Hindawi Publishing Corporation
ISRN Machine Vision
Volume 2013, Article ID 138057, 10 pages
http://dx.doi.org/10.1155/2013/138057

Research Article

Active Object Recognition with a Space-Variant Retina

Christopher Kanan

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA

Correspondence should be addressed to Christopher Kanan; ckanan@caltech.edu

Received 6 October 2013; Accepted 24 October 2013

Academic Editors: H. Erdogan, O. Ghita, D. Hernandez, A. Nikolaidis, and J. P. Siebert

When independent component analysis (ICA) is applied to color natural images, the representation it learns has spatiochromatic properties similar to the responses of neurons in primary visual cortex. Existing models of ICA have only been applied to pixel patches. This does not take into account the space-variant nature of human vision. To address this, we use the space-variant log-polar transformation to acquire samples from color natural images, and then we apply ICA to the acquired samples. We analyze the spatiochromatic properties of the learned ICA filters. Qualitatively, the model matches the receptive field properties of neurons in primary visual cortex, including exhibiting the same opponent-color structure and a higher density of receptive fields in the foveal region compared to the periphery. We also adopt the "self-taught learning" paradigm from machine learning to assess the model's efficacy at active object and face classification, and the model is competitive with the best approaches in computer vision.

1. Introduction

In humans and other simian primates, central foveal vision has an exceedingly high spatial resolution (acuity) compared to the periphery. This space-variant scheme enables a large field of view, while allowing visual processing to be efficient. The human retina contains about six million cone photoreceptors but sends only about one million axons to the brain [1]. By employing a space variant representation, the retina is able to greatly reduce the dimensionality of the visual input, with eye movements allowing fine details to be resolved if necessary. The retina's space-variant representation is reflected in early visual cortex's retinotopic map. About half of primary visual cortex (V1) is devoted solely to processing the central 15 degrees of visual angle [2, 3]. This enormous overrepresentation of the fovea in V1 is known as cortical magnification [4].

Neurons in V1 have localized an orientation sensitive receptive fields (RFs). V1-like RFs can be algorithmically learned using independent component analysis (ICA) [5–8]. ICA finds a linear transformation that makes the outputs as statistically independent as possible [5], and when ICA is applied to achromatic natural image patches, it produces basis functions that have properties similar to neurons in V1. Moreover, when ICA is applied to color image patches, it produces RFs with V1-like opponent-color characteristics, with the majority of the RFs exhibiting either dark-light opponency, blue-yellow opponency, or red-green opponency [6–8].

Filters learned from unlabeled natural images using ICA and other unsupervised learning algorithms can be used as a replacement for hand-engineered features in computer vision tasks such as object recognition. This is known as self-taught learning when the natural images that the filters are learned from are distinct from the dataset used for evaluating their efficacy [9]. Methods using self-taught learning have achieved state-of-the-art accuracy on many datasets (e.g., [9–12]).

Previous work has focused on applying ICA to square image patches of uniform resolution. Here, we use ICA to learn filters from space-variant image samples acquired using simulated fixations. We analyze the properties of the learned filters, and we adopt the self-taught learning paradigm to assess their efficacy when used for object recognition. We review related models in the discussion.

2. Space-Variant Model of Early Vision

Our model consists of a series of subcomponents, which are depicted in Figure 1. We first describe the space-variant

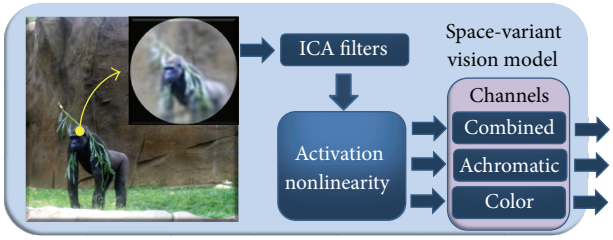


Figure 1: A cartoon schematic of our space-variant visual feature model. The model samples a region of the image in a space-variant manner, and this representation is fed into a bank of ICA filters. A learned activation nonlinearity modulates the activity of the filters. Finally the filters are subdivided into multiple channels, which project to an object recognition model.

representation we use, and then how we learn the space-variant ICA filters.

2.1. Cone-Like Representation. When our model of space-variant vision fixates a region of an image, it converts the image from standard RGB (sRGB) colorspace to LMS colorspace [13], which more closely resembles the responses of the long, medium, and short wavelength cone photoreceptors in the human retina. Subsequently, we apply a cone-like nonlinearity to the LMS pixels. This preprocessing helps the model cope with large-scale changes in brightness [6, 10, 14], and it is related to gamma correction [15]. The formulation we use is given by

$$F_{\text{cone}}(z) = \max\left(\frac{\log(\gamma+1) - \log(F_{\text{LMS}}(z)+\gamma)}{(\log(\gamma+1) - \log(\gamma))(\gamma-1)} + 1, 0\right), \tag{1}$$

where γ controls the normalization strength. In our experiments $\gamma = 0.01$. The nonlinearity is shown in Figure 2.

2.2. A Space-Variant Representation. We use Bolduc and Levine's [16, 17] log-polar model of space-variant vision. Log-polar representations have been used to model both cortical magnification [18] and the retina [17]. Unlike other log-polar models (e.g., [18]), Bolduc and Levine's model does not have a foveal blind spot. Moreover, it incorporates overlapping RFs, which produces images of superior quality [19], and the RFs in the fovea are of uniform size. Each unit in this representation can be interpreted as a bipolar cell, which pools pixels in a cone-like space. The mammalian retina contains at least 10 distinct bipolar cell types [20], and most of them are diffuse; that is, they pool the responses of multiple cones.

We briefly describe Bolduc and Levine's [16, 17] model. The full derivation is given in [17]. A log-polar mapping is governed by equations for the eccentricity of each ring of RFs from the center of the visual field and the spacing between individual RFs, that is, the grid rays. Bolduc and Levine's model uses separate equations for the foveal region and the

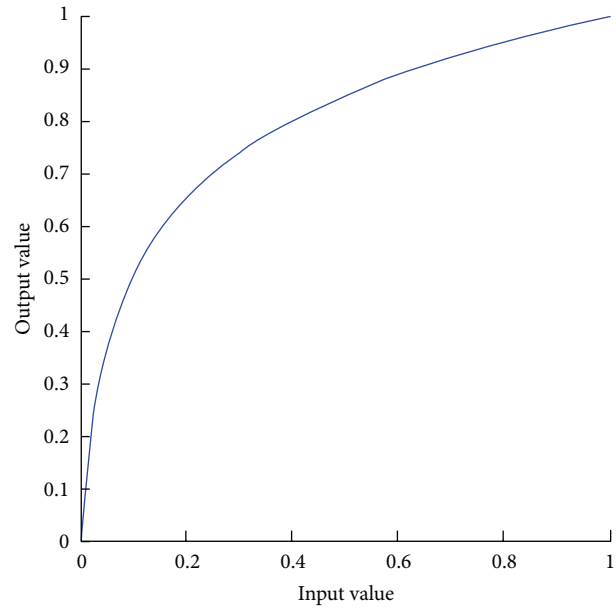


Figure 2: The cone nonlinearity plotted with $\gamma = 0.01$.

periphery. The ray spacing angle formula outside of the fovea is given by

$$\theta(\alpha, \omega) = 2\pi\left(\text{round}\left(2\pi(\arccos(z))^{-1}\right)\right)^{-1}, \tag{2}$$

where α is the ratio of the RF size to eccentricity, $z = 1 - 0.5\alpha^2(1-\omega)^2$, and ω is the amount of RF overlap. The use of the round function ensures an integer number of grid rays. The eccentricity ε of each peripheral ring $s \in \{0, 1, \ldots, L\}$ is given by

$$\varepsilon(\alpha, \omega, s, f) = \left(-\frac{\alpha(1-2\omega)+d}{\alpha-2}\right)^s f, \tag{3}$$

where f is the radius of the fovea, $d = \sqrt{4+\alpha^2((1-2\omega)^2-1)}$, and L is the total number of peripheral layers. The radius of peripheral RFs at eccentricity ε is given by

$$r(\alpha, \omega, s, f) = \frac{\alpha}{2}\varepsilon(\alpha, \omega, s, f). \tag{4}$$

Foveal RFs are all constrained to be the same size as the inner most ring of the periphery; that is,

$$r_{\text{fovea}}(\alpha, \omega, 0, f) = \frac{\alpha}{2} f. \tag{5}$$

Constraining foveal RFs to be the same size means that there are a decreasing number of RFs in each foveal ring as the center of the retina is approached, in contrast to peripheral rings, which each contains the same number of RFs. The eccentricity of foveal ring κ is given by

$$\varepsilon_{\text{fovea}}(\kappa) = (\varepsilon(\alpha, \omega, 1, f) - f)\kappa. \tag{6}$$

The ray spacing angle formula between RFs in foveal ring κ is given by $\theta_{\text{fovea}}(\kappa) = \theta(\alpha f / \varepsilon_{\text{fovea}}(\kappa), \omega)$.

We use normalized circular RFs for the retina, which act as linear filters. A retina RF j at location (x_j, y_j) with radius r_j is defined as follows:

$$H_j\left(x', y'\right) = \frac{h_j\left(x', y'\right)}{\int\int_{-\infty}^{\infty} h_j\left(x, y\right) dx\, dy}, \tag{7}$$

where

$$h_j\left(x, y\right) = \sqrt{\max\left(r_j^2 - \left(x - x_j\right)^2 - \left(y - y_j\right)^2, 0\right)}. \tag{8}$$

The retina we used in experiments is shown in Figure 3. We set $\alpha = 0.2$ and used a RF overlap of 50%, that is, $\omega = 0.5$, which are biologically plausible values [17]. We set the fovea's radius to 7 pixels and we used 15 peripheral layers. These settings yield a retina with a radius of 35 pixels that reduces the dimensionality from 3749 pixels to 1304 retina RFs (296 in the fovea, 1008 in the periphery).

Our images are resized, so that their shortest side is 160 pixels, with the other side rescaled to preserve the image's aspect ratio. If this canonical size is altered, then the fovea's radius should be changed as well. This change will not alter the total number of RFs.

To use our retina with color images, we sample each color channel independently. After sampling a region of an image with the retina, we subtract each color channel's mean and then divide the by the vector's Euclidean norm. Sampling the image with our retina yields $\mathbf{r}$, a 3912-dimensional unit length vector of retinal fixation features (1304 dimensions per color channel).

2.3. Learning a Space-Variant Model of V1. We learned ICA filters from 584 images from the McGill color image dataset [21]. Each image is randomly fixated 200 times, with each fixation location chosen with uniform probability. The images are not padded, and fixations are constrained to be within images.

Prior to ICA, we first reduce the dimensionality of the fixation data from 3912 dimensions to 1000 dimensions using principal component analysis (PCA), which preserves more than 99.4% of the variance. We then learn ICA filters using the Efficient Fast ICA algorithm [22]. We denote the learned ICA filters using the matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]^T$, with the rows of $\mathbf{U}$ containing the ICA filters. The learned ICA basis functions are shown in Figure 4.

2.4. ICA Filter Activation Function. For object recognition, the discriminative power of ICA filters can be increased by taking the absolute value of the responses and then applying the cumulative distribution function (CDF) of generalized Gaussian distributions to the ICA filter responses [10, 12]. We pursue a similar approach, but we use the CDF of the exponential distribution instead. The CDF of the exponential distribution is computationally more efficient to calculate, and it is easier to fit since it has only one parameter. For each ICA filter $\mathbf{u}_i^T$ (the ith row of $\mathbf{U}$), we fit an exponential distribution's rate parameter λ_i to the absolute value of the filter responses to the fixations extracted from the McGill dataset [21]. Fitting was done using MATLAB's "fitdist" function. The final ICA activation nonlinearity is given for each ICA filter by

$$g_i = 1 - \exp\left(-\lambda_i \left|\mathbf{u}_i^T \mathbf{r}\right|\right), \tag{9}$$

where g_i is the ith element of the vector $\mathbf{g}$.

3. Analysis of Learned Receptive Fields

We fit Gabor functions to the ICA filters to analyze their properties. Gabor functions are localized and oriented band-pass filters given by the product of a sinusoid and a Gaussian envelope [23], and they are a common model for V1 RFs. To do this, we represent the ICA filters in Cartesian space and convert them to grayscale using the Decolorize algorithm [24], which preserves chromatic contrast. In general, Gabor functions were a good fit to the learned filters, with a median R^2 value of 0.81; however, 70 of the 1000 fits were poor ($R^2 < 0.5$) and we did not further analyze their spatial properties.

Figure 6 shows a scatter plot of the peak frequencies and orientations of the Gabor filter fits, revealing that they cover a wide spectrum of orientations and frequencies. While the orientations are relatively evenly covered irrespective of the filter's location, most of the filters sensitive to higher spatial frequencies are located in the foveal region. We also found that there was a greater number of ICA filters in the foveal region compared to the periphery (see Figure 5), with the RFs getting progressively larger outside of the fovea (see Figure 7).

4. Image Classification with Gnostic Fields

4.1. Gnostic Fields. A gnostic field is a brain-inspired object classification model [26], based on the ideas of the neuroscientist Jerzy Konorski [27]. An overview of the model is given in Figure 8. Gnostic fields have been shown to achieve state-of-the-art accuracy at image classification using color SIFT features. We use a gnostic field with our space-variant ICA features. We briefly provide the details necessary to implement gnostic fields here, but see [26] for additional information.

A gnostic field's input is segregated into one or more channels [26], which helps it cope with irrelevant features. We used three channels: (1) all 1000 ICA filters, (2) the 744 achromatic ICA filters, and (3) the 256 color ICA filters. We let $\mathbf{g}_c$ be a vector that denotes features from channel c, which is a subset of the dimensions of $\mathbf{g}$.

Whitened PCA (WPCA) [5] is applied to each channel independently to learn a decorrelating transformation that normalizes that channel's variance; that is,

$$\mathbf{W}_c = \left(\mathbf{D}_c + \xi \mathbf{I}\right)^{-1/2} \mathbf{E}_c^T, \tag{10}$$

where $\mathbf{I}$ is the identity matrix, the columns of the matrix $\mathbf{E}_c$ contain the eigenvectors of the channel's covariance matrix calculated using the fixations from the McGill dataset, $\mathbf{D}_c$ is

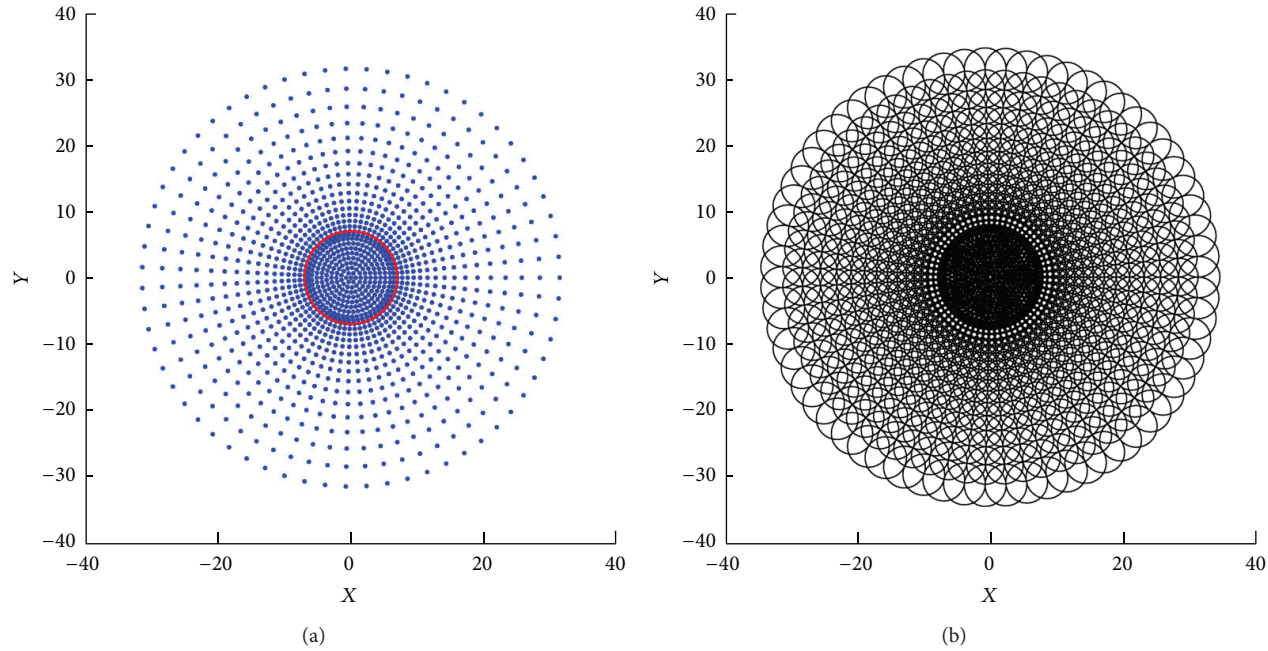


Figure 3: (a) The center of each retina RF, with a red circle drawn around the fovea. (b) A depiction of the retina's RF sizes. Each RF operated on between 1 (fovea) and 32 pixels.

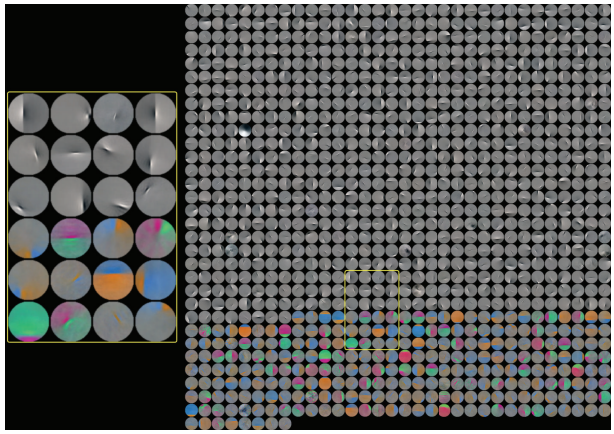


Figure 4: The 1000 V1-like basis functions learned using ICA. Two distinct populations of 256 chromatic and 744 achromatic filters were learned. The learned features are Gabor like, and they share the dark-light, red-green, and blue-yellow opponency characteristics of V1 neurons [6–8].

the diagonal matrix of eigenvalues, and ξ is a regularization parameter, with $\xi = 0.01$ in experiments. The output is then made unit length, which allows measurements of similarity using dot products [28]. At each time step t, this yields whitened and normalized vector $\mathbf{f}_{c,t}$, that is,

$$\mathbf{f}_{c,t} = \frac{\mathbf{W}_c \mathbf{g}_{c,t}}{\left\| \mathbf{W}_c \mathbf{g}_{c,t} \right\|}. \tag{11}$$

Let $\mathbf{x}_t = [x_t \;\; y_t \;\; 1]^T$ denote the (x_t, y_t) location of the fixation, with the coordinates normalized by the image size to be between -1 and 1. To incorporate this location information into the unit length features, we normalize $\mathbf{x}_t$ to unit length and weight it by δ; that is, $\widehat{\mathbf{x}}_t = \delta(\mathbf{x}_t / \|\mathbf{x}_t\|)$, with δ controlling the strength of the fixation location's influence. The $\widehat{\mathbf{x}}_t$ vector is concatenated to $\mathbf{f}_{c,t}$, which is then renormalized to unit length, yielding $\widehat{\mathbf{f}}_{c,t}$. In our experiments, $\delta = 0.01$.

A gnostic field is made up of multiple gnostic sets, with one set per category. Each gnostic set contains neurons that assess how similar the fixation features are to previous observations from the category. For each gnostic set, the activity of a neuron j for category k and from channel c is given by the dot product

$$a_{c,k,j}\left(\mathbf{f}_{c,t}\right) = \mathbf{v}_{c,k,j} \cdot \widehat{\mathbf{f}}_{c,t}, \tag{12}$$

where $\mathbf{v}_{c,k,j}$ is the neuron's weight vector.

The output of the gnostic set for category k and channel c is given by the most active neuron:

$$\varphi_{c,k}\left(\mathbf{f}_{c,t}\right) = \max_j a_{c,k,j}\left(\mathbf{f}_{c,t}\right). \tag{13}$$

Max pooling enables the gnostic set to vigorously respond to features matching the category's training data.

Spherical k-means [29] is an unsupervised clustering algorithm for unit length data that is used to learn the localized $\mathbf{v}_{c,k,j}$ units for each of the K gnostic sets and C channels [26]. The number of units in a gnostic set depends on the number of fixations from that category, albeit with fewer units being recruited as the number of fixations increases. To implement this, the number of $\mathbf{v}_{c,k,j}$ units learned for a category k from channel c is given by

$$m(k,c) = \min\left(\left\lceil b\left(\log\left(n_{k,c}\right) + 1\right)^2 \right\rceil, n_{k,c}\right), \tag{14}$$

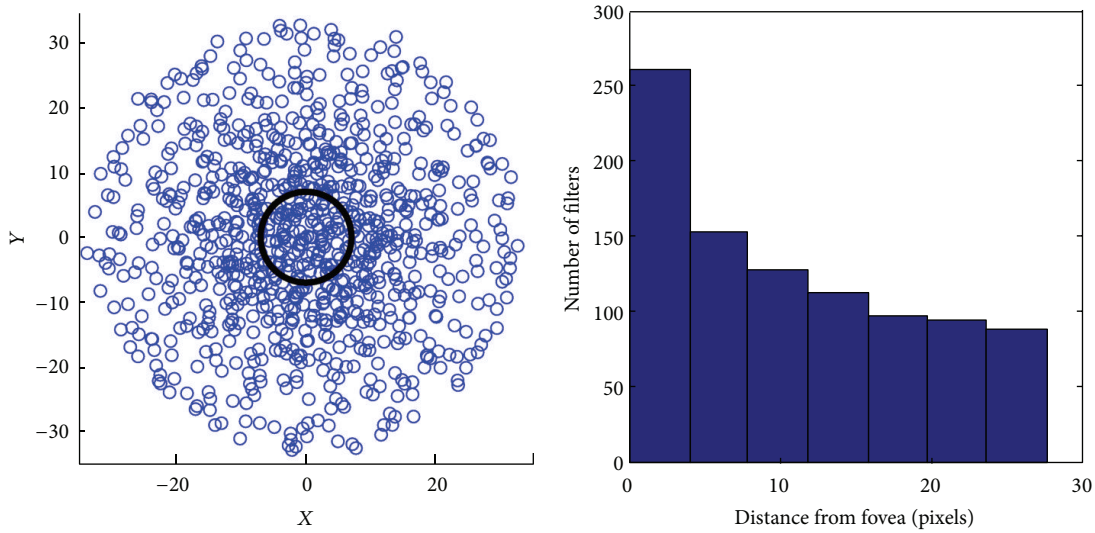


Figure 5: (a) The center location of the Gabor functions fit to the ICA filters. The fovea is contained within the black circle. (b) A histogram of the Gabor function centers as a function of the distance from the fovea, which reveals that the number of filters is decreasing farther from the fovea.

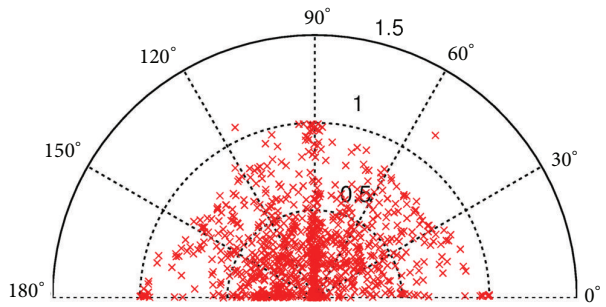


Figure 6: Scatter plot of the peak frequencies and orientations of Gabor functions fit to the ICA filters. The filters cover a wide spectrum of orientations and frequencies.

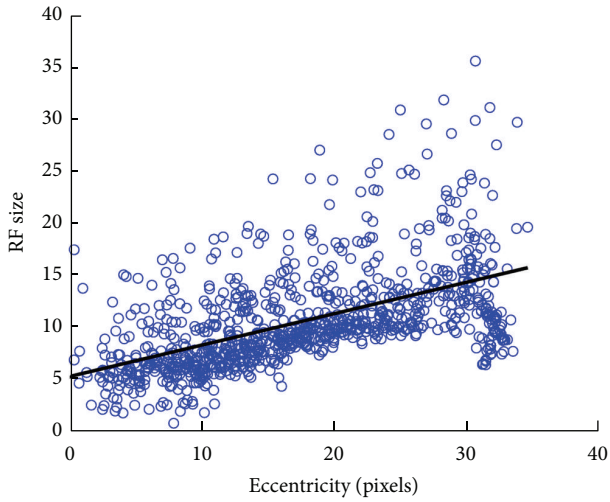


Figure 7: Plot of ICA filter/RF size as a function of eccentricity, along with a linear regression line. The size of the RFs was taken to be the area of the Gaussian envelope at full width at half maximum of the Gabor functions fit to the ICA filters. Like neurons in V1, RF size increases with eccentricity [25].

where $n_{k,c}$ is the total number of fixations from category k and b regulates the number of units learned ($b = 10$ in our experiments). This equation is plotted in Figure 9.

Inhibitive competition is used to suppress the least active gnostic sets. This is implemented for the K gnostic sets by attenuating their activity using

$$q_{c,k}\left(\mathbf{f}_{c,t}\right) = \max\left(\varphi_{c,k}\left(\mathbf{f}_{c,t}\right) - \theta_{c,t}, 0\right), \tag{15}$$

with the threshold $\theta_{c,t} = (1/K)\sum_{k'} \varphi_{c,k'}(\mathbf{f}_{c,t})$. Subsequently, the nonzero responses are normalized using

$$\beta_{c,k}\left(\mathbf{f}_{c,t}\right) = \nu_{c,t} q_{c,k}\left(\mathbf{f}_{c,t}\right), \tag{16}$$

with

$$\nu_{c,t} = \frac{\sum_{k'} q_{c,k'}\left(\mathbf{f}_{c,t}\right)}{\left(K^{-1} + \sum_{k'} q_{c,k'}\left(\mathbf{f}_{c,t}\right)^2\right)^{3/2}}, \tag{17}$$

acting as a form of variance-modulated divisive normalization [26].

As fixations are acquired over time, the gnostic field accumulates categorical evidence from each channel

$$\psi_{c,k}\left(\mathbf{f}_{c,1}, \ldots, \mathbf{f}_{c,T}\right) = \sum_{t=1}^{T} \beta_{c,k}\left(\mathbf{f}_{c,t}\right). \tag{18}$$

Subsequently, the responses from all of these evidence accumulation units are combined across all categories and

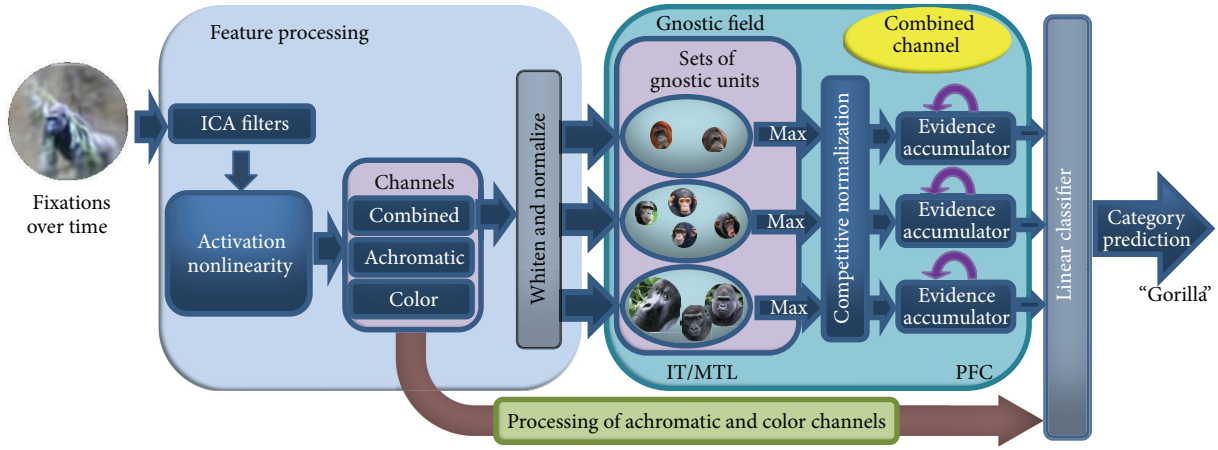


FIGURE 8: A high-level depiction of a gnostic field for classifying nonhuman apes using our space-variant ICA filters. The model splits the ICA filter output into chromatic, achromatic, and combined channels. This visual information projects to a gnostic set for each category, with units in the gorilla gnostic set responding strongest. The output of each gnostic set is given by the most active unit, and subsequent competitive normalization adjusts the activity to suppress the output of the chimpanzee and orangutan sets. Finally, evidence from the current fixation is added to the model's beliefs, and information from all categories and channels is combined using a linear classifier.

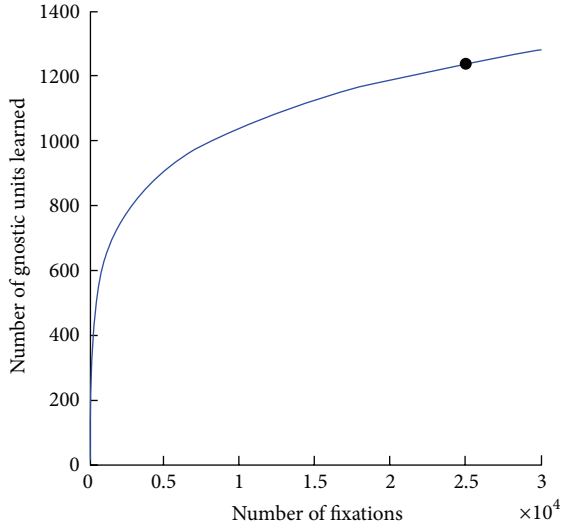


FIGURE 9: The total number of units learned for a gnostic set as a function of the number of fixations extracted from images labeled with the gnostic set's category. The black dot indicates the number of units learned for each Caltech-256 category with 50 training images per category and 500 fixations per image, so each gnostic set contained 1239 units learned from 25000 fixations.

channels into a single vector $\boldsymbol{\Psi}$. This vector is then made mean zero and normalized to unit length.

A linear multicategory classifier decodes the activity of these pooling units. This allows less discriminative channels to be downweighted and it helps the model cope with confused categories. The model's predicted category is given by $\widetilde{k} = \text{argmax}_k \mathbf{w}_k \cdot \boldsymbol{\Psi}$, where $\mathbf{w}_k$ is the weight vector for category k. The $\mathbf{w}_k$ weights were learned with the LIBLINEAR toolbox [30] using Crammer and Singer's multiclass linear support vector machine formulation [31], with a low cost parameter (0.0001).

4.2. Face and Object Recognition Experiments. We assess performance of the space-variant ICA features using two computer vision datasets: the Aleix and Robert (AR) face dataset [32] and Caltech-256 [33]. Training and testing consisted of extracting 500 fixations per image from random locations without replacement. We did not attempt to tune the number of fixations.

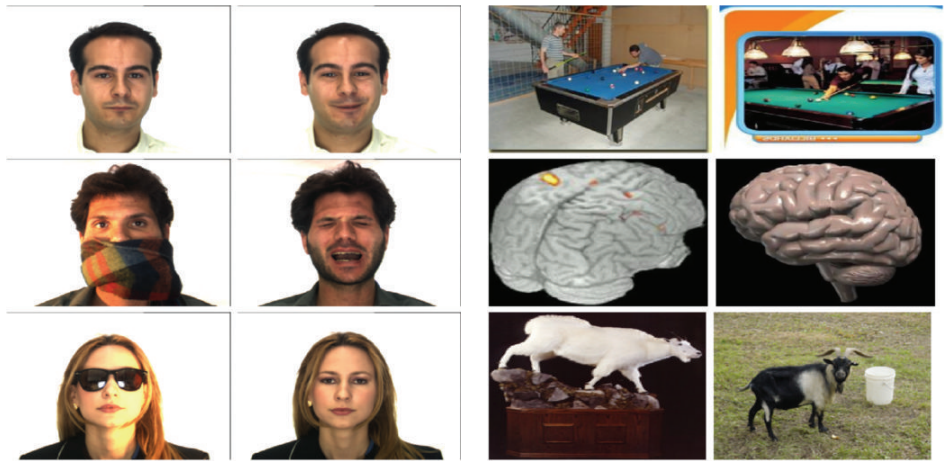
AR contains 4,000 color face images under varying expression, dress (disguise), and lighting conditions. We use images from 120 people, with 26 images each. Example images are shown in Figure 10(a). Results are shown in Figure 11. Our model performs slightly better than the best algorithms.

Caltech-256 [33] consists of images found using Google image search from 256 object categories. Example Caltech-256 images are shown in Figure 10(b). It exhibits a large amount of interclass variability. We adopt the standard Caltech-256 evaluation scheme [36]. We train on a variable number of randomly chosen images per category and test on 25 other randomly chosen images per category. We report the mean per-class accuracy over five cross-validation runs in Figure 12.

We performed an additional experiment on Caltech-256 to assess the impact of omitting the location information in the fixation features. Omitting it caused performance to drop by 3.6% when using 50 training images per category.

To examine how well gnostic fields trained using each channel individually performed compared to our main results using the multichannel model, we performed another experiment with Caltech-256 using 50 training instances per category. The multichannel approach performed best, and the chromatic filters alone worked comparatively poorly. These results are shown in Table 1.

We conducted additional experiments to examine performance as a function of the number of fixations used during testing. These results are shown in Figure 13. For both



(a) (b)

Figure 10: (a) Two example images from three of the models in AR. (b) Two example images from three Caltech-256 categories.

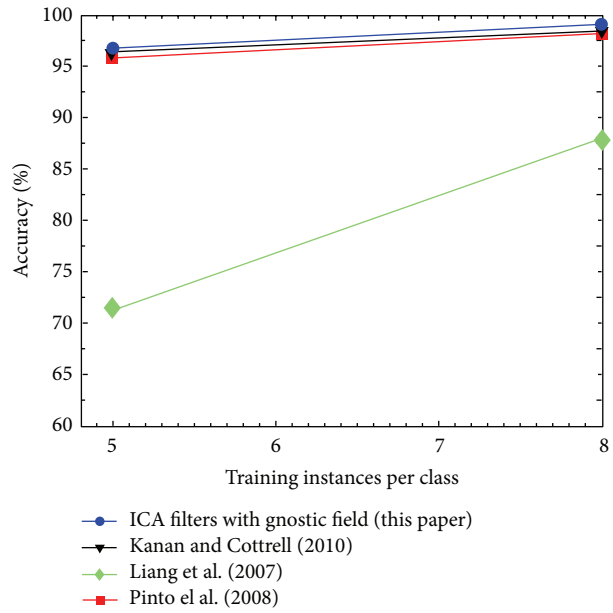


Figure 11: Mean per-class accuracy on the AR Face dataset of our approach compared with the methods of [10, 34, 35]. Chance performance is 1/120. Kanan and Cottrell [10] used a nonfoveated model of active vision (see discussion). Pinto et al. [35] used V1 + Gabor features with a linear SVM.

datasets, performance quickly rises; however, Caltech-256 appears to need more fixations to approach its maximum performance. In both cases, it is likely that choosing fixations in a more intelligent manner would greatly decrease the number of fixations needed (see Section 5).

Table 1: Mean per-class accuracy on Caltech-256 using 50 training instances per class for each channel specific gnostic field, along with the multichannel approach that combines all three channels.

Achromatic	Chromatic	All	Multichannel
45.6 ± 0.5	31.4 ± 0.4	48.4 ± 0.5	50.8 ± 0.5

5. Discussion

We applied ICA to spatially-variant samples of chromatic images. Our goal was to analyze the properties of the learned filters and to assess their efficacy at object recognition using the self-taught learning paradigm.

Our fixation-based approach to object recognition is similar to the NIMBLE model [10]. NIMBLE used a square retina, which pooled ICA filter responses learned from square patches. Instead of a Gnostic Field, NIMBLE used a Bayesian approach to update its beliefs as it acquired fixations. NIMBLE was unable to scale to large datasets because it compared

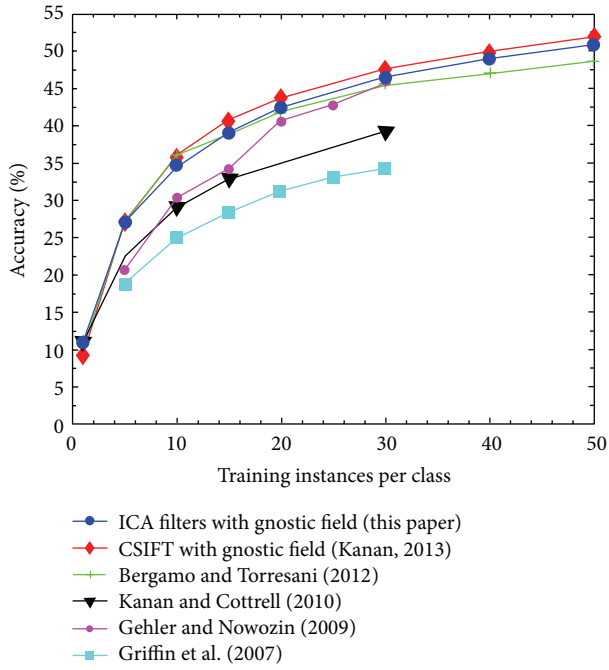


Figure 12: Mean per-class accuracy for our approach on Caltech-256 as a function of the number of training instances compared to the methods of [10, 26, 33, 36, 37]. Chance performance is 1/256. Kanan [26] used a gnostic field with color SIFT features, and our space-variant ICA filters achieve almost the same accuracy (slightly more for one training instance), despite being a self-taught approach. Bergamo and Torresani [37] combined five kinds of features (color GIST, oriented HOG, unoriented HOG, SSIM, and SIFT) into a metadescriptor using spatial-pyramid histograms. Gehler and Nowozin [36] used five types of engineered features (PHOG, SIFT, LBP, V1+ Gabors, and region covariance) and used multiple kernel learning to combine 39 different kernels. Kanan and Cottrell [10] used a nonfoveated model of active vision (see discussion). Griffin et al. [33] provides baseline results.

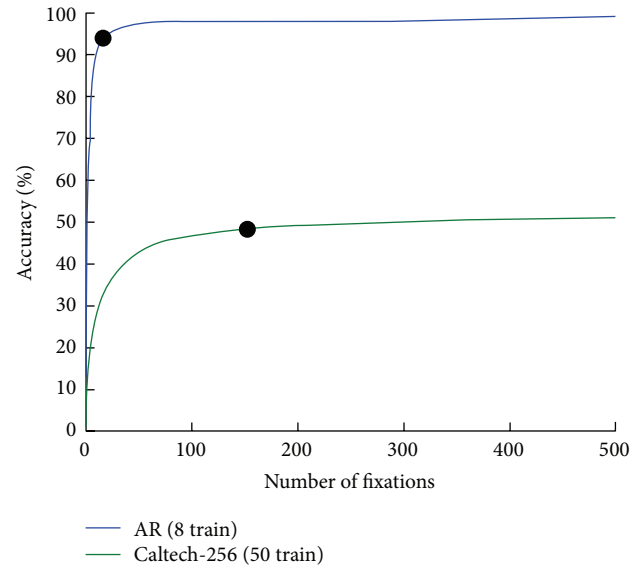


Figure 13: Our approach's mean per-class accuracy on AR and Caltech-256 as a function of the number of test fixations. For AR, 8 training instances per category were used, and for Caltech-256, 50 training instances per category were used. The results are averaged over five cross-validation runs. The black dots indicate the number of fixations needed to achieve 95% of the maximum accuracy (17 for AR and 152 for Caltech-256).

new fixations using nearest neighbor density estimation to all stored fixations for each category. For example, on Caltech-256 with 500 training fixations per image and 50 training instances per category, NIMBLE would store 25000 high dimensional fixation features per class, whereas a gnostic field would only learn 1239 gnostic units. This allows gnostic fields to be faster and more memory efficient, while also being more biologically plausible.

Like us, Vincent et al. [38] learned filters from a space-variant representation, but instead of ICA they used an unsupervised learning algorithm that penalized firing rate. Their algorithm also learned Gabor-like filters. They found that RF size increases away from the fovea, and that more filters are learned in the fovea compared to the periphery. While they were primarily interested in the RF properties, it would be interesting to examine how well their filters work for object recognition.

Log-polar representations can be made rotation and scale tolerant with respect to the center of a fixation [39], since changes in rotation and scale consist of "spinning" the retina or having it "zoom" in or out. Exploiting this could lead to improved object recognition performance, although if used in all situations it is likely to cause a loss of discriminative power (see [40] for an extensive discussion of the discriminative power-invariance tradeoff).

We are currently exploring avenues for developing a better controller for choosing the location of fixations. In our experiments we randomly chose the locations of fixations, but it is likely that significant gains in performance could be obtained by using a smarter controller that chose the next fixation location based on evidence acquired during previous fixations. The controller could also manipulate the rotation and size of the retina, potentially allowing it to increase its tolerance to changes in scale and rotation. One approach to learning a controller is to use reinforcement learning [41], with the reward function being crafted to reduce uncertainty about the object being viewed as quickly as possible. An alternative to reinforcement learning for fixation control was proposed by Larochelle and Hinton [42]. They developed a special kind of restricted Boltzmann machine that accumulated evidence over time. Their model learned a controller that selected among fixation locations on a $m \times m$ grid ($m \leq 7$ in their experiments), with the controller trained to choose the grid location most likely to lead to the correct label prediction.

A better controller would allow us to compare the model's simulated eye movements to the eye movements of humans when engaged in various visual tasks. We could also explore how changes in the retinal input might impact the way the controller behaves. For example, we could induce an artificial scotoma into our retinal model. Scotomas are regions of

diminished visual acuity, which are caused by diseases such as retinitis pigmentosa and age-related macular degeneration. Inducing an artificial scotoma would allow us to examine how the scotoma alters the acquired policy and if the changes are consistent with eye tracking studies in humans that have similar scotomas.

6. Conclusions

Here, for the first time, ICA was applied to a spatially-variant input, and we showed that this produces filters that share many spatiochromatic properties with V1 neurons, including eccentricity properties. Further, we showed that when these features are used with an object recognition system, they rival the best hand-engineered features in discriminative performance, despite being entirely self-taught.

Acknowledgments

The author would like to thank Akinyinka Omigbodun and Garrison Cottrell for feedback on earlier versions of this paper. This work was completed, while the author was affiliated with the University of California San Diego. This work was supported in part by NSF Science of Learning Center Grants SBE-0542013 and SMA-1041755 to the Temporal Dynamics of Learning Center.

References

[1] C. A. Curcio and K. A. Allen, "Topography of ganglion cells in human retina," *Journal of Comparative Neurology*, vol. 300, no. 1, pp. 5–25, 1990.

[2] R. F. Dougherty, V. M. Koch, A. A. Brewer, B. Fischer, J. Modersitzki, and B. A. Wandell, "Visual field representations and locations of visual areas v1/2/3 in human visual cortex," *Journal of Vision*, vol. 3, no. 10, pp. 586–598, 2003.

[3] S. A. Engel, G. H. Glover, and B. A. Wandell, "Retinotopic organization in human visual cortex and the spatial precision of functional MRI," *Cerebral Cortex*, vol. 7, no. 2, pp. 181–192, 1997.

[4] P. M. Daniel and D. Whitteridge, "The representation of the visual field on the cerebral cortex in monkeys," *The Journal of Physiology*, vol. 159, pp. 203–221, 1961.

[5] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[6] M. S. Caywood, B. Willmore, and D. J. Tolhurst, "Independent components of color natural scenes resemble V1 neurons in their spatial and color tuning," *Journal of Neurophysiology*, vol. 91, no. 6, pp. 2859–2873, 2004.

[7] T. W. Lee, T. Wachtler, and T. J. Sejnowski, "Color opponency is an efficient representation of spectral properties in natural scenes," *Vision Research*, vol. 42, no. 17, pp. 2095–2103, 2002.

[8] T. Wachtler, E. Doi, T. W. Lee, and T. J. Sejnowski, "Cone selectivity derived from the responses of the retinal cone mosaic to natural scenes," *Journal of Vision*, vol. 7, no. 8, article 6, 2007.

[9] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 759–766, June 2007.

[10] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2472–2479, June 2010.

[11] Q. V. Le, M. A. Ranzato, R. Monga et al., "Building high-level features using large scale unsupervised learning," in *Proceedings of the International Conference on Machine Learning (ICML '12)*, pp. 81–88, 2012.

[12] H. Shan and G. W. Cottrell, "Looking around the backyard helps to recognize faces and digits," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[13] M. D. Fairchild, *Color Appearance Models*, Wiley Interscience, 2nd edition, 2005.

[14] C. Kanan, A. Flores, and G. W. Cottrell, "Color constancy algorithms for object and face recognition," in *Advances in Visual Computing*, vol. 6453 of *Lecture Notes in Computer Science*, no. 1, pp. 199–210, 2010.

[15] C. Kanan and G. W. Cottrell, "Color-to-grayscale: does the method matter in image recognition?" *PLoS ONE*, vol. 7, no. 1, Article ID e29740, 2012.

[16] M. Bolduc and M. D. Levine, "A real-time foveated sensor with overlapping receptive fields," *Real-Time Imaging*, vol. 3, no. 3, pp. 195–212, 1997.

[17] M. Bolduc and M. D. Levine, "A review of biologically motivated space-variant data reduction models for robotic vision," *Computer Vision and Image Understanding*, vol. 69, no. 2, pp. 170–184, 1998.

[18] E. L. Schwartz, "Spatial mapping in the primate sensory projection: analytic structure and relevance to perception," *Biological Cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[19] M. Chessa, S. P. Sabatini, F. Solari, and F. Tatti, "A quantitative comparison of speed and reliability for log-polar mapping techniques," in *Computer Vision Systems*, vol. 6962 of *Lecture Notes in Computer Science*, pp. 41–50, 2011.

[20] R. H. Masland, "The fundamental plan of the retina," *Nature Neuroscience*, vol. 4, no. 9, pp. 877–886, 2001.

[21] A. Olmos and F. A. A. Kingdom, "A biologically inspired algorithm for the recovery of shading and reflectance images," *Perception*, vol. 33, no. 12, pp. 1463–1473, 2004.

[22] Z. Koldovský, P. Tichavský, and E. Oja, "Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1265–1277, 2006.

[23] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.

[24] M. Grundland and N. A. Dodgson, "Decolorize: fast, contrast enhancing, color to grayscale conversion," *Pattern Recognition*, vol. 40, no. 11, pp. 2891–2896, 2007.

[25] R. Gattass, C. G. Gross, and J. H. Sandell, "Visual topography of V2 in the Macaque," *Journal of Comparative Neurology*, vol. 201, no. 4, pp. 519–539, 1981.

[26] C. Kanan, "Recognizing sights, smells, and sounds with gnostic fields," *PLoS ONE*, vol. 8, no. 1, Article ID e54088, 2013.

[27] J. Konorski, *Integrative Activity of the Brain*, University of Chicago Press, Chicago, Ill, USA, 1967.

[28] M. Kouh and T. Poggio, "A canonical neural circuit for cortical nonlinear operations," *Neural Computation*, vol. 20, no. 6, pp. 1427–1451, 2008.

[29] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, no. 1-2, pp. 143–175, 2001.

[30] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[31] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[32] A. M. Martinez and R. Benavente, "The AR face database," Tech. Rep. 24, CVC, 1998.

[33] G. Griffin, A. D. Holub, and P. Perona, "The Caltech-256 object category dataset," Tech. Rep. CNS-TR-2007-001, Caltech, Pasadena, Calif, USA, 2007.

[34] Y. Liang, C. Li, W. Gong, and Y. Pan, "Uncorrelated linear discriminant analysis based on weighted pairwise Fisher criterion," *Pattern Recognition*, vol. 40, no. 12, pp. 3606–3615, 2007.

[35] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?" *PLoS Computational Biology*, vol. 4, no. 1, article e27, 2008.

[36] P. Gehler and S. Nowozin, "On feature combination for multiclass object classificationpages," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV '09)*, pp. 221–228, IEEE Computer Society, Los Alamitos, Calif, USA, 2009.

[37] A. Bergamo and L. Torresani, "Meta-class features for large-scale object categorization on a budget," in *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR '12)*, 2012.

[38] B. T. Vincent, R. J. Baddeley, T. Troscianko, and I. D. Gilchrist, "Is the early visual system optimised to be energy efficient?" *Network: Computation in Neural Systems*, vol. 16, no. 2-3, pp. 175–190, 2005.

[39] V. Javier Traver and A. Bernardino, "A review of log-polar imaging for visual perception in robotics," *Robotics and Autonomous Systems*, vol. 58, no. 4, pp. 378–398, 2010.

[40] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision (ICCV '07)*, October 2007.

[41] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Mass, USA, 1998.

[42] H. Larochelle and G. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proceedings of the 24th Annual Conference on Neural Information Processing Systems 2010 (NIPS '10)*, December 2010.