An effective binarization method for readability improvement of stain-affected (degraded) palm leaf and other types of manuscripts

Lalit Prakash Saxena*

Department of Computer Science, University of Mumbai,
Vidyanagari, Santacruz (E), Mumbai 400 098, India

A stain is any colour change perceived as lying within the manuscript fibres. Eliminating stains is considered as more significant, since it leads to improved manuscript image quality and hence improved readability. The communication describes an improved binarization method for removing stains from severely degraded manuscript images where the text is unclear. The method is tested on manuscript images on different media having different scripts: palm leaf (Grantha), rock (Brahmi), and paper (Modi, Newari, Persian and Roman), and Document Image Binarization Contest datasets. In this work, we obtained 66.27%, 92.15%, 97.90%, 56.23%, 78.62% and 98.91% readability for rock, palm leaf and paper manuscripts respectively.

Keywords: Binarization, degradation, manuscripts, readability improvement, stains.

A stain is any colour change perceived as lying within the manuscript fibres rather than on top of them[1]. There are several causes for the stains found on manuscript surfaces, captured through a camera or scanner[2]. Stains often affect folio edges, but when they spread to the interior portions and affect the text content the readability is affected, especially when the R, G, B (red, green, blue) values of the stains coincide with the R, G, B values of the text. Diminished readability of the folios and folio images leads to corrupted or incomplete knowledge.

The main cause of stains on the surface of old manuscripts is fungal growth[2]. Most of the stains, smears and spots are bluish, greenish, blackish and brownish in colour. Dirt, dust particles and insect remnants (faecal matter and dead bodies) also cause stains on the manuscripts. Improper and inadequate manual handling also adds to stains. Sometimes the thread used for tying the folios together and cloth used to protect palm leaves also leave stray marks that contribute to stains.

For researchers working with images of fragile and degraded documents, the elimination of stains is considered to be more significant[3], since stain removal leads to improved manuscript image quality and hence improved readability. In this situation, the text on the stained portions needs to be recovered through improved binarization (conversion of greyscale to black and white image).

Binarization is an essential process used in many image processing and segmentation applications. This communication describes an improved binarization method for removing stains from severely degraded and stained manuscript images where the text is unclear.

The binarization technique described below divides the image into windows of size ranging from 3 × 3 to 15 × 15 and calculates a local threshold value for each window by finding the mean $m(x, y)$ and standard deviation $s(x, y)$ of the grey values in an image. The window size is determined by a trade-off between size of the image and processing time, since smaller the window size, and hence more number of windows for a given image size, more will be the processing time. (A simple calculation: Suppose N is the number of pixels in the image and M is the number of pixels in a window. Then calculating the mean takes $O(N)$ time, calculating the standard deviation takes $O(L)$ time, where L is the number of greyscale values, total time per window is $O(N + L)$, and there are N/M windows; so time per image is $O(N(N + L)/M))$. Therefore, smaller the window size M for a given image size N, more will be the processing time.)

The proposed method suits the degraded manuscript images, where scanning is not possible; only camera photographing is permissible. It will be seen to be effective in segmenting degraded manuscript images into readable texts images. In particular, this method is suitable for segmentation of document images with dark and complex backgrounds.

Since a global threshold does not produce acceptable binarization results for images with complex backgrounds[4], local thresholding methods have been proposed[5,6]. Niblack's method calculates local thresholds $t(x, y)$ using the formula $t(x, y) = m(x, y) + k \cdot s(x, y)$, where $m(x, y)$ is the local mean and $s(x, y)$ is the standard deviation in the pixel neighbourhood, and k takes values from 0.2 for light backgrounds to –0.2 for darker backgrounds.

Niblack's method needed enhancement for images with low contrast between foreground and background pixel intensities. The method adds noise around the text, making it unreadable or unrecognizable to Optical Character Recognition (OCR). Thus, the threshold value obtained by this method is not recommended for applications performing character recognition. An increase and decrease in the window size from 15 × 15, noise gets reduced and increased but the text gets fainter and thicker respectively, both of which are unfit for recognition purposes. In the proposed method, this factor is controlled by adding 1 to the constant $k = [-0.2 \text{ to } +0.2]$. This removes additional noise in the binarization results, making the characters legible enough for OCR applications.

Sauvola modified Niblack's equation to

$$t(x,y) = m(x,y)\left[1 + k\left(\frac{s(x,y)}{R} - 1\right)\right]$$

*e-mail: lasaxmail@aol.com

with $R = 128$, where R is the dynamic range standard deviation, and constant k ranges between 0.2 and 0.5 depending on the image. This served the purpose of separating the background and foreground having similar pixel intensities. It reduces the extra noise produced in Niblack's method, but stroke connectivity is not fully maintained. This is because of the value of R was fixed at 128 and $k = 0.5$ was used in this method. It was insufficient for reducing stain effects and maintaining the stroke connectivity of the textual contents.

In Sauvola's method ($R = 128$), the dynamic range of the standard deviation has minimal effect on the binarization results for 8-bit monochrome images. To control this factor, in the proposed method, the value of R to 255, sufficient for both 8-bit monochrome and colour images. The window size and constant $k = 0.5$ affect the binarization procedure; small value of k results in thick and overlapping characters, and small window size produces thin characters.

To retain positive effects of both Niblack and Sauvola methods, but remove the deficiencies of the latter, they have modified Sauvola's method by increasing the value of R to 255. This choice of R maintains stroke connectivity and also removes the dark and complex background of the stain-affected manuscript images. The modified binarization method is given by

$$t(x, y) = \frac{1}{4}\left[1 + 2(k+1)s(x, y) - \left(\frac{m(x, y)}{R}\right)\right].$$

In the present method, the constant of proportionality ¼ gives satisfactory results with Grantha and other manuscript images. This value is not fixed for all types of images and can be changed according to the required result. Also, the value of k varies between –0.5 and +0.5. In the proposed method k does not play a major role, since the value of $2(k + 1)s(x, y)$ will always be a positive value. However, the effect of window size is observable in terms of processing time, smaller the window size, the more will be taken for time processing. The factor $(m(x, y)/R)$ leads to improvement in terms of preserving the width of the strokes, maintaining the shape and connectivity, which is solely needed for recognition purposes. This is not possible in both the Niblack or Sauvola methods. This solution to the issue of stains on manuscript images is simple, quick and cost-effective.

Su *et al.*[7] have controlled the contrast by employing local optimum grey values for the normalization of the image gradient of greyscale document images. The technique does not perform satisfactorily in the case of low-resolution scanned document images and 'bleed-through' in folios with writing on both sides. They proposed that within a local neighbourhood window, pixel intensities on either side of the character stroke boundaries should be used to determine maxima and minima for thresholding of old images.

Lu *et al.*[8] propose a polynomial smoothing surface estimating the shading variations and degradations to produce a rough estimation of uniform illumination in the document image. The global thresholding technique is then applied over the compensated document images. This method is robust and tolerant to the text size and document contrast variations, but more complex shading variations are its limitation. For text documents the model work as well, it produces incorrect shading variation estimation for graphics documents.

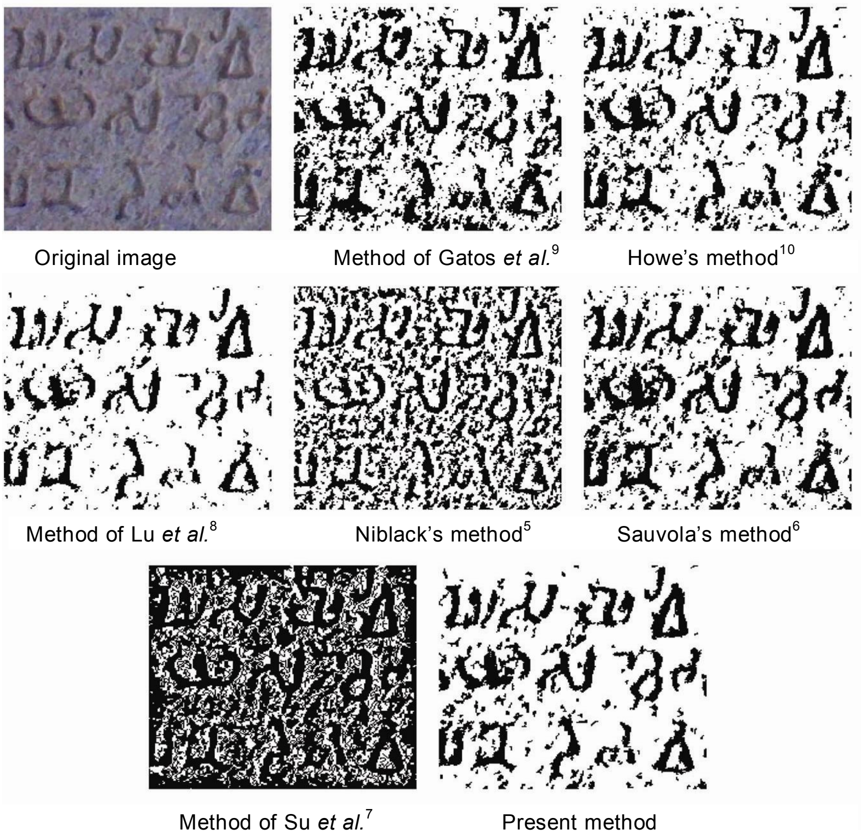
Gatos *et al.*[9] proposed an adaptive thresholding method to enhance degraded document images that interpolate neighbouring background intensities for background surface estimation. This method improved the recognition accuracy in the case of low-resolution greyscale images. This process emphasized on enhancing low-illumination, uneven background and details of the regions of interest towards the background and foreground separation.

Howe[10] proposed an automatic parameter-tuning algorithm for binarization of the document images. In all, six parameters were considered in this algorithm; two are of importance and need automatic tuning, while four are less influential and they can be set to be constant for the remaining processing. The algorithm uses stability criterion and simple heuristics in choosing these two parameter values. The method is efficient as it takes about one-eighth of the simple binarization method with fixed parameters.

The proposed method is tested on manuscript images on different media and containing different scripts: palm leaf (Grantha), rock (Brahmi) and paper (Modi, Newari, Persian and Roman), and Document Image Binarization Contest (DIBCO) datasets. Figures 1–9 show a comparison of the proposed method with the state-of-the-art methods available in the literature.

The proposed method handles the low contrast and low illumination of the Brahmi manuscript image effectively, at least for most of the image region. The resultant image contains clean symbols with no background noise and good intra-connectivity, whereas the noise at some regions obscures the symbols. There is a nominal presence of blurred regions in the corners of the thresholded image obtained by the proposed method, and the corner symbols are thicker than the symbols present in the original image. Further enhancement and fine-tuning are required in the proposed algorithm to get complete readability of the image symbols. The present method successfully removed all the stain effects. Also intra-connectivity is maintained in this method.

The present binarization method cleans the stains from degraded manuscript folio images and makes them sufficiently legible for meaningful interpretation. Although the method is not perfect, i.e. it does not reduce every stain effect from manuscript images, much improvement can be observed in the experimental results. The text symbols are clearly visible and more readable. As a result,



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

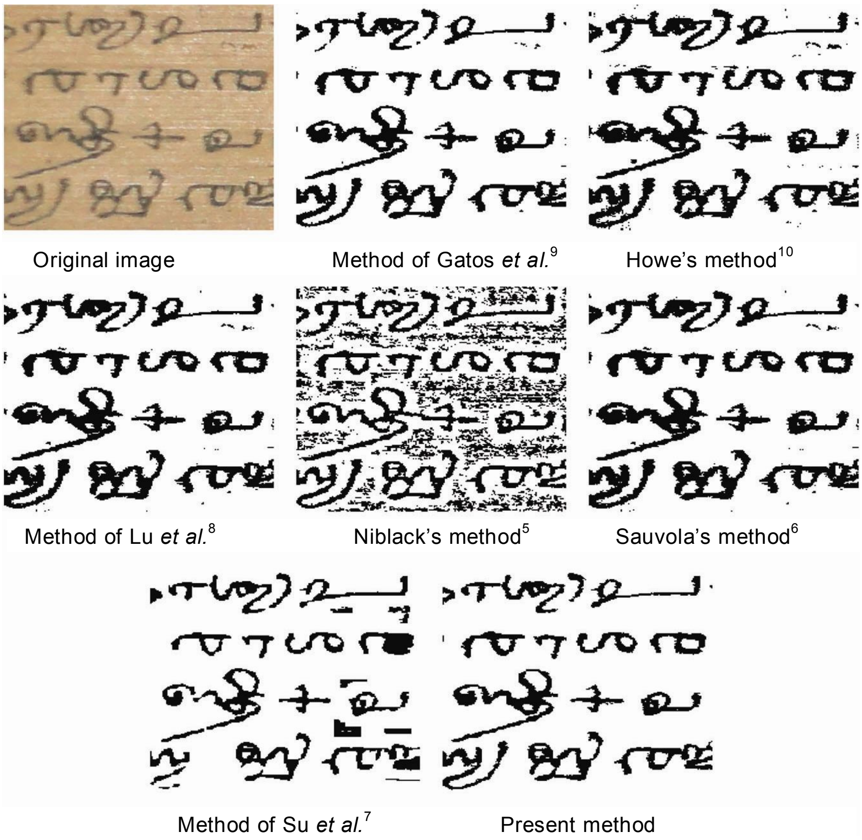
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 1. Brahmi rock inscriptions.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

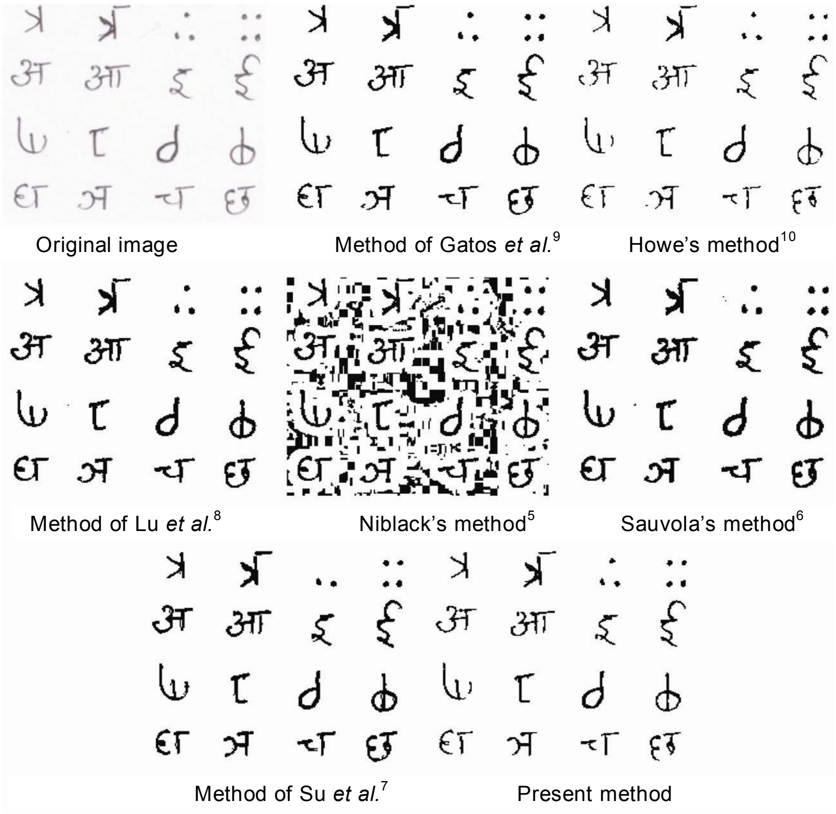
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 2. Grantha palm-leaf manuscript images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

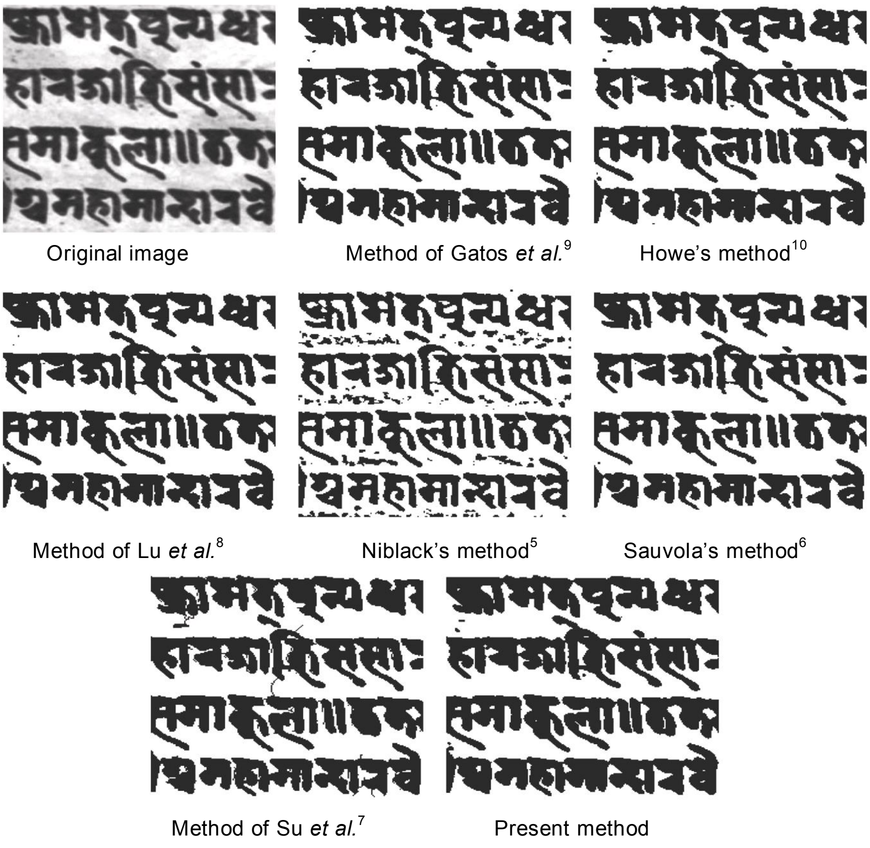
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 3. Modi paper document images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

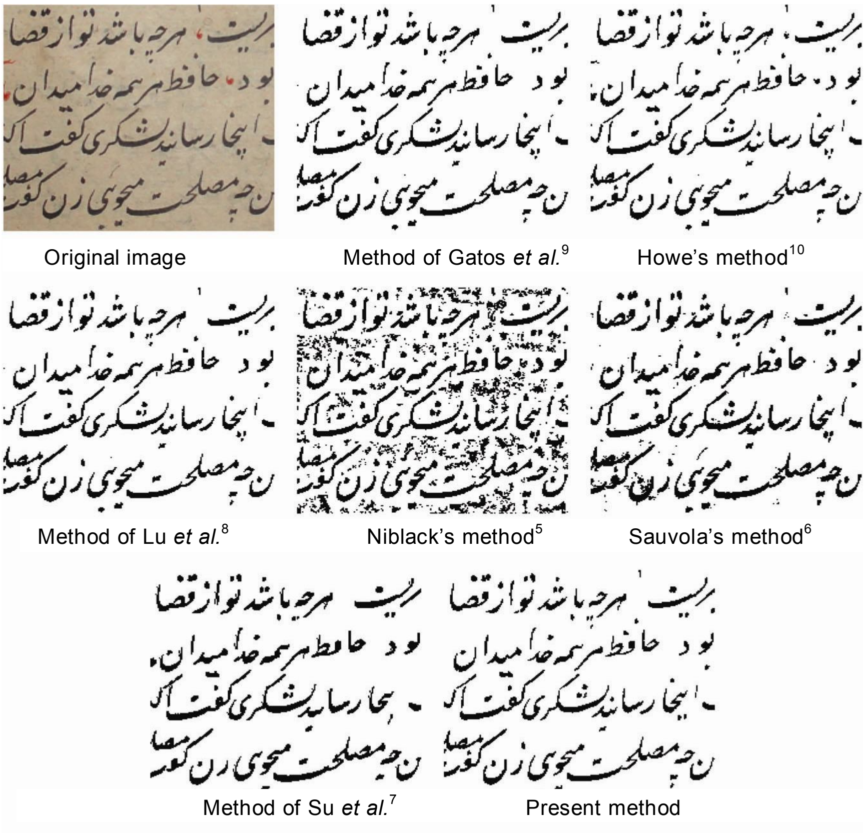
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 4. Newari old paper manuscript images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

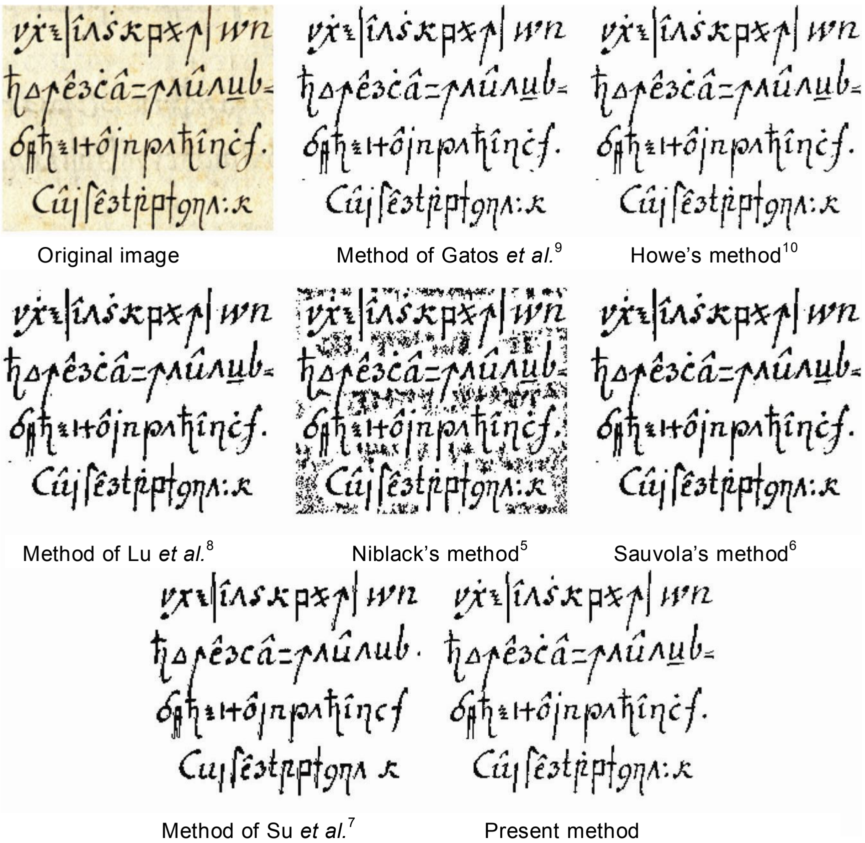
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 5. Persian old paper manuscript images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

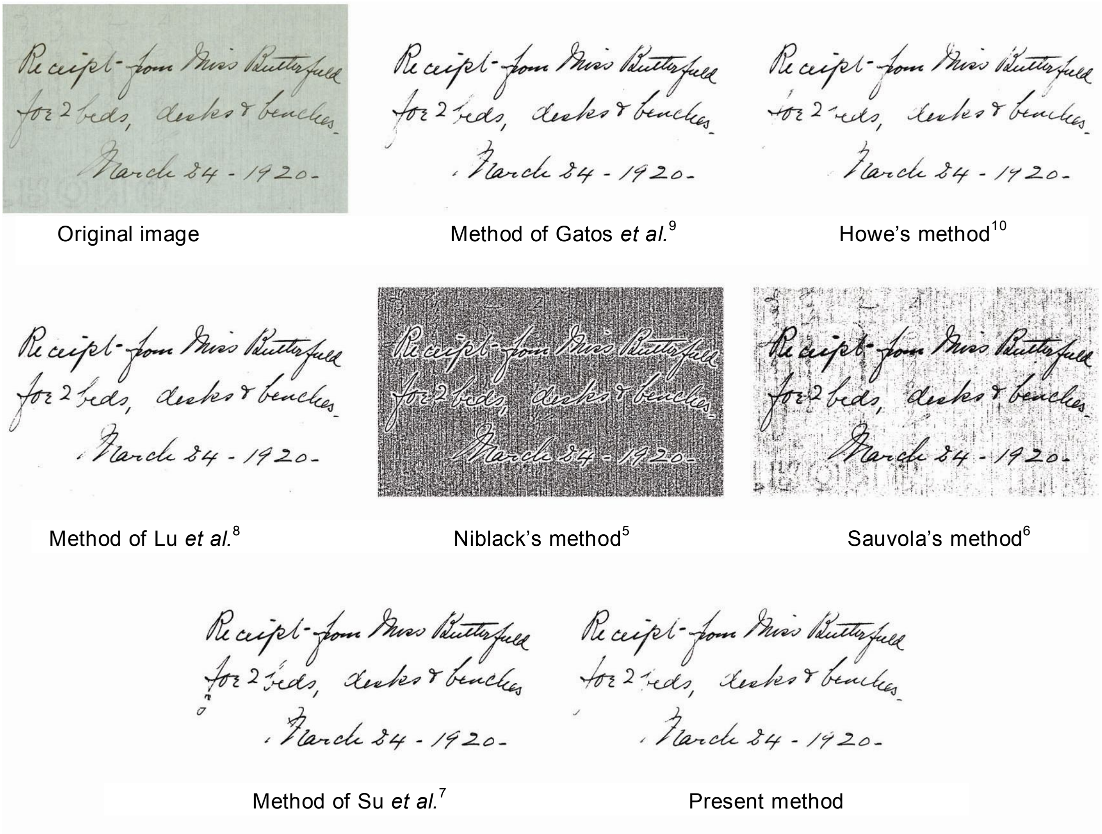
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 6. Roman old paper manuscript images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

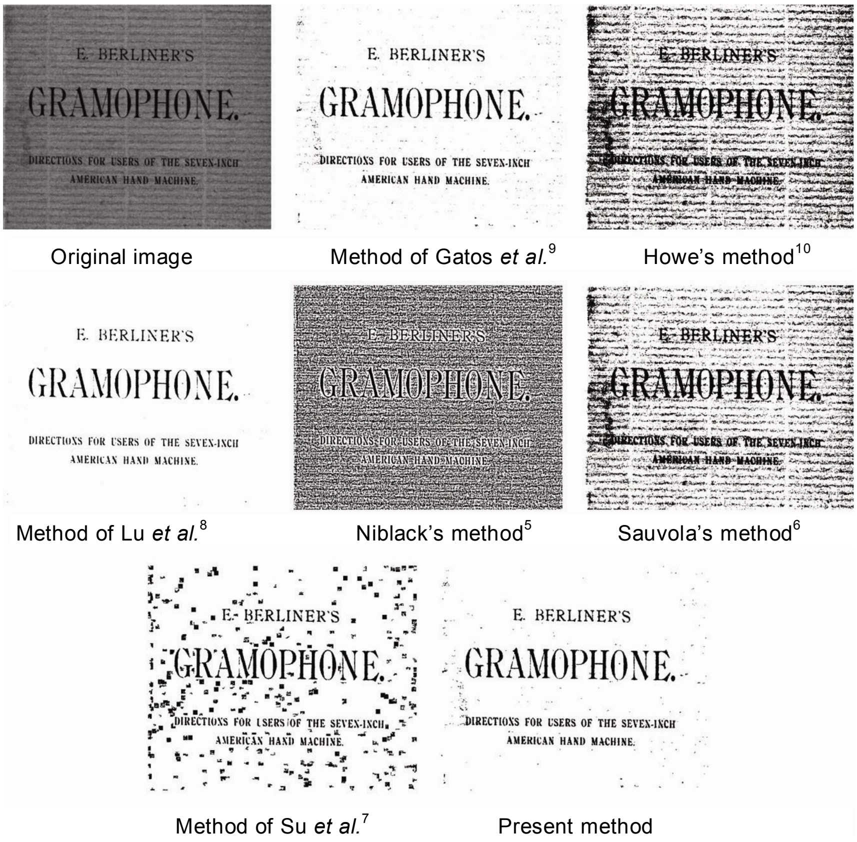
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 7. DIBCO 2011 handwritten paper document images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

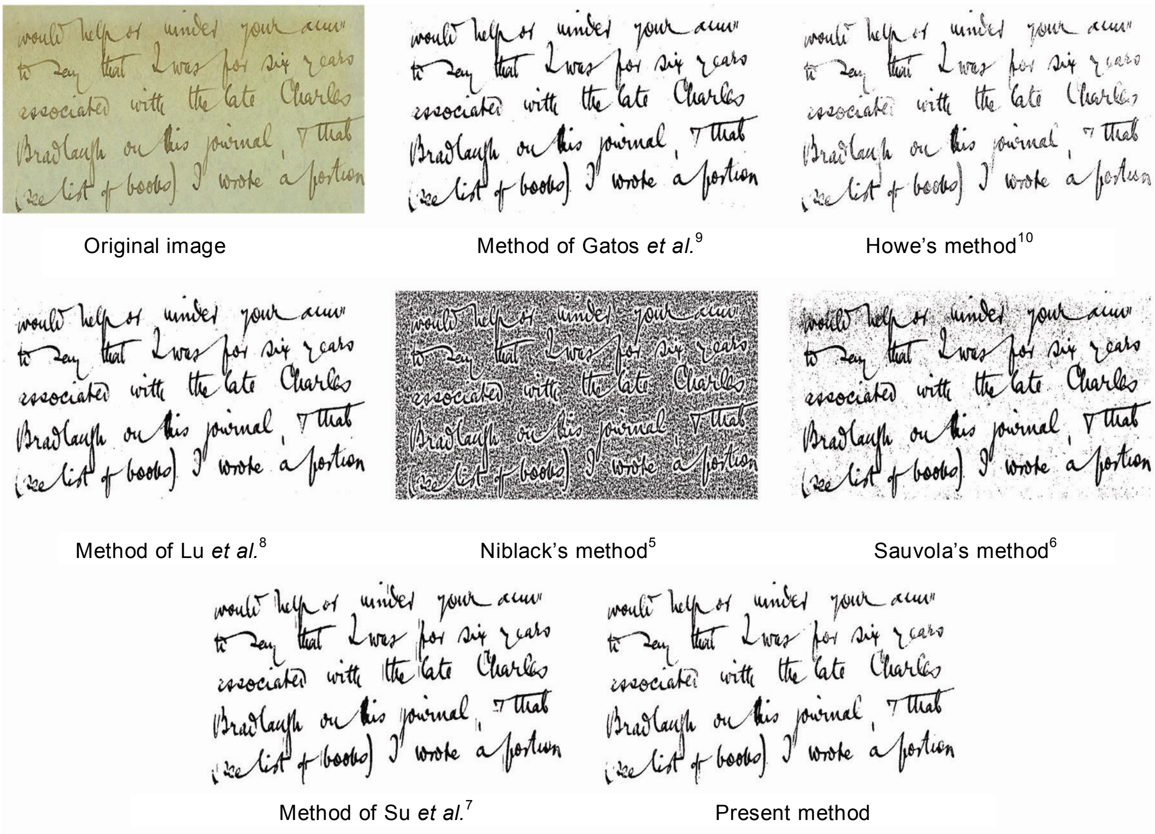
Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 8. DIBCO 2011 machine-printed paper document images.



Original image

Method of Gatos *et al.*[9]

Howe's method[10]

Method of Lu *et al.*[8]

Niblack's method[5]

Sauvola's method[6]

Method of Su *et al.*[7]

Present method

Figure 9. HDIBCO 2012 handwritten paper document images.

the manual effort to recognize stain-affected symbols is much reduced. The method satisfies two concerns: maintains the shape of symbols and preserves their intra-connectivity.

In particular, for Grantha symbols on palm-leaf manuscripts (for which the present method was proposed) it obtained 92.2% recognition, while the scanned images of Modi and Roman manuscripts had near-ideal recognition rates of 97.9% and 98.9% respectively. On the other hand, it obtained 66.3% recognition for images of rock inscriptions in Brahmi, 56.2% for Newari manuscripts and 78.6% for Persian manuscripts. An explanation of these relatively low recognition rates is that images of rock inscriptions have low contrast even with enough illumination, while the use of scanners on degraded documents was absolutely forbidden, and for taking photographs sunlight or other high-illumination sources cannot be used. In this context, the figures would be more acceptable.

Attempts to eliminate holes, tears and lost areas in the Grantha manuscripts caused decreased intra-connectivity of the text symbols. No ink or other high-contrast substance had been used for writing the symbols on the Brahmi rock inscriptions.

Enhanced images of the manuscripts are ready for interpretation and transliteration to a modern script or a script of one's choice. On the basis of expert comments, the enhanced image can be used as the source for transliteration for 90–95% of the text, and it only need to take recourse to the original image only for the remaining symbols. This practice is not recommended, especially for already degraded manuscripts, because manual handling of the manuscripts, even if it is carefully done, adds to their deterioration and subsequent loss of text content. Working on the cleaned images is a way to avoid regular handling, thus minimizing degradation of the original manuscripts.

An enhanced manuscript image is suitable for character recognition without any stains or degradations affecting the text. The proposed recognition system solely depends on the script type. The writing style of the script is the most important aspect of any recognition system. For example, Grantha script characters are written distinctly – a type of non-cursive style of writing. Here, the bounding box algorithm can easily separate the characters and template matching will be effective to recognize the characters.

Work is under way on Modi and Persian script character recognition, using chain-code-based hybrid correlation and hough transform for their recognition systems respectively.

1. Agrawal, O. P., *Conservation of Manuscripts and Paintings of Southeast Asia*, Butterworths, London, 1984.
2. Alahakoon, C. N. K., Identification of physical problems of major palm leaf manuscripts collections in Sri Lanka. *J. Univ. Libr. Assoc. Sri Lanka*, 2006, **10**, 54–65.

3. Menon, S. and Williams, G. M., Novel, cost-effective method of archiving manuscripts. *Curr. Sci.*, 1999, **76**, 1299–1301.
4. Sezgin, M. and Sankur, B., Survey over image thresholding techniques and quantitative performance evaluation. *J. Electron. Imag.*, 2004, **13**, 146–165.
5. Niblack, W., *An Introduction to Digital Image Processing*, Prentice-Hall, New Jersey, 1986, pp. 115–116.
6. Sauvola, J. and Pietikäinen, M., Adaptive document image binarization. *Pattern Recogn.*, 2000, **33**, 225–236.
7. Su, B., Lu, S. and Tan, C. L., Binarization of historical handwritten document images using local maximum and minimum filter. In Proceedings of DAS'10, The Eighth IAPR International Workshop on Document Analysis Systems, Cambridge, MA, USA, 9–11 June 2010, pp. 159–166.
8. Lu, S. J. and Tan, C. L., Binarization of badly illuminated document images through shading estimation and compensation. In IEEE ICDAR 2007, Ninth International Conference on Document Analysis and Recognition, Parana, Brazil, 23–26 September 2007, vol. 1, pp. 312–316.
9. Gatos, B., Pratikakis, I. and Perantonis, S. J., Adaptive degraded document image binarization. *Pattern Recogn.*, 2006, **39**, 317–327.
10. Howe, N. R., Document binarization with automatic parameter tuning. *Int. J. Doc. Anal. Recogn.*, 2013, **16**, 247–258.

Received 4 April 2014; revised accepted 18 June 2014

Expression analysis of drought-induced genes in wild tomato line (*Solanum habrochaites*)

Ranjit Singh Gujjar[1], Moin Akhtar[1], Ashutosh Rai[2] and Major Singh[1,*]

[1]Division of Crop Improvement,
Indian Institute of Vegetable Research, Varanasi 221 305, India
[2]Department of Biochemistry, Faculty of Science,
Banaras Hindu University, Varanasi 221 005, India

Many plant genes are regulated in response to abiotic stresses such as drought, high salinity, heat and cold, and their gene products function in stress response and tolerance. The whole process of plant adaptation to these environmental stresses is controlled by orchestration of complex molecular networks. In the present study, eight genes showing significant difference of expression on exposure to artificial drought stress in tomato, were selected from the previously performed microarray experiment. Expression analysis of the genes was done semi-quantitatively as well as quantitatively under artificially imposed drought stress and the results were almost similar to those of microarray experiment. Tissue-specific analysis of the genes, performed on tolerant line, revealed fairly a similar pattern of expression in root, stem and leaf with notable differences in flower, which experienced the least influence of drought. The results confirmed that *SlPRP16*, *SlCYP51-17*, *SlMCPI19* and *SlGDSL20* were downregulated in both the lines with stronger downregulation in sensitive line. *SlWRKY4* was downregulated in both the lines with more folds of downregulation in tolerant line. *SlEFH12* and *SlSNF4-15* were upregulated in tolerant line. *SlUSPA9* was upregulated in both the lines with relatively more folds of upregulation in sensitive line.

Keywords: Abiotic stress, drought, gene expression, tomato, transcription factors.

TOMATO (*Solanum lycopersicum*), a major horticultural crop consumed all over the world, suffers heavy losses due to drought. Water deficit causes various physiological and biochemical effects on plant populations. In response, plants utilize a number of protective mechanisms to maintain normal cellular metabolism and prevent damage to cellular components. Tolerance to water stress in plants is generally associated with maintenance of plant water status. This is achieved through closing of stomata to reduce transpiration, enhancing the capacity of roots to extract more water from soil and osmotic adjustment by accumulating low molecular weight molecules. Drought response, being a complex signalling network, leaves a number of genes with upregulated expression and an equal number of genes with downregulated expression. Most of these upregulated and downregulated genes are directly or indirectly linked to each other.

WRKY transcription factors, earlier identified as key regulators of biotic stress, have been reported to impart abiotic stress tolerance in plants[1,2]. The role of WRKY transcription factors as negative regulators of abiotic stresses was revealed by constitutive expression of *BcWRKY46* gene in transgenic tobacco, under the control of the *CaMV35S* promoter, which conferred susceptibility of transgenic tobacco to freezing, ABA (abscisic acid), salt and dehydration stresses[3]. EF-hand proteins, with a helix–loop–helix Ca^{2+} binding motif, are one of the largest protein families involved in modulation of intracellular Ca^{2+} levels in response to various signals, including hormones, light, mechanical disturbances, abiotic stress and pathogen elicitors[4–7]. USP (universal stress protein) family proteins, first identified in prokaryotes, appear to play an active role in abiotic stress response, but their function remains largely unknown in plants. A USP gene (*SpUSP*), cloned from wild tomato (*Solanum pennellii*) and functionally characterized in cultivated tomato, exhibited increased expression under dehydration stress, salt stress, oxidative stress and phyto-hormone ABA treatment[8]. SNF1 (sucrose non-fermenting 1)/SNF1-related kinases/AMPKs (adenosine monophosphate-activated protein kinases) are evolutionary conserved sensors found in all eukaryotic organisms from simple unicellular fungi

*For correspondence. (e-mail: singhvns@gmail.com)