Learning Symbolic Names for Perceived Colors*

Johan M. Lammens & Stuart C. Shapiro
Computer Science Department
226 Bell Hall, State University of New York at Buffalo
Buffalo, NY 14260
e-mail {lammens|shapiro}@cs.buffalo.edu
(To appear in proc. of AAAI-93 Fall Symposia)

1 Introduction

We are working on a computational model of color perception and color naming, which can be seen as an instance of symbol grounding [4] in the domain of color, or as an attempt to provide an artificial intelligent agent with embodied concepts of color. This effort is part of a larger one in the field of intelligent autonomous agents [8, 5], and will provide a perceptual grounding for some of the symbolic representations in the SNePS Knowledge Representation formalism [13], especially those referred to as "sensory nodes" [12]. The implemented model will allow an agent to *name* colors in its environment, *point out* examples of named colors in its environment, and *learn* new names for colors. Our research draws on work in the neurophysiology of color perception, particularly [3], in semantic universals for natural languages, particularly [1], and other work in AI and Cognitive Science.

We discuss two areas of the model where learning is used: learning a non-linear mapping between two color spaces, and learning a relation between color space coordinates and a set of symbolic color names. We have used a traditional error back-propagation learning algorithm for the first problem, and are considering several different learning paradigms for the second problem, ranging from traditional clustering techniques to an experimental "space warp" method. Using learning gives us a relatively easy way to determine a non-linear transformation of spaces in the first case, and increases the flexibility of the approach with respect to different application needs (and languages) in the second case. We discuss the learning methods used or considered and the problems encountered.

In general terms, our model has to explain (and reproduce) a signal-to-symbol transition, going from light entering a sensor to symbols representing the corresponding perceived color. To make our problem manageable, we make some simplifying assumptions. We are only concerned with single-point determination of color, thus disregarding spatial interactions in color perception. We only take context into account to the extent that it is necessary for this determination. We assume foveal cone photoreceptors as sensors, and we restrict the problem to any given fixed state of adaptation of the vision system, thus avoiding issues of color constancy [2, 14]. We are developing a physical implementation of an agent, using a color camera for sensing device and a robot arm for pointing device [8], and there we will deal with these issues to some extent, but that does not concern us here. Conceptually, we break the model up into two parts. The first part takes us from the visual stimulus to color space coordinates, the second from color space coordinates to a set of color names.

2 Learning a color space transform

We hypothesize that the nature of the color perception mechanism underlying human color naming determines to some extent the existence and the nature of the semantic universals of color (this is what we refer to as the *embodiment* of color concepts), hence we want to use a color space which is based on neurophysiology. Based on data published in [3], we have reconstructed 3D models of the response of 4 types of color-opponent cells and 2 types of non-opponent cells in the Macaque LGN.[1] Response is a function of both spectral composition and radiance of the stimulus. For each of the 6 cell types and each of the 12 sampled wavelengths in [3], we fitted sigmoid functions of input radiance (in log units) to the 3 data points given, minimizing the RMS error of fit:

$$S(x) = (1 + e^{\frac{h-x}{t}})^{-1}, t \geq 0.1 \quad (1)$$

The constraint on t is to prevent the slope of the sigmoid function from becoming too steep, which sometimes happened when using Mathematica's FindMinimum function to do the fit. The parameter h represents the half-response radiance, and is optimized together with t. We assumed a total response range of 3 log units from threshold response to saturation, which accords well with known data on photoreceptors [9]. We further estimated the maximum stimulus radiance used in [3] to be about 1 log unit below saturation. We added zero-radiance data points to constrain the interpolation better. To arrive at the response for an arbitrary radiance and wavelength, we constructed arrays of these sigmoid functions indexed by wavelength, and interpolated between seven values given by the functions closest

*This work was supported in part by Equipment Grant No. EDUD-US-932022 from SUN Microsystems Computer Corporation.

[1] The lateral geniculate nucleus, a body of cells midway between the retina and the primary visual cortex. The Macaque visual system is very similar to that of humans.

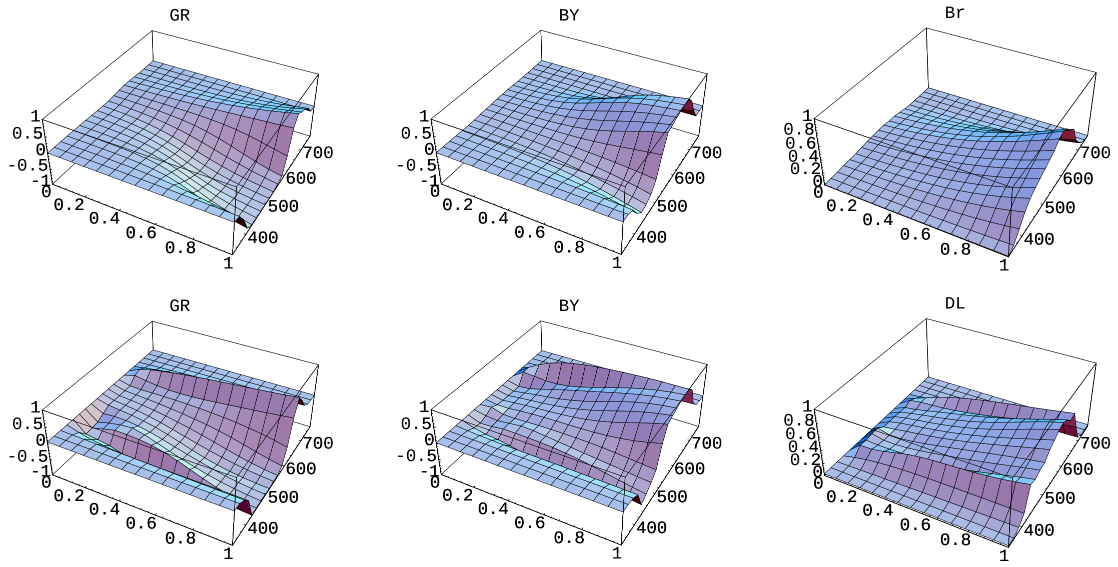

Figure 1: Basis functions for the Valois color space: directly modeled on physiological data (upper), and the result of a learned XYZ to Valois transform (lower). From left to right: Green-Red (GR), Blue-Yellow (BY), and Brightness (Br). X axis: radiance in log units relative to the threshold response; Y axis: wavelength; Z axis: activation.

to the desired wavelength (using Mathematica's standard 3rd order polynomial Interpolation). This gives us the 6 reconstructed neural response functions. By pairwise adding these functions, we have derived new 3D response functions that we take as the basis of an opponent color space: Green-Red and Blue-Yellow opponent functions and a non-opponent Brightness function (Fig. 1, upper half). We will refer to this newly defined space as the Valois color space. We have reasons to believe that existing opponent color models as used in computer vision and computer graphics (for instance [6]) are not accurate as models of color perception, but we will not discuss that here.

Most work in Color Science [14] and color computer vision is based on the CIE XYZ color space and derivatives thereof, so we want a transformation from XYZ coordinates to Valois space. We also need this to transform RGB camera output to Valois coordinates. Since we could not determine a linear transform between XYZ and Valois which has a reasonable margin of error, we used the error back-propagation algorithm to determine such a transform, as commonly used in artificial neural networks research [11]. In contrast with many applications of the backpropagation technique we did not use the network as a classifier, but rather to learn three simultaneous functions of three real variables. To construct a training set we made a regular grid of wavelengths and radiances and computed both XYZ and Valois coordinates for each point. The XYZ coordinates functioned as inputs, the Valois coordinates as teaching outputs. We normalized both the XYZ and the Valois coordinates to the $[0,1]$ range, which requires adding an offset of 0.5 to the latter, since the original range comprises both negative and positive numbers. The order of presentation of the training set was not varied. The training set was linearized by concatenating segments corresponding to fixed radiances, in increasing order of radiance. Some optimizations were done for momentum, temperature, and learning rate parameters. We experimented with several network topologies: one, two, and three layer with varying numbers of nodes in the hidden layers. We settled on a $3 \times 9 \times 9 \times 3$ topology, loosely based on Kolmogorov's result (described in [10]) that states that a three layer net with $N(2N+1)$ nodes (not counting the input nodes) using continuously increasing non-linearities can compute any continuous function of N variables, hoping that the same number of nodes would suffice for three simultaneous functions of three variables. After a large number of iterations (up to 260,000 for a learning rate of 0.03 and a training set of 473 I/O vector pairs), this net converged to a reasonable RMS error of about 3% of the output range over the complete training set. The transformation reduces to an equation of the form

$$\mathbf{o} = s(\mathbf{M_3} \cdot (s(\mathbf{M_2} \cdot (s(\mathbf{M_1} \cdot \mathbf{i}))))) \qquad (2)$$

where $\mathbf{M}_j$ represents the weight matrix for stage j, $\mathbf{i}$ represents the vector of XYZ input coordinates, $\mathbf{o}$ the vector of Valois output coordinates, and s represents the usual sigmoid activation function, essentially the same as equation 1. The result is shown in the lower half of Figure 1.

Using the error back-propagation algorithm to learn the transform is convenient since it is hard to determine non-linear transforms in general, but it is not easy to determine the right morphology and other parameters of the network for the particular problem at hand. There was considerable trial and error involved in getting the net to converge to a reasonable error level. The data representation is also important in this respect. The RMS error measure is probably not the best possible one to use to measure the goodness of fit, but we don't know of a more useful one. As long as the weight matrices are square, the transform is invertible by computing the inverse of the sigmoid activation function and of the weight matrices, which then amount to linear transforms (this is not the case with the $3 \times 9 \times 9 \times 3$ topology we used).

Now that we have defined the Valois space and a mapping into it from CIE XYZ coordinates, we can study its properties with respect to the existing body of color perception research, in addition to using it as a basis for the naming part of our model. We speculate that the Valois space is a good candidate for systematizing and/or explaining experimental color perception results, but at this point we have not yet investigated this any further. We can derive measures for the psychophysical variables of hue, saturation, and brightness from the Valois space.

3 Learning color names

For the second part of our model, relating Valois coordinates to a set of color names, we are experimenting with a different kind of learning. Basically we need to partition a 3D space into a set of volumes or regions, each corresponding to a named color concept (or category). From a model-theoretical (logical) point of view, we can consider these regions to be the extensional referents of the color terms involved. From the work of [1] and others we know that there exists a set of semantic universals in the domain of color, known as basic color categories (BCC), each corresponding to a "basic color term" (BCT). We restrict our attention to these. According to [1], they are categories of *perception* rather than of *language*. Different languages may have lexicalized different numbers of these categories as basic color terms, but nevertheless they are equally distinguishable and identifyable to speakers of any language. More precisely, the *foci* or "best examples" of the categories are universal, but there is some disagreement about whether or not the extent (size) of the regions are the same across languages. They speculate that these categories reflect certain properties of the human color perception mechanism, which are shared by all people regardless of language. We want our naming algorithm to learn the BCTs of a set of different languages (one at the time) corresponding to (a subset of) the universal BCCs. Depending on the requirements we impose on the shape of the BCC regions of our color space, we can create binary valued or continuous valued "characteristic functions" for the regions, and have the regions overlap or not. We could interpret the BCCs as fuzzy sets [15, 7], for instance. From the work of [1] and others we know that human BCC characteristic functions are continuous (or at least non-binary) valued, and that the regions may or may not overlap, perhaps depending on the particular experimental paradigm used.

From the literature (and partly from measurements) we have estimates of the coordinates of the BCC foci in CIE XYZ space, and using the transform from Section 2 we can transform them into Valois space coordinates. There are several learning approaches we could use to learn the relation between Valois coordinates and BCTs. We could use error back-propagation again, but we have chosen not to do so, because we consider other approaches more revealing in this case. An alternative might be to use well-known clustering techniques, i.e. determining some kind of center of mass of a number of input data points representing examples of a particular category. Our work is still in an exploratory stage in this area, but one of the approaches we favor is the following. Assuming that there are indeed particular properties of the human color perception mechanism responsible for the semantic universals, we are hoping that our Valois color space captures those properties well enough to enable us to use them in a name learning algorithm which would (very schematically) work as follows: 1) determine the Valois coordinates of the stimulus; 2) find the nearest "distinguished location" in the space, 3) attach the given name to that location. A naming algorithm would work in a similar way: 1) determine the Valois coordinates of the stimulus; 2) determine the nearest named distinguished location; 3) return the name associated with it. A "distinguished location" in the color space is one that stands out in one way or another, in terms of the shape or properties of the space itself, not in terms of the stimulus. There are two possible sources of constraints on the location of BCTs: environmental and physiological. The environment determines which surfaces and light sources are around, and hence which visual stimuli can be perceived, and the physiology of color perception applies its own representations and computations to the perceived stimuli. Given that the physiology of perception has evolved in a particular environment, it is likely attuned to the latter in some way.

We can represent all physically possible surface spectral reflectance functions[2] in a solid known as the Object Color Solid [14]. The surface of this solid represents the limit of physically realizable surface colors, known as the Optimal Color Stimuli, and can be generated by computing the response of a given set of sensors to a continuum of a particular kind of artificial spectra. Assuming a flat-spectrum light source (with a spectrum given by $\sigma(\lambda) = 1$), we have computed the shape of the surface of the OCS solid. Fig. 2 represents the Optimal Color Stimuli surface in CIE XYZ, CIE L*a*b* and the Valois color space. Note the different shapes of the solid, and how it is "warped" in the L*a*b* and Valois spaces, relative to XYZ. Because of the non-linearities in the Valois model, the shape of the surface changes with changing radiance of the stimuli, which is not the case for the other two color spaces. The OCS surface in the Valois space is displayed for a relative radiance of 0.66, or $\frac{2}{3}$ of the maximum.

To investigate the warpedness of the Valois space, we used a crude measure of how the local density of the space has changed with respect to the corresponding region in the XYZ space, computing the ratio of a volume in XYZ space to the volume of its image in Valois

[2] Or transmittance functions, which makes no difference for our purpose.

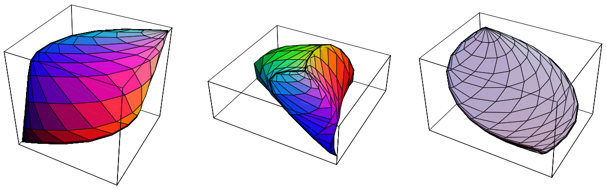

Figure 2: The Optimal Color Stimuli solid in the CIE XYZ space (left), the CIE L*a*b* space (middle) and the Valois space (right). On the X,Y,Z axes (arbitrary units): CIE X,Y,Z coordinates; CIE a*,b*,L* coordinates; Valois Green-Red, Blue-Yellow, and Brightness coordinates.

space:

$$w(r) = \frac{\sum_{i \in P} \sqrt{\sum_{d=1}^{3} (f^{-1}(C_i^d) - f^{-1}(C_r^d))^2}}{\sum_{i \in P} \sqrt{\sum_{d=1}^{3} (C_i^d - C_r^d)^2}} \qquad (3)$$

where P is a set of points around the reference point r for which the function value is computed (we used six points equidistant from the reference, along the three Valois axes), d ranges over the dimensions of the space: Green-Red, Blue-Yellow, and Brightness, f is the function that transforms XYZ to Valois, C_i^d is the d-th coordinate of point $i \in P$, and C_r^d is the d-th coordinate of the reference point r.[3] This doesn't actually measure a volume of course, but the bigger the set P is, the better the approximation will be. A large set of random points in a sphere around r would probably work best. If $w(r) < 1$ the Valois region around r is expanded relative to XYZ, if $w(r) > 1$ the region is shrunk, and if $w(r) = 1$ it is unchanged. Figure 3 shows some contour plots of w at planes of constant Brightness value.[4] We are still investigating the relation of the warpedness of the Valois space to the spacing of basic color categories. For some preliminary observations, consider Figure 4. The locations and relative spacing of the (foci of) the 11 basic color categories for English (*white, black, red, green, yellow, blue, brown, purple, pink, orange, grey*) after Von Kries adaptation are different in the different color spaces. We have to analyze the data further in the hope of discovering some interesting properties of this spacing, e.g. with respect to saturation or other psycho-physical measures. Because of the non-linearities in the XYZ to Valois transform, the spacing actually varies with radiance in the Valois space (Figure 5). Note how the categories near the far upper left corner (e.g. white, yellow, and pink) get squeezed together as radiance increases. In psychophysical terms that means among other things that their saturation decreases and they become increasingly difficult to distinguish from each other, which accords well with

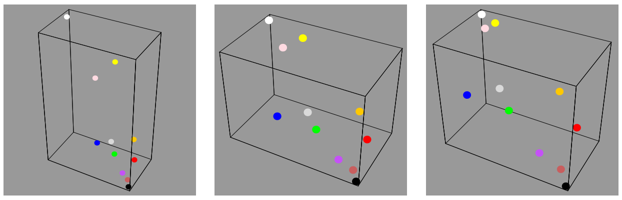

Figure 5: Spacing of the basic color categories in the Valois space varies with radiance. From left to right: relative radiance of 0.33, 0.66, and 0.99.

visual experience (at very high relative radiance levels everything tends to look white, for instance when stepping out into the bright mid-day sunlight after working all night long on a paper in a dark windowless dungeon – a.k.a. office). None of the other color spaces currently in use have this property, to our knowledge. It may explain the well-known hue shifts with changing radiance too (the Bezold-Brücke effect), but this remains to be investigated in detail.

4 Discussion and conclusion

We have presented two areas in our work where learning is or can be used: learning a transform from one color space to another, and learning names for basic color categories. In the first case we used an error backpropagation method to learn a non-linear transform with reasonable success, in the second case we are investigating a "space warp" learning method. Both procedures are part of our signal-to-symbol transformation going from visual stimuli to symbolic color names. The advantage of the backpropagation learning is that it allows us to determine a nonlinear transform which would be hard to do otherwise. The main advantage of learning the color space to color names relation is that it allows flexible adaptation of the model to different tasks, e.g. different languages.

References

[1] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Berkeley CA (1991 edition), 1969.

[2] R. M. Boynton. *Human Color Vision*. Holt, Rinehart and Winston, 1979.

[3] R. L. De Valois, I. Abramov, and G. H. Jacobs. Analysis of response patterns of LGN cells. *Journal of the Optical Society of America*, 56(7):966–977, July 1966.

[4] S. Harnad. The symbol grounding problem. *Physica D*, 42(1-3):335–346, 1990.

[5] H. Hexmoor, J. Lammens, and S. Shapiro. Embodiment in GLAIR: A grounded layered architecture with integrated reasoning for autonomous agents. In D. D. Dankel, editor, *Proceedings of the 6th Florida AI Research Symposium*, pages 325–329. Florida AI Research Society, 1993.

[3] This assumes that that f is a topological transform, which is a reasonable assumption to make.

[4] Unfortunately, Mathematica does not have Jello graphics built-in. A 3D Jello solid with locally varying density or color could be used to visualize the warp function values over the entire space at a glance.

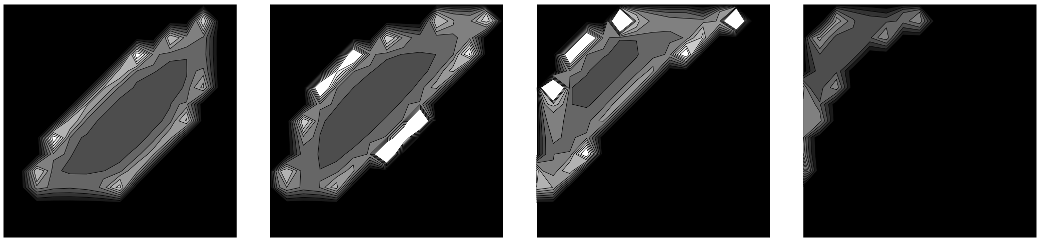

Figure 3: Some contour plots of the warp function w at constant Brightness values of 0, 0.2, 0.4, and 0.6 (see text). X axis: GR, Y axis: BY Valois dimension. Lighter color means higher values (greater warp), black areas represent non-real valued areas of f^{-1}.

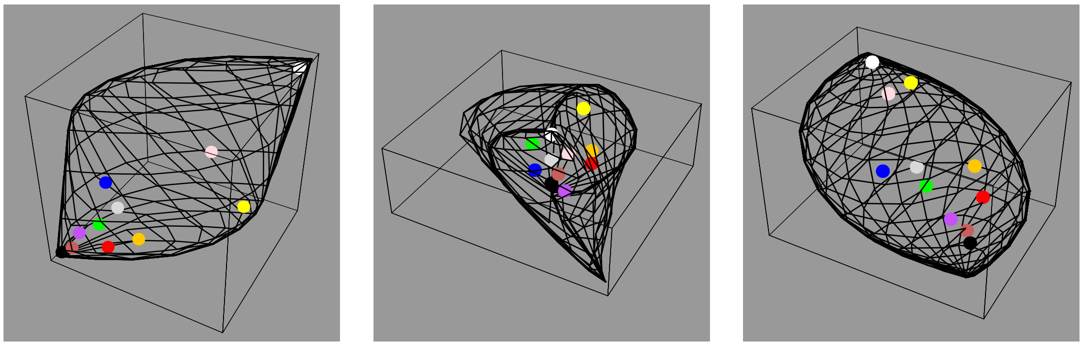

Figure 4: Location of the 11 Basic Color Categories for English relative of the Optimal Color Stimuli surface in XYZ, L*a*b*, and Valois spaces. (The dots inside the wire frame volumes are actually colored according to the basic color category they represent. The online PostScript version of this paper is in color; it is available from AAAI or by anonymous ftp to ftp.cs.buffalo.edu (128.205.38.1) in pub/tech-reports, or from the waters server on wais.)

[6] D. Jameson and L. M. Hurvich. Some quantitative aspects of an opponent-colors theory: I. chromatic responses and spectral saturation. *Journal of the Optical Society of America*, 45(7):546–552, 1955.

[7] P. Kay and C. K. McDaniel. The linguistic significance of the meaning of basic color terms. *Language*, 54(3):610–646, 1978.

[8] J. M. Lammens, H. H. Hexmoor, and S. C. Shapiro. Of elephants and men. In *Proceedings of the NATO ASI on Intelligent Autonomous Agents (Trento, Italy)*. Springer Verlag (To appear. Preprint available as CS TR-93-13, SUNY/Buffalo [anonymous@ftp.cs.buffalo.edu:/pub/tech-reports]).

[9] K. N. Leibovic. Vertebrate photoreceptors. In K. N. Leibovic, editor, *Science of Vision*, chapter 2, pages 16–52. Springer, New York, 1990.

[10] G. G. Lorentz. The 13th problem of Hilbert. In F. E. Browder, editor, *Mathematical Developments Arising from Hilbert Problems*. American Mathematical Society, 1976.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing*, chapter 1, pages 319–362. MIT Press, Cambridge MA, 1986.

[12] S. C. Shapiro and W. J. Rapaport. SNePS considered as a fully intensional propositional semantic network. In N. Cercone and G. McGalla, editors, *The knowledge frontier: essays in the representation of knowledge*, pages 262–315. Springer, New York, 1987.

[13] S. C. Shapiro and W. J. Rapaport. The SNePS family. *Computers Math. Applic.*, 23(2–5):243–275, 1992.

[14] G. Wyszecki and W. S. Stiles. *Color Science*. Wiley, New York, 1982.

[15] L. A. Zadeh. Quantitative fuzzy semantics. *Information Sciences*, 1971.