iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 3 **Issue:** Issue I **Month of publication:** May 2015

**DOI:**

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3, Special Issue-1, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

# Constitution Text File Image Binarization Technique for Disgrace Text File Images

S. Madhuri[1], Miss. K. Sudha Mayee[2]
*M.Tech(DECS), Associate professor, Department of ECE, BCETFW, KADAPA*

***Abstract— Segmentation of textis done in between the document background and the foreground text of different document images. In this paper, we propose a novel document image binarization technique that addresses the adaptive image contrast. The combination of local image contrast and local image gradient is called adaptive image contrast that tolerant the different degraded document text. In the proposed technique, an adaptive contrast map is first constructed for an input degraded document image. The contrast map is then binarized and combined with Canny's edge map to identify the text stroke edge pixels. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. The proposed method is simple, robust, and involves minimum parameter tuning.***
***Index Terms— Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification.***

## I. INTRODUCTION

Binarization is performed in that aims to segment the foreground text from the document background. It is very important to ensure the different tasks of OCR that is Optical character recognition.
Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem due to the high inter/intra-variation between the text stroke and the document background across different document images. As illustrated in Fig. 1handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. In addition, historical documents are often degraded by the bleed-through as illustrated in Fig. 1(a) and (c) where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts as illustrated in fig.e


Fig. 1. Five degraded document image examples (a)Ð(d) are taken from DIBCO series datasets and (e) is taken from Bickley diary dataset.

## II. LITERATURE SURVEY

A new method is presented for adaptive document image binarization, where the page is considered as a collection of subcomponents such as text, background and picture. The problems caused by noise, illumination and many source type-related degradations are addressed. Two new algorithms are applied to determine a local threshold for each pixel. The performance evaluation of the algorithm utilizes test images with ground-truth, evaluation metrics for binarization of textual and synthetic images, and a weight-based ranking procedure for the "nal result presentation. The proposed algorithms were tested with images

including different types of document components and degradations.

## III. RELATED WORK

For segmenting the text from the document the local image contrast and the local image gradient are very useful features because the document text usually has certain image contrast to the neighboring document background.for this many techniques are there one of thE BERNSEN'S method says as shown

Local image contrast is as follows

$$C(i, j) = I_{max}(i, j) - I_{min}(i, j) \quad (1)$$

where $C(i, j)$ denotes the contrast of an image pixel $(i, j)$, $I_{max}(i, j)$ and $I_{min}(i, j)$ denote the maximum and minimum intensities within a local neighborhood windows of $(i, j)$, respectively.but complex back ground it cannot work properly.

We have earlier proposed a novel document image bina-rization method by using the local image contrast that is evaluated as follows

$$C(i, j) \quad \frac{I_{max}(i, j) - I_{min}(i, j)}{I_{max}(i, j) + I_{min}(i, j) + €} \quad (2)$$

where €_is a positive but inÞnitely small number that is added in case the local maximum is equal to 0. Compared with Bernsens.inthis local image contrast introduces a normalization factor (the denominator) to compensate the image variation within the document back-ground.

Local image gradient: An image gradient is a directional change in the intensity or color in an image. Image gradients may be used to extract information from images.In graphics software for digital image editing, the term gradient or color gradient is used for a gradual blend of color which can be considered as an even gradation from low to high values, as used from white to black in the images to the right. Another name for this is color progression.

## IV. PROPOSED METHOD

Document image bina-rization techniquesis described in our proposed method . Given a degraded document image, an adaptive contrast map is first constructed and through the combination of the bina-rized adaptive contrast map and the canny edge map text stroke edges are detected. By using local threshold the text is segmented based on the the estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

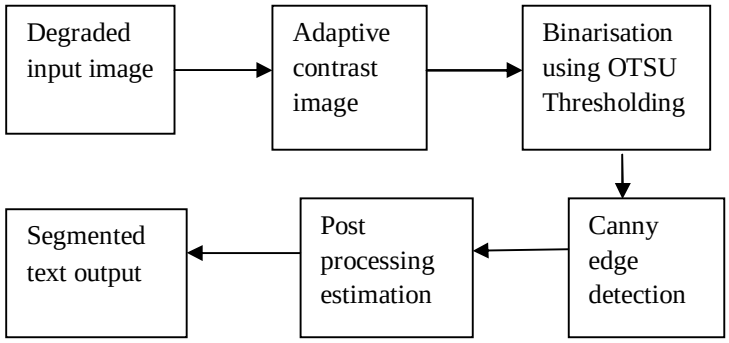



FIG 3.1: Block Diagram Of Proposed System

### A. *Contrast Image Construction*

It can detects many non-stroke edges from the background of degraded document that often contains certain image variations due to

noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to com-pensate the image variation within the document background.

The local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 2. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient, suppresses the image variation within the document background by using normalization facor. For image pixels within bright regions, it will produce a large normal-ization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the image pixels within dark regions, it will produce a small denominator and accordingly result in a relatively high image contrast. However, the image contrast in Equation 2 has limitation that it cannot handle text file images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small.


Fig3. 2. Contrast Images constructed using (a) local image gradient (b) local image contrast , and (c) our proposed method of the sample document images

In 3.2a suppress the noise at the upper left of image by normalization factor and (b)shows the contrast map it can remove the high intensity of the text and then finally our proposed methed adaptive image contrast shown in (c)can produce proper contrast maps for document images with different types of degradations.

*B. Adaptive Image Contrast*

To overcome this over-normalization prob-lem, we combine the local image contrast with the local image gradient and derive an adaptive local image contrast as follows:

$$C_a(i,j) = \alpha C(i,j) + (1-\alpha)(I_{max}(i,j) - I_{min}(i,j)) \quad (3)$$

where $C(i,j)$ denotes the local contrast in Equation 2 and $(I_{max}(i,j) - I_{min}(i,j))$ refers to the local image gradient that is normalized to [0, 1]. The local windows size is set to 3 empirically. $\alpha$ is the weight between local contrast and local gradient that is controlled based on the document image statistical information. Ideally, the image contrast will be assigned with a high weight (i.e. large $\alpha$) when the document image has signiÞcant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. The proposed binarization technique relies more on image gradient and avoid the over normalization problem of our previous method [5].

We model the mapping from document image intensity variation to $\alpha$ by a power function as follows:

$$\alpha = \frac{Std}{128}^{\gamma} \quad (4)$$

where *Std* denotes the document image intensity standard deviation, and $\gamma$ is a pre-deÞned parameter.

*C. Text Stroke Edge Pixel Detection*

We therefore detect the text stroke edge pixel candidate by using OtsuÕs global thresholding method.

The binary map can be further improved through the combination with the edges by CannyÕs edge detector because CannyÕs edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image.

CANNY EDGE DETECTOR: It can detect as many real edges as possible it is

Good detector: it can mark as many as real edges

Good localization: it can mark as close as possible

Minimal response: edge should be marked once and where it possible

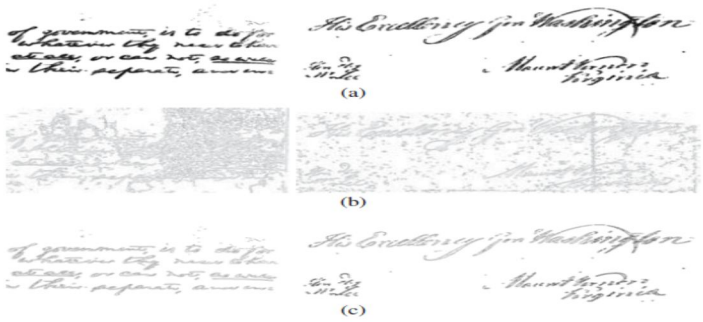


Fig. 3.3 (a) Binary contrast maps, (b) canny edge maps, and their (c) combined edge maps of the sample document images

Fig. 3.3(a) shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Fig. 3.3(b) without tuning the parameter manually. In the combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map. The combination helps to extract the text stroke edge pixels accurately as shown in Fig. 3.3(c).

*D. Local Threshold Estimation*

The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

$$R(x,y) = \begin{cases} 1 & I(x,y) \le E_{\text{mean}} + \frac{E_{\text{std}}}{2} \\ 0 & \text{otherwise} \end{cases}$$

where E mean and E std are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window W , respectively stroke width of the document image under study, E W , which can beestimated from the detected stroke edges as stated in Algorithm 1.

Algorithm 1 Edge Width Estimation

Require: The Input Document Image I and Corresponding Binary Text Stroke Edge Image Edg

Ensure: The Estimated Text Stroke Edge Width E W

1: Get the width and height of I

2: for Each Row i = 1 to height in Edg do

3: Scan from left to right to find edge pixels that meet the

following criteria:

a) its label is 0 (background);

b) the next pixel is labeled as 1(edge).

4: Examine the intensities in I of those pixels selected in Step 3, and remove those pixels that have a lowerintensity than the following pixel next to it in the same row of I .

5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.

6: end for

7: Construct a histogram of those calculated distances.

8: Use the most frequently occurring distance as the estimated stroke edge width E W .

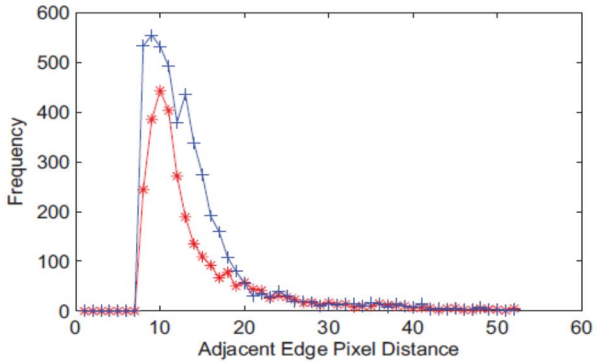



Fig. 4.1 Histogram of the distance between adjacent edge pixels. The "+++"line denotes the histogram of the image in Fig. 1(b). The "***" line denote the histogram of the image in Fig. 1(d).

*E. Post- Processing*

binarization result can be further improved that can be described in Algorithm 2. First, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely.Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foregroundtext).

Algorithm 2 Post-Processing Procedure

Require: The Input Document ImageI, Initial Binary Result B and Corresponding Binary Text Stroke Edge Image Edg

Ensure: The Final Binary Result B f

1: Find out all the connect components of the stroke edge pixels in Edg.

2: Remove those pixels that do not connect with other pixels.

3: for Each remaining edge pixels (i, j ): do

4: Get its neighborhood pairs: (i − 1, j ) and (i + 1, j); (i, j − 1) and (i, j + 1)

5: if the pixels in the same pairs belong to the same class (both text or background) then

6:Assign the pixel with lower intensity to foreground class (text), and the other to background class.

7: end if

8: end for

9: Remove single-pixel artifacts [4] along the text stroke boundaries after the document thresholding.

10: Store the new binary result to B f

## V. EXPERIMENTS AND DISCUSSION

The proposed technique is then tested and compared with state-of-the-art method over on three well-known competition datasets: DIBCO 2009dataset H-DIBCO 2010 dataset and DIBCO 2011dataset Finally, the proposed technique is further evaluated over a very challenging Bickley diary dataset . In this experiment, we compare our proposed method with other techniques on DIBCO 2009,H-DIBCO 2010 and DIBCO 2011 datasets. These methods include Otsu's method (OTSU) Sauvola'smethod(SAUV),Niblack's method (NIBL)Bernsen'smethod (BERN) Gatos et al.'s method (GATO) and our previous methods (LMM BE ). TheDIBCO 2009 dataset contains ten testing images that consist of five degraded handwritten documents and five degraded printed documents. The DIBCO 2011 dataset contains eight degraded handwritten documents and eight degraded printed documents. In total, we have 36 degraded document images with ground truth.

TABLE I: EVALUATION RESULTS OF DATASET OF DIBCO 2009

| methods | Fmeasure % | PSNR | NRM | MPM | Rank score |
|---|---|---|---|---|---|
| OTSU | 78.72 | 15.4 | 5.6 | 13.3 | 196 |
| SAUV | 85.41 | 16.3 | 6.4 | 3.45 | 177 |
| NIBL | 55.82 | 9.2 | 16.3 | 61.3 | 251 |
| BERN | 52.48 | 8.7 | 14.7 | 113.2 | 313 |
| GATO | 85.25 | 16.5 | 10 | 0.7 | 176 |
| LMM | 91.06 | 18.0 | 7 | 0.3 | 126 |
| Proposed method | 93.5 | 19.6 | 3.2 | 0.43 | 100 |

TABLE II: EVALUATION RESULTS OF DATASET OF DIBCO 2011

| methods | Fmeasure % | PSNR | DRD | MPM | Rank score |
|---|---|---|---|---|---|
| OTSU | 82.2 | 15.67 | 8.4 | 15.6 | 412 |
| SAUV | 82.3 | 15.2 | 8.0 | 9.2 | 403 |
| NIBL | 68.8 | 12.7 | 28.6 | 26.9 | 362 |
| BERN | 47.2 | 7.2 | 82.3 | 136.3 | 664 |
| GATO | 82.3 | 16.9 | 5.4 | 7.2 | 353 |
| LMM | 85.3 | 16.5 | 6.2 | 6.4 | 516 |
| Proposed method | 87.8 | 17.56 | 4.84 | 5.85 | 307 |

Volume 3, Special Issue-1, May 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

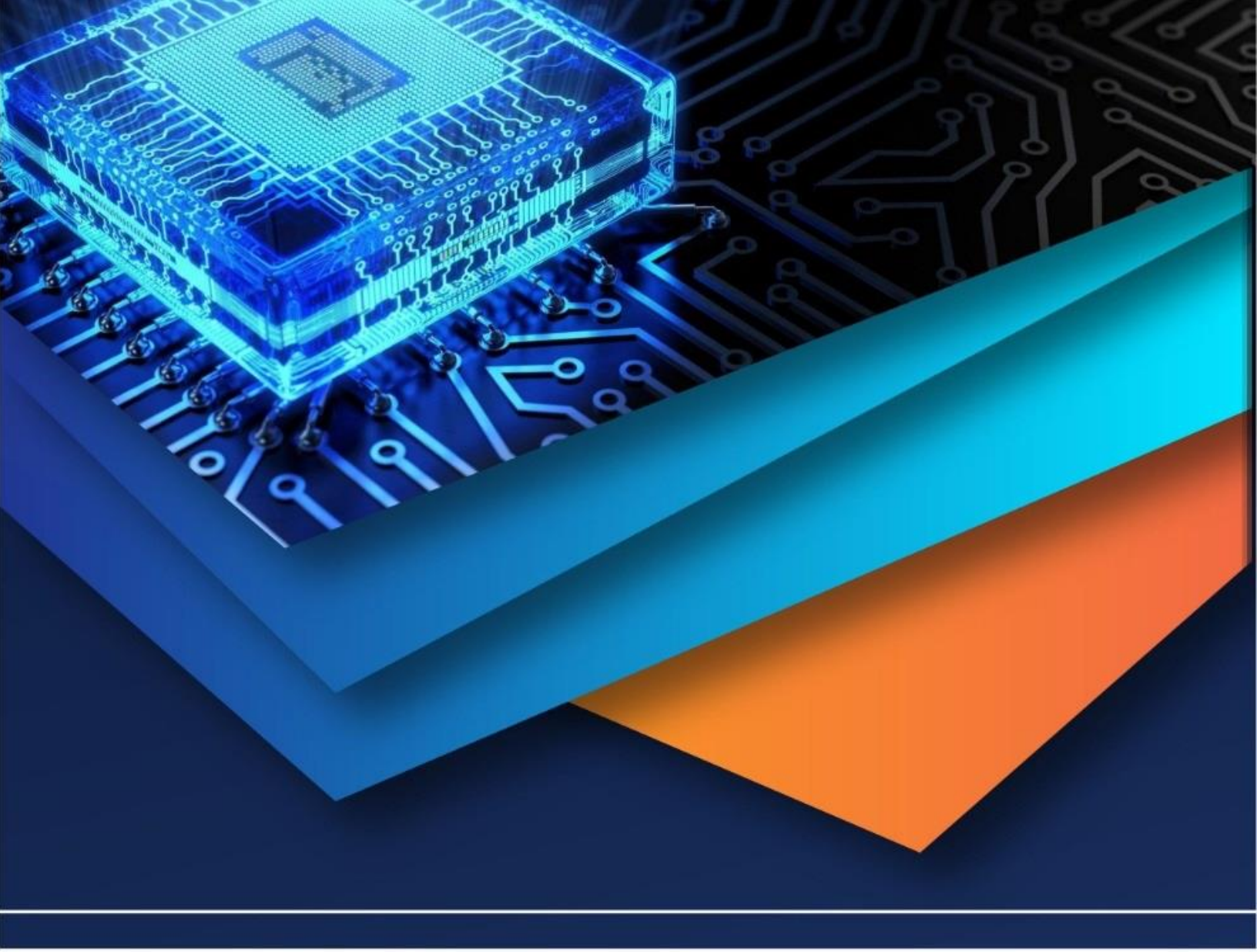

Fig. 5.1 Binarization results of the sample document image in Fig. 1(a)produced by different methods. (a) OTSU (b) SAUV (c) NIBL d) BERN (e) GATO [21]. (f) LMM (g) BE (h) Proposed.

## VI. CONCLUSION

This paper presents document image binarization technique based on the adaptive image contrast that is tolerant to different types of document degradation such as uneven illumination and document spot. The proposed technique is simple and robust, only few parameters are involved. Moreover, it works for different kinds of degraded text file images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measure, PSNR, NRM, MPM and DRD.

## REFERENCES

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 1375–1382.
[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 hand written document image binarization competition," in Proc. Int. Conf Frontiers Handwrit. Recognit., Nov. 2010, pp. 727–732.
[3] S. Lu, B. Su, and C. L. Tan, "Document image binarization using back ground estimation and stroke edges," Int. J. Document Anal. Recognit. vol. 13, no. 4, pp. 303–314, Dec. 2010.
[4] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," in Proc. Int. Workshop Document Anal. Syst., Jun. 2010, pp. 159–166.
[5] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images," in Proc. Int. Conf. Document Anal.Recognit., vol. 13. 2003, pp. 859–864.
[6] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," J. Electron. Imag., vol. 13,no. 1, pp. 146–165, Jan. 2004.
[7] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 12, pp. 1191–1201, Dec. 1995.
[8] O. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 3, pp. 312–315,Mar1955



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY