A Survey on Document Processing By Color Form Dropout

Ankush D. Kadu **Dr. P. R. Deshmukh**

Abstract — **Color dropout refer as dropping out of color form backgrounds from images of completed forms to Obtain color form dropout images retaining only the respondent information. Color dropout is performed by associating darker non-dropout colors with information that is entered in the form a needs to be preserved. At least one non-dropout color is selected and transformed to RGB or a Luminance-Chrominance space. Color dropout process includes scanning of image, conversion to RGB or a Luminance-Chrominance space, distance calculation, dropout threshold detection, finally storing converted black and white image. Processing may be performed in RGB or a Luminance-Chrominance space, such as YCbCr. Color dropout is obtained by converting pixels that have color within the tolerance sphere of the non-dropout colors to black and all others to white in RGB or a Luminance-Chrominance space. This approach needs an ideal FPGA platform which lends itself to high speed hardware implementation with low memory requirements. This is done using VHDL coding. The color space conversion from RGB to YCbCr is to be done by matrix transformation multiplication formula and the dropout filter implementation is similar in both cases. Color dropout processing result may be either display or view in RGB or YCbCr space.**

Index Terms—Color dropout, Color space conversion,FPGA, MATLAB, Threshold detection, VHDL

I. INTRODUCTION

Color forms constitute a large number of documents that are scanned using high-speed scanners. When electronically processing a document such as review form or the like having respondent information entered, so there is a need to remove or dropout the background of the document from a scanned image of the document thereby facilitating minimum storage requirement of image. Color dropout is nothing but the image processing function whose objective is to convert the scanned color document to a binary image where the color form backgrounds are turned to white and the text colors are turned to black. To develop this we need to differentiate between the colors of the background and the colors of the entered text so the image is converted from a full-color form to black and white. By removing background significantly file gets compress and reduces the storage requirements for the resulting document files. The main advantages of color dropout during optical character recognition that information to be read separate from the background information, such as line, boxes and other textual instruction and by means of this minimizes line interference with the text characters, and may reduce complications during character recognition. This process results in the elimination of all but the desired information.

Color processing has two approaches as RGB space or Luminance/Chrominance color space. Color dropout based on luminance/chrominance processing involves all the steps that are used in RGB processing, as well as one color space transformation from RGB to YCbCr color space. To accomplish this we need to distinguish between the colors of the background and the colors of the entered text. Idea to develop an algorithm using MATLAB & VHDL programming for Document Processing for Automatic Color Form Dropout on FPGA platform to get impressive speed of operation increased by using hardware instead of software. Developed VHDL coding for distance coding to be tested using a Xilinx FPGA. The core provides an excellent amount of processing performance given the FPGA space requirements. However, it also gives excellent system scalability for much greater performance. The method presented in this paper is designed to operate in a fully automatic environment and is implemented in hardware.

II. LITERATURE REVIEW

In image processing, there is a need to extracttextual information from an image that has color content inthe background. The removal of the color content is useful in specific applications, such as forms processing, where thecolor content on the form, used to facilitate data entry, addsno value to subsequent data processing. So we are going to propose Color dropout techniques which help us to reduces the image file size, eliminates extraneous information,represent the aspects of significant interest to the end user,less memory is requiredthe invention reduces the information extraction process time.

B. Yu and A. Jain [1] presented Color dropout methods based on digital processing methodsdescribes a generic system for form dropout when the filled-in characters or symbols are either touching or crossing the form frames. We propose a method to separate these characters from form frames whose locations are unknown. Since some of the character strokes are either touching or crossing the form frames, they address the following three issues: 1) localization of form frames; 2) separation of characters and form frames; and 3) reconstruction of broken strokes introduced during separation.

J. Mao and K. Mohiuddin [2] presented the distance transformation and its gradient flow are employed to remove form lines. Form templates are pre-processed off-line to obtain their distance transforms and gradient flows. They demonstrate that various components in the form dropout algorithm can derive benefit from rich geometric information about the form template which are made explicit in the distance transform and its gradient flow Such approaches may work for specific cases,

but require significant computational effort and are very expensive to implement in real-time hardware that are used in high-speed scanners .

Another approach to color dropout, originally developed in the context of optical character recognition. In this work, the average RGB dropout colors in color patches are determined and used in a dropout filter that can be implemented using electronic hardware. The filter bandwidth is adjusted to accommodate for color variations between forms. The advantage of this approach is that the presence of noise, e.g. black specs, does not significantly affect the average color in the color patch considered, and consequently does not affect the final color dropout result[3].

Y. Murai and T. Amagai [4] presented another approach in proposes scanning a blank form, extracting the dropout colors from the blank form, and using them to perform color dropout when scanning other forms.

In this dropout method is based on image subtraction and line elimination for distorted images. The location, rotation and magnification are modified for distorted form images. Character patterns and short ruled lines are eliminated by subtraction of bitmap template images. Long ruled lines are extractedand direct eliminated by using run data [5].

Vote counting accuracy has become a well-known issue in the vote collection process. Digital image processing techniques can be incorporated in the analysis of printed election ballots. This paper explores methods of voting between the results of the different mark extraction methods to improve recognition. To provide diversity a simple image subtraction technique is paired with a distance transform and a morphology based algorithm. The result has a higher detection rate and a lower false alarm rate [6].

III. PROPOSED WORK

The document or input image is scanned using high speed scanners which usually in RGB space. In document image processing there is a need to extract textual information from an image that has color content is useful in the background lines which are not of any practical use has been eliminated.Color dropout is the image processing function who converts the scanned color document to a binary image where the color form backgrounds are turned to white and the text colors are turned to black.

Our objective is to develop an algorithm using MATLAB & VHDL programming forDocument Processing for Automatic Color Form. Implementation of Color space conversion. Implementation of FPGA which is an ideal platform for image processing engine.Develop VHDL coding and threshold detection.To be tested using Xilinx FPGA.Implementation of interface system to allow easy manipulation of processing parameter, input pixel and read the result value.Performance evolution parameters of the proposed algorithm to be observed using MATLAB

A. *Color dropout architecture*

In this Color Dropout Architecture is used as shown in Fig. 1 which consists of three main steps as follows:

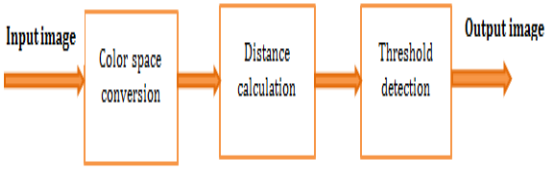


Figure 1: color dropout architecture

1. Color space conversion

The input image is scanned using scanner which is in RGB Color Space and to be read this image using MATLAB & Separate out its R,G,B pixels, these image pixels are input Color Space Conversion.Most case used space is RGB space, but it is device dependent and color differences are not exactly the same, it is desirable to convert RGB space into Luminance/Chrominance (YCbCr) color spaces 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.
Hence we use the YCbCr color space, which consists of Luminance Y, Blue Chrominance Cb, and Red Chrominance Cr. YCbCr has much better characteristics than RGB and only a matrix multiplication is required for the color space conversion based on the following transformation formulae:

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

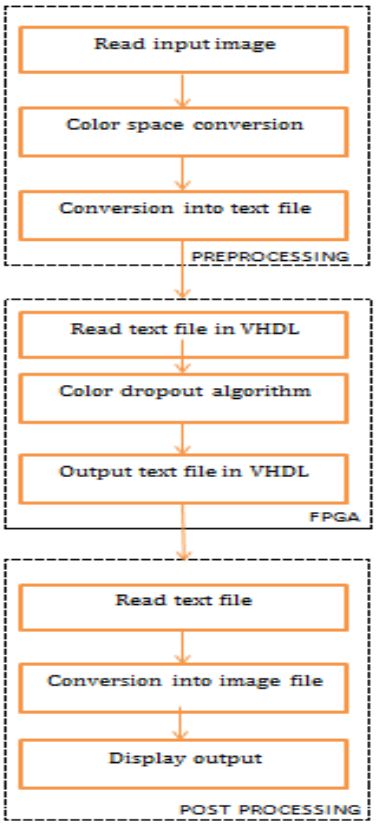
2. Distance calculation using VHDL Coding

Then the second step is to compute the distance in which select a Non Dropout Color compare it with original image pixels which comes from matrix multiplication as this is done using VHDL Coding as follows:

- Finding the distance between the colored pixels of interest.
- Each of the distances is compared with dropout values.
- If the distance is less than threshold value, the pixel belongs to a non-dropout color, and it is turned to black.
- Otherwise it is turned to white.

3. Dropout Threshold Detection using VHDL Coding

Next step will be the dropout threshold detector is mechanism that determines whether or not the pixel falls within the threshold of either dropout color. If the image color matches one of the colors of interest within specified range, i.e. threshold, the output color is set to black, otherwise the output color is set to white i.e. black and white image. This is done using VHDL Coding, means output is black &white image. Eachconverted pixel is compared to threshold value, which allows for color tolerances due tostability, printing andother variations. 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.



MATLAB involves preprocessing and post processing operation. In preprocessing operation, input image file which is being obtained by scanner need to be convert this file in into text file since we are developing VHDL coding, only able to read text file.Text file contain each pixel value which is then processed as per algorithm design to perform color dropout processing. While in post processing the exact opposite process is to be executed to recover original input image.

CONCLUSION

Color dropout methods is digital image processing methods sometimes attempt to remove the form background information from the scanned gray scale image. Color Dropout algorithm has been developed, which is one ofautomatic environment and is implemented in hardwarethat may reduce the text contrast, is suppressed.The processing is done on individual pixels without the requirement of observing any neighborhood pixel. Thus, there is no needs to buffer the image, and no additional memory requirements are required. Highperformance can easily be achieved by simply using a newer technology FPGA.It significantly reduces the storage requirements for the resulting document files which is the dropout image, reducesprocess time and improves image transmission time.

Future work will involve approach and to concentrate on all parts of image rather than text and background and also processing in other uniform color spaces with some supervise learning.

REFERENCES

[1] B. Yu and A. Jain, "A Generic System for Form Dropout," IEEE Trans. PAMI, 1998
[2] J. Mao and K. Mohiuddin, "Form Dropout using Distance Transformation," Proc. ICASSP'95, 1995, pp. 328-331.
[3] P. Rudak, "Automatic Detection and Selection of a dropout color using zone calibration in conjunction with optical character recognition of preprinted forms," US Patent 5014329, 1991.
[4] Y. Murai and T. Amagai, "Image processing apparatus with function of extracting visual information from region printed in dropout color on sheet," US Patent 5,664,031, 1997.
[5] Shima, Y. ; Ohya, H. ; Yasuda, M. " A form dropout method based on line-elimination and image-subtraction", Eighth International Conference IEEE, 2005
[6] Smith,E.H.B. ; Goyal,S. ; Scott,R. ; Lopresti,D. "Evaluation of voting with Form Dropout Techniques for Ballot Vote Counting",in Document Analysis and Recognition (ICDAR), International Conference on IEEE ,2011, pp-473-477
[7] Shuli Sun ; Lihua Xie ; Wendong Xiao ; Nan Xiao, " Optimal Filtering for Systems With Multiple Packet Dropouts" , IEEE Trans, 2008,Pp: 695 - 699 .
[8] A.Savakis and J.Madigan," Automatic Color Form Dropout using Luminance/ Chrominance Space Processing," U.S. Patent Number 6035058, 2000.
[9] Gonzalez, R. C., Woods R. E. 2003, Digital Image Processing, Pearson Education.