Visual Recognition using a Combination of Shape and Color Features

Sepehr Jalali (tmssj@nus.edu.sg)
Cheston Tan (cheston-tan@i2r.a-star.edu.sg)
Joo-Hwee Lim (joohwee@i2r.a-star.edu.sg)
Jo-Yew Tham (jytham@i2r.a-star.edu.sg)
Sim-Heng Ong (eleongsh@nus.edu.sg)
Paul James Seekings (mmrl@nus.edu.sg)
Elizabeth A. Taylor (tmsohe@nus.edu.sg)

National University of Singapore, Singapore 119077
Institute for Infocomm Research, A*STAR,Singapore 138632

Abstract

We develop and implement a new approach to utilizing color information for object and scene recognition that is inspired by the characteristics of color- and object-selective neurons in the high-level inferotemporal cortex of the primate visual system. In our hierarchical model, we introduce a new dictionary of features representing visual information as quantized color blobs that preserve coarse, relative spatial information. We run this model on several datasets such as Caltech101, Outdoor Scenes and Underwater Images. The combination of our color features with (grayscale) shape features leads to significant increases in performance over shape or color features alone. Using our model, performance is significantly higher than using color naively, i.e. concatenating the channels of various color spaces. This indicates that usage of color information *per se* is not enough to produce good performance, and that it is specifically our biologically-inspired approach to color that results in significant improvement.

Keywords: Visual recognition; Color; HMAX; Biologically inspired; Visual cortex; Image classification

Introduction

Many models are inspired by the hierarchical organization of the visual cortex, such as Fukushima (1980) and Riesenhuber and Poggio (1999). Most of these models focus on grayscale information and ignore color information. While the broad use of color information in the primate visual system is well-known, the details are still under active investigation (Conway et al., 2010). Nonetheless, in this paper, we attempt to utilize what is currently known about the use of color to enhance object and scene recognition by computer algorithms. In this paper we utilize the HMAX model (Riesenhuber & Poggio, 1999), but this approach can be extended to other computational models.

In our experiments, we use the HMAX model (Riesenhuber & Poggio, 1999) in concatenation with our color model in order to evaluate the use of both shape and color. HMAX is a biologically-inspired model which focuses on the shape processing capabilities of the ventral visual pathway, and has been used to perform classification tasks (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007).

We focus on extending the model by modelling the high-level usage of color by incorporating insights from cognitive psychology and neuroscience. The broad intuitive inspiration for our model follows from the fact that colors are recognized categorically just as object classes are, even though color discrimination and matching is continuous (Palmer, 1999). Interestingly, people of different races (Boynton & Olson, 1987), as well as chimpanzees (Matuzawa, 1985), organize colors into the same basic color categories, such as red, blue, yellow, green.

More importantly for object and scene recognition, the categorical recognition of color suggests that, if color information is incorporated into object and scene classification, then fine-grained color information (e.g. precisely specified hue) may not be necessary. For example, a beach scene might be recognized from the blue (sky and sea) and brown (sand) regions. It may not be important exactly how blue the sky/sea or how brown the sand grains are. In fact, it may be important to disregard such details in order to perform classification that is tolerant to variations in lighting, and so on.

In addition, the coarse relative spatial position of such color regions may be important. A blue region above a yellow-brown region might suggest a beach scene. If the relative positions are reversed, then the image is probably not a beach scene (or might be an upside-down one). Not only is the detailed spatial information unnecessary, it may be crucial to discard it and only retain coarse spatial information, since the exact spatial relations will depend on factors such as the precise shape of the beach and the camera angle.

Overall, our model can be loosely described as performing object and scene classification by reducing a given image to a "coarse arrangement of categorical color blobs", similar to the idea of spatial aggregation of visual keywords (Lim, 1999), but with realization on the HMAX model. This is different from approaches that utilize color information in a low-level fashion, although the two types of approaches are not mutually exclusive. Crucially, our biologically-inspired approach outperforms the naive use of color, where an image is decomposed into separate color channels that are processed independently until the final classifier stage.

Related Work

First, we go beyond the intuitive motivation for our approach and review the biological evidence that the primate visual system utilizes color information in a manner that is broadly consistent with our model. Specifically, we review studies of color processing in the high-level visual area of the primate

brain known as infero-temporal cortex (IT), which is commonly associated with invariant object recognition.

In the broadest terms, IT is known to play an important role in color discrimination. A majority of IT neurons are color-selective (Desimone, Schein, Moran, & Ungerleider, 1985) and two independent studies estimated this proportion to be roughly 70% (Komatsu, Ideura, Kaji, & Yamane, 1992; Edwards, Xiao, Keysers, Földiák, & Perrett, 2003). Contrary to the theory that color processing occurs after more rapid luminance-only processing, no evidence was found that colored images evoke responses that are delayed relative to achromatic images (Edwards et al., 2003). More direct evidence for the role of IT comes from findings that color discrimination is severely disrupted by lesions (Heywood, Shields, & Cowey, 1988) or cooling (Horel, 1994).

Color-selective neurons in IT are found in clusters, suggesting that they may form a segregated and independent processing network (Conway, Moeller, & Tsao, 2007). As further evidence of this, one color cluster in IT received projections from a color cluster from another part of IT, suggesting that these clusters of color-processing neurons form reciprocally-connected modules within a distributed network (Banno, Ichinohe, Rockland, & Komatsu, 2011).

IT neurons are selective for both hue and saturation (Komatsu, 1993). Different cells have different preferred hues, and as a population, the cells' preferred color spans most of the color space (Conway et al., 2007). The colors for which IT neurons are selective for tend to correspond to the basic color names (Komatsu, 1998). Komatsu (1998) proposed that IT has templates corresponding to color categories and may be involved in determining color category by finding the best match over these categories. More recently, the distribution of color-selective neurons found in IT seems to correspond to the three to four most basic colors (Stoughton & Conway, 2008). The largest peaks align with red, green, and blue, in order of size of peak, with a smaller peak corresponding to yellow. These peaks roughly correspond to colors perceived by humans. Prior to this, neural representation of such unique hues (Hurvich, 1981) had not been found (Valberg, 2001). Note that in the low-level primary visual cortex, the axes defined by cone opponency should more accurately be denoted bluish-red/cyan and lavender/lime opponency (Stoughton & Conway, 2008), rather than red-green and blue-yellow opponency.

Finally, the region of IT where color-selective neurons are found is coarsely retinotopic (Yasuda, Banno, & Komatsu, 2010), meaning that spatial information is maintained in a coarse manner, rather than completely discarded or maintained with high fidelity. Overall, these studies are broadly consistent with our proposed "coarse arrangement of categorical color blobs" model of high-level color processing in the primate visual system.

In contrast, most computer vision algorithms utilize color information in a relatively low-level manner. The simplest color extension of a non-color algorithm would be to apply it independently to the R, G and B channels, and then concatenate the features from all 3 channels just before the final classifier stage. Most algorithms are variants of this basic idea, either using some other color space, or fusing the channels before the classifier stage (usually at the dictionary or keyword learning stage). For example, SIFT features can be computed separately for each channel in HSV color space (Bosch, Zisserman, & Muñoz, 2008), while Brown and Susstrunk (2011) do this for RGB space, along with an NIR (near infra-red) channel. Besides SIFT features, other algorithms use (non-orientation based) histograms in the HSV (Tang, Miller, Singh, & Abbeel, 2012), Gaussian opponent color (Burghouts & Geusebroek, 2009), normalized RGB or opponent color spaces (Gevers & Stokman, 2004). What these algorithms have in common is that in terms of the biology of color vision, they correspond to at most the level of color-opponent cells in the primary visual cortex, the lowest level in the hierarchically-organized visual cortex.

CQ-HMAX

In this section, we describe our new biologically-inspired model, CQ-HMAX (Color Quantization Hierarchical Max), which uses color information in a hierarchical organization of simple and complex cells. HMAX is a hierarchical model which uses Gabor filters to find simple and complex shapes in the images. Our model has a similar hierarchical structure. However, we use color quantization cores and not Gabor filters, hence our model encodes color information. When combined with HMAX, the overall model includes both color and shape information.

Our color model has a hierarchical structure of simple and complex cells, as can be seen in Fig. 1. We first introduce the model briefly, followed by a more detailed description of each layer. An image pyramid is created in YIQ color space. The Y channel represents luminance information, while the I and Q channels represent chrominance information. The pyramid has 10 scales, with each neighboring scale different by a ratio of $1/(2^{1/4})$. In order to evaluate the use of color information in our model, we determined that the YIQ color space produced the best results in comparison with HSV and RGB color spaces. A set of representative values from each color channel is selected as color cores and used to find the best matching unit to each individual pixel value in the pyramid. The $S1$ layer is created on 10 scales indicating the index of the best matching YIQ core to each pixel in the image pyramids. At the $C1$ layer, a local max pooling is computed over $\pm 10\%$ spatial neighborhoods of approximately 6×6 on ± 1 neighbor scales to find the most frequent color core in each neighborhood. A dictionary of features is sampled randomly from the $C1$ layer of images. The distance of each dictionary feature to all patches in a neighborhood of that dictionary feature is calculated to create the $S2$ layer and the best response to each dictionary feature in each image is chosen as the $C2$ layer to be fed to the SVM layer for classification. We describe each layer in more detail below.

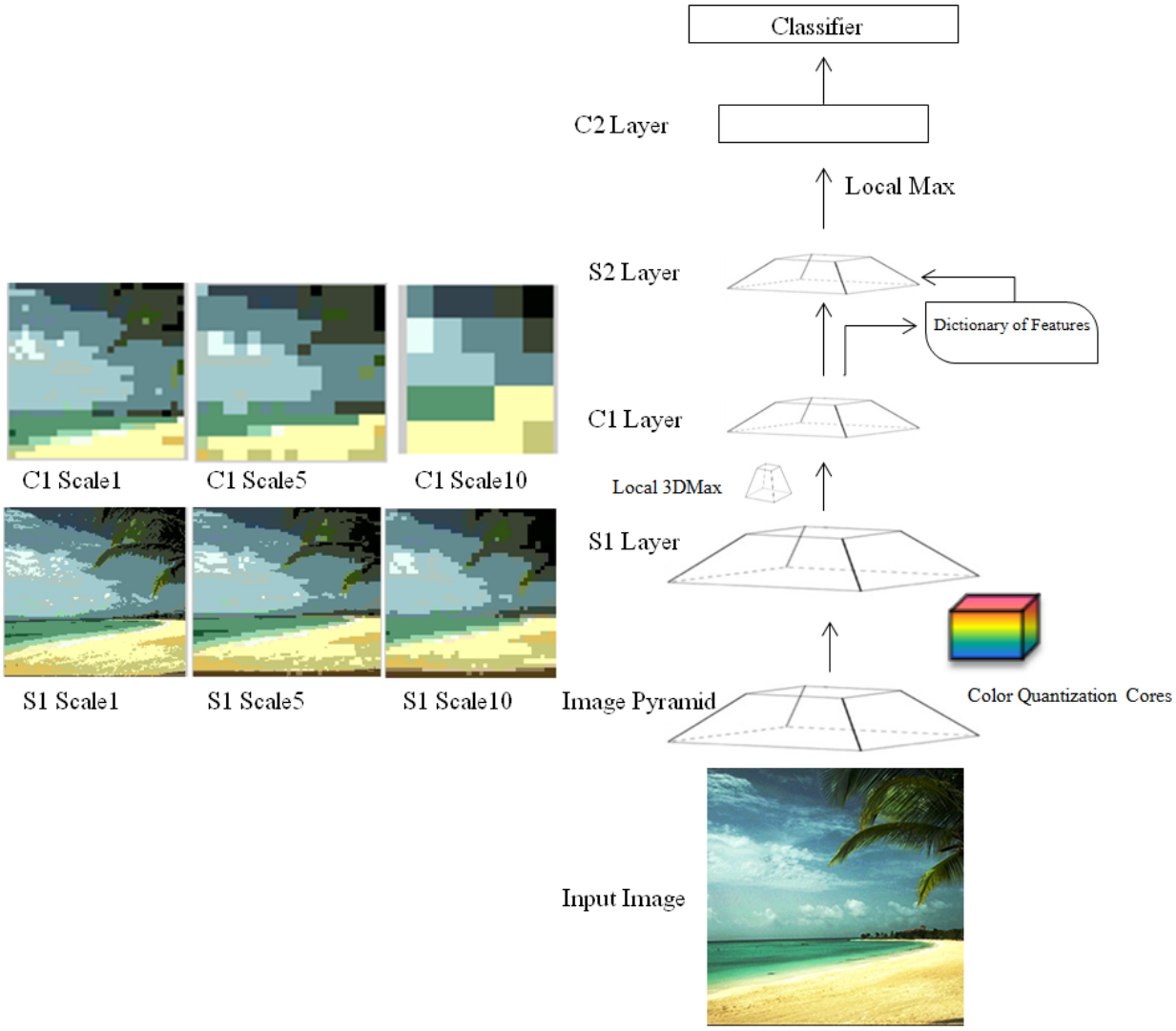

Figure 1: The CQ-HMAX model and the processing of an example beach image.

$S1$ Layer and Quantization Cores

The input images are first converted into YIQ color space and a pyramid of 10 scales with a ratio of $2^{1/4}$ is created, with the first scale having the shorter side set to 140 pixels, maintaining the aspect ratio of the original image. This image pyramid is then used as the input to the $S1$ layer. A series of YIQ quantized "color cores" over YIQ channels are created to be used as filters for this layer. We experimented with different numbers of quantization values per color channel, and chose 5 per channel as the optimal number (which results in $5 \times 5 \times 5 = 125$ cores). In order to choose the optimal cores, 500 images were randomly selected and the color range of these images in YIQ color space was calculated after normalization to the range $[0,1]$. The values of YIQ channel are mostly in the range $[0,1]$, $[0.4,0.7]$ and $[0.4,0.6]$ respectively. These ranges were selected and divided into 5 bins. The quantized values of Y, I and Q after normalization to $[0,1]$ were therefore chosen as follows: $Y = (0,0.25,0.5,0.75,1)$, $I = (0.4,0.47,0.55,0.63,0.7)$, $Q = (0.4,0.47,0.5,0.53,0.6)$. Using these values results in better classification performance than using the full range $[0,1]$ in each YIQ channel. The outputs at the $S1$ layer are the index values (i.e. 1,2,...,125) of the best-matching color core for each element in the image pyramid.

$C1$ Layer

The $C1$ layer provides local invariance to position and scale as it pools nearby $S1$ units, and as a result, subsamples $S1$ to reduce the number of units. The $S1$ pyramid is convolved with a $3D$ max filter to set the $C1$ layer size of the bottom of the pyramid to 25×25 and the highest layer of the pyramid to 5×5 accordingly. The max is calculated over $\pm 10\%$ spatial neighborhood on ± 1 neighbor scales in the middle of the pyramid and -2 on the highest level and $+2$ on the lowest layer of the pyramid (hence it is called a $3D$ max, as it takes the max over $2D$ spatial distribution and over ± 1 scale).

This layer provides a model for $V1$ complex cells. Fig. 1 also shows an example image of $S1$ and $C1$ layer. $S1$ and $C1$ layers have a distribution of quantization cores from coarse to fine. The higher layers of the $S1$ pyramid are taken from smaller scales of the images in the input pyramid and respectively the higher levels of $C1$ layer are computed by taking a $3D$ max over higher levels of $S1$ layer. As can be seen in Fig. 1, the higher levels of the pyramid in the $S1$ and $C1$ layers represent less detailed information from the image. All levels in the $C1$ intermediate layer are used for sampling a dictionary of features.

Dictionary of Features and Distance Table

Once the $C1$ layer is created, sampling is performed by centering patches of size 4×4 at random positions and scales using a normalized random number generator function. A distance table is created to store the actual weighted Euclidean distances of the indices from YIQ quantization cores. Since the values of the Y channel are normally distributed between $[0,1]$, but the values of I and Q channels fall in the approximate range of $[-0.6,+0.6]$ and $[-0.5,+0.5]$ respectively, and as in most of the images the actual values of these two latter channels fall between $[-0.1,+0.2]$ and $[-0.1,+0.1]$ (before normalization to $[0,1]$) we weighed the distances to have an equal effect in the distance calculation. The distance table weights are calculated as:

$$DistanceTable(i,j) = \sqrt{D(1) + \gamma D(2) + \beta D(3)}$$
$$\text{Where } D(k) = (YIQCore(i,k) - YIQCore(j,k))^2 \quad (1)$$

with $\gamma = 3.3$ and $\beta = 5$. In Jalali, Lim, Ong, and Tham (2010) and Jalali, Lim, Tham, and Ong (2012) various clustering methods in the creation of the dictionary of features were implemented and it is shown that by use of random sampling in HMAX model, relatively good results can be achieved with a lower computational cost in comparison with clustering of features.

$S2$ Layer

Once the dictionary of features and the distance table are created, each entry in the dictionary of features is used as a filter to be convolved on $C1$ patches of size 4×4 on the neighbor scales of the dictionary feature in the pyramid. The responses $V(d,p)$ of each dictionary feature, d to all of the neighbor patches of the same size in ± 1 scale and $\pm 10\%$ in position, p are calculated using a Euclidean distance equation as:

$$V(d,p) = \exp\left(-\frac{\| d-p \|^2}{2\sigma^2\alpha}\right) \quad (2)$$

where d is a feature in the dictionary and p is a patch in the image $C1$ pyramid. σ and α are set to 0.5 and 1 respectively as in Mutch, Knoblich, and Poggio (2010).

$C2$ Layer

Once the $S2$ layer is generated, the maximum values for each patch in the dictionary are taken as the $C2$ output. This layer outputs a vector of the same size as the dictionary of features. We chose different sizes for the dictionary of features and in most cases a dictionary of size 10000 was chosen which results in slightly better performances than smaller sizes of about 1000 dimensions.

Classification Layer

The $C2$ vectors are classified using a multi-class one-versus-rest linear kernel support vector machine. The algorithm used to train the classifier is weighted regularized least-squares after the data is sphered and the mean and variance of each dimension are normalized to zero and one respectively as in Mutch and Lowe (2008).

Use of HMAX for Encoding Shape Information

For shape information, we used the HMAX model implementation of Mutch and Lowe (2008). In HMAX, the maximum response of the $S2$ layer is chosen as the $C2$ layer to be fed to the classifier. An N-dimensional vector is calculated as the output of the $C2$ layer, where each element is the maximum response (everywhere in the image in Serre, Oliva, and Poggio (2007) and in a spatial neighborhood of each dictionary feature in Mutch and Lowe (2008)) over image patches for each dictionary feature where N is the number of features in the dictionary.

Let V_i^j be the response of the image patch p_i to the dictionary feature d_j calculated using Eq. 2. The response of the $C2$ layer is calculated as:

$$C2(j) = \max(V_i^j) \text{ for } \forall i \in M \text{ for } j = 1,...,N \quad (3)$$

where M is the number of valid patches in each image and N is the size of the dictionary of features. This is consistent with the recent HMAX models (Mutch & Lowe, 2008; Serre, Oliva, & Poggio, 2007; Jalali et al., 2010; Theriault, Thome, & Cord, 2011).

Experimental Results

First we examine the naive use of color by computing various color spaces (RGB, HSV, YIQ) on the Caltech101 dataset (Fei-Fei, Fergus, & Perona, 2004) and compare the results with grayscale images. The Caltech 101 dataset, includes 101 classes of objects plus a background category. Each class contains between 31 to 800 color images of different sizes. The size of each image is approximately 300×200 pixels on average. We used 30 randomly chosen images for training from each class and the rest of the images were used in the test phase. We first divide the images into three channels and feed them to the unmodified HMAX (Mutch & Lowe, 2008) directly and evaluate the classification performance.

As can be seen in Table 1, the use of three different channels and concatenating the $C2$ vectors of all channels to the SVM provides only marginal improvement. Since the YIQ color space gives the best overall results, we use this color space in our color model. In the rest of this section, we

Color Component	Caltech101	Scenes
Y channel (i.e. gray scale)	54.65	71.48
I channel	35.20	54.62
Q channel	26.86	50.75
YIQ channels concatenated	55.06	72.66
RGB channels concatenated	26.53	73.81
HSV channels concatenated	31.32	73.69

Table 1: Results (percentage accuracy) for the naive use of various color channels and color spaces.

evaluated our model on three datasets: Caltech101, Outdoor Scenes and Underwater Images.

Caltech101 Dataset

The results of using CQ-HMAX on Caltech 101 are shown in Table 2. All experiments are performed 8 times on random splits of training and test sets and the average performance is reported. As can be seen, the use of our color model in this dataset does not outperform HMAX. However, when the $C2$ features of the color model are concatenated with $C2$ features of HMAX, the classification results are improved by more than 6% over HMAX alone. HMAX is a computationally expensive model as Gabor filter responses over different orientations in $S1$ layer are calculated. However, CQ-HMAX is relatively faster than HMAX as it performs a quantization with 125 cores in the $S1$ layer instead of Gabor filters.

Model	Caltech101	Scenes	UWI
HMAX (i.e. shape)	54.65	71.48	92.93
CQ-HMAX (i.e. color)	38.11	69.21	94.03
CQ-HMAX + HMAX	**61.09**	**78.97**	**96.23**

Table 2: Results (percentage accuracy) on the Caltech101, Outdoor Scenes and Underwater Images (UWI) datasets.

Outdoor Scenes Dataset

This dataset contains 8 outdoor scene categories: *coast, mountain, forest, open country, street, inside city, tall buildings and highways* (Oliva & Torralba, 2001). There are 2600 color images of size 256×256 pixels. We used 100 random images per category for training and the rest (236 on average) for testing. As can be seen in Table 2, the combination of shape and color significantly improves performance.

Underwater Images Dataset

We also evaluated CQ-HMAX on the Underwater Images dataset (Jalali, Tan, Lim, Tham, Ong, Seekings, & Taylor, 2013). This dataset is made of 1664 images of around 740 x 420 pixels from 13 different categories. We used 30 randomly selected images per category for training and the rest for testing. These underwater images contain small objects of various shapes and color against a varied seabed background. The main challenge with these images is in light absorption by the water, and the existence of particles that limit visibility and result in scattering and reflection of light. In this experiment, we created a set of images using both grayscale and color cameras and compared the performance of CQ-HMAX on color images and HMAX on grayscale images. As seen in Table 2, the classification accuracy increases when color and shape information are combined.

Conclusions

In this paper, we introduced a new biologically-inspired approach to image classification which uses color in a manner consistent with high-level visual cortex processing by incorporating insights from cognitive psychology and neuroscience. We ran this model on several datasets such as Caltech101, Outdoor Scenes and Underwater Images. The combination of our color features with (grayscale) shape features led to significant increases in performance over shape or color features alone. Using our model, performance is significantly higher than using color naively, i.e. concatenating the channels of various color spaces.

Currently, our model quantizes the YIQ color space into cubed-shaped "color cores" at the $S1$ layer. Following the work of Shahbaz Khan et al. (2012) and Van De Weijer and Schmid (2006), learning the color values that correspond to semantic color names such as "orange", "brown", could also further improve performance. Alternatively, color cores can be learnt through unsupervised clustering, in which more frequent colors in each dataset are chosen as color cores.

Our model emulates color processing in the high-level IT cortex. Interestingly, the combination of our features with those of Zhang, Barhomi, and Serre (2012) – a biologically-inspired model that emulates the lower-level cortex – results in classification performance as good (or better) than the state-of-the-art on several benchmark datasets (Jalali, Tan, Lim, Tham, & Ong, 2013).

References

Banno, T., Ichinohe, N., Rockland, K. S., & Komatsu, H. (2011). Reciprocal connectivity of identified color-processing modules in the monkey inferior temporal cortex. *Cerebral Cortex*, *21*(6), 1295–310.

Bosch, A., Zisserman, A., & Muñoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(4), 712–27.

Boynton, R. M., & Olson, C. X. (1987). Locating basic colors in the OSA space. *Color Research & Application*, *12*(2), 94–105.

Brown, M., & Susstrunk, S. (2011). Multi-spectral SIFT for scene category recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 177–184). IEEE.

Burghouts, G. J., & Geusebroek, J.-M. (2009). Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, *113*(1), 48–62.

Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., et al. (2010). Advances in color science: from retina to behavior. *Journal of Neuroscience*, *30*(45), 14955–63.

Conway, B. R., Moeller, S., & Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron*, *56*(3), 560–73.

Desimone, R., Schein, S. J., Moran, J., & Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research*, *25*(3), 441–52.

Edwards, R., Xiao, D., Keysers, C., Földiák, P., & Perrett, D. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *Journal of Neurophysiology*, *90*(2), 1245–56.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *IEEE CVPR 2004 Workshop on Generative-Model Based Vision* (Vol. 2).

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Gevers, T., & Stokman, H. (2004). Robust histogram construction from color invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(1), 113–118.

Heywood, C. A., Shields, C., & Cowey, A. (1988). The involvement of the temporal lobes in colour discrimination. *Experimental Brain Research.*, *71*(2), 437–41.

Horel, J. A. (1994). Retrieval of color and form during suppression of temporal cortex with cold. *Behavioural Brain Research*, *65*(2), 165–72.

Hurvich, L. M. (1981). *Color Vision*. Sutherland, MA.: Sinauer Associates Inc.

Jalali, S., Lim, J., Ong, S., & Tham, J. (2010). Dictionary of features in a biologically inspired approach to image classification. In *Internationl Conference on Neural Information Processing (ICONIP)* (pp. 541–548). Springer.

Jalali, S., Lim, J., Tham, J., & Ong, S. (2012). Clustering and use of spatial and frequency information in a biologically inspired approach to image classification. In *International Joint Conference on Neural Networks* (pp. 1–8).

Jalali, S., Tan, C., Lim, J., Tham, J., & Ong, S. (2013). CQ-HMAX: A New Biologically Inspired Color Approach to Image Classification. *(Manuscript under preparation)*.

Jalali, S., Tan, C., Lim, J., Tham, J., Ong, S., Seekings, P., et al. (2013). Encoding Co-occurrence of Features in the HMAX Model. In *Proceedings of the Annual Conference of the Cognitive Science Society.*

Komatsu, H. (1993). Neural coding of color and form in the inferior temporal cortex of the monkey. *Biomedical Research*, *14*, 7–13.

Komatsu, H. (1998). Mechanisms of central color vision. *Current Opinion in Neurobiology*, *8*(4), 503–8.

Komatsu, H., Ideura, Y., Kaji, S., & Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience*, *12*(2), 408–24.

Lim, J. (1999). Learning visual keywords for content-based retrieval. In *International conference on multimedia computing and systems* (Vol. 2, pp. 169–173).

Matuzawa, T. (1985). Colour naming and classification in a chimpanzee. *Journal of Human Evolution*, *14*, 283 – 291.

Mutch, J., Knoblich, U., & Poggio, T. (2010). *CNS: a GPU-based framework for simulating cortically-organized networks* (Tech. Rep. No. MIT-CSAIL-TR-2010-013 / CBCL-286). Cambridge, MA: Massachusetts Institute of Technology.

Mutch, J., & Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision*, *80*(1), 45–57.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *29*(3), 411–426.

Shahbaz Khan, F., Anwer, R. M., Weijer, J. Van de, Bagdanov, A. D., Vanrell, M., & Lopez, A. M. (2012). Color attributes for object detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3306–3313).

Stoughton, C. M., & Conway, B. R. (2008). Neural basis for unique hues. *Current Biology*, *18*(16), R698–9.

Tang, J., Miller, S., Singh, A., & Abbeel, P. (2012). A textured object recognition pipeline for color and depth image data. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3467–3474).

Theriault, C., Thome, N., & Cord, M. (2011). HMAX-S: deep scale representation for biologically inspired image classification. In *International Conference on Image Processing* (pp. 3–6).

Valberg, A. (2001). Unique hues: an old problem for a new generation. *Vision Research*, *41*(13), 1645–57.

Van De Weijer, J., & Schmid, C. (2006). Coloring local feature extraction. *European Conference on Computer Vision (ECCV)*, 334–348.

Yasuda, M., Banno, T., & Komatsu, H. (2010). Color selectivity of neurons in the posterior inferior temporal cortex of the macaque monkey. *Cerebral Cortex*, *20*(7), 1630–46.

Zhang, J., Barhomi, Y., & Serre, T. (2012). A new biologically inspired color image descriptor. In *European Conference on Computer Vision* (Vol. 7576, pp. 312–324). Springer.