A Generative Model of Retinal Ganglion Cells Learns Separate Channels from Color Images

Daniel von Poschinger-Camphausen *, Cornelius Weber, Stefan Wermter

University of Hamburg
Knowledge Technology, Dept. of Informatics
Vogt-Kölln-Straße 30
D - 22527 Hamburg, Germany
www.informatik.uni-hamburg.de/wtm

Abstract

The retina has remarkable properties of efficiently encoding visual information. In particular, various aspects of the visual stream are separated into distinct channels, which manifest in specific retinal ganglion cell (RGC) types. Nevertheless, the exact principles are neither well understood nor technically exploited; in particular an account of how learning contributes to the formation of these channels is missing. Therefore, we simplified an established computational model of self-organizing RGC receptive fields (RFs) by removing one parameter, extended it to process color images, and trained it with images of natural scenes. The results show that localized color opponent RFs emerge in chromatically distinct channels. Each channel covers the entire visual space. The opponent structure of the resulting RFs captures biologically plausible chromatic contrasts compared to filters derived from statistical methods.

Keywords: Retina — Color vision — Retinal ganglion cells — Energy efficiency — Chromatic contrast — Unsupervised learning

1. Introduction

The retina of the human eye is often compared to a photo-sensor of a digital camera. Such a comparison neglects that several retinal neuronal layers perform pre-processing of the image stream such as compression, noise removal and contrast enhancement (Kolb, 2004), (Weber and Triesch, 2009). As opposed to the output of luminance and color information of a CCD sensor, the human retina has at least seventeen different types of retinal ganglion cells (Field and Chichilnisky, 2007), indicating that the 'output' of the retina is far more complex and richer in features than that of a simple camera.

In human perception, color appears to be relatively stable to chromatic changes in the illumination of a scene. The robustness of color perception is referred to as color constancy (Brainard et al., 1997), (Purves et al., 2002). By contrast, the color information of a camera is unreliable. Even a subtle change in the illumination of a scene can lead to drastic changes in the color-values sampled by a camera sensor. How the visual system establishes this perceptual phenomenon at the cellular level, is still not fully understood and subject to discussion (Shapley and Hawken, 2011). However, chromatic contrast, or color opponency, has been identified as the fundamental building block of color constancy (Shapley and Hawken, 2011). Therefore, computational models learning chromatic contrast from images of digital cameras appear to be of value. Cells providing color opponency have been found in the ganglion cell layer of

* Corresponding author.
Email: *daniel.poschinger.camphausen@gmail.com*

the retina and in the primary visual cortex (V1).

Neuroanatomical studies have shown that the mammalian retina consists of "many parallel, anatomically equipotent microcircuits" (Masland, 2001) forming discrete pathways, which process different aspects of the visual stream. These pathways emanating from the retina manifest in distinct retinal ganglion cell (RGC) types, which are classified by their physiology, morphology, dendritic connectivity and to which regions their axons project (Field and Chichilnisky, 2007). In each class of morphologically distinct RGC types the receptive fields (RFs) cover the entire retinal space uniformly with constant overlap, in some cases with no overlap. This uniform mosaic structure, see Fig. 1A, is assumed to enable a regular sampling of the visual field (Field and Chichilnisky, 2007). Five cell types account for 75% of all RGCs: ON and OFF *parasol*, ON and OFF *midget*, and *small bistratified* cells (Field et al., 2010). It has been observed, that a particular RGC is tuned to a specific contrast of spatial frequencies in the image (Nelson, 2007). The contrast sensitivity of RGCs, together with the de-correlation into ON and OFF channels, allows later stages of neural computation, e.g. the visual cortex, to define precise edges (Kolb, 2004).

Correlations in the signals of cone types are removed by the filtering through chromatic contrast selective RGCs (Gegenfurtner, 2003) and as such frees the visual cortex having to process these less informative components of images. By pitting the signals of different cone types in the same area of the visual field against one another, the neural circuits of the retina establish chromatic contrast (Gouras, 2009). We differentiate here between luminosity or intensity contrast, which does not discriminate between different cone types, and chromatic contrast, which is the result of any of the three different cone types, see Fig. 1B. Chromatic contrast can be separated into two classes.

First, *single-cone* contrast resulting solely from ON and OFF connectivity to a single cone type. For red (L), green (M) and blue (S) cones we have L/-L, M/-M and S/-S contrasts. If viewed in RGB color-space for example L/-L contrast corresponds to a vector of red $[1, 0, 0]$ with a counterpart of cyan $[0, 1, 1]$.

Second, *multi-cone*, color opponent contrast resulting from comparing the ON and OFF contributions of different cone types. The majority of *midget* RGC are selective to color opponent L/-M and -L/M *multi-cone*, and also S/-S *single-cone* contrast. Other cone opponent contrasts are less commonly found in primate retinas, see Table A.3. If viewed in RGB color-space for example L/-M contrast corresponds to a vector of red $[1, 0, 0]$ and a magenta $[1, 0, 1]$ counterpart.

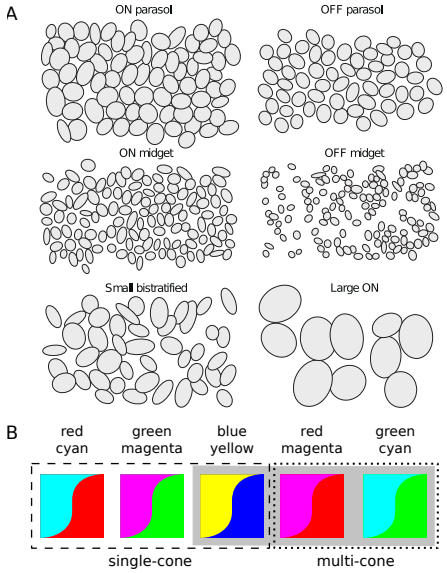

Figure 1: (A) mosaics of RFs of six RGC types in the primate retina (Field and Chichilnisky, 2007). Each mosaic was obtained from a single 512-electrode recording from isolated peripheral primate retina, in which each ellipse shows the 1.3 SD contour of a Gaussian fit to the RF of a single cell. (B) Different types of chromatic contrast. Inside the dashed line: *single-cone* contrast along the red, green and blue axis in RGB color-space. Inside the dotted line: *multi-cone* or color opponent contrast of red/magenta and green/cyan. Inside shaded area: chromatic contrast for which the majority of *midget* RGC are selective, see Table A.3 in the appendix.

As we have described, each major particular RGC-type manifests in a separate channel. The RFs of such a channel form a regular mosaic spanning the entire visual field. Additionally, the antagonistic Difference of Gaussian (DOG) RFs texture of each RGC type is tuned to a specific contrast of the visual input. Therefore, the question

we are addressing here is whether a model can predict these channels and contrasts by learning from images of natural scenes. In the following, we will describe our computational RGC model, followed by the results of fitting and clustering the training results. After that, we compare the training results with biological findings, compare our model with related methods and give an outlook on how our model can be extended further.

2. Methods

The retina, compared to V1 simple cells, has not yet attracted much attention among computational modelers, regardless of its functional role of producing chromatic contrast. Numerous developmental models of learning cortical V1 simple cells exist, based on ICA, e.g. (Bell and Sejnowski, 1997) or sparse coding, (Olshausen and Field, 1997); (Vincent et al., 2005); (Weber and Triesch, 2008); (Doi et al., 2012). For learning retinal ganglion cells, a model based on synaptic energy efficiency was proposed by (Vincent and Baddeley, 2003).

2.1. Generative model of retinal ganglion cells

Vincent and Baddely utilized a simplified linear auto-encoder in order to explore the effect of minimizing metabolic costs while learning optimal filters to represent natural scenes. The model consists of a single hidden layer, where units in the visible layer functioning as photoreceptors feed image patches of natural scenes to the hidden layer. In the absence of a non-linear transfer function (Vincent and Baddeley, 2003) showed that by solely constraining the connection weights of the two layers, localized DOG shaped RFs emerge, organized in a weakly overlapping mosaic, resembling those of RGCs.

2.2. Extended generative RGC model

We extended the model of (Vincent and Baddeley, 2003): instead of processing gray scale images, the model processes RGB images, see Fig. 2. Furthermore, the original model applied the weight constraint to a RF only if the sum of the absolute weight values of this RF were above a specific threshold, leaving the RF unchanged otherwise. We removed this threshold mechanism, always applying the constraint to every receptive field.

The algorithm for learning the connection weights W in the generative model consists of five steps: For an input vector x and a matrix W, the hidden activity is computed in Eq. (1). The input is reconstructed in Eq. (2) and the reconstruction error, Eq. (3), is used in the Hebbian learning rule, Eq. (4), to update W. Afterwards the weight constraint, Eq. (5), is applied to the RF of each hidden neuron j. The full algorithm is as follows:

$$y = \max(0, W \cdot x) \tag{1}$$
$$z = W^T \cdot y \tag{2}$$
$$e = x - z \tag{3}$$
$$\Delta W = \eta\, e \odot y \tag{4}$$
$$\Delta W_j = \eta\, (-k\ \mathrm{sgn}(W_j)\ |W_j|^{p-1}) \tag{5}$$

Here, $\cdot$ denotes the inner product, $\odot$ denotes the outer product, W_j is the synaptic weight vector of hidden unit j. Notably, Eq. (1) effectively introduces a non-linear transfer function, also known as rectified linear unit $ReLU$ (Krizhevsky et al., 2012). Clipping hidden unit activity has a biological motivation, since neurons are not able to have negative responses. Additionally, compared to binary hidden units, with a sigmoid transfer function $ReLU$s have been reported to improve the learning of feature detectors in generative models (Nair and Hinton, 2010). Apart from the number of visible units vis and the number of hidden units hid, the model has only three hyperparameters: strength k, shape p of the metabolic weight-constraint and learning rate η.

During training the combined cost function $\mathrm{E} = \mathrm{E}_{rec} + k\ \mathrm{E}_{syn}$, consisting of the quadratic reconstruction error term, Eq. (6), and the synaptic energy term, Eq. (8), is minimized. By convergence, the shape of the resulting RFs stabilizes, which indicates that the cost function is minimal. The partial derivatives Eq. (7) and Eq. (9) con-

form to the already given Eq. (4) and Eq. (5).

$$\mathrm{E}_{rec} = \frac{1}{2}\,(x - W^T \cdot y)^2 \quad = \frac{1}{2}\,e^2 \quad (6)$$

$$-\frac{\partial}{\partial W}\,\mathrm{E}_{rec} = (x - W^T \cdot y)\; y \quad = e \odot y \quad (7)$$

$$\mathrm{E}_{syn} = \sum_{j}^{hid} \frac{1}{p}\;\mathrm{sgn}(W_j)\;|W_j|^p \quad (8)$$

$$-k\,\frac{\partial}{\partial W_j}\,\mathrm{E}_{syn} = -k\;\mathrm{sgn}(W_j)\;|W_j|^{p-1} \quad (9)$$

2.2.1. Input

The input of the model is a vector of m units, which is derived from a patch cut from a training image *. A vector is created from the following steps:

1. A random image from the training set is selected; from this image a patch of fixed dimension and of random location is cut. To avoid unnatural image statistics patches are never cut near the borders of an image.
2. In 50% chance the axes of the patch are interchanged. We experienced that by not interchanging axes, localized RFs had an unrealistic mostly long vertical shape, presumably resulting from statistics of vertical lines, e.g. trees, grass, in the training images.
3. In one setting of the model the mean of the patch is subtracted, which yields negative and positive pixel values (see Section 3.1.3).
4. The patch is vectorized, its two-dimensional structure is transformed to a flat vector. A square $(n \times n)$ patch of n pixels results in a flattened vector of $m = n^2$ units.

* The training set contained 293 RGB-images of various natural scenes. All images originated from *McGill Calibrated Colour Image Database* http://tabby.vision.mcgill.ca.

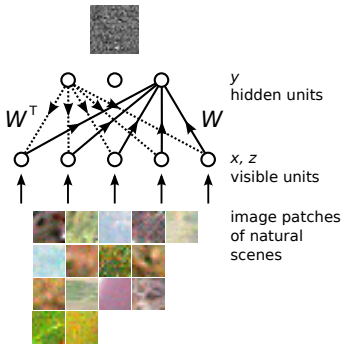

Figure 2: The RGC model extended to color images. The visible units together represent an image patch, the hidden units represent RGCs. Solid arrows indicate connections from visible to the hidden units. Dotted arrows indicate generative connections from hidden to visible units.

2.2.2. Color

In order to operate on non-scalar pixel values, the elements of RGB pixel vectors have to be flattened into a single vector. Thus, processing patches of $n \times n$ pixels size results in an input-vector of m units: $m = n^2 * r$ with input channels r, where $r = 1$ for grayscale images and $r = 3$ for RGB color images.

2.2.3. Shape of the weight constraint

The parameter p sets the shape of the weight constraint, defining how weights of an individual RF are affected:

- $p = 1$: the corresponding cost function resembles a L1-norm, effectively reducing all weights of a RF equally. This affects in particular the small weights and thus leads to a sparse distribution.
- $p = 2$: the L2-norm in the cost function penalizes weights proportionally to their values, due to its parabolic / spherical shape. But in contrast to L1 will not change the shape of a neuron's weight vector.
- $p = 1.5$: We found that this compromise led to the best results.

2.2.4. Strength of the Weight constraint

The parameter k defines the strength of the weight constraint. If not constraining the con-

nection weights, the resulting RFs become non-localized and cover the whole input (Vincent and Baddeley, 2003). The application of the weight constraint has the effect of enforcing the input-reconstruction with minimal weights of each individual receptive field. This, in turn, has the desired effect of each unit learning distinct aspects, or features, of the presented input statistics, whereas spatial localization is the most prominent. If k has a large value all weights of a unit's RF can be reduced to zero. Such a unit is considered dead (see Section 3.1.2). Dead units do not occur with the threshold mechanism of the original model.

3. Results

3.1. Tuning of hyper parameters

3.1.1. Under- and overcompleteness

The completeness of a model simply takes into account the number of visible units (vis) in relation to the number of hidden units (hid). A model which employs more visible than hidden units ($vis > hid$) is said to be *undercomplete*, encoding the presented statistics in a compressed, possibly incomplete, manner (Baldi and Homik, 1995). Conversely a model which has more hidden units than visible ($vis < hid$) is said to be *overcomplete* resulting in a neural code with some degree of redundancy (Olshausen and Field, 1997), see Fig. 3B. The size and spatial location of an individual RF depend on the sizes and locations of all other RFs contributing to the same aspect of the learned input statistic. The model has the tendency, in certain bands in the parameter space, to converge to a distribution in which all RFs are of near equal size, uniformly and collectively covering the input. Thus, the minimal RF size can be approximated:

$$\mathrm{RF}_{size} = \begin{cases} 1 & \text{if (over-)complete} \\ vis/hid & \text{if undercomplete} \end{cases} \quad (10)$$

We experienced that an *overcomplete* model converges in localized RFs with a center of one pixel and no surround, see Fig. 3B. Therefore for the emergence of localized DOG-shaped RFs with a significant antagonistic surround, the model needs to be *undercomplete*. With $vis/hid > 1.2$, we observed that our model produces RFs with noticeable antagonistic surround. We used only weakly *undercomplete* coding, because for computational efficiency we cannot afford large input area sizes. Note that we require computationally expensive all-to-all connectivity, since localization of the RFs should be a result of learning.

3.1.2. Dead hidden units

If all connection weights of a unit's RF are close to zero a unit is considered dead. Being in that dead state, a unit is no longer activated by any input and neither contributes to the reconstruction of the input. Three factors cause unit death during the training process:

Large value of k: If over a sufficient number of training epochs, the weights of an individual unit grow slower than they are reduced by the weight constraint, such a unit trivially ends up being dead. This effect can be observed drastically if the constraint parameter k is set to a very large value: After a few epochs the connection matrix is filled with dead RFs while the remaining RFs are of large shape, since very few alive RF have to reconstruct the input, see Fig. 3A.

Overcompleteness: Even if the value of k is not too large, individual units not contributing much to the input-reconstruction over a longer time eventually die off, see Fig. 3B. A small reconstruction error occurs when there are many hidden units in case of *overcompleteness*. Consequently, a reconstruction error of minimal magnitude results in minimal growth of a unit's RF weights. If, over time, the magnitude of growth is smaller than the constraint which reduces the weights in every epoch, this particular unit eventually dies. As dead units reduce the effective number of hidden units to learn the input statistics, the model is moved towards being *complete* during the training process. This in turn results in compact neural codes with low redundancy, in which all alive hidden units participate in input reconstruction.

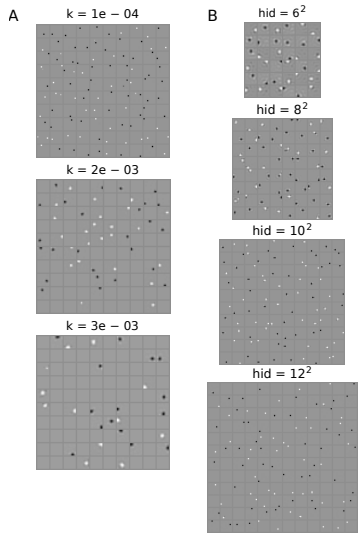

Figure 3: (A) The effect of a large value of k in relation to η causing dead hidden units. All connection maps are trained with the same set of parameters and unconstrained, i.e. linear, hidden units: $vis = 10^2$, $hid = 10^2$, $r = 1$ (luminosity), $p = 1.5$ and $\eta = 0.01$. (B) The effect of model *completeness* in relation to RFs size: All parameters are as in (A), except the number of hidden units *hid* and $k = 5e - 04$. As the model becomes *overcomplete* ($hid = 10^2$ and larger) the RF size does not shrink further, but more RFs die off.

The rectification of hidden unit activity: Using $ReLU$s is known to produce results in which about 40 % of of all hidden units may die off during the training process (Krizhevsky et al., 2012). Thus, this property of $ReLU$s additionally contributes to the existence of dead units in the converged connection matrix.

3.1.3. Constraining input values

We trained the model in two settings. One, denoted as $x\pm$, allowing positive and negative input values and one, denoted as $x+$, allowing solely positive input values.

In setting $x\pm$, we subtract the mean of the input patch x. This results in values of the input ranging from -0.5 to 0.5. Subtracting the mean has a biological motivation since it emulates the differences of intensity propagated by horizontal and bipolar cells (Gouras, 2009). As a result from an information theoretic perspective, the model is freed from learning the meaningless mean of the patch, which may be large compared to the pixel variance. If the data contains structure beyond the mean values of particular patches, the model can learn this structure faster.

In setting $x+$, we do not subtract the mean of the input patch, which results in feeding scaled normalized RGB values into the model. The values of the input then range from 0 to ≈ 1.0. Biologically, this resembles the responses of the photoreceptors to light.

3.2. Training results

The training results show that most RGC neurons have developed localized, centre-surround RFs. These have distinguished properties, primarily defined by their color and size, which appear to fall into distinct classes. We determined these classes by fitting each RF with a DOG function and then clustering the obtained DOG parameters, see Appendix for details.

3.3. Initial parameters

For computational efficiency we trained the model with a relatively small number of input units, resulting in a very small photoreceptor to RGC density of the hidden units, see Table 1. Because of dead hidden units during the training process, the number of hidden units is set initially to three times the number of visible units, which renders the model *complete*. After the training, the model is reasonably *undercomplete* with $vis/hid = 1.11$ in setting $x\pm$ and strongly *undercomplete* with $vis/hid = 1.54$ in setting $x+$, see Table 1. The comparably high value of vis/hid in setting $x+$ reflects the higher loss of hidden units during the training process.

To produce localized RFs, we found a suitable set of parameters for setting $x\pm$ first. Training setting $x+$ with similar parameters did not converge at all, which suggested to increase k. We experienced, while training in setting $x\pm$, that a larger value of k resulted in faster localization of RFs. With increased k, training in setting $x+$ converged as expected, but resulted in non-localized RFs, which are the result of a small number of surviving hidden units. Therefore, values of k need to be large enough for RFs to localize,

but small enough for a sufficient number of hidden units to survive. For this reason, compared to setting $x\pm$, we increased the initial number of hidden units and also subtly increased k, see Table 1.

setting	vis	hid	hid alive	alive%	$\frac{\text{vis}}{\text{hid}}$	k	p	η
$x+$	$3*13^2$	42^2	659	37%	1.54	2e-05	1.5	0.01
$x\pm$	$3*13^2$	39^2	916	60%	1.11	7e-06	1.5	0.03

Table 1: Model parameter values. From left to right: name of the setting, number of visible units, hidden units, hidden units alive after training, the *undercompleteness* of the model, values of k, η and p.

3.4. Emergence of distinct channels

Plotting the results of the fitting process clusterwise reveals that in all settings distinct channels emerge, see Fig. 4. The RFs of each cluster appear to cover the entire input space with minimal overlap. Setting $x\pm$ develops six distinct complete channels of uniform RF signature. The ratio of the number of units in each channel in relation to the number of alive hidden units fluctuates around $1/n$ for n channels, see Table 2.

Interestingly, setting $x+$ develops six distinct complete channels of uniform RF signature, of which two appear to cover non-chromatic or luminosity aspects, whilst the remaining four cover chromatic aspects of the input. The other setting $x\pm$ has been extensively trained, even with large weight constraints producing nonuniform RF morphologies, but separate achromatic channels did not appear in the clustered data. Therefore, the property of separating the input into chromatic and luminous aspects appears to be attributed to $x+$ exclusively.

setting	white	black	red	green	blue	cyan	magenta	yellow
$x+$	25 %	23 %	12 %	10 %		17 %	13 %	
$x\pm$			17 %	15 %	18 %	14 %	19 %	17 %

Table 2: Sizes of clusters in percentage: the number of units per cluster in relation to the total number of alive hidden units. Meaningful names are given, reflecting the prototype color of each cluster.

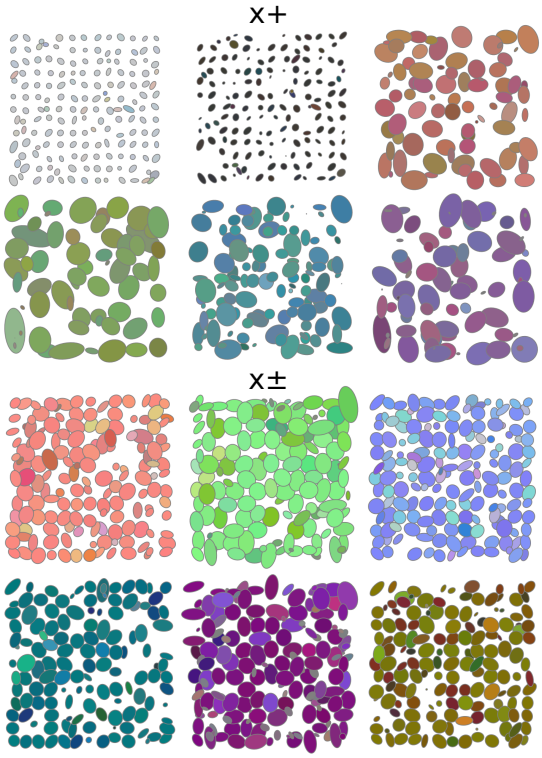

Figure 4: The result of the fitting and clustering process; training of setting $x+$ (above) results in two luminosity and four chromatic channels. Setting $x\pm$ (below) results in six chromatic channels and does not develop separate luminosity channels. Each alive RF is present in one of the subplots for each setting. The RF of a hidden unit is drawn as an ellipse with the parameters obtained from the best fit of the spatio-chromatic DOG model. The color in which an ellipse is drawn reflects the value of the center weight of a particular receptive field. For clarity, only the center ellipse of a fit is shown.

3.5. Prototype RFs

Manually selected prototypical RFs for each channel reveal that in setting $x+$ prototype RFs of the four chromatic channels are several times larger than RFs of the two luminous channels, see Fig. 5. In contrast, prototypes of setting $x\pm$ are of all similar size and appear to be more localized and sharper, compared to chromatic prototypes of setting $x+$.

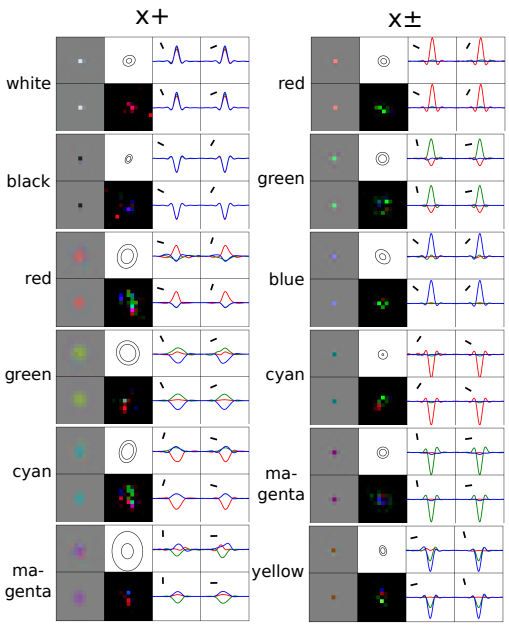

Figure 5: A box of eight tiles shows the prototype RF and its best fit (1st column), the fit's center surround ellipses and the reconstruction error (2nd column). Interpolated cuts of the prototype RF along the primary axis (above, 3rd column) and secondary axis of the best fit ellipse (above, 4th column), likewise the DOG reconstruction of the RF (bottom, 3rd and 4th column). The direction of the axis is denoted by the small line in the top left corner.

Furthermore, prototype RFs of both settings differ in the texture of the antagonistic surround: In setting $x+$ the large surround of chromatic prototypes contains a mixture of sub-blobs resulting in *multi-cone* and *single-cone* contrast. The smaller surround of the remaining two non-chromatic prototypes show a luminosity contrast. In contrast, in setting $x\pm$ the small surround of all prototypes show a uniform *single-cone* contrast.

3.6. Convolution filters

The deviation of both settings is also visible in the results of filtering each obtained prototype RF with a RGB test image. The test image contains luminous and chromatic contrasts, the former, visible in the background and in the claws of the parrot, the latter, in the remaining areas of the parrot, see Fig. 6.

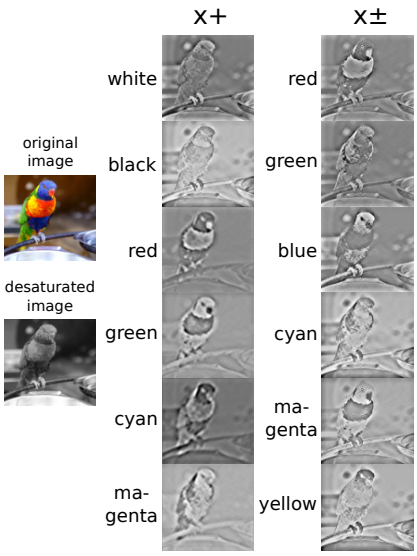

Figure 6: Convolution results of filtering a test image with the obtained prototype RFs.

The results indicate that $x+$ prototypes discriminate stronger than $x\pm$ prototypes among chromatic and luminous aspects of the input: The claws of the parrot, luminous contrast, are only visible in the output of black and white prototype filters of $x+$ and are invisible in the output of the remaining prototype filters. On the contrary, the claws are visible in the output of all prototype filters of $x\pm$, indicating no separation of luminous and chromatic aspects in setting $x\pm$. Moreover, in the filtered images the sharpness of the background indicates the amount of luminous information captured by the corresponding prototype. All results of setting $x\pm$ prototypes show a similar sharp background. On the contrary, the results of $x+$ prototypes are varying; two show a sharp, whilst the remaining four show a rather blurred background.

3.7. Clipping hidden units

The presented results were obtained with clipped activities of *ReLU*s, which cannot become negative. We also experimented with linear hidden units which allow negative values. In this case, training results in a RF distribution in which for each input channel, one complete output channel emerges, i.e. three channels for RGB images. A particular channel is composed of a weakly overlapping mosaic of mixed ON and OFF units, see

Figure 3. In contrast, clipped activities result in a distribution of RFs in which for each input channel two complete output channels emerge, one of which compensates for the missing negative activities, i.e. six channels for RGB images. A channel is uniformly composed of weakly overlapping mosaic of either ON or OFF units.

4. Discussion

The original work of (Vincent and Baddeley, 2003) showed that by enforcing input reconstruction with synaptic efficient weights, spatially localized RFs emerge which cover the entire visual field. The results demonstrate that our extended model replicates this property of emerging localized RFs also in the color domain.

Due to the property of our model being *undercomplete* a compressed representation of the input data with minimal reconstruction error is learned. This property is of high benefit in applications using high dimensional input, such as processing RGB images with CNN, since todays cameras have resolutions of many megapixels, which renders the application of deep NN intractable in directly processing each input pixel. For example, learning from the pixels of an Atari video game (210×160 RGB video) required reducing the input dimensionality to 84×84 units (Mnih et al., 2013). Correspondingly, the retina compresses the information of approximately 6.4 million cone photoreceptors plus an even larger number of rod photoreceptors, to one million RGC (Kolb, 2004), indicating a high rate of compression of visual information.

4.1. Constraining the input activations

The results show that in setting $x+$, luminous and chromatic aspects of the input are separated by the resulting prototype filters. Additionally, prototypes in this setting also capture biologically more plausible *multi-cone* chromatic contrast. In contrast, setting $x\pm$ produces filters in which luminous and chromatic aspects are not separated. The resulting filters are capable of capturing solely *single-cone* chromatic contrast. This suggests that setting $x+$ is more realistic in terms of biological plausibility than $x\pm$. It also indicates that modeling the full process of retinal image processing rather than a partial aspect leads to better results: an input of solely positive values, like in setting $x+$, resembles the output of photoreceptor cells. Subtracting the mean of an input patch, like in setting $x\pm$, results in an input of positive and negative values, resembling the ON and OFF signals of the combined output of photoreceptor and bipolar cells (Masland, 2001).

4.2. Formation of RFs mosaic and RGC types

The self-organization of localized RF mosaics that emerge during the training process of our model is unanimously observed in biology. Our model simulates synaptic activity patterns, which are attributed to drive the formation of ON and OFF ganglion cell types during retinal development (Chalupa and Günhan, 2004). Also, local synaptic activity patterns between neighbor cells of similar type are found to cause the formation of RF mosaics (Reese, 2012). This locality also occurs in our model: During training, similar localized RFs compete in reconstructing a similar (spatially) local aspect of the input. Competing hidden units therefore indirectly interact in a local manner, through their RF overlaps.

By learning from natural images, our model suggests visual experience to contribute to the formation RF mosaics and RGC types. However, RF mosaics are found also to evolve without visual stimulation (Anishchenko et al., 2010). In prenatal retinas, before the onset of visual experience, RF mosaics and corresponding RGC types have been observed to evolve (Wong, 1999), (Elliott and Shadbolt, 1999). However, during retinal maturation RF mosaics are found to be refined after the onset of visual experience (Tian, 2004) indicating a contribution of visual experience in the formation of RF mosaics.

4.3. Hidden unit death

The weight-constraint without a threshold results in a competition among the hidden units. Weakly participating units eventually die off, which appears to be a useful property: The competition results in a compact neural code, in which our

model, being *overcomplete*, automatically adjusts the number of hidden units towards an undercomplete ratio of visible to hidden units. Moreover, the competition of hidden units resembles the competition of dendrites and ganglion cell death during the development of the retina (Linden and Perry, 1982), (Linden, 1993).

4.4. Deviations from biology

Comparison of the emerged RFs of our model with the biological findings reveals quantitative deviations in some aspects, which can be explained by the simplicity of our model: Our model assumes an equally spaced, grid-like spatial distribution of cones, which obviously diverges from the concentric organization of cells around the fovea (Weber and Triesch, 2009). Learning from RGB images also implies that all cones exist in equal numbers, which ignores that e.g. blue cones are absent in the fovea region and that blue cones only amount to 15 % of all cones (Masland, 2001). The grid-like input statistics can also be suspected to result in our model to produce distributions of ON and OFF sub-types of similar number and similar diameter, see Fig. 4 and Table 2. This differs from biology in which *midget* ON-center RGC occur roughly 3 times more often than corresponding OFF types, see Appendix table A.3. Also, ON-center *midget* and *parasol* RGC show about 30 % to 50 % larger RF diameters than their OFF-center counterparts (Dacey and Petersen, 1992).

In setting $x+$, our model converges to a distribution in which RFs of luminosity selective hidden units are of small localized shape, whilst the RFs of color selective hidden units are several times larger, see Fig. 4. This ratio is found inversely in the morphology of ganglion cells in primate retinas: Near the fovea, luminosity selective *parasol* RGCs exhibit two to three times larger RFs than color selective *midget* RGCs, whilst in peripheral areas, the *parasol* RFs may even be up to 10 times larger than *midget* RFs (Nelson, 2007). We expect this contradicting result to disappear if a metabolic cost is imposed on a cell for having faster conduction velocities. In primate retinas *parasol* RGC transmit information faster than *midget* RGC, which comes at a cost of a larger soma and thicker axons of *parasol* RGC (Baden et al., 2014). Subsequently, imposing such a metabolic cost would result in fewer faster cells with larger RFs and conversely, in more slower cells with smaller RFs. However, our simple model neglects any temporal aspects, since it learns from still RGB images.

Compared to actual recordings of ganglion cell RFs, the RF mosaics that our model produces appear less regularly spaced, see Fig. 1A and 4. The regularity in actual recordings could be due to the practice that for each RGC type, a separate retina is sampled and fitted with the elliptical parametric model. In contrast, our model produces a single "retina" for all RGC types which allows less degrees of freedom in fitting the parametric model, resulting in the parametric fitting to appear less regular.

4.5. Future work

Different options exist for extending and exploring our model. A thorough comparison with the denoising auto-encoder model would be of interest, since being trained on MNIST data and not on images of natural scenes, center surround RFs among other forms emerge (Vincent et al., 2008).

It would be possible to extend our model in the temporal domain together with constraining the transmission speed of the hidden units, in which faster transmission of information comes with a higher metabolic cost. Such a model would be trained with videos or short image sequences instead of static images. It is expected that such a model will converge in a more biologically plausible distribution of RGC types, in which units tuned to achromatic aspects will be of a small number but with larger RFs compared to units tuned to chromatic aspects. Instead of training our model on RGB image patches, our model can be extended to be trained with higher dimensional input, such as RGB-D data, which contains additional depth information or fMRI data.

Another path in extending our model is to include scotopic, low light and night, vision. It is completely unknown how such an extended auto-

encoder would learn these very divergent image statistics. The auto-encoder would have its entire daylight vision input near zero when a scotopic image patch is presented and conversely its night vision input entirely saturated when a daylight image patch is presented. Despite lacking exact knowledge of how *amacrine* cells mediate rod (night) and cone (day) input onto the same ganglion cells, the larger RFs size of achromatic RGC could also be hypothesized to be the result of being required to sample sufficient light of a large number of rods. This gives a different argument of why RFs sizes of achromatic RGC types are larger compared to color selective RGC types, than the argument of being the result of faster transmission of information at a higher metabolic cost.

And finally, our model could be possibly of use in simulating some aspects of color blindness since our model can be trained in various input modes (see Section 2.2.2), in which some simulate dichromatic vision. This might be of relevance since training our model with less than three input channels results, compared to training with RGB images, in the emergence of a lesser number of RGC channels.

5. Conclusion

Color opponent feature detectors have been manually constructed or derived statistically from RGB images (Yang et al., 2013), (Brown et al., 2011). The novelty in our work is the application of an auto-encoder network for directly learning optimal filters from natural images. Setting $x\pm$, in which the input is positive and negative, produces prototypes capturing *single-cone* chromatic contrast, reproducing the result of deriving optimal filters by means of statistical methods from RGB images of natural scenes (Brown et al., 2011). However, setting $x+$, in which the input is solely positive, produces prototypes capturing also *multi-cone* chromatic contrast, which is a novel property of computational models learning directly from RGB images.

Moreover, chromatic contrast information is essential for the task of border detection (Yang et al., 2013), which is the foundation of many basic tasks in computer vision, such as image segmentation and object recognition. Consequently, tasks which learn features from RGB images could benefit by preprocessing the visual information in a biologically inspired manner.

AppendixA.

AppendixA.1. Spatio-chromatic model

The excitatory center and inhibitory surround organization of retina ganglion cell RFs is commonly modeled with a Difference of Gaussians (DOG) function (Shapley and Hawken, 2011). In its simplest case, a circular two-dimensional Gaussian function has a center position μ_x μ_y, a radius r and an amplitude a. A circular DOG function, simply the sum of two Gaussians sharing the same center position, needs two more parameters: a surround radius r_s and the amplitude a_s of the surround part.

Moreover, the circular DOG function can be made elliptical which adds a rotation parameter and splits the radius variable into the spread of the two main axes of the ellipsoid. The parametric model needs to reconstruct RFs trained in a arbitrary *mode*, thus it has to be extended to at least three dimensions. Subsequently, the elliptical DOG model has a set of 13 parameters: The elliptical Gauss function G has center μ_x μ_y, spread σ_x σ_y and rotation θ.

$$\mathrm{G}(x,y) = e^{-\,\mathrm{a}(x-\mu_x)^2\,+\,2\,\mathrm{b}(x-\mu_x)(y-\mu_y)\,+\,\mathrm{c}(y-\mu_y)^2} \tag{A.1}$$

$$\mathrm{a} = \frac{1}{2\,\sigma_x^2}\cos^2(\theta) + \frac{1}{2\,\sigma_y^2}\sin^2(\theta) \tag{A.2}$$

$$\mathrm{b} = -\frac{1}{4\,\sigma_x^2}\sin(2\,\theta) + \frac{1}{4\,\sigma_y^2}\sin(2\,\theta) \tag{A.3}$$

$$\mathrm{c} = \frac{1}{2\,\sigma_x^2}\sin^2(\theta) + \frac{1}{2\,\sigma_y^2}\cos^2(\theta) \tag{A.4}$$

The DOG function has ratio of center to surround γ and scale of the surround part relative to the center part k_s.

$$\mathrm{DOG}(x,y) = \mathrm{G}(x,y) - k_s\,\gamma\;\mathrm{G}(x,y) \tag{A.5}$$

The chromatic part of the model with additional parameters bias b and direction d for each color channel:

$$\mathrm{DOG}_{rgb}(x, y) = [\mathrm{red}, \mathrm{green}, \mathrm{blue}] \tag{A.6}$$

$$\mathrm{red} = b_r + d_r\ \mathrm{DOG}(x, y) \tag{A.7}$$
$$\mathrm{green} = b_g + d_g\ \mathrm{DOG}(x, y) \tag{A.8}$$
$$\mathrm{blue} = b_b + d_b\ \mathrm{DOG}(x, y) \tag{A.9}$$

AppendixA.2. Parametric fitting

A single run of the fitting algorithm emits a set of parameters which reconstruct the given RF with the smallest error. To avoid solutions stuck in local minima, a sequence of fitting attempts is executed, each initialized with slightly varied start parameters. Subsequently, from this sequence, the best fit, with the smallest reconstruction error, is chosen. We are using the SLSQP (Sequential Least Squares Programming) algorithm of the *scipy.optimize* package to fit RFs to the parametric model.

AppendixA.3. Clustering receptive fields

Upon the fitted data, the receptive fields can be further analyzed: knowing the spatial position of the *center* of the ellipse, the value of its corresponding weight can trivially be extracted from the RF vector. The center weight, representing its particular receptive field, is used as an observation to be fed, along with the center weights of all other RFs, into a clustering algorithm. We are using the *k-means* algorithm of the *scipy.cluster.vq* package.

This is also true for a simple approximation of the values of the *antagonistic surround* of a receptive field: By extracting the weights at the four locations where the minor and the major axis of the ellipse cross the outer curve, a mean of the four weights gives a simple approximation. If used, the surround value together with the center value is concatenated into a single observation vector.

Interpreting the observations as color values the distribution of the RFs of a learned connection map are clustered over a color space. Moreover, the resulting clusters are ordered by their particular prototype color, the mean of all center weights belonging to a cluster, resulting in the familiar order of colors: red, green, blue, cyan, magenta, yellow. Clustering solely upon the center-weight value of a RFs avoids overfitting of the clustering algorithm.

By utilizing the results of the fitting and clustering process, the RFs of an original trained map can be ordered by their particular center positions and belonging to a specific cluster. Linearizing the center position $[x, y]$ of a parametric fit into a scalar $x + y * pw$, (pw is the pixel-size of an input patch), gives a partial order in which the RFs can be sorted spatially and grouped by their proximity to a specific cluster.

AppendixA.4. Prototype filters

A prototypical RF represents best the morphology of a particular cluster. It can be selected manually by specifying one RF for each channel. The resulting prototypical RFs can then be used in *convolutional neural networks* (CNN) for preprocessing RGB input.

References

Anishchenko, A., Greschner, M., Elstrott, J., Sher, A., Litke, A. M., Feller, M. B., Chichilnisky, E. J., 2010. Receptive field mosaics of retinal ganglion cells are established without visual experience. Journal of Neurophysiology 103 (4), 1856–1864.

Baden, T., Nikolaev, A., Esposti, F., Dreosti, E., Odermatt, B., Lagnado, L., 2014. A synaptic mechanism for temporal filtering of visual signals. PLOS Biology 12 (10), doi: 10.1371/journal.pbio.1001972.

Baldi, P. F., Homik, K., July 1995. Learning in linear neural networks: A survey. IEEE Transactions on Neural Networks 6 (4), 837–858.

Bell, A. J., Sejnowski, T. J., 1997. The 'independent components' of natural scenes are edge filters. Vision Research 37, 3327–3338.

Brainard, D. H., Brunt, W. A., Speigle, J. M., 1997. Color constancy in the nearly natural image. Optical Society of America 14 (9), 307–325.

Brown, M., Süsstrunk, S., Fua, P., 2011. Spatio-chromatic decorrelation by shift-invariant filtering. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 27–34.

Chalupa, L. M., Günhan, E., 2004. Development of on and off retinal pathways and retinogeniculate projections. Progress in Retinal and Eye Research 23 (1), 31–51.

Dacey, D. M., Petersen, M. R., 1992. Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. Proceedings National Acadamy of Science USA 89, 9666–9670.

cell class	count	%	assumed RGC Type
1. Colour-opponent concentric, (61%)			
M+/L-	95	20.79	Midget
M+/(S+L)-	4	0.88	
M-/L+	30	6.56	Midget
M-/(S+L)+	3	0.66	
L+/M-	76	16.63	Midget
L+/(S+M)-	3	0.66	
L-/M+	27	5.91	Midget
L-/(S+M)+	2	0.44	
(M+L)+/S-	6	1.31	Midget
(M+L)-/S+	10	2.19	Midget
S+/(M+L)-	17	3.72	Midget
S-/(M+L)+	4	0.88	Midget
2. Colour-opponent, non-concentric (2%)			
S+/(M+L)-	5	1.09	
S-/(M+L)+	1	0.22	
L+/M-	3	0.66	
3. Broad-band, non-opponent (24%)			
ON/OFF	69	15.1	Parasol
OFF/ON	41	8.97	Parasol
4. Broad-band, colour-opponent (4%)			
(M+L)+ /L-	11	2.41	Small Bistratifie
(M+L)+ /M-	4	0.88	Small Bistratifie
(S+M+L)+/M-	2	0.44	
(M+L)- /L+	2	0.44	Small Bistratifie
(M+L)- /M+	1	0.21	Small Bistratifie
5. Non-concentric. phasic (6%)			
ON	10	2.19	
OFF	3	0.66	
ON-OFF	14	3.06	
6. Non-concentric, motion-sensitive (3%)			
unidentified	14	3.06	

Table A.3: Retinal distribution of 460 RGC units, of which 457 could be classified, from the rhesus monkey retina. Data are from (Demonasterio and Gouras, 1975). For readability, we simplified the notation of the cell classes and give the cell counts and the percentages as sums over all eccentricities.

Demonasterio, F. M., Gouras, P., 1975. Functional properties of ganglion cells of the rhesus monkey retina. The Journal of Physiology, 167–196.

Doi, E., Gauthier, J., Field, G., Shlens, J., Sher, A., Greschner, M., Machado, T., Jepson, L., Mathieson, K., Gunning, D., Litke, A., Paninski, L., Chichilnisky, E. J., Simoncelli, E., 2012. Efficient coding of spatial information in the primate retina. The Journal of Neuroscience 32 (46), 16256–16264.

Elliott, T., Shadbolt, N. R., 1999. A neurotrophic model of the development of the retinogeniculocortical pathway induced by spontaneous retinal waves. The Journal of Neuroscience 19 (18), 7951–7970.

Field, G. D., Chichilnisky, E. J., 2007. Information processing in the primate retina: Circuitry and coding. Annual Review of Neuroscience 30, 1–30.

Field, G. D., Gauthier, J. L., Sher, A., Greschner, M., Machado, T., Jepson, L. H., Shlens, J., Gunning, D. E., Mathieson, K., Dabrowski, W., Paninski, L., Litke, A. M., , Chichilnisky, E., 2010. Functional connectivity in the retina at the resolution of photoreceptors. Nature 467 (10), 673–677.

Gegenfurtner, K. R., 2003. Cortical mechanisms of color vision. Neuroscience Nature Reviews 4, 563–572.

Gouras, P., 2009. Color Vision. Webvision: The Organization of the Retina and Visual System [Internet], http://www.ncbi.nlm.nih.gov/books/NBK11537/.

Kolb, H., 2004. How the retina works. American Scientist 91, 28–35.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (Book), 1097–1105.

Linden, R., 1993. Dendritic competition in the developing retina: ganglion cell density gradients and laterally displaced dendrites. Neuroscience 10 (2), 313–337.

Linden, R., Perry, V., 1982. Ganglion cell death within the developing retina: a regulatory role for retinal dendrites? Neuroscience 7 (11), 2813–2840.

Masland, R. H., 2001. The fundamental plan of the retina. Nature Neuroscience 4 (9), 877 – 886.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning, arXiv:1312.5602.

Nair, V., Hinton, G. E., 2010. Rectified linear units improve restricted Boltzmann machines. Proceedings of the 27th International Conference on Machine Learning, 807–814.

Nelson, R., 2007. Visual Responses of Ganglion Cells. Webvision: The Organization of the Retina and Visual System [Internet]. Salt Lake City (UT): University of Utah Health Sciences Center; 1995-., http://www.ncbi.nlm.nih.gov/books/NBK11550/.

Olshausen, B. A., Field, D. J., 1997. Sparse coding with an overcomplete basis set - a strategy employed by V1? Vision Research 37 (23), 3311–3325.

Purves, D., Lotto, R. B., Nundy, S., 2002. Why we see what we do. American Scientist 90 (3), 236–242.

Reese, B. E., 2012. Retinal mosaics: pattern formation driven by local interactions between homotypic neighbors. Frontiers in Neural Circiuts 6 (24), doi: 10.3389/fncir.2012.00024.

Shapley, R., Hawken, M. J., 2011. Color in the cortex: single- and double-opponent cells. Vision Research 51, 701–717.

Tian, N., 2004. Visual experience and maturation of retinal synaptic pathways. Vision Research 44 (28), 3307–3316.

Vincent, B. T., Baddeley, R. J., 2003. Synaptic energy efficiency in retinal processing. Vision Research 43, 1283–1290.

Vincent, B. T., Baddeley, R. J., Troscianko, T., Gilchrist, I. D., Jun-Sep 2005. Is the early visual system optimised to be energy efficient? Network: Computation in Neural Systems 16 (2), 175–190.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008, 1096–1103.

Weber, C., Triesch, J., 2008. A sparse generative model of V1 simple cells with intrinsic plasticity. Neural Computation 20, 1261–1284.

Weber, C., Triesch, J., 2009. Implementations and implications of foveated vision. Recent Patents on Computer Science 2009 (2), 75–85.

Wong, R. O., 1999. Retinal waves and visual system development. Annual review of neuroscience 22 (1), 29–47.

Yang, K., Gao, S., Li, C., Li, Y., 2013. Efficient color boundary detection with color-opponent mechanisms. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2810–2817.