International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 1 72 - 75

A Review on Automatic Color Form Dropout for Document Processing

Ankush D. Kadu
M.E. Student, Department of Electronics and Tele-communication,
Sipna C.O.E.T, S.G.B. Amravati University,
Amravati(Maharashtra State),India.
ankush.kadu20@gmail.com

Dr. P.R.Deshmukh
Professor, Department of Electronics and Tele-communication,
Amravati(Maharashtra State),India
pr_deshmukh@yahoo.com

***Abstract*—**Color Dropout converts documents like color forms to black and white images by deleting the specific color which is maintained only the information entered in the form. The successful color dropout simplifies the task of extracting textual information from the image for the reader. 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 done 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 uses 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 transformation from RGB to YCbCr involves a matrix multiplication and the dropout filter implementation is similar in both cases. Color dropout processing result may be either represented in RGB or YCbCr.

Keywords*—Color dropout, Color space conversion, FPGA, MATLAB,VHDL*

*****

I. INTRODUCTION

Color forms constitute a large number of documents that are scanned using high-speed scanners. In document image processing, there is a need of extracting textual information rather than the document background and document lines which are not of any practical use has been eliminated. Some examples where color dropout is important is in the field of optical character recognition (OCR), business forms which are typically printed with some background, etc. When performing color dropout, it is desirable to eliminate the color background of the form, and at least one non dropout color to be selected and transform to luminance chrominance space. Each pixel of image is then converted into black if the distance from the non dropout is less than or equal to threshold value, otherwise converted to white then this converted black and white pixel are stored.

The main advantages of color dropout during 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 errors during character recognition. By removing background, lines significantly file gets compress and reduces the storage requirements for the resulting document files. There are numerous advantages of the present invention including, but not limited to color removal is performed by evaluating local image content without access to the entire image, less memory is required than for other techniques,, improves image transmission time, and the color or colors retained represent the aspects of significant interest to the end user, the process does not require buffering the entire image, an operator is not required to set parameters for each image or image type, the invention reduces the information extraction process time.

Color processing has two approaches as RGB 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. This adds some cost in terms of processing time and complexity

To accomplish this we need to distinguish between the colors of the background and the colors of the entered text. Color dropout may be viewed as a form of color image rendering, since the image is converted from a full-color form to black and white by means of MATLAB and VHDL programming. Developed VHDL coding for distance coding to be tested using a Xilinx FPGA.

II. LITERATURE REVIEW

B. Yu and A. Jain [1] presented Color dropout methods based on digital processing methods describes 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. The form frame is automatically located by finding long straight lines based on the block adjacency graph. Form frame separation and character reconstruction are implemented by means of this graph. Proposed system includes form structure learning and form dropout. First, a form structure based template is automatically generated from a blank form which includes form frames, preprinted data areas and skew angle. With this form template, our system can then extract both handwritten and machine-typed filled-in data. Experimental results on three different types of forms show the performance of our system. Further, the proposed method is robust to noise and skew that is introduced during scanning

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. We 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. Amaai [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 extracted and 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. Current image processing techniques in the vote collection process are heavily dependent on the anticipated, geometric positioning of the vote. These techniques don't account for markings made outside of the requested field of input. Using various form dropout techniques, however, every mark on the form can be extracted and used by the machine to make an intelligent decision. Most methods will still miss a few marks and result in a few false alarms. 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].

Shuli Sun ; Lihua Xie ; Wendong Xiao ; Nan Xiao [7] is concerned with the optimal filtering problem for discrete-time stochastic linear systems with multiple packet dropouts, where the number of consecutive packet dropouts is limited by a known upper bound. Withoutresorting to state augmentation, the system is converted to one with measurement delays and a moving average (MV) colored measurement noise. An unbiased optimal filter is developed in the linear least-mean-square sense. Its solution depends on the recursion of a Riccati equation and a Lyapunov equation.

III. PROPOSED WORK

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.

Our objective to develop an algorithm using MATLAB & VHDL programming for Document Processing for Automatic Color Form Dropout. Implementation of FPGA which is an ideal platform for image processing engine. Developing VHDL coding for distance coding and threshold detection and to be tested using a Xilinx FPGA. Performance evolution of the proposed algorithm to be observed using MATLAB.

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

1. Color space conversion
2. Distance calculation using VHDL Coding,
3. Dropout Threshold Detection using VHDL Coding.

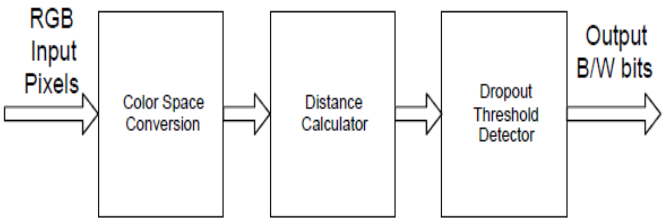


Figure 1. Color Dropout Algorithm Architecture

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 Color Space Conversion. In step of Color Space Conversion RGB Color Space is converted to YCbCr Color Space with the help of matrix multiplication. It has much better characteristics than RGB and only a matrix multiplication is required for the color space conversion based on the following transformation.

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

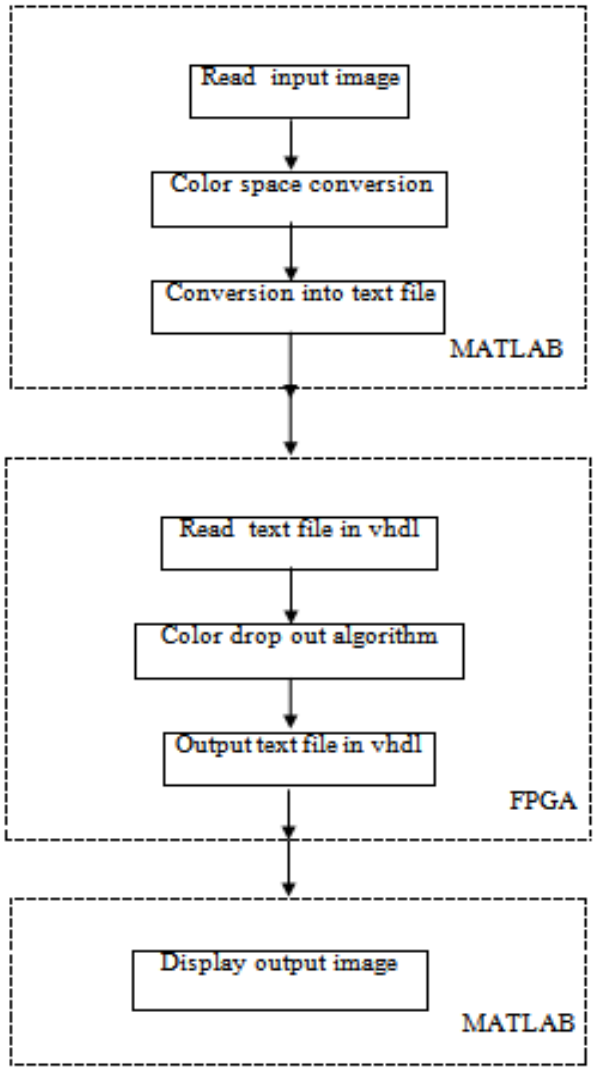
Then algorithm is to compute the distance in which select a Non Dropout Color compare it with original image pixels which comes from matrix multiplication as this is done using VHDL Coding as follows

- Find the distance between the colored pixels of interest.
- Each of the distances is compared with the associated 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

Next step will be the dropout threshold detector is a simple 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 tolerances, i.e. threshold, the output color is set to black, otherwise the output color is set to white. This is done using VHDL Coding, means output is black &white image.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.

Here we use two platforms with their compatible tool and OS's software which are describe as follows.

1. MATLAB
2. FPGA



IV. CONCLUSION

In this paper, Color Dropout algorithm 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.

Idea to develop an algorithm using MATLAB & VHDL programming for Document Processing for Automatic Color Form Dropout on FPGA platform to get impressive the speed of operation 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, reduces the information extraction process time and improves image transmission time. Future work will involve processing in other uniform color spaces with some supervise learning approach.

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.

[10] Li-jun Zhang , Li-Xin Yang , Li-Dong Guo and Jun Li "Optimal Estimation for Multiple Packet Dropouts Systems Based on Measurement Predictor" Sensors Journal,IEEE,2011,Pp:1943-1950.

[11] www.Xilinx.com.