Practical Use of Color Imaging in Automatic Classification

By N. Ohsumi, Tokyo

SUMMARY

Some proposals are presented for utilizing color information in data analysis in computer graphics environments. The possibilities presented by using color in data analysis and the problems involved are presented first, then the systematization of color models in data analysis is dealt with. A proposal is submitted regarding the user interface software in color graphics system design. Examples of data analysis using color are provided showing color imaging patterns of data matrices, color plots of principal component scores, and so on.

KEYWORDS: Color models (RGB model, HLS model), color monitor display, principal component analysis, colored pattern matrix, automatic classification, colored vector, areal clustering system

1. Introduction

Research and application of graphical representation techniques as a method of visual communication of statistical information is now being carried on actively. Notable improvements have been achieved in the hardware and software environments supporting such techniques and a large number of statistical graphing program packages have become easily available. These advances in hardware and software environment have given great impetus to the use of color in the preparation of statistical graphs and maps.

In the area of data analysis, too, a notable increase is observed in proposals and experiments on the use of color. However, the use of color is still open to question because color science is inexact and there is no firmly established, systematic color theory yet. Nevertheless, the use of color graphics in data analysis is quite tempting to many researchers because it is natural for us to see things in color.

This paper presents some proposals for exploratory methods that can be applied in using color to examine classification problems. All ideas advanced in this paper are practical, interactive ones that can be realized *with a microcomputer and a low-cost color monitor display,* and focus on the importance of color handling software engineering to data analysis.

2. Interface between color models and data analysis

The use of color must be examined and developed with a clear recognition of the differences between the perceptional model and the descriptive model. The proposals made in this paper are intended to open up a new approach to the use of color in data analysis, with due consideration to the points enumerated in past reports.

Compstat 1986

The color models used in computer graphics are descriptive models built on the principle of color harmony, advocated by Ostwald and Munsell. Color handling carefully based on these color models is a minimal condition for the use of color in data analysis. The representative color models cited in past reports dealing with color graphics are summarized in Fig. 1. The HSV and HLS color models descriptionally represent human color perception, using three aspects of color (H = hue, L = lightness or V = value, S = saturation). Thus there are best suited to data analysis methods because they are intuitive, yet algorithmic. The latest advanced color graphics workstations can display as many as 16 million colors, but in data analysis the freedom of color handling is more important than number of displayable colors. It is quite troublesome in the RGB model to specify the mixture rates of various colors to obtain the desired color. Specifying the colo using the HSV or HLS models suits human perception better allowing the user to select a color on the basis of various intuitive parameters such as hue, intensity, brightness, lightness, tints, shades and tones.

This leads us to the conclusion that it is necessary to develop color transformation algorithms for linking the RGB model which is suited to color monitor control, and the HSV/HLS models. Figure 2 shows such an architecture, which can be fully realized with a microcomputer and a graphics workstation. Even more important is an easy-to-use program that provides a user-friendly interface. From this viewpoint, we have developed new programs for utilizing color models along the lines shown in Fig. 2. Each of these programs services as a basic tool for data analysis using color, and all ideas described in the following presuppose the use of these color systems and programs. These algorithmic color systems do not accurately represent the human color perception system, but are certainly more realistic and practical than the use of graphical representations or statistical graphs in which meaningless coloring is added.

3. Use of color information in automatic classification

Graphical representations are often used in automatic classification to display data, such as symbolic scatter plots, dendrograms and other tree representations, minimum spanning trees, cluster ellipses. Classification involves grouping the analytical data, which leads directly to coloring. However it presupposes that the process of coloring is free of subjective judgement. Thus main objective of this paper is to evaluate the effectiveness of using color in automatic classification by means of a system comprising color transformation algorithms linking predetermined color models and automatic classification programs.

3.1 Color imaging patterns of multivariate data

Classification of objects using a given multivariate data matrix for graphical representation of the results must answer to the following questions.

1) Which of the variables actually served for discrimination of objects?

2) Do we interpret the relationships among the variables?
3) Do we interpret the influence on the classification results caused by differences in clustering methods?
4) Do we avoid possible misinterpretation of the results based on the method used for drawing the dendrogram?
5) Do we know the influence of the selected slimialirty/dissimilarity or standardization on the clustering result.

To solve these problems, reexamination and idenfication of the clustering result against the original data set must be conducted. It is in this process of evaluation that the use of color will prove highly effective. The following basic procedure for color utilization in this process is quite simple.

Step 1: The objects of a given n objects by p variables multivariate data matrix X = (X_{ij}) (i = 1, 2 ... n; j = 1, 2 ... p) are grouped.

Step 2: The objects of the data matrix X are reordered according to the clustering results.

Step 3: Reordering of variables is conducted for representing the relationships (similarity or dissimilarity) among variables in color. Variables reordered can be conducted by the principle component analysis described below.

Step 4: The matrix obtained by reordering the objects and the variables is defined by a matrix X*.

Step 5: The matrix X* is converted to the *colored pattern matrix* Y represented by three aspects of color (H, L, S) using a transformation method computable with the color model. The colored pattern matrix Y thus obtained is transmitted through the RGB model to the color monitor display as *color patterns*.

The simultaneous two-way clustering method of the data matrix explained above offers the following advantages.

1) The characteristics of the multivariate data matrix resulting from its clustering can be grasped at a glance.
2) The characteristics of the original data matrix subjected to neither value transformation nor data modification can be directly observed in color.
3) Moreover, the method is easy and simple.

Some important points in connection with the above steps are explained below.

3.1.1 Conversion of reorganized data matrix X* to colored pattern matrix Y

Conversion of the reorganized matrix X* to the colored pattern matrix Y using the HLS (or HSV) color model is expressed as

$$X^* = (X_{ij}^*) \xrightarrow[\text{color model}]{\text{HLS}} Y = \{Y_{ij};\ H_j,\ L_{ij},\ S_j\} \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p) \tag{1}$$

where

H_j : hue, $0 \leq H_j \leq 360$; L_{ij} : lightness, $0 \leq L_{ij} \leq 1$

S_j : Saturation (see 3.1.3).

3.1.2 Coloring dispersion within a variable - Ulitization of lightness -

The dispersion of the observed values X_{ij}^* (i = 1, 2, ... , n) of a variable j is subjected to linear rescaling so that it will correspond to the lightness and come

within the interval [0, 1].

$$L_{ij} := (X_{ij}^* - \min_i X_{ij}^*) \,/\, (\max_i X_{ij}^* - \min_i X_{ij}^*) \qquad (2)$$

$$(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, p)$$

If necessary, a counter variable is generated according to the significance of the characteristics, so that the lightness and the observed values will correspond, i.e., the lightness will increase or decrease according to increasing values of the data.

$$X_{ij}^* := \max_i X_{ij}^* + \min_i X_{ij}^* - X_{ij}^* \text{ (for some variable } j) \qquad (3)$$

3.1.3 Reordering variables - Utilization of saturation and hue -

A variety of cases can be conceived for coloring variables according to their intended use.

- Variables resembling (or expected to resemble) each other in characteristics are alloted colors with similar hues in close-by positions.
- The complement of a hue is taken for variables differing from each other in characteristics.
- The hue is fixed and the characteristics of each variable are expressed only by a difference of lightness.

This freedom to select the interrelation among variables is important, but it is equally important to provide a procedure for objective determination of the relationships among variables.

Option 1: The saturation is fixed, and the hue only is varied for each variable as shown below.

$$S_j := r \text{ (normally } r = 1)$$
$$H_j := \theta j,\ \theta j = 2\pi\,(j - 1)/p \quad (j = 1, 2 \ldots p) \qquad (4)$$

Taking the saturation at $r = 1$ is identical to using only a certain specific pure color on the surface of the cone of HLS color model (see Fig. 4).

H_j of the variable j can be given so as to form equal angles to the number of variables p, but it can also be given according to subjective judgement of the similarity/dissimilarity among variables or on the basis of any information available in advance. In other words, the variable ordering need not be the same as that of the original data matrix. It may be also conceivable to use the variable ordering on the dendrogram obtained from hierarchical clustering of variables.

Option 2: Determinaing saturation and hue based on the similarity among variables

Free setting of the saturation and hue offers various advantages, but may also cause confusion when interpreting. For this reason, a method of automatic conversion of interrelationships to saturation and hue based on the principal component analysis is proposed. It satisfies the following conditions.

- Variables are reordered with consideration given to their relationships, and are made to correspond to the changes in hue.

- The contribution of each variable is expressed as changes in tint, for example, variables with a high contribution are expressed more brightly.

The process operates as follows:

Step 1: A given data matrix X is subjected to principal components analysis and converted to a component loadings matrix.

Step 2: The configuration of component loadings obtained by choosing two components ℓ, ℓ' is made to correspond a holizontal slice of the HLS color model, as shown in Figs. 3 and 4. Next, the fundamental colors are arranged consecutively on the periphery of the unit circle (Fig. 3), with red on the right side followed counterclockwise by yellow, green, cyan, blue and magenta.

Step 3: H, L and S are given at the following transformed values for the observed values X_{ij} of a variable j.

$$H_j := \theta_j,\ 0 \leq \theta \leq 360\ ;\ S_j := d_j,\ 0 \leq d_j \leq 1 \qquad (5)$$

L_{ij}: = Same way as that of Eq. (2)

The distance d_j from the origin (Fig. 3) is defined as the saturation for a variable j, and θj as the horizontal angle in the counterclockwise direction. Ordering of variables selected consecutively in the counterclockwise direction from an arbitrary baseline is adopted. H, L and S thus determined are distributed along the dotted line in Fig. 4 for the variable j.

With this method, the color effectively brings out the significance of the multivariate data. Moreover, the data can be represented by a 3-dimensional color space. Colors acquire the following significance.

1) Variables with similar hues have similarity relationships while those with complementary hues are dissimilar.
2) Variables approaching grey have small component loadings and those with highly saturated hues have a high contribution. In other words, the degree of contribution of variables can be visually observed as changes in tints.
3) The shades of a color represent the dispersion and spread of observed values. Outliers especially appear conspicuously.

3.2 Example of a simple experiment

Below is the experimental result of colored pattern matrix based on Roger de Piles' data used for clustering by Davenport et al ([2]). The original data was a collection of aesthetic judgements of 56 painters, less two who did not give some scores.

3.2.1 Determination of hue and saturation by principal component analysis

The results of principal component analysis can all be represented on the graphics display. Figure 5 indicates the first and second principal components, and shows the configuration of component loadings, order of variables, and values of hue and saturation. The order of variables was obtained by specifying a baseline, with

X_2 (drawing) coming first, followed by X_4 (expression), X_1 (composition) and X_3 (coloring). The differences in hue show that $\{X_2\}$ and $\{X_3\}$ are separated from $\{X_4, X_1\}$. It is possible to see approximately which hue represents each variable, e.g., X_3 is close to green. The saturation shows a large value along both the first two axes, so that all four variables present a bright hue.

3.2.2 Sorting of objects by automatic classification

Conventional techniques such as Ward's method and the k-means method can be applied for classifying. After the objects are classified and the variables reordered, the original data matrix is converted to a reorganized matrix X*, which is then transformed to a colored pattern matrix Y. Figure 6 is a monochrome copy of the colored pattern matrix obtained by Ward's method. The colored pattern thus makes it possible to obtain a firmer grasp of the features of each cluster than is possible from dendrograms or plots of principal component scores alone.

3.2.3 Coloring principal component scores

Each multivariate vector of the original data matrix is represented by a colored belt, which have called a *colored vector*. Each colored vector is then located at the coordinates of the corresponding principal component score plotted along the specified principal components ℓ, ℓ'. Thus, the values for each multivariate vector, in this experiment four, can be seen at glance at the location of the composite score for the multivariate vector and the original four dimensions are reduced to two. When compared with the dendrogram obtained by Ward's method, the colored pattern matrix makes it possible to obtain an immediate characteristics of each painter.

The important features of the color imaging method are:

- The order of variables can be fixed by a single calculation worked out for principal component analysis. However, the order of objects varies from one clustering method to another, so that the degree of clustering requires visual inspection.
- The color assigned to each element of the matrix X* remains the same even if the clustering method is changed, eliminating confusion in color judgement and evaluation.
- When other principal components are chosen, the variable relationships in the configuration of such new components change in relation to the hue and saturation, so that these changes can be observed as significant color changes.
- In addition to the relationships among variables, the interaction between variables and the variables that are contributed in clustering the objects can be observed as the variation in lightness.

4. More sophisticated applications

The development of color graphics systems and supporting software can open up greater possibilities of data analysis using color. Using our color handling system, we have conducted a variety of practical analyses, including, area

information analysis using the *areal clustering system* ([5] , [6]).
The system can represent multivariate data having coodinates on a map as changes in color and makes it possible to grasp such changes intuitively. It is expected that the concept of the system will be applied in multicolor display of choropleth maps, trend surfaces and contour mapping in the area of quantitative geography as well as in coloring of various output data in the research of computational geometry.

5. Conclusion

Although color science still involves many aspects awaiting further study, the use of color is progressing rapidly because it promises to enrich visual information both quantitatively and qualitatively. It is incumbent on us to develop a new graphics system suited to data analysis by introducing the mechanism of color perception and the knowledge of color models so far made clear. We earnestly hope that this paper will prove helpful in paving the way in this direction.

References

[1] Becker, R.A. (1983), Integrating Color into Statistical Software, Computer Science and Statistics: Proc. of the 14th Symposium of the Interface, 220-223.

[2] Davenport, M. et al. (1972), The Statistical Analysis of Aesthetic Judgment: An Exploration, Applied Statistics, 21, 324-333.

[3] Foley, J.D., Van Dam, A. (1982), Fundamentals of Interactive Computer Graphics, Addison-Wesley.

[4] Nicholson, W.L., Littlefield, R.J. (1983), Interactive Color Graphics for Multivariate Data, Computer Science and Statistics, Proc. of the 14th Symposium of The Interface, 211-219.

[5] Ohsumi, N. (1983), Practical Techniques for Areal Clustering, Data Analysis and Informatics III (eds., E. Diday and others), North-Holland.

[6] Ohsumi, N., Mizuno, K. (1984), Microcomputer Graphics Clustering Techniques for Urban Experimental Evaluation, Proc. of the XIIth International Biometric Conference, 148-156.

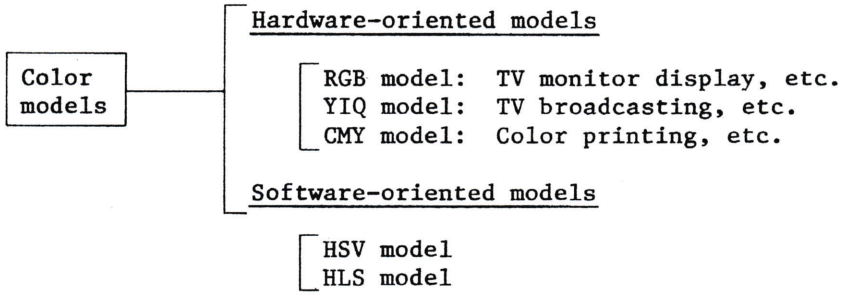

Fig. 1 Color Models for Graphics

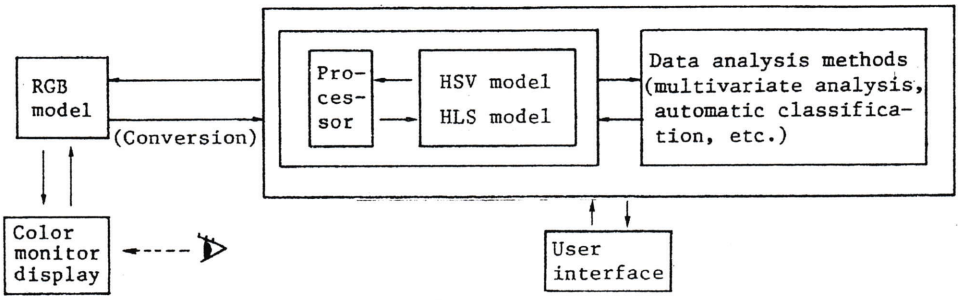

Fig. 2 Schematic of Color Handling System for Data Analysis

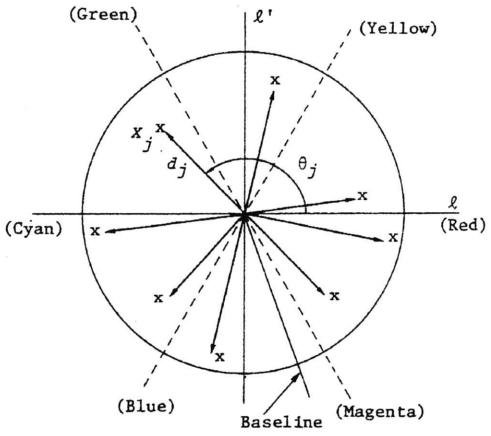

Fig. 3 Configuration of Component Loadings

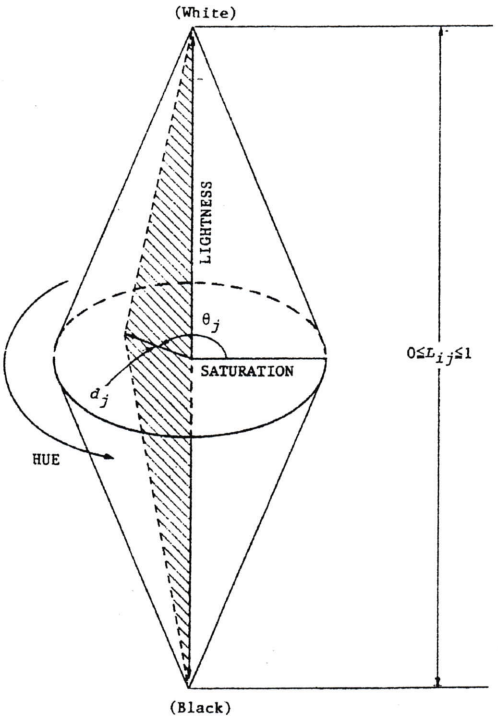

Fig. 4 Relationship between HLS Color Model and Data

```
X & Y ? : 1 2 ← Indicate the axes
                of the principal
                components

Order : Var.  :    Hue     : Sat.

   1       2       341.69     0.89
   2       4        20.78     0.91
   3       1        33.60     0.90
   4       3       124.24     0.94
Do you want to print ?(y/n) : Y
```

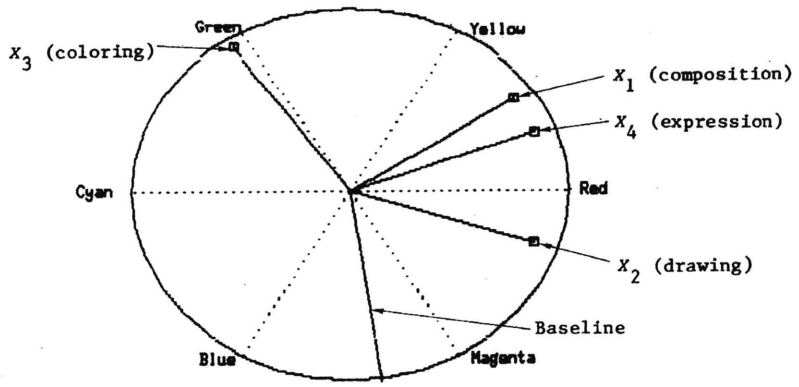

Fig. 5 Output Example of Piles' Data

By giving the baseline shown in the figure, the order of variables is determined: X_2 (drawing), X_4 (expression, X_1 (composition), and X_3 (coloring).

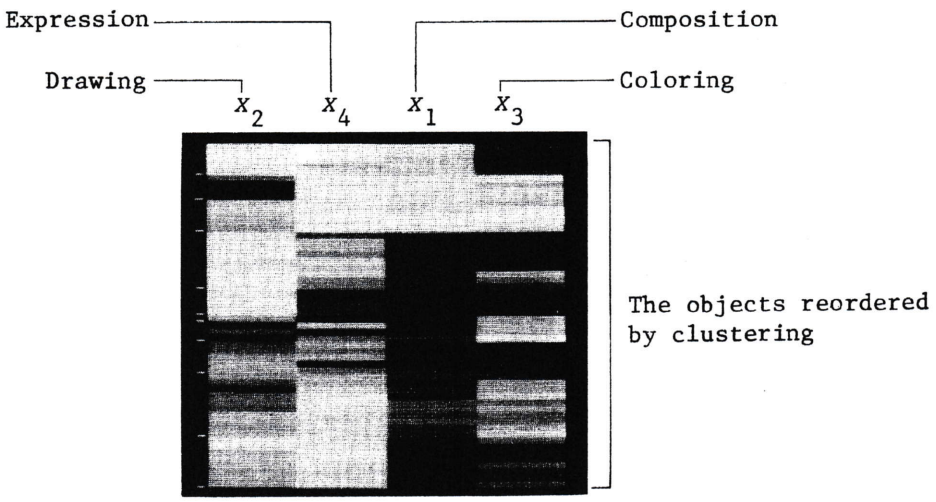

Fig. 6 Example of Colored Pattern Matrix