Blue-Print Document Analysis for Color Classification

Gabriel Marcu and Satoshi Abe*

Array Corporation, 3-32-11 Yoyogi, Shibuya-ku, Tokyo-151, Japan
* Department of Information Science, Fac. of Science, Univ. of Tokyo, Japan

Abstract. The blue print copies result in common designing process by copying the half-transparent white paper hand drawn in black ink. Over the blue print papers, very often, the designer marks in different colors some area or symbols of interest. This paper presents the particularities of analysis of such kind of documents in order to extract the colored lines as well as the original blue lines resulted during the copying process. The procedure is based on processing for clusterization of the colormap of the original image, and on post processing of the resulted run-length files for each classified color. It runs for any size of the scanned image and is independent on the contents of the image. For A1 size document, the color classification and bynary coding takes 20 min for 400MB image data.

1. Introduction

The blue print copies result in common designing process by copying the half-transparent white paper hand drawn in black ink. Over the blue print papers, very often, the designer marks in different colors some area or symbols of interest. The drawn colors on a very bluish and not uniform background are different from the drawn colors on a white background. A small sample of a blue print image is presented in figure 1. The images to be processed for color classification have the following particularities.

First, the blue-print scanned images are very large, usually A0 and A1, and are scanned with 16 dots/mm and 8 bits/color components, resulting in a large files. An A1 image stored in RGB raw format is 9000 x 12000 pixels and it conducts to about 400 MBytes image data, if each R,G,B channel is coded with 8 bits/component. A high speed processing algorithm is required in order to implement a practically usable procedure. Second, the background of the blue-print copy is bluish and locally not uniform. The bluish background color varies in a wide range from sample to sample, from very light to dark blue. Additionally, in time, the paper color changes from bluish to grey-yellowish, dependent on the exposure of paper to the sun light. Third, the color range of the document is narrow, enabling quantization of the image in hundred of colors.

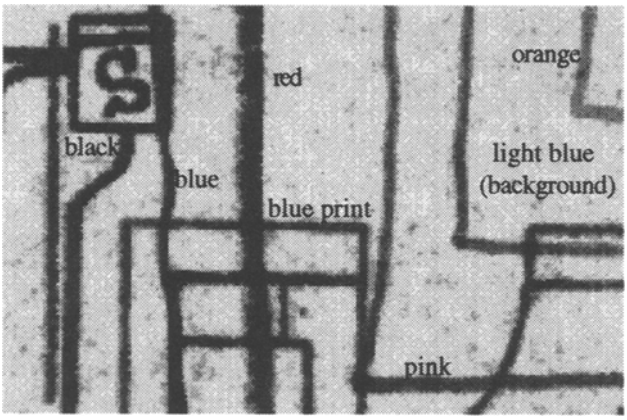


Fig.1. Sample of the original image

Over the blue print papers, the designer marks in different colors some area or symbols of interest. Pens or pencils are used for manual drawing. The colors non-uniformities and color deviations must be compensated in the color classification algorithm.

This paper presents the particularities of analysis of such kind of documents in order to extract the colored lines as well as the original blue lines resulted during the copying process. The paper proposes a fast classification method for color images, that can be effectively used for particular blue-print images.

For a high speed color classification, processing the image using spatial information of pixels [1,2] results in unacceptable time performance. As an example, the segmentation method for color images proposed by Yan and Hedley[1], can process a 336x336 pixels image in 6 seconds, but for 9000x12000 image size, the algorithm takes about 2 hours. In case of classification method proposed by Tominaga[2], the processing time increases more due to the color transformations required to perform an iterative process of detection of significant peaks and valleys of the histogram of main color components computed in Lab color space. Other methods based on spatial pixel information processing are not applicable due to limitation of processing time.

The procedure proposed in this paper is based on processing for clusterization of the colormap of the original image, and on post processing of the resulted run-length files for each classified color. The color classification method based on colormap clusterization has proved to be a very fast procedure, limited only by the access time to image data, and with acceptable results. The clusterization procedure runs optimally in Lab color space in order to compute the color distances between clusters, based on perceptually color distances between colors. The procedure runs for any size of the scanned image and is independent on the contents of the image. Results and diagrams illustrating the procedure are included.

2. Global Processing Procedure

The procedure proposed in the paper is based on few fast steps, as it is illustrated in figure 2.

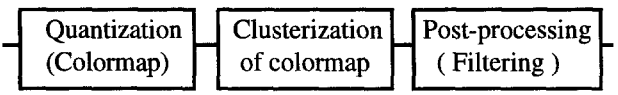


Fig.2. Diagram of fast color classification procedure

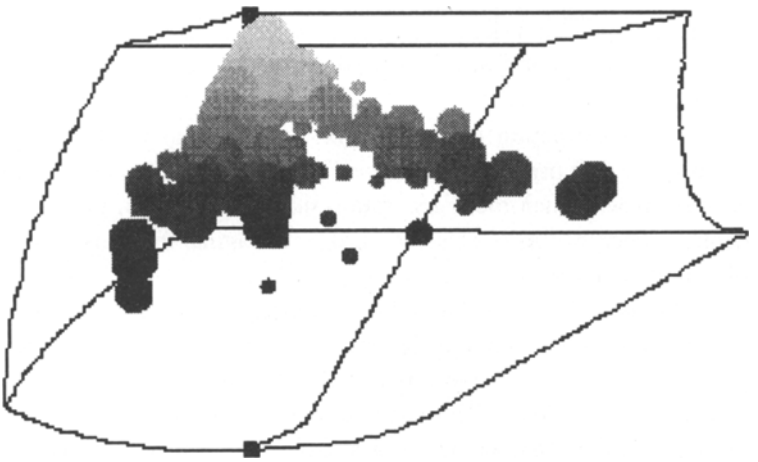
First step consists of extraction of the colormap of original image by a color quantization procedure. We used a modified version of Wu algorithm [3], due to time performance, but in case when more accurate results are required, the algorithms oriented on color properties can be selected [4,5]. The original image is transformed from true color to pseudo color (pixelmap image).
The second step consists of clusterization of the colormap. The clusterization process enables to derive a classification color table for each color class. The advantage of the an individual classification color table for each class is that in a single class can be included more that a single cluster. This can be usefully when in a class are included also the clusters resulted by overlapping the class color with other color classes.
The third step consists of derivation of binary color classes files, by passing the pixelmap image through the LUTs defined by each classification table.
The last step consists of post-processing the classes files for elimination of noise and transition colors.

3. Colormap Clusterization Procedure

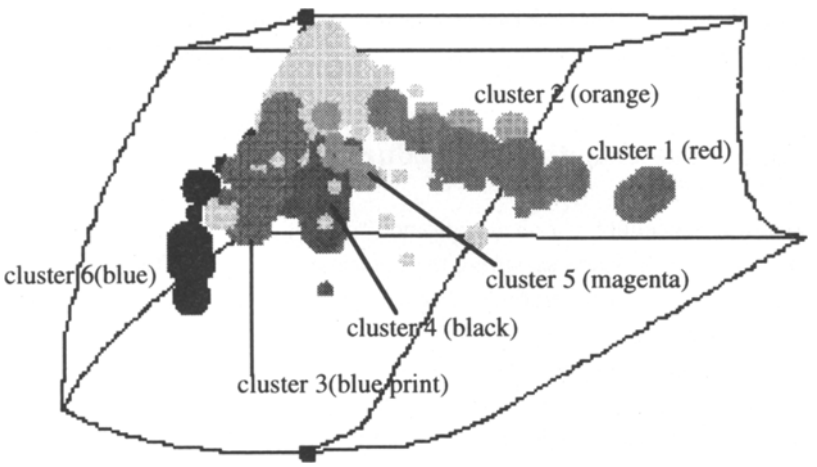
The clusterization procedure uses an agglomerative algorithm based on nearest neighbor principle[6,7]. It starts assigning the colormap colors in different clusters. The clusterization is performed iteratively. For each clusterization step, the algorithm finds the closest colors, or clusters, to be merged, in order to reduce the clusters number with 1. The algorithm stops at a threshold number of clusters. At one step, the algorithm selects a pair of two elements of the current classified colormap that have to be merged. The pair of elements have the minimum distance over the current classified colormap. The elements can be two colormap colors, a colormap color and a cluster or two clusters. The distance, D_{clust}, between two elements, referred as 1 and 2 (and named as colormap color or cluster), is: $D_{clust} = D - R_1 - R_2$, where D is the euclidean distance between the elements center (color components or cluster mean value components) and R_1, R_2 have the significance dependent on the elements involved in distance computation. For a color, R_i is:

$R = A.\log((a.x+b)/x_{max})+B$, if $R_{min} < R < R_{max}$,

$R = R_{max}$, if $R > R_{max}$, and $R = R_{min}$, if $R < R_{min}$,
where A, B, a, b, R_{min}, R_{min} are constants, and x_{max} is the maximum value of the histogram. For a cluster, R_i represents the radius of cluster, that is maximum Mahalanobis distance[8] of the cluster elements to its center, such that the cluster volume includes 80% of its elements.



(a) 3-D visualization of color map including information about histogram values



(b) result of clusterization procedure for 6 clusters

Fig.3. The clusterization procedure takes advantage of the 3-D visualization parameters for the definition of the cluster distance criteria

We select for clusterization the Lab color space, determined by the color transformation described bellow. The RGB color components of each colormap, based

on scannr calibration, are transformed to XYZ device independent coordinates. The XYZ device independent components are then transformed in L*a*b* components, corresponding to a more uniform color space. Using other color transformations, also other color spaces can be investigated. The L*a*b* space was selected in order to take advantage of the color difference formula that models better than other spaces the particularities of the human observer. Figure 3 illustrates the colormap clusterization procedure in Lab color space.

The detection of the clusters can be investigated over specific region of the color space, in order to increase the accuracy of classification algorithm, and in the same time to detect overlapped colors.

The clusterization of the colormap enables to build the color look-up table that is used to associate to each pixel value a color class. The image file is passed through the look-up table and the files resulted for each color are saved in run-length format. Effective clusterization of colormap it takes seconds but processing of the large image file is determined by the access time to image data on storage media. Figure 3 illustrates the clusterization procedure. Figure 3a represents the 3-D colormap and histogram visualization. Figure 3b depicts the 6 clusters identified by the algorithm.

4. Post-Processing the Classes Files

The colormap classification procedure may conduct to transition colors, represented as thin lines describing the edge of some hand drawn lines, dependent on the color and their relative position in the color space. For these cases, a post-processing procedure is required. The post-processing procedure takes advantage of the binary format of the files coding each color class. The post-processing procedure filters all the segments having the width smaller than a threshold value. The procedure is very fast since it uses only boolean operations on a binary image, avoiding the manipulation of 3 components pixels that is more time consuming procedure. The size of run length files depends on the contents of the image, but it was observed that can vary from 500KB to 5-6 MB for very complex diagrams and it takes less than 1 minute for A1 document size. For improved efficiency, the postprocessing procedure is applied simultaneously with run-length coding procedure.

5. Results and Conclusions

The procedure was implemented and tested on IBM PC with Pentium processor, at 90 Mhz. The procedure requires about 20 minutes for extraction of 7 colors from an A1 size document. Figure 4 depicts three samples of classified colors from image in figure 1.

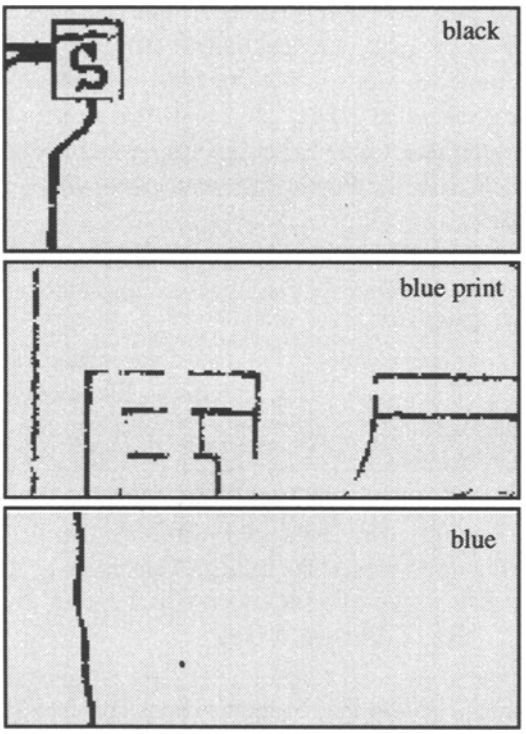


Fig.4. Samples of 3 classified colors

References

[1].M.Hedley, H.Yan, "Segmentation of Color Images using Spatial and Color Space Information", *Journal of Electronic Imaging*, Vol.1, N.4, October 1992.

[2].S.Tominaga, Color Classification of Natural Images, *Color Research and Applications*, V17, N4, 1992.

[3]. X.Wu, Efficient Statistical Computations for Optimal Color Quantization, *Graphics Gems III*, edited by D.Kirk, Academic Press, 1992.

[4].T.Orchard, A.Bouman, Color Quantization of Images, *IEEE Trans. on Signal Processing*, V39, N12,1991.

[5].R.Balasubramanian, C.Bouman, J.Allebach, Sequential Scalar Quantization of Color Images, *Journal of Electronic Imaging*, V3., N1., January 1994.

[6]. R.O.Duda, P.E.Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[7].B.V.Dasarathy, Nearest Neighbor Pattern Classification Techniques, IEEE Comp. Society Press, 1990.

[8].J.M.Jolion, P.Meer, S.Batauche, Robust Clustering with Applications in Computer Vision, *IEEE Trans. on PAMI*, V13, N8, 1991.