset for as much variety as possible, and images converted to grayscale. For each image, 500 samples of output from the 36 log-Gabor filters were recorded for a total of 200,000 36-dimensional data vectors. This data set then formed the input to subsequent projection pursuit via ICA to minimize statistical dependency across individual filters. As mentioned earlier in this section, it is worth noting that if standard Gabor filters were used in place of log-Gabor filters, this guarantees some loss in the degree of information (or entropy) captured by the filter ensemble. For standard Gabor filters, units are driven disproportionately by lower frequency patterns within the effective frequency band for a given filter. It is not possible in practice to recover this compressive loss in resolution for higher frequencies via ICA.

Given the relatively compact form of this input, we employ ICA based on the joint approximal diagonalization (JADE) algorithm

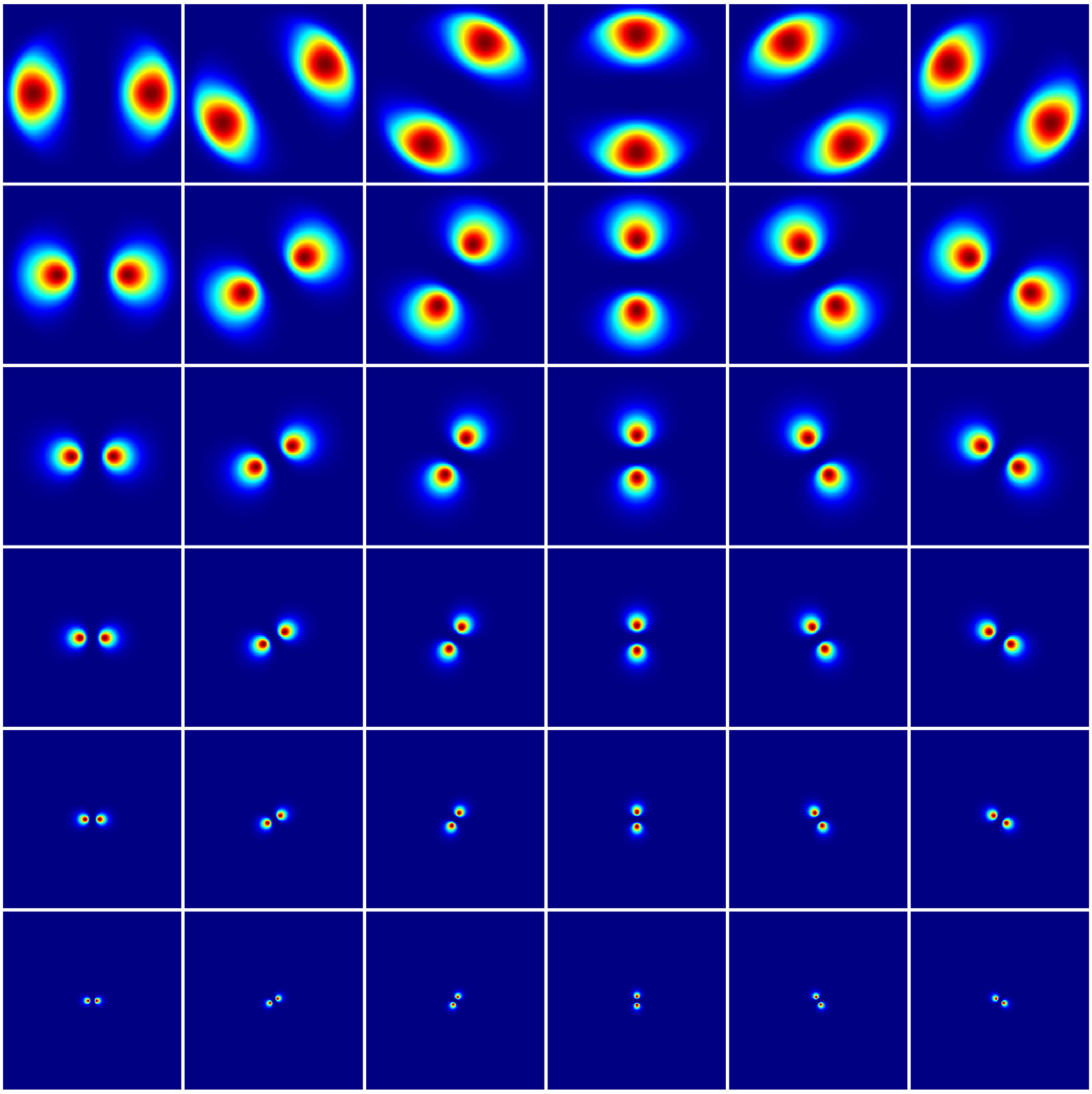

Fig. 7. Spectral profiles of the filters that comprise the base log-Gabor filter bank. These provide a base case for comparison with hybrid analytic-statistical encoding, combining log-Gabor outputs and statistical methods for feature learning.

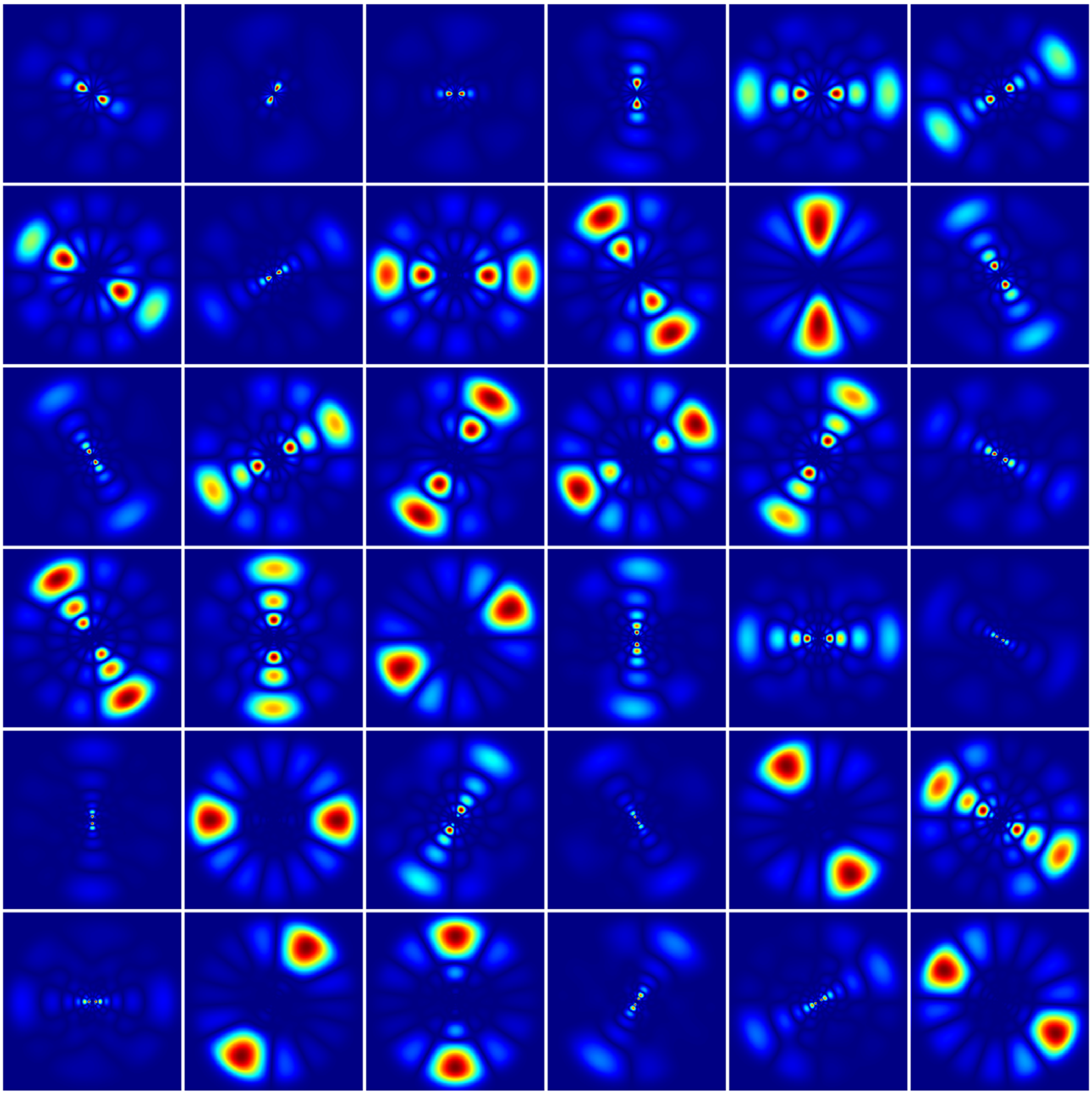

Fig. 8. Spectral profiles for hybrid filters. Note the anisotropic profiles with respect to angular frequency, varying radial frequency bandwidth, and dependencies across frequencies with the representation of higher spatial frequency content sometimes coupled to lower frequency patterns.

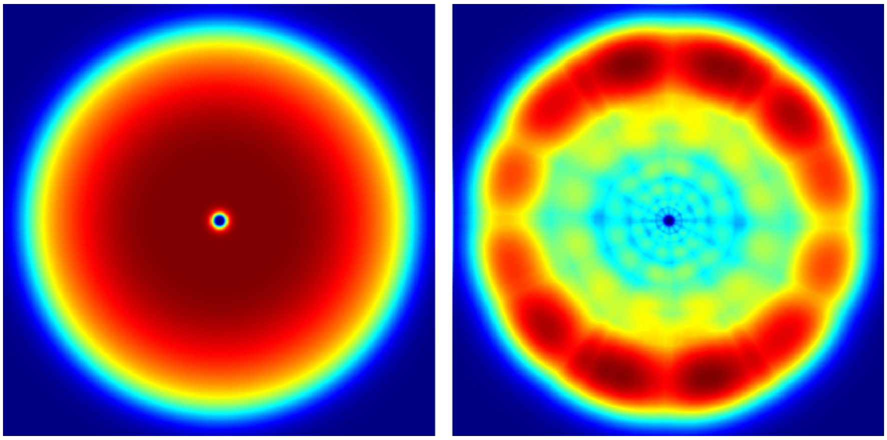

Fig. 9. Combined spectral coverage for all filters in the standard Gabor filter bank (left), and Hybrid filter bank (right). For the hybrid filters, there is a correspondence to the prevalence of spectral structure in natural images. Spectral energy within the filter bank is inversely related to prevalence of combined angular and radial frequency statistics. This implies a relatively even response across the set of filters in representing typical image patterns.

[7]. In our prior work we have observed that this ICA algorithm results in a relatively better alignment with properties of V1 cells, in particular in the direction of axes that define color opponency [5]. It is notable that this ICA algorithm also carries a relatively low ceiling on dimensionality that is feasible to consider.

4.2. Results

Fig. 7 demonstrates the magnitude spectra of even log-Gabor filters that exhibit approximately even spectral coverage from centre frequencies of 3-42.6 pixels with a factor of 1.7 separating the peak frequency of each scale. While this defines a relatively standard oriented bandpass filter bank, there are nuances associated with the transfer function of each unit, relative peak frequencies, and angular span in frequencies inspired by some of the initial *optimality* considerations discussed. In contrast, Fig. 8 demonstrates the magnitude spectra of analytic-statistical hybrid filters (referred to as hybrid filters from hereon for brevity). This latter set of spectral profiles is generated based on the linear combination of log-Gabor filters dictated by the transformation derived from ICA. In this case, dependencies across filters are minimized and there are some notable properties in the coverage of the resulting filter bank. With respect to orientation, filters sensitive to vertically and horizontally oriented intensity variation are associated with a narrower angular span in frequency. There are also instances of filters that span a broader range of frequencies (radial) within the high frequency range. Moreover, oblique orientations also elicit stronger dependencies across orientation. While filters than encode low spatial frequency intensity variation are in many cases specific to only a range of low frequencies within a specific angular frequency band, many filters sensitive to high frequencies also exhibit some sensitivity to lower frequency bands. The dependence of bandwidth on peak frequency is typically not one of the primary dimensions for analysis in human studies, however one does observe cells that are characterized by a more graded drop-off from high to low frequency patterns versus steeper drop-off in tuning for units with a lower center frequency [17]. There is also considerable evidence of recurrent interaction between relatively faster magnocellular pathways sensitive to lower spatial frequencies, and parvocellular streams sensitive to higher frequency bands [6] which may provide an alternative mechanism for diminishing redundancy within the ensemble of neurons representing observed stimuli. This observation also once again hints at the importance of considering the broader set of units that encode visual information beyond V1, and their potential implications for an optimal encoding within V1.

As discussed, one motivation for the strategy assumed in this investigation, lies in constraining the space of possibilities that may result from statistical machinery (ICA in this case) in determining an optimal feature set. There are evidently qualitative differences in the filter set resulting from log-Gabor filters as a starting point, as opposed to image patches. This supports the notion that leveraging known properties of cells in visual cortex in combination with statistical methods and constrained optimization, one may observe greater variety in the end product of available options for feature learning. This is of value both in further understanding computation in visual cortex, but also for artificial vision systems. Fig. 9 demonstrates the spectral coverage of the standard filter bank (left) and the hybrid filter bank (right) based on the sum of the magnitude spectra of filters. Total energy for the hybrid filters appears to be inversely related to the prevalence of patterns typical of natural images. In combination with the more even range of frequencies driving individual cells, this is consistent with expectations for optimality in an information theoretic sense. It is also worth noting that the specific configuration of units across frequency bands has some consistency with hypothesized configurations of units within area V2 [47]. Fig. 10 depicts configurations consistent with V2 neuron properties as proposed by Wilkinson and Wilson [47] that are consistent with some of the hybrid filter profiles in their pairing of parallel edge sensitive filters spanning different frequency bands.

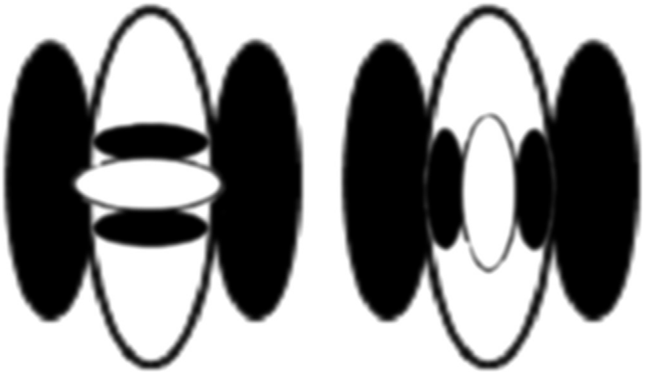

Fig. 10. Proposed structure of units in visual area V2, marked by combinations of simpler Gabor configurations [47]. There exists similarity between hybrid filter response profiles, and assumed models of response properties of neurons implicated in early visual processing.

5. Discussion

We have examined several cases of encoding of V1 type features in the presence of additional specific data or model structure beyond what is typically considered. This includes a specific focus on contrast polarity, inclusion of depth information, or using an existing analytic representation as a starting point. This yields a number of interesting observations:

1. Relatively specific nuances of computation in visual cortex may have a significant impact on the nature of representation or computation that is performed. This may imply significant implications for understanding neural coding for vision, or corresponding utility for applications. This is especially apparent in considering the results on contrast polarity that have been presented.
2. Auxiliary sources of input may alter a representation that relies on a subset of such input. The results presented on coding in the presence of depth hint at this notion. In practice, there is a large gap in information between stationary images and the sensory input that the human visual system faces. It is therefore prudent to exercise caution in drawing functional conclusions from simulation or empirical studies given that involvement of additional input, or interaction among distinct regions within visual cortex may have implications for neuronal properties within any localized region.
3. The proposed hybrid filters present an alternative strategy to producing features as a foundation for computational vision and in artificial vision systems. Moreover, this line of investigation hints at some of the potential gains that may be had in imposing additional structure on strategies for feature learning. The methods presented allow for precise control over frequency bands, and scale-space coverage while also deriving benefits of a basis or feature ensemble that subscribes to information theoretic criteria for optimal visual encoding.

Combining known properties relating to visual function in the brain with statistical methods may provide alternative means for producing new strategies for encoding visual input of value to artificial vision systems. In particular, selectively choosing where to impose structure, and where to leave slack for parameter learning may also allow for advantageous properties in the feature ensemble, and new insight on principles that drive neural encoding in humans. This also notably includes both properties of individual cells, but also interaction within and between visual areas through mechanisms such as normalization or recurrence.

6. Conclusion

The results in this paper demonstrate a variety of interesting properties that emerge in learning features for early representation of image patterns. These include focus on units with very specific properties, the impact of including auxiliary information in feature learning, and imposing established properties of neural information processing in human vision on statistical methods for deriving features.

While each of the individual lines of investigation presented are interesting in their own right, a much grander objective of this work is towards encouraging greater variety in strategies for deriving features for the computational understanding of vision or applications in artificial vision. This appeals to the value of imposing additional and carefully chosen constraints on learned features based on hypothesized properties of individual neurons or cell assemblies that have strong theoretical or experimental support. While there are many different definitions that drive feature learning based on entropy, sparsity, and higher order statistics, there is a relative paucity of efforts that enforce more specific (and possibly non-linear) structure on the resulting feature assembly, and evident value to a more targeted emphasis on the latter set of strategies based on the results presented in this paper.

Acknowledgment

This research was supported by a grant from the Office of Naval Research (N00178-14-Q-4583) and the Natural Sciences and Engineering Research Council of Canada Discovery Grants Program.

References

[1] E.H. Adelson, J.R. Bergen, Spatiotemporal energy models for the perception of motion, J. Opt. Soc. Am. A 2 (1985) 284–299.
[2] P. Bayerl, H. Neumann, Disambiguating visual motion through contextual feedback modulation, Neural Comput. 16 (2004) 2041–2066.
[3] A.J. Bell, T.J. Sejnowski, The independent components of natural scenes are edge filters, Vis. Res. 37 (1997) 3327–3338.
[4] A. Borst, F.E. Theunissen, Information theory and neural coding, Nat. Neurosci. 2 (1999) 947–957.
[5] N.D. Bruce, J.K. Tsotsos, Saliency, attention, and visual search: an information theoretic approach, J. Vis. 9 (2009) 5.
[6] J. Bullier, Integrated model of visual processing, Brain Res. Rev. 36 (2001) 96–107.
[7] J.-F. Cardoso, High-order contrasts for independent component analysis, Neural Comput. 11 (1999) 157–192.
[8] F.S. Chance, S.B. Nelson, L. Abbott, Complex cells as cortically amplified simple cells, Nat. Neurosci. 2 (1999) 277–282.
[9] A. Coates, A.Y. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: International Conference on Artificial Intelligence and Statistics, 2011, pp. 215–223.
[10] R.L. De Valois, D.G. Albrecht, L.G. Thorell, Spatial frequency selectivity of cells in macaque visual cortex, Vis. Res. 22 (1982) 545–559.
[11] R.L. De Valois, E. William Yund, N. Hepler, The orientation and direction selectivity of cells in macaque visual cortex, Vis. Res. 22 (1982) 531–544.
[12] G.C. DeAngelis, B.G. Cumming, W.T. Newsome, Cortical area mt and the perception of stereoscopic depth, Nature 394 (1998) 677–680.
[13] A. Dobbins, S.W. Zucker, M.S. Cynader, Endstopped neurons in the visual cortex as a substrate for calculating curvature, Nature 329 (1987) 438–441.
[14] M.P. Eckert, G. Buchsbaum, Efficient coding of natural time varying images in the early visual system, Philos. Trans. R. Soc. Lond. Ser. B: Biol. Sci. 339 (1993) 385–395.
[15] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11 (2010) 625–660.
[16] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells, J. Opt. Soc. Am. A 4 (1987) 2379–2394.
[17] K. Foster, J.P. Gaska, M. Nagler, D. Pollen, Spatial and temporal frequency selectivity of neurones in visual cortical areas v1 and v2 of the macaque monkey, J. Physiol. 365 (1985) 331–363.
[18] R.M. Gray, Vector quantization, ASSP Mag. IEEE 1 (1984) 4–29.
[19] N.S. Harper, D. McAlpine, Optimal neural population coding of an auditory spatial cue, Nature 430 (2004) 682–686.
[20] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[21] G.E. Hinton, R.S. Zemel, Autoencoders, minimum description length, and Helmholtz free energy, Advances in Neural Information Processing Systems, 1994, pp. 3–10.
[22] P.O. Hoyer, A. Hyvärinen, Independent component analysis applied to feature extraction from colour and stereo images, Netw.: Comput. Neural Syst. 11 (2000) 191–210.
[23] A. Hyvärinen, P. Hoyer, Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces, Neural Comput. 12 (2000) 1705–1720.
[24] A. Hyvärinen, P.O. Hoyer, A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images, Vis. Res. 41 (2001) 2413–2423.
[25] A. Hyvärinen, P.O. Hoyer, M. Inki, Topographic independent component analysis, Neural Comput. 13 (2001) 1527–1558.
[26] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, vol. 46, John Wiley & Sons, 2004.
[27] P.D. Kovesi, MATLAB and Octave functions for computer vision and image processing, in: Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia. Available from: ⟨http://www.csse.uwa.edu.au/~pk/research/matlabfns/⟩.
[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[29] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, S. Yan, Depth matters: influence of depth cues on visual saliency, in: Computer Vision–ECCV 2012, Springer, Berlin, Heidelberg, 2012, pp. 101–115.
[30] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: The Handbook of Brain Theory and Neural Networks, 1995, p. 3361.
[31] H. Lee, C. Ekanadham, A.Y. Ng, Sparse deep belief net model for visual area v2, in: Advances in Neural Information Processing Systems, 2008, pp. 873–880.
[32] T.-W. Lee, M. Girolami, T.J. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources, Neural Comput. 11 (1999) 417–441.
[33] J. Ngiam, Z. Chen, S.A. Bhaskar, P.W. Koh, A.Y. Ng, Sparse filtering, in: Advances in Neural Information Processing Systems, 2011, pp. 1125–1133.
[34] I. Ohzawa, G.C. Deangelis, R.D. Freeman, Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors, Science 249 (1990) 1037–1041.
[35] B.A. Olshausen, D.J. Field, Natural image statistics and efficient coding*, Netw.: Comput. Neural Syst. 7 (1996) 333–339.
[36] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1? Vis. Res. 37 (1997) 3311–3325.
[37] B.A. Olshausen, et al., Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[38] N. Qian, Computing stereo disparity and motion with known binocular cell properties, Neural Comput. 6 (1994) 390–404.
[39] T.O. Sharpee, H. Sugihara, A.V. Kurgansky, S.P. Rebrik, M.P. Stryker, K.D. Miller, Adaptive filtering enhances information transmission in visual cortex, Nature 439 (2006) 936–942.
[40] E.P. Simoncelli, B.A. Olshausen, Natural image statistics and neural representation, Annu. Rev. Neurosci. 24 (2001) 1193–1216.
[41] L.C. Sincich, J.C. Horton, The circuitry of v1 and v2: integration of color, form, and motion, Annu. Rev. Neurosci. 28 (2005) 303–326.
[42] G. Tkačik, P. Garrigan, C. Ratliff, G. Milčinski, J.M. Klein, L.H. Seyfarth, P. Sterling, D.H. Brainard, V. Balasubramanian, Natural images from the birthplace of the human eye, PLoS One 6 (2011) e20409.
[43] J. Touryan, G. Felsen, Y. Dan, Spatial structure of complex cell receptive fields measured with natural images, Neuron 45 (2005) 781–791.
[44] D.Y. Ts'o, A.W. Roe, C.D. Gilbert, A hierarchy of the functional organization for color, form and disparity in primate visual area v2, Vis. Res. 41 (2001) 1333–1349.
[45] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[46] W.E. Vinje, J.L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision, Science 287 (2000) 1273–1276.
[47] F. Wilkinson, H.R. Wilson, Global processes in form vision and their relationship to spatial attention, in: Vision and Attention, 2001, Springer, New York pp. 63–81.
[48] A. Yazdanbakhsh, M.S. Livingstone, End stopping in v1 is sensitive to contrast, Nat. Neurosci. 9 (2006) 697–702.

Neil D.B. Bruce is currently an Assistant Professor in the Department of Computer Science at the University of Manitoba, Canada. He has been a Postdoctoral Fellow at the Centre for Vision Research at York University, and at INRIA Sophia Antipolis, in France. He holds a Ph.D. in computer science (York University, 2008), M.A.Sc. in system design engineering (University of Waterloo, 2003), and an Honours B.Sc. with Majors in computer science and mathematics (University of Guelph, 2001). His research interests include human and computer vision, computational neuroscience, visual attention, HCI and visualization.

Shafin Rahman is currently pursuing his M.Sc. degree in the Department of Computer Science at the University of Manitoba. Previously, he received his B.Sc. degree from the Department of Computer Science and Engineering at Bangladesh University of Engineering and Technology, in 2011. His current research interests include visual attention, computer vision and computational neuroscience as points of emphasis.

Diana Carrier graduated with a BCS Honours from the University of Manitoba, in 2014. Since then she has been a freelance software developer, a video game developer, and is currently working as a front-end web developer at a small startup company. Her research interests include artificial intelligence, computer vision, and robotics.