

# A Form Dropout Method based on Line-elimination and Image-subtraction

Yoshihiro Shima<sup>1</sup> Hiroshi Ohya<sup>1</sup> Michio Yasuda<sup>2</sup>

*1 Faculty of Physical Sciences and Engineering, Meisei University*

*2 Faculty of Informatics, Meisei University*

*{shima, ohya}@ee.meisei-u.ac.jp yasuda@ei.meisei-u.ac.jp*

## Abstract

*A method of preprint elimination for form images is proposed. Form images have fixed parts, such as character strings and lines. These fixed parts are called as preprint, which are not used for data entry. 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 directly eliminated by using run data.*

## 1. Introduction

Processing for paper forms such as contact application form, payment request form, tax bill form is the essential procedure in insurance, banking and government fields. The needs of form image processing are strong, that is by optical scanning the surface of paper forms is converted to image data and is handled like electric forms [1].

The purpose of this paper is to propose a data compression method for form images by removing fixed and common preprint parts from filled-in forms. This method is based on digital image processing, and the redundancy parts such as ruled lines and preprint character lines are removed. This data compression method is called as digital dropout. This digital dropout method is useful for preprocessing of character recognition as well as data compression.

In form processing, filled-in data have important information. On the other hand fixed portion, namely, preprint, has redundant information, because preprint portion is commonly printed for the same type of forms.

Some conventional lossy compression methods are known, which are based on removing preprint of forms that is redundant for data entry. These lossy compression methods are classified into three types, namely, (a) cut-and-paste method for allotted rectangular space to be filled in, (b) subtraction

method for form image by using template bitmap, (c) form element elimination method for preprint character line and ruled line. In the above method (a), the rectangular location of allotted space is registered and filled-in sub image is cut from the entered form image and is pasted to bitmap image of which background color is white [3][4]. In the method (b), the template bitmap image is captured from the paper form that has not been filled in. The location, rotation and magnification are modified for entered form images. From modified form images, template bitmap image is subtracted pixel by pixel so that preprint portions are cleaned and filled-in characters are left [5]. In the method (c), preprints of ruled lines and character lines are extracted directly from the entered form images and those preprints are eliminated [6][7][8]. To discriminate preprinted character lines and filled-in character lines, the location and size of preprinted character lines is registered in advance by scanning the form image that has not been filled in.

The problem of method (a) is that the character is eliminated from the form image, which is filled outside the allotted rectangular region. Because the location of filled-in space, that is registered in advance, is fixed. The method (b) has the problem that the part of filled-in character, which is near or touching with the preprint, is lacked by subtraction from template bitmap image because of location shift, distortion, and magnification error [10][11].

In the method (c), it is not easy to discriminate between the stroke of filled-in character and the short ruled-line. There is a problem that the stroke of filled-in character is disappeared if the short ruled-line of which length is the same as the character stroke is eliminated. In this paper, a new digital dropout method is proposed, that is based on the combination of the above method (b) and method (c), namely, subtraction method for modified form image by using template bitmap, and form element elimination method for preprint. The MMR compression method [2] can be applied to the form images, of which preprints are removed by the proposed method.

## 2. Problems of preprint elimination from form images

Figure 1 shows an example of form images. The problems of removing redundant preprints from form images are classified in three types, namely, (1) problems of the diverse printing style of forms, (2) problems of image distortion caused by scanning of paper form surfaces, and (3) problems of the diverse way of writing characters. Figure 2 shows the sub images of forms for explanation of these problems.

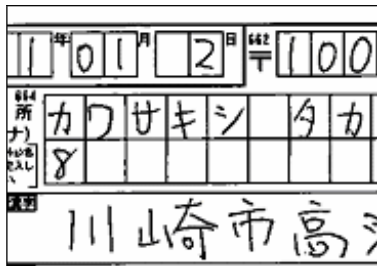
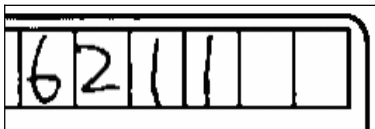
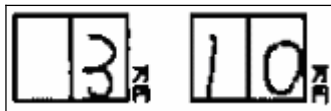


Figure 1. An example of form images



(a) Discrimination short lines from character strokes



(b) Lack of character stroke touched with the lines



(c) Check mark overwrite on guided circles

Figure 2. Sub images of forms to explain the problems of preprint elimination

## 3. Digital dropout for form images

The Digital dropout method is a kind of lossy compression and decompression for form images. Figure 3 shows a system for form image processing. In branch offices entered form images are captured and preprint portions are eliminated from form images by using the template bitmap image. The form images, of which preprints are removed, are converted to the coded data by using lossless MMR (Modified Modified Read) compression. MMR coded images are

transferred to the center office. In the data entry center, after coded images are expanded, the template bitmap image and form image without preprint are overlapped and displayed on the screen to check the entered form by human eyes.

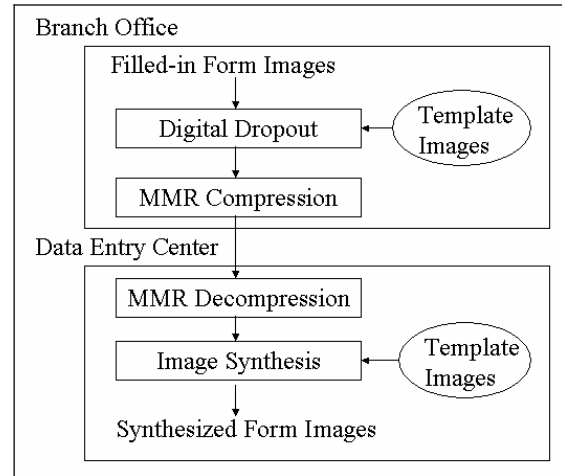


Figure 3. A system for form image processing

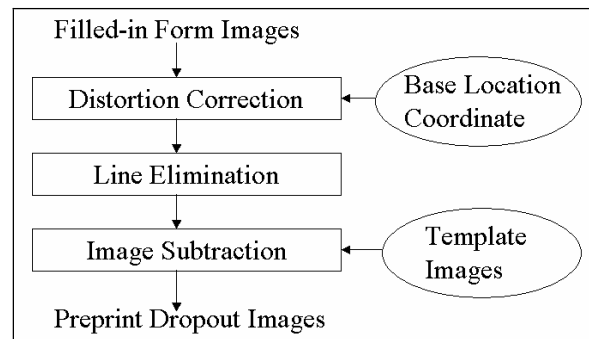


Figure 4. A flow of digital dropout

Figure 4 shows the flow of the proposed digital dropout. Digital dropout is the process of elimination of preprints, from the entered form images, which are common to the same type of forms. In advance, the template bitmap images are created semi-automatically from the paper form that is not filled-in. The template images are binary and are fattened to overcome the location shift error. From template image, long lines and overwrite checkmark guides are eliminated not to disappear the filled-in mark.

Form images captured by optical scanners have some distortion, location shift and elasticity. So for inputted form images, first, the nonlinear distortion, location shift, rotation and magnification are corrected by using the basis of the template images. Secondly, ruled-lines are extracted and eliminated in some condition from

corrected form images. Thirdly, template images are subtracted from form images. The result of image subtraction is the filled-in portions to be extracted. The color of the filled-in portions is black and the one of the removed preprints and background is white. The subtracted image are coded by lossless MMR compression and transferred to the center office.

In form images, some rotation, magnification, shear distortion and location shift are occurred. These distortions are corrected and are adjusted to the template images.



Figure 5. Extraction of based lines

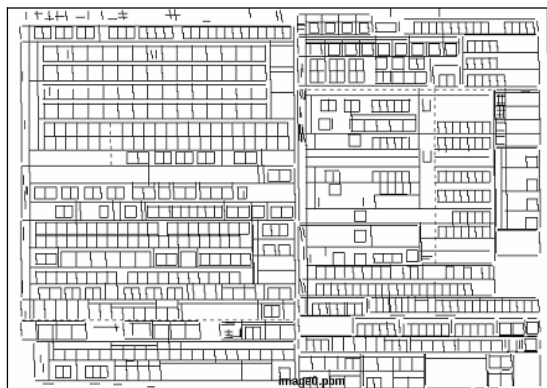


Figure 6. Result of shift correction

This rotation process has two steps. First step is the detection of rotation angle. Second step is the skew correction for form images. First, horizontal lines are extracted and the skew angles of these lines are detected. Secondly, the skew of the images is corrected by using the run data of the images.

This magnification process contains three steps, namely, based line extraction, magnification rate calculation, and image expansion and contraction. Two horizontal lines and two vertical lines are in advance registered as the based lines of the template image. In inputted images, these lines are searched and

extracted. Figure 5 shows the result of extraction of four based lines.

The shear distortions of form images are corrected. To detect the shear angle, the vertical based lines are extracted and the slant angle of those lines are detected, that is regarded as the shear angle. By using run data, shear distortion is corrected. Let the coordinate of image be  $(x,y)$  and the corrected image be  $(x',y')$ . Let the shear angle be  $\beta$ . The coordinate of the shear corrected image is shown as

$$x' = x + y \tan \beta ,$$

$$y' = y .$$

Horizontal and vertical shifts are corrected by extracting the based lines after correcting the shear distortion. Location shift is executed by using the run data. Figure 6 shows the result of shift correction.

After correcting the distortion images, ruled lines are extracted and eliminated. Long lines are directly extracted from corrected form images and eliminated by using run data. On the other hand, short lines are subtracted by using the template bitmap. To overcome the difficulty of discrimination of short lines and character strokes, short lines are eliminated by using template images.

Form images are subtracted by template images after distortion correction and line elimination. The minuend image is the form image and the subtracter image is the template bitmap. Subtraction process is based on the run data.

In image subtraction process, the minuend run is accessed from left to right on each scanning line. At the same time, the subtracter run is accessed from left to right on the corresponding scanning line. The location coordinates of the start and end points of both run data are compared to check the overlap of the horizontal scanning line, which is shown in Figure 7. The remainder run are created according to the overlap of the minuend run and the subtracter run.

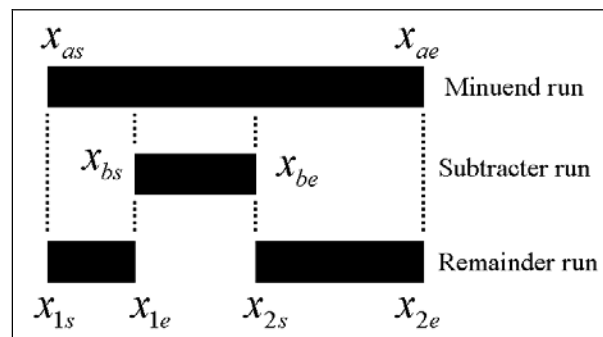


Figure 7. Image subtraction by using run data

#### 4. The experiment of digital dropout

Form images are white and black binary images. The resolution is 200dpi. Experimental programs are consisted of image capturing, image distortion correction, line elimination, template image subtraction, and image output. Program language C is used. The total step of digital dropout process is 6.5 K steps.

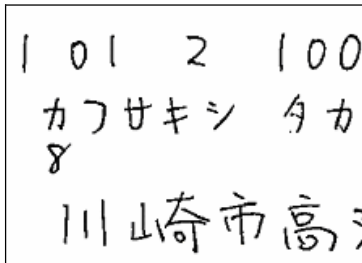


Figure 8 Result of digital dropout

Figure 8 shows the result of digital dropout for the form images shown in Figure 1. Pattern quality is good for almost the characters. Some lack of stroke is occurred for the character touching the short line.

Figure 9 shows another example of form images. Result of preprint elimination from the above form images is shown Figure 10. Figure 11 shows an example of the result of form image synthesis. Two colors are used for the display to check and correct the entry data. The template image is grey(blue) and the result image of dropout is black.

Table 1 shows the effect of data compression for 10 samples of form images. Captured raw image data is 486k Byte. By applying MMR compression to raw image, the amount of data is 78.1k Byte in average. By applying proposed lossy dropout method and MMR compression, the amount of data is 14.1k Byte. The processing time of digital dropout is 6.9 seconds for A4 form images ,where CPU clock is 160M Hz.

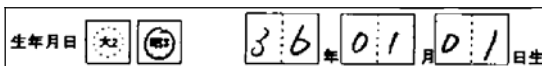


Figure 9 Another example of form images (part)

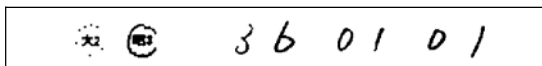


Figure 10 Result of preprint elimination (part)

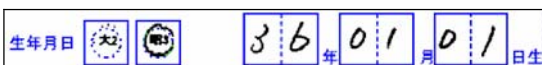


Figure 11 Result of form image synthesis (part)

Table 1 The effect of Data compression (measured)

Sample No.	Raw(K byte)	Compressed by MMR(K byte)	Compressed by Digital Dropout & MMR(K byte)
1	486	77.38	12.12
2	486	80.11	15.41
3	486	78.09	13.14
4	486	77.38	14.20
5	486	76.14	13.44
6	486	76.38	13.54
7	486	78.77	15.16
8	486	78.07	14.40
9	486	78.83	15.23
10	486	79.77	14.70
Average	486	78.09	14.13
Compression rate	1	6.2	34.4

#### 5. Conclusion

Preprints are eliminated for entered form images to decrease the amount of form image data. By subtraction of the template image and elimination of long lines, preprint portions are removed and disappeared. The elimination and subtraction process is based on run data. The effectiveness of lossy data compression is ascertained by the experiment of digital dropout for small number of form images.

#### 6. References

- [1]S. Gopisetty, R. Lorie, J. Mao, M. Mohiuddin, A. Sorin and E. Yair, "Automated forms-processing software and services," IBM J. Res. Develop., Vol.40, No.2, pp.211-230, March 1996
- [2]R. Hunter and A. H. Robinson, "International Digital Facsimile Coding Standards," Proc. IEEE, Vol.68, No.7, pp.854-867, 1980
- [3]Suzanne Liebowitz Taylor, Richard Fritzon, and Jon A. Pastor, "Extraction of Data from Preprinted Forms," Machine Vision and Application , No.5, pp.211-222, 1992
- [4]L.Simoncini and Zs. M. Kovacs-V., "A System for Reading USA Census '90 Hand-Written Fields," Proc. of 3th Int. Conference on Document Analysis and Recognition, pp.86-91, August 1995
- [5]Richard Casey, David Ferguson, K. Mohiuddin, and Eugene Walach, "Intelligent Forms Processing System," Machine Vision and Application , No.5, pp.143-155, 1992
- [6]Dacheng Wang and Sargur N. Srihari, "Analysis of Form Images," Proc. of 1th Int. Conference on Document Analysis and Recognition, pp.181-191, September 1991
- [7]Theo Pavlidis and Jiangying Zho, "Page Segmentation and Classification," CVGIP: Graphical Models and Image Processing, Vol.34, No.6, pp.484-496, November 1992
- [8]Bin Yu and Anil K. Jain, "A Form Dropout System," Proc. of 13th Int. Conference on Pattern Recognition, pp.701-705, August 1996

[9]Bin Yu and Anil K. Jain ``A Generic System for Form Dropout,"IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.18, No.11, pp.1127-1134, November 1996

□

[10]Yuan Y. Tang and Ching Y. Suen, ``Nonlinear Shape Restoration by Transformation Models," Proc. of 10th Int. Conference on Pattern Recognition, Vol.2, pp.14-19, June 1990

[11]Yun Weng and Qiuming Zhu, ``Nonlinear Shape Restoration for Document Images,"Proc. IEEE Comput. Soc. Conf. Comput. Vis Pattern Recognit., pp.568-573, 1996