2011 International Conference on Software and Computer Applications
IPCSIT vol.9 (2011) © (2011) IACSIT Press, Singapore

Color Quantization in Document Images Using Biogeography Based Optimization

Surbhi Gupta [1], Deepti Bhardwaj [2], Parvinder S. Sandhu [1 +]

1Senior Lecturer, Department of Computer Science,
Rayat Institute of Engineering & Information Technology
2 Student, Department of Computer Science,
Rayat Institute of Engineering & Information Technology
1+ Professor, Department of Computer Science & Engg.,
Rayat & Bahra Institute of Engineering & Biotechnology, Mohali, India
1 royal_surbhi@yahoo.com, 1+ parvinder.sandhu@gmail.com

Abstract. A color Quantization in document images algorithm based on Biogeography Based Optimization (BBO) is developed in this paper. BBO is a population-based optimization algorithm primarily focused on the distribution of species among neighboring islands and follows similar steps as evolutionary algorithms to find near-optimal solutions. The purposed module for Color Quantization has the purpose of reducing the number of colors to decrease its storage requirement and this reduction must not affect the quality of the image, so that human eye cannot differentiate between the original and the modified image. This work uses BBO technique to determine the close colors and far off colors are dropped such that the modification is least perceptible by the user. In the previous survey of color reduction this optimization approach is missing. Taking every possible candidate and selecting one out of them certainly improves the solution. The aim of this work is to optimize the results of color quantization based on the theory of immigration. The far colors are taken as the residents of islands with low HSI & close colors are termed as high HSI islands. We need to migrate the far color to more close or average close islands based on their color distance i.e. CMC color distance for L*a*b color model. This color difference takes the advantage of characteristics of human vision.

Keywords: Biogeography, Evolutionary Algorithms, High Suitability index (HSI), Suitability Index Variables (SIV).

1. Introduction

Color Quantization is the process of reducing the number of color in the document by preserving the most important color information and compromise with the other. It helps in reducing the memory requirements of the images. It has a lot of application in digital imaging. The process can be applied to any colored image or colored text documents. To accomplish the color quantization the first need is to find out the actual number of colors presents in the image. One can choose any color model like RGB, LAB, CMY etc for the same. Then the color distance or their closeness is studied using different types of distance formula techniques. Then the mapping of close colors is done and the far off colors are dropped. Based on particular requirement we can decide the final number of colors in the image and stop the further reduction of colors. The process of quantization should be such that it must not cause the loss of visual information from the image but reduces its memory requirements. The existing algorithms of color map design may be categorized into two classes, splitting algorithm and clustering-based algorithm.

1.1. The Splitting Algorithms

The splitting algorithms split the color space of original images into two disjoint cells according to some criteria. Then, the splitting procedure is iterated until the desired number of cells is obtained. Finally, the

cluster heads of the cells are chosen to be the representative colors in a color map (Scheunders, 1997). The first example is the popularity algorithm which generates the color map by finding the popular color regions in the image. In this approach the histogram is studied & the color with highest occurrences in the histogram (peaks in histogram denotes the popularity or high occurrence) is selected as the final members of selected color list. Then there is median-cut algorithm in which the color space is divided into two. The orientation of cutting plane is normal to one of the coordinate axes with a largest range of image pixels. It considers the median of the color space for division. The operation is completed when desired number of color cells is left (Heckbert, 1982). The center-cut algorithm is similar to the median-cut algorithm. It uses the center point instead of median point to cut the plane (Joy and Xiang, 1993). In other variance-based algorithm the further partition of the cell is based on the weighted variances of color distribution. The cutting plane is chosen to be perpendicular to the coordinate axis where the expected variance is most reduced (Wan et al., 1998). One another approach is the octree algorithm relies on a tree structure. The root of the octree is an entire cell and the terminal nodes are the actual individual colors. The number of colors is reduced by taking the average of the color in the terminal node (Gervautz and Purgathofer, 1990). Yang and Tsai algorithm uses a moment preserving threshold technique in a quantization process to preserve information of input images (Yang and Tsai, 1998).Cheng and Yang algorithm repeatedly divides the color space into smaller cells. It uses mean of the color cells to decide the far off colors (Cheng and Yang, 2001).

1.2. The Clustering-Based Algorithms

In these techniques first some clusters are identified and then updated common techniques are C-means (Shafer and Kanade, 1987) and Fuzzy C-means (Lim and Lee, 1990; Shafer and Kanade, 1987), K-means (Verevka, 1995) and Kohonen SOFM (Dekker, 1994; Papamarkos, 1999; Papamarkos et al., 2002). The LBG algorithm is a method used in vector quantization (Patane and Russo, 2001; Linde et al., 1980). The pair wise clustering algorithm iterates a local optimization technique to reduce the number of image colors (Velho et al., 1997; Velho et al., 1998). Xiang introduces a minimizing the maximum intercluster distance algorithm for color quantization (Xiang, 1997). An agglomerative clustering algorithm restricts the maximum quantization errors (Xiang and Joy, 1994). In addition to clustering techniques, Ozdemir and Akarun apply a fuzzy technique in color quantization (Ozdemir and Akarun, 2002. Scheunders proposed the genetic Lloyd-Max, hierarchy competitive learning (HCL) and genetic C-means (GCMA) algorithms (Scheunders, 1996; Scheunders, 1997). Other examples of clustering-based algorithms are based on artificial neural networks (Gonzalez and Grana, 1997).

After color map has been designed, the next step is pixel mapping. Each pixel in original images is mapped to the representative color in color map. It is basically done by nearest-neighbor search i.e. finding the nearest representative color in color map. There are several nearest-neighbor search methods like MSE, Euclidean distance, CMC distance etc.

In the present work we are dealing with complex color documents such as cover books or Journal covers raises some challenging difficulties. Here the document text is overlapped with images, graphics, and line-drawings and for these reasons it is very difficult to extract the text regions from background. Generally, a color segmentation algorithm for text information extraction applications must be able to perform its task without over-segmenting the characters and with no fusion with the background.

Additionally, it is desirable to merge low contrast—non-text objects with their background and create large compact areas. This will result to a smaller number of connected components and improved form of the document image which will simplify the document segmentation and text extraction procedure. In this article, a new powerful technique for color reduction-segmentation of document images is proposed that satisfies the aforementioned criteria and can be applied to color and gray-scale documents.

Initially, to improve the quality of noisy color document images and the performance of the entire color reduction procedure, an edge preserving smoothing filter (EPSF) is applied as a preprocessing procedure. After this, the RGB color distribution of the image is effectively approximated by taking advantage of an important property of the image edge map. Specifically, the image is sub sampled by selecting only those pixels which are local minima in the 8-neighborhood on the edge map image. This ensures that the samples are not edge points, which guarantees that they are located in the interior region of the objects. Thus, fuzzy

points on the transition areas between objects are avoided. Also, all objects' colors (regardless of their size) are represented in the obtained sampling set. These samples are used in the next stage to initially reduce the colors of the input image to a relatively large number of colors (usually no more than 100). The extracted image at this stage is over-segmented. That is, the objects consist of at least one connected component. After this, a mean-shift operation procedure is applied on the obtained color centers that lead to the final color centers in the RGB color space (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Cheng, 1995). The final number of colors is small(less than 15 colors in most of the cases), and the final document image obtained has solid characters and uniform local backgrounds.

2. Proposed Quantization Technique and Algorithm

The proposed document color quantization technique consists of the following main stages:

Stage 1 Noise Removal for smoothing using filters

Stage 2 Edge Detection for text

Stage 3 Color Quantization in L*a*b color space using BBO approach

2.1. Noise Removal for smoothing using filters

Noise Removal is done for removing noise from images by blurring them with a weighted mean or a Gaussian filter. Through these processes noise reduction is achieved, but unfortunately, valuable information is lost and the details of object boundaries are deformed. Adaptive mean filter is a good solution where the amount of blurring for each pixel is determined after gathering local information in a specified n * n neighborhood (Perona and Malik, 1990; Harwood et al., 1987).

2.2. Edge Detection for Text

The next challenge in the current problem is to subsample the image but to retrain the edge information of the text. The common edge detection schemes are

- **The Marr-Hildreth Edge Detector**

It was a very popular gradient based edge operator. It uses the Laplacian to take the second derivative of an image. The idea is that if there is a step difference in the intensity of the image, it will be represented by in the second derivative by a zero crossing.

- **The Canny Edge Detector**

It is widely considered to be the standard edge detection algorithm in the industry. And it outperforms many of the newer algorithms that have been developed. The main steps are: Smooth the image with a two dimensional Gaussian, then take the gradient of the image. This shows changes in intensity, which indicates the presence of edges. This actually gives two results, the gradient in the x direction and the gradient in the y direction. Edges will occur at points the where the gradient is at a maximum.

- **Color Edge Detection Using Euclidean Distance and Vector Angle**

It uses two operators: Euclidean Distance and Vector Angle to detect edges in regions with high color variations & is thus selected for solving the current problem. Euclidean Distance is a good operator for finding edges based on intensity and the Vector Angle is a good operator for finding edges based on hue and saturation. The detector applies both operators to the RGB color space of an image, and then combines the results from each based on the amount of color in a region.

The algorithm for finding edges in the image is as follows:

- For each pixel in the image, take the 3x3 window of pixels surrounding that pixel.
- Calculate the saturation-based combination of the Euclidean Distance and Vector angle between the centre point and each of the eight points around it.
- Assign the largest value obtained to the centre pixel.
- When each pixel has had a value assigned to it, run the results through a threshold to eliminate false edges (Nadernejad et al., 2008)

2.3. Color Quantization in L*a*b color space using BBO approach

- **L*a*b color model**

CIELUV and CIELAB are the spaces two color recommended by CIE. These spaces have three-dimensional coordinates which approximately relate the tristimulus values with the perceived lightness, chroma, and hue for a color stimulus. However, the main purpose for developing these color space is to promote the uniformity of color difference formulas.

The CIE Lab color space is derived from CIE XYZ as follows:

X = 0.412453*R' + 0.35758 *G' + 0.180423*B'

Y = 0.212671*R' + 0.71516 *G' + 0.072169*B'

Z = 0.019334*R' + 0.119193*G' + 0.950227*B'

L = 116. * (Y/Yn)1/3. - 16 for Y/Yn > 0.008856

L = 903.3 * (Y/Yn)1/3. for Y/Yn 0.008856

a = 500. * [f(X/Xn)-f(Y/Yn)]

b = 200. * [f(Y/Yn)-f(Z/Zn)]

where

f(t) = t1/3. - 16 for t > 0.008856

f(t) = 7.787*t + 16/116 for t 0.008856

Here Yn = 1.0 is the luminance, and Xn = 0.950455, Zn = 1.088753 are the chrominance for the D65 white point. The values of the L component are in the range [0..100], a and b component values are in the range [-128..127].

- **CMC distance**

In color quantization, it is important to search appropriate color space and metrics which translate as faithfully as possible perceptual color similarity, i.e. that produce distances avoiding inaccurate values which may lead to merging dissimilar colors or to separating perceptually really close colors. It is noteworthy that, whichever the color space, those errors tend to happen either in quite specific conditions or in some occasional cases. As such the choice of an appropriate color space along with a measure respecting color perceptual similarity will sacrifice processing time to reduce the probability of incoherent color reduction. Unfortunately, given the quantity of color distance measures to be computed within a color image, we need to make a tradeoff between processing speed and the respect of perceptual similarity. To assess computational cost, Alain Pujol & Liming Chen benchmarked several color distances, including CIE94, CMC, CIE2000 and euclidean distance, on a set of pixel wise distance measures between two different images of the same size (both converted to CIE Lab as previously stated). This experiment reveals that these distances have significant different computational time, with CIE2000 being significantly slower than the two other dvanced metrics, which are themselves quite slower (about twice as slow) than the euclidean distance. Therefore any choice of an advanced color metrics has to be motivated either by performance and/or by a guarantee that a limited number of measurements will be taken. On the other hand, as shown by a study on perceptual color distances by X. Haisong and Y. Hirohisa, the CMC distance formula, shows convincing results on its property to better characterize perceptual color similarity : it is only beaten by the much more complex CIE2000 distance and as such represents an interesting compromise between accuracy and complexity. For two colors of respective CIELab components (L1, a1, b1) and (L2, a2, b2), CMC metrics define three components for the distance measure as follows:

Chroma difference:

$$\triangle C = \sqrt{a_1^2 + b_1^2} - \sqrt{a_2^2 + b_2^2}$$

Lighting difference :

$$\triangle L = L_1 - L_2$$

Hue perceptual difference :

$$\triangle H = \sqrt{(\triangle a)^2 + (\triangle b)^2 - (\triangle C)^2}$$

With the global distance given by :

$$\triangle E = \sqrt{\left(\frac{\triangle H}{Sh}\right)^2 + \left(\frac{\triangle L}{l \cdot Sl}\right)^2 + \left(\frac{\triangle C}{c \cdot Sc}\right)^2}$$

l and c are application dependent coefficients. Typical respective values are 1:1 for perceptual threshold and 2:1 for acceptability threshold.

In this paper this BBO technique is used to determine the best substitute for a less popular color & replaces it with a more popular color such that the modification is least perceptible by the user. In the previous survey of color reduction this optimization approach is missing. Taking every possible candidate and selecting one out of them certainly improves the solution. The aim of this work is to optimize the results of color reduction based on the theory of immigration. The colors with less popularity are taken as the residents of islands with low HSI & colors with more popularity are termed as high HSI islands. We need to migrate the less popular color to more popular or average popular islands based on their color distance i.e. CMC color distance for L*a*b color model.

Migration: Suppose that we have a problem and a population of candidate solutions that can be represented as vectors of integers. Each integer in the solution vector is considered to be an SIV. Further suppose that we have some way of assessing the goodness of the solutions. In this case the goodness of the solution is assessed by the number of pixels present in the solution color. Based on number of pixels in a particular color the suitability or goodness is assessed. Those solutions that are good are considered to be habitats with a high HSI, and those that are poor are considered to be habitats with a low HSI. Immigration rate for low HSI solution is high & emigration rate is low and vice versa. We use the emigration and immigration rates of each solution to probabilistically share information between habitats. With probability, we modify each solution based on other solutions. If a given solution is selected to be modified, then we use its immigration rate to probabilistically decide whether or not to modify each suitability index variable (SIV) in that solution. If a given SIV in a given solution is selected to be modified, then we use the emigration rates of the other solutions to probabilistically decide which of the solutions should migrate a randomly selected SIV to solution. We have categorized all colors in three categories. Every category is assigned a immigration & emigration rate & migration is done accordingly.

- **Color Reduction Algorithm**

1. Initialize the BBO parameters. Initialise every single color as a different habitat containing a number of pixels (species). Initialize a random set of habitats, each habitat corresponding to a potential solution to the given problem.
2. Select the color represented by maximum number of colors as the max and color represented by minimum number of colors as min for defining a range to categorize the habitats.
    - Habitats with number of pixels lying in the range min to 1/3*max are habitats with low HSI
    - Habitats with number of pixels lying in the range 1/3*max to 2/3*max are habitats with moderate HSI
    - Habitats with number of pixels lying in the range 2/3*max to max are habitats with high HSI
3. The immigration rate and the emigration rate is decided based on the HSI value. High HSI habitats have low emigration but high immigration, whereas a low HSI habitat has high emigration but low immigration rate. For the migration of pixels the habitat RGB color value is first converted to LAB color value & then a pixel is migrated if

- Its habitat allows emigration
- The second habitat allow immigration
- CMC color distance between two habitats is less than 0.4

4. Probabilistically use immigration and emigration rates to modify each non-elite habitat as discussed i.e. perform migration from low HSI to high HSI.
5. For each habitat, update the immigration & emigration probability of its species count using (2). Increase the CMC color distance threshold by 0.1 for every iteration. Iterate till the required number of habitats left.

3. Conclusion and Results

The proposed scheme is an effort to improve the quantization process in color document images. The BBO approach is an optimization technique & thus optimizes the color reduction to find out the best color match from the colormap

Fig. 1: Orignal image (Book.jpg)

Fig. 2: New image

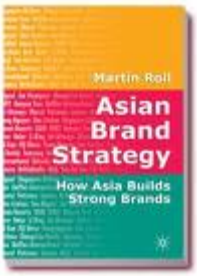

Fig. 3: Original image

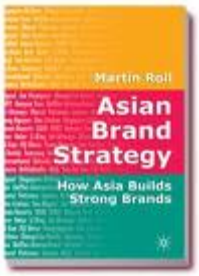

Fig. 4: New image

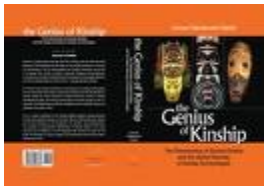

Fig. 5: Orignal image

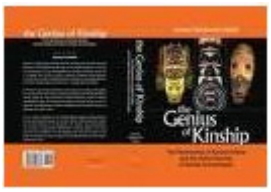

Fig. 6: New image

4. References

[1] Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell 17, 790–799.

[2] Cheng, S.C., Yang, C.K., 2001. A fast and novel technique for color quantization using reduction of color space dimensionality, Pattern Recognit Lett. 22, 845–856.

[3] Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell. 24, 603–619

[4] Dekker, A.H., 1994. Kohonen neural networks for optimal color quantization, Network Computer Neural System 5, 351–367.

[5] Fukunaga, K., Hostetler, L.D., 1975.The estimation of the gradient of a density function, with applications in

pattern recognition, IEEE Trans Inform Theory 21, 32–40

[6] Gervautz, M., Purgathofer.W., 1990 ‘‘A simple method for color quantization: Octree quantization,’’ In Graphics gems, A.S. Glassner (Editor) Academic Press, San Diego, pp. 287-293

[7] Gonzalez, A.I., Grana, M., 1997. Competitive neural networks as adaptive algorithms for non-stationary clustering: Experimental results on the color quantization of image sequences. In: Internat. Conf. on Neural Networks 1997, vol. 3. pp. 1602–1607.

[8] Harwood, D., Subbarao, M., Hakalahti. H., Davis, L.S., 1987. A new class of edge-preserving smoothing filters, Pattern Recognit Lett 6, 155–162.

[9] Heckbert, P., 1982. Color image quantization for frame buffer display, Computer Graphics 16, 297–307.

[10] Joy, G., Xiang, Z., 1993. Center-cut for color-image quantization.Visual Comput. 10 (1), 62–66.

[11] Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design. IEEE Trans. Comm. COM 28 (1), 84–95.

[12] Nadernejad, E., Sharifzadeh, S., Hassanpour, H., 2008. Applied Mathematical Sciences 2, 1507-1520.

[13] Ozdemir, D., Akarun, L., 2002. A fuzzy algorithm for color quantization of images, Pattern Recognit 35, 1785–1791.

[14] Papamarkos, N., 1999.Color reduction using local features and a SOFM neural network, Int J Imaging Syst Technol 10, 404– 409.

[15] Papamarkos, N., 2002 .A. Atsalakis, and C. Strouthopoulos, Adaptive color reduction, IEEE Trans Syst Man Cybern B Cybern 32 , 44–56.

[16] Patane, G., Russo, M., 2001. The enhanced LBG algorithm. Neural Networks 14, 1219–1237.

[17] Perona, P., Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion, IEEE Trans Pattern Anal Mach Intell 12, 629–639.

[18] Scheunders, P., 1996. A genetic Lloyd-Max image quantization algorithm. Pattern Recognition Lett. 17, 547–556.

[19] Scheunders, P., 1997b. A genetic c-means clustering algorithm applied to color image quantization. Pattern Recognit. 30(6), 859–866.

[20] Shafer, S.A., Kanade, T., 1987. Color vision. In: Shapiro, S.C., Eckroth, D. (Eds.), Encyclopedia of Artificial Intelligence. Wiley, New York, pp. 124–131.

[21] Velho, L., Gomes, J., Sobreiro, M.V.R., 1997. Color image quantization by pairwise clustering. In: X Brazilian Sympos of computer Graphics and Image Processing. IEEE Computer Society, Los Alamitos, CA, pp. 203–210.

[22] Velho, L., Gomes, J., Sobreiro, M.V.R., 1998. Visualization of color image quantization using pairwise clustering. In: Annual Sympos. on Computational Geometry, Proc. Fourteenth Annual Sympos on Computational Geometry, Minneapolis, Minnesota, United States. pp. 407–408

[23] Verevka,O., 1995. The local K-means algorithm for color image quantization, M. Sc. thesis, University of Alberta.

[24] Wan, S.J., Wong, S.K.M., Prusinkiewicz, P., 1988. An algorithm for multidimensional data clustering. ACM Trans. Math. Software 14 (2), 153–162.

[25] Xiang, Z., Joy, G., 1994. Color image quantization by agglomerative clustering. IEEE Comput. Graph Appl. 14 (3), 44–48.

[26] Xiang, Z., 1997. Color image quantization by minimizing the maximum inter cluster distance. ACM Trans. Graphics 16 (3), 260–276.