Automatic Color Form Dropout to Achieve Faster Document Processing

Shital A. Dhanfule[1], Prashant N. Pusdekar[2], Vinaya V. Gohokar[3]

[1] PG, Student, Department of Electronics and Telecommunication Engineering, SSGMCE, Shegaon,
[2] Professor, Department of Electronics and Telecommunication Engineering, PRPCET, Amravati,
[3] Professor, Department of Electronics and Telecommunication Engineering, SSGMCE, Shegaon,

[1]sdhanfule@rediffmail.com
[2]pusdekar.wardha@gmail.com
[3]vvgohokar@rediffmail.com

Abstract-- **Color Dropout converts documents such as color forms to black and white images by deleting the specific color which is intended for the background or the structure of the form. After successful dropout, only the relevant information by means of Black/Blue ink or Pencil is retained. The color dropout filter parameters include the color values of the non-dropout colors, color space conversion, distance calculation, dropout threshold detection. Color dropout is accomplished by converting pixels that have color within the tolerance sphere of the non-dropout colors to black and all others to white. This is done using VHDL coding. Processing may be performed in RGB or a Luminance-Chrominance space, such as YC_bCr. The color space transformation from RGB to YC_bCr involves a matrix multiplication and the dropout filter implementation is similar in both cases. Result for color dropout processing in YCbCr space is presented.**

Keywords--Color Dropout, Chrominance Euclidean distance, Threshold detection, Luminance,.

I INTRODUCTION

The document is scanned using high speed scanners. In document image processing there is a need to extract textual information from an image that has color content is useful in the background. The removal of the color content is useful in specific applications, such as forms processing, where the color content on the form used to facilitate data entry adds no value to subsequent data processing. Basic assumption is with ink color i.e., darker colors, such as black & dark blue & lighter colors as the part of document background. Color dropout is the image processing function whose purpose is to convert the scanned color document to a binary image where the form background colors are turned to white and the text colors are turned to black. Color dropout reduces the image file size, eliminates extraneous information, & simplifies the task of extracting textual information from the image.

Business forms are typically printed with some background; color for example, a pastel color. One way of eliminating this background color is to use an optical filter in the electronics, matched to the background color to be eliminated. Color dropout may be accomplished using optical or digital methods. Optical filters have been used when the document form involves a single dropout color. However, optical filters cannot be used with multiple dropout colors, and it is difficult to adjust the optical filter parameters of the optical filters to match nonstandard colors. Color dropout methods based on digital processing methods sometimes attempt to remove the form lines and background information from the scanned image. The main advantages of color dropout are the removal of the form lines minimizes interference with the text characters, and may reduce errors during character recognition. Another advantage is that the uncompressed file size is reduced by a factor of 24, since the color image consisting of 24 bits per pixel is converted to a binary image with only one bit per pixel. Textual information of interest is enhanced, because it is rendered black, while the background color that reduces the text contrast is suppressed. This fact significantly reduces the storage requirements for the resulting document files. Today there are many different Color Dropout algorithms, for use in various applications. All these algorithms differ in several important features. This paper aims to develop an algorithm using MATLAB & VHDL programming for Document Processing for Automatic Color Form Dropout.

In this paper Color Dropout Algorithm Architecture is used as shown in Fig. 1 which consists of three main steps. Document is scanned using scanner; the input image is in RGB color space. Read input image using MATLAB code, then perform Color Space Conversion i.e., in

YCbCr color space, Distance calculation &Dropout Threshold Detection using VHDL Coding. Finally, programs with different VHDL codes will be run, after that output image will be seen using MATLAB, which is nothing but a Color Dropout image.

II METHOD

Color Dropout Algorithm Architecture

A document is scanned to provide a digital image. Representative documents of this type are medical forms, insurance forms, census forms etc. Color dropout convert scanned color document to a binary image where the form background colors are turned to white and text colors are turned to black, since the image is converted from a full color form to black and white. At least one non-dropout color is selected and transformed to a Luminance-Chrominance space. Each pixel of the scanned image is converted to the Luminance-Chrominance space and the distance of each of the image pixels from the non-dropout color is determined.

Each of the image pixels is converted to black if the distance from the non-dropout color is less than or equal to a threshold value, and converted to white if the distance is greater than the threshold value. The converted black and white pixels are then stored. Advantages of color dropout are removal of the form lines minimizes interference with the text characters and may reduce error during character recognition, Color image consisting of 24 bits per pixel is converted to binary image with only one bit per pixel, Reduces the storage requirement There are numerous advantages of the present invention including, but not limited to: an operator is not required to set parameters for each image or image type; color removal is performed by evaluating local image content without access to the entire image; less memory is required than for other techniques; the process does not require buffering the entire image; the invention reduces the information extraction process time; improves image transmission time; and the color or colors retained represent the aspects of significant interest to the end user.

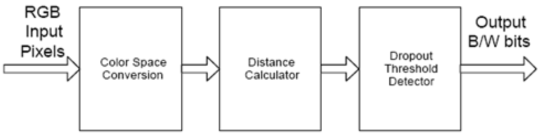


FIG 1. Color Dropout Algorithm

Architecture

As shown in above figure I , the input image is scanned using scanner which is in RGB Color Space. Read this image using MATLAB & Separate out its R, G, B pixels, these image pixels are input to Color Space Conversion. In Color Space Conversion RGB Color Space is converted to YCbCr Color Space with the help of matrix multiplication explain in section (A) This is done using VHDL Coding , then next step is Distance Calculation in which select a Non Dropout Color (40,40,40) compare it with original image pixels which comes from matrix multiplication as explain in section (B) This is done using VHDL Coding. Next step is Threshold Detection apply a threshold on distance , If the distance is less than threshold then output is white otherwise output is black as explain in section (C). This is done using VHDL Coding, means output is black &white image. Again output image is seen using MATLAB.

A. Color Space Conversion

RGB color space is the most widely used, but it is device dependent and color differences are not perceptually the same throughout the space. In this approach RGB image data is converted into YCbCr color space because YCbCr is more uniform color space, as compared to others. It is possible to transform the RGB values to one of the Luminance/Chrominance color spaces, such as CIE Lab. Here we use the YCbCr color space, which consists of Luminance Y, Blue Chrominance Cb, and Red Chrominance Cr . Even though YCbCr is not perfectly uniform, it has much better characteristics than RGB and only a matrix multiplication is required for the color space conversion based on the following transformation, this is done using VHDL coding.

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ 0.439 & -0.368 & -0.071 \\ -0.148 & -0.291 & 0.439 \end{bmatrix} \cdot \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix}$$

Y = 16 + (0.257*R + 0.504*G + 0.098*B)
= 16+(1/256)*[256*(0.257*R+0.504*G+ 0.098*B)]
= 16 + (1/256) * [65.792*R + 129.024*G + 25.088*B]
= 16 + (1/256) * [66*R + 129*G + 25*B
= (1/256) * [16*256 + (66*R + 129*G +25*B)]
= (1/256) *[(16*256 + 129*G)+(66*R + 25*B)]

Similar Equations for Cb & Cr respectively

For implementation of above three equation i,e. for Y, Cb, Cr in VHDL, We need three multiplier & three adder for Y. & for implementation of Cb needs three multiplier, two adder & one substractor, Cr needs three multiplier. One adder and two substractor as shown in the following fig.2

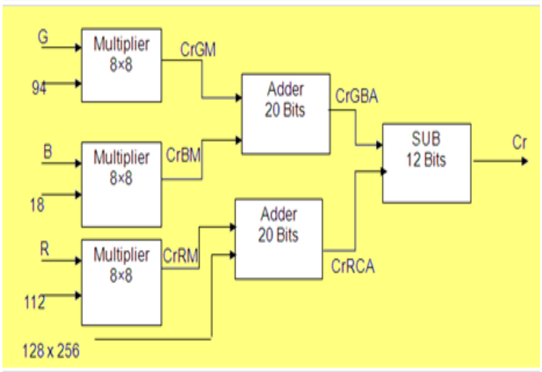


Fig 2. Implementation of Cr

Similar Implementation for Y & Cb.

The RGB variables take values in (0-255) and the resulting ranges are (16-235) for Y, and (16-240) for C. The quantities that need to be transformed during the system initialization are the centers of the non-dropout color spheres. During the actual processing of the document, only the RGB values of the pixel under consideration need to be transformed. The decision of whether or not the pixel color is inside a non-dropout sphere is made in luminance/chrominance space by performing comparisons that are similar in nature to those in RGB processing. Since the non-dropout colors are turned black and all other colors white, there is no need to perform the inverse transformation from YCbCr to RGB.

B Distance Calculation

Color dropout filter parameters includes:-

RGB values of the non-dropout color. Default parameters of the dropout filter are black (RGB = 40,40,40) & dark blue (RGB = 30,30,30).

How to know color is a non-dropout color?
1. Find the distance between the colored pixels of interest.
2. Each of the distances is compared with the associated dropout values.
3. If the distance is less than threshold value, the pixel belongs to a non-dropout color, and it is turned to black.
4. Otherwise it is turned to white.

Example Code For Distance Calculation

```
if (Y >= Y_DROPOUT1 )THEN
 Y_D1 <= Y - Y_DROPOUT1;
elseif  (Y < Y_DROPOUT1 )THEN
 Y_D1 <=  Y_DROPOUT1 - Y;
 end if;
 if (Y >= Y_DROPOUT2 )THEN
  Y_D2 <= Y - Y_DROPOUT2;
  elseif  (Y < Y_DROPOUT2 )THEN
   Y_D2 <=  Y_DROPOUT2 - Y;
   end if
```

Similar approach for Cb & Cr.

In the working space in this embodiment, in the Luminance-Chrominance space, each color component of the image pixel is allowed a variation for ink choice, printing variation, dye stability, and noise due to paper texture. A threshold value, shown as a radial distance in is chosen to determine the space containing the non-dropout color or colors. A determination is made by comparing each individual image pixel to the threshold value, and if a distance to each pixel is greater than the threshold value, the pixel is converted to white. If the distance to each pixel is less than the threshold value, the pixel is converted to black.

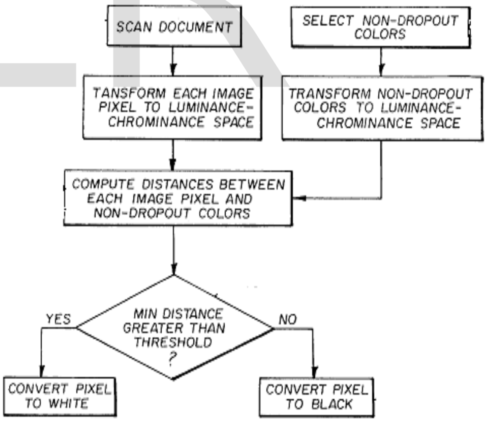

SCAN DOCUMENT
SELECT NON-DROPOUT COLORS
TANSFORM EACH IMAGE PIXEL TO LUMINANCE-CHROMINANCE SPACE
TRANSFORM NON-DROPOUT COLORS TO LUMINANCE-CHROMINANCE SPACE
COMPUTE DISTANCES BETWEEN EACH IMAGE PIXEL AND NON-DROPOUT COLORS
MIN DISTANCE GREATER THAN THRESHOLD ?
YES
NO
CONVERT PIXEL TO WHITE
CONVERT PIXEL TO BLACK

Fig.3 Flow chart for distance calculation & threshold detection

The processing was done in YCbCr space, where the black color was retained while all other colors were dropped. In some applications, a plurality of non-dropout colors may be chosen, for example, blue and black. Each non-dropout color of interest is stored in memory and is used to evaluate each image pixel against it. Each image pixel is evaluated in a raster fashion and is classified as follows.

C Threshold Detection

If the image color matches one of the colors of interest within specified tolerances, i.e. threshold, the output color is set to black, otherwise the output color is set to white. Since only the colors of interest are stored and used, it is not necessary to add information specific to a particular form or image scanned therefore eliminating the need to define many forms or templates used to match patterns against to determine which image elements to retain or eliminate. This is shown schematically in fig.4 where a first non-dropout color 50, a second non-dropout color 52, and a threshold 54 is established around these points and image pixels outside the threshold spheres are converted to white, and images inside the threshold spheres are converted to black.

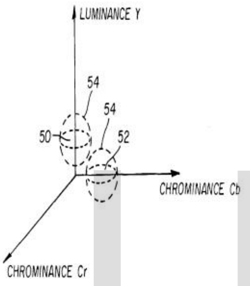


FIG. 4. Threshold Detection

In another embodiment of the invention, if each image pixel is less than the threshold value, it is converted to a first grayscale image rather than being converted to black. If the image pixel is greater than the threshold value, it is converted to a second grayscale image rather than to white. This gives the user the opportunity to select an output which may be in printed form in non-standard format.

Example Code For Threshold Detection

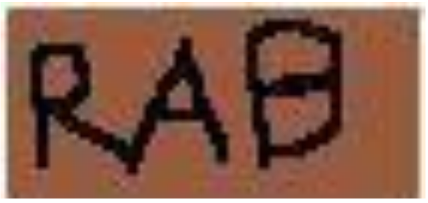
```
if (Y_DIST1<=Y_THRSHOLD1 AND Cb_DIST1 <=Cb_THRSHOLD1
                              AND Cr_DIST1 <=Cr_THRSHOLD1 )
                    OR                 (Y_DIST2<=Y_THRSHOLD2 AND
Cb_DIST2 <=Cb_THRSHOLD2                             AND Cr_DIST2
<=Cr_THRSHOLD2 ) THEN

            RESULT <= 1;
            else
            RESULT <= 0;
            end if;
```

- Output Black & White bits.

III RESULT

a. Test Input Image



b. R, G, B Matrices

Red values using MATLAB

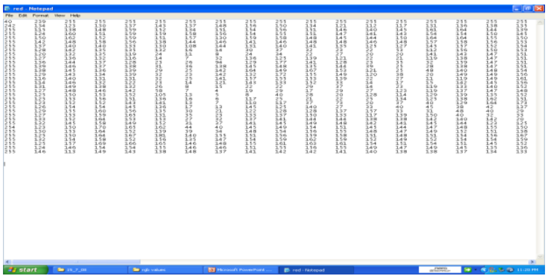
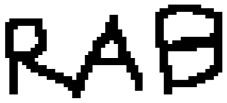


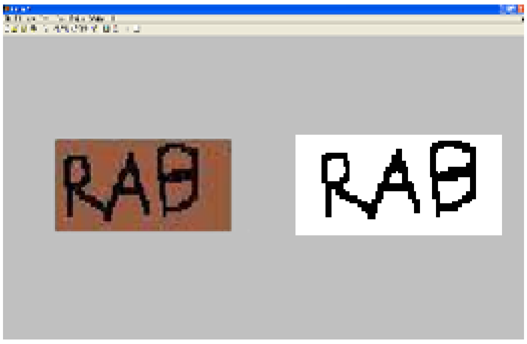
Fig.5. r coefficient Matrix

Similarly Green and Blue Values Matrix are obtained using MATLAB.

c Color Dropout Image



D. Comparison of Input & Output Image

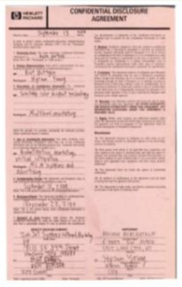


Discussion

In this project a VHDL coding for Color Dropout has been developed, which is one of the initial step for image compression & the textual information of interest is enhanced because it is rendered black, while the background color, that may reduce the text contrast, is suppressed. In addition, the removal of form lines minimizes interference with the text character that may reduce errors during character recognition. Also, the speed of operation can be increased by using hardware instead of software. Uncompressed file size is reduced by a factor of 24, since the color image consisting of 24 bits per pixel is converted to binary image with only one bit per pixel It significantly reduces the storage requirements for the resulting document files which is the dropout image.

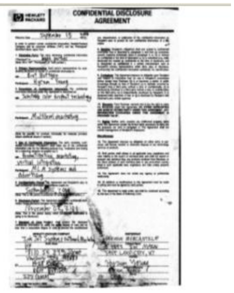
Examples of Non Dropout Image & Dropout

Images



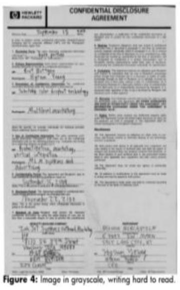
CONFIDENTIAL DISCLOSURE AGREEMENT

Figure 2: Original image, writing hard to read.



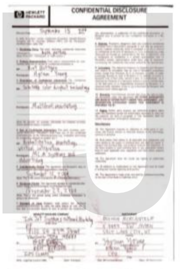
CONFIDENTIAL DISCLOSURE AGREEMENT

Figure 3: Image in 1-bit w/o color dropout, writing illegible.



CONFIDENTIAL DISCLOSURE AGREEMENT

Figure 4: Image in grayscale, writing hard to read.



CONFIDENTIAL DISCLOSURE AGREEMENT

Figure 5: Image in 1-bit with color dropout, writing legible.

References

[1] Bhaskar, J. 2007, VHDL Primer, Pearson Education, 386

[2] Gonzalez, R. C., Woods R. E. 2003, Digital Image Processing, Pearson Education, 793
.
[3] Savakis A. E. and Brown C. R.,“Document Processing for Automatic Color Form Dropout”

[4] Link B. A., Lee Y., 2002: “System and methods for image processing by automatic color dropout”, Patent Application Publication, US 2002/ 0136447 A1

[5] Savakis A. E., 2000:”Automatic color dropout using luminance- chrominance space processing”, United States Patent No.: 6035058