

Evolution of Degraded Historical Manuscripts Binarization Techniques - A Review

Aniketh Shenoy¹, Sachin S. Bhat², Shrinivasa³, Chiranth⁴, Ashith⁵ and Bibek Kumar⁶

¹⁻⁶Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, India

Email: research@anikethshenoy.in

Abstract—In the modern world, physical documents are increasingly being converted into digital formats. This process includes storing and preserving large volumes of documents through electronic scanning. These documents mainly include ancient handwritten palm leaf manuscripts, old legal records, scientific literature, historic letters and newspaper articles. Particularly, the degradation of ink and the writing medium makes ancient and historical manuscripts difficult to comprehend. Degraded document binarization has received extensive study recently, and a number of strategies have been created to address its problems and difficulties. The obstacles and difficulties encountered throughout the image binarization process are thoroughly reviewed in this study, and then insights into the various methods employed are provided. This paper also explores cutting-edge techniques for document augmentation that boost deteriorated documents' quality during binarization. We further discuss the efficacy and reliability of current approaches, and the scope to build a hybrid strategy that can effectively and efficiently handle the binarization of degraded historical documents.

Index Terms— Binarization, Print and Handwritten Document, Degraded Document Image Binarization, Performance Evaluation.

I. INTRODUCTION

Across the world, there are huge collections of ancient and historic documents which have both scientific and cultural importance. These documents are significant as the information is unavailable in the contemporary trends and are typically handwritten. As these documents are preserved for a long period of time, they begin to degrade making them hard to interpret and read. Initially, manual labor was used to rewrite these degraded documents but after the invention of electronic scanning these documents have now been converted into digital forms for better accessibility and quality preservation.

Today, there are multiple methods to convert handwritten text into digital form. For instance, characters from the document images are read using Optical Character Recognition (OCR), which turns them into digital text. The quality of documents is a major limitation for application of OCR programs. Prehistoric handwritten documents typically suffer from a variety of degradations that include uneven lighting, varying image contrast, ink seepage, background noise and others that affect the legibility of such documents. Additionally, different stroke patterns, pressure on the surface and exposure of the document to adverse environmental conditions for extended periods of time lead to document degradation which can render OCR systems ineffective in digitizing such documents.

In order to overcome these degradation issues of historic document various image processing techniques have

been developed which employ filtering techniques to remove the noise and enhance the document quality. One such method is image binarization, which is used to separate the foreground characters from the degraded background. This helps in the application of OCR to degraded documents and convert them into digital format. Majority of researchers today aim at increasing the efficiency, accuracy and effectiveness of the degraded image binarization techniques. This paper reviews the issues in historical documents, methods employed to overcome the issues and the future scope for degraded document binarization techniques.

II. ISSUES IN HISTORICAL DOCUMENTS

Ancient manuscripts are the source to obtain accurate knowledge about scientific and cultural history. These documents are generally stored in old libraries and departments. As shown in Fig. 1, ancient manuscripts suffer from various degradation issues so it is necessary to convert these documents into digital form to preserve them. The process of image digitization includes conversion of handwritten text documents into digital text using OCR tools but due to the degradation the OCR tools suffer from recognising the handwritten text. In order to overcome this issue, the degradations are corrected using document binarization techniques for accurate text extraction using OCR tools.

Before we review the binarization techniques for degraded documents, it is necessary to understand the nature of degradation of historical documents. The following section discusses frequently seen degradation and defects in historical documents.

A. Faded Ink or Faint Characters

The ink used in printed and handwritten historical documents gradually fade/faint due to repeated use, ageing, degradation of paper due to exposure to sunlight, stains, discoloration and tears. When the images of such degraded documents are taken, the characters of the document appear blurry making it hard to read.

B. Contrast Variation

Contrast is basically the difference in brightness of low and high intensity pixels in an image. The variation in contrast arising from exposure to sunlight, illumination and noisy environment results in image analysis algorithms to face difficulties in separating foreground and background of an image using threshold-based methods.



Figure 1. Historical manuscripts with different degradations taken from DIBCO and H-DIBCO datasets. (a) Uneven lighting (b) Variation of Contrast (c) Ink bleed degradation (d) Faded Ink Characters (e) Smear effect (f) Motion Blur (g) Weak Text (h) Deteriorated Documents

C. Blur

There are two major types of blurs in degraded document images. First, motion blur, which occurs due to relative motion between the object and camera. Second, out-of-focus blur, where certain part of the image is in focus

while some points of the image is not. Various tools have been developed recently to assess the blur and provide the user with reduced response to capture new images which generate better OCR results.

D. Smear or Show Through

The digitised historical document generally has noises and artefacts appearing in the image due to degradation of the document over time and exposure to environmental elements. As majority of the ancient manuscripts have been written on both sides, the ink from one page appears on the other, making it hard to read. Various image processing techniques are employed to remove the show-through issue and obtain a clear background.

E. Uneven Illumination

In document analysis, the uneven lighting conditions severely affect the process of recognition of characters from the image of the historical document using an OCR and degrade the efficiency of text recognition. Generally OCR performs better when the document images are first binarized and then processed to extract text. But the uneven illumination results in incorrect extraction of text from the artefact region.

F. Bleed-Through Degradation

When a document has pages written on both sides and the ink bleeds through to the other side of the paper, rendering the text unreadable, this is termed as bleed-through degradation.

G. Thin or Weak Text

Most of the historical handwritten documents were written using ink or drawn with paint. They begin to deteriorate as the ink begins to shrink because of the quality of the ink or paint used to create them. Weak or thin text makes it challenging to binarize the image and apply OCR algorithms. This is a result of the usage of low-quality ink and the type of paper that was used.

H. Deteriorated Documents

Traditionally, paper was the medium to document association, information and evidence. These documents have historical, cultural and legal significance. Poor storage, environmental condition, inherent instability and inadequate handling results in severe degradation of the historical document.

III. REVIEW OF DOCUMENT BINARIZATION METHODS

Document binarization is a process of separating the foreground objects like charters, graphics and words from the background. The fundamental principle of binarization is the conversion of a grayscale image into a binary image for further processing. Due to the previously mentioned issues, image binarization algorithms struggle when dealing with degraded historical documents. A typical degraded document binarization process, as shown in Fig 2, involves pre-processing of the image followed by application of filters and threshold. The process results in a clearer image which can be sent to OCR systems for extracting the texts. The various approaches to binarize degraded ancient and historical documents is discussed in the following subsections.

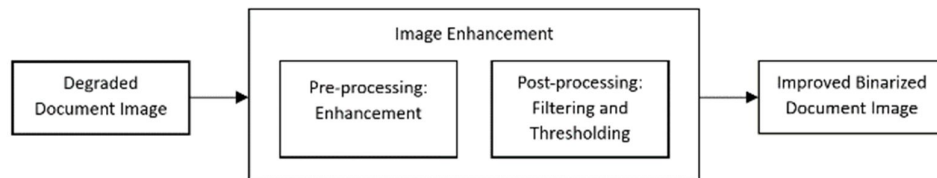


Fig 2. General Architecture of Degraded Document Binarization methods

A. Global Thresholding-based Binarization Methods

Application of a single threshold value across the entire image is known as Global Thresholding. The classified group of pixels are grouped and based on the threshold value they are either included or excluded. One of the most popular global thresholding algorithms is proposed by Otsu [1], where the pixels are classified into the foreground and background classes. This is applied when the background of the image is uniform and has no parameters to tune but in case of degraded historical manuscripts background is not homogeneous, with some background pixels being darker than the foreground pixels. Hence, for degraded documents global thresholding does not perform well.

B. Local/Adaptive Thresholding based Binarization Methods

Local/Adaptive thresholding works by calculating the threshold limit from each pixel or a group of pixels unlike the application of a single threshold value across the entire image as in global thresholding. Niblack [2] presented a basic local thresholding method in which the threshold is determined by means and standard deviation and a local rectangular sliding window is centred on the pixel of interest. Sauvola [3] proposed a solution to the problem faced by Niblack method using background only windows by tuning the parameters which resulted in fewer background pixels to be set to foreground. Wolf [4] proposed an extension to Sauvola by normalising the local statistics by considering the global statistics resulting in better efficiency of binarizing images with low contrast and grayscale intensity. AdOtsu [5] is a generalisation of Otsu's approach that is adaptive and without parameters; it determines the threshold based on the local window rather than the global histogram.

C. Edge based Binarization Methods

Based on the pixel border, various methods have been devised to separate an image's foreground and background. Lu [6] proposed a method where intensity of edge pixels is used to determine the threshold for local windows. The foreground pixel in this case needs to be close to a minimal number of edge pixels, and its intensity needs to be lower than the average intensity of the nearby edge pixels. Su [7] developed a novel method to measure local contrast to filter edges detected by Canny edge detection [8]. A binarization technique based on the concept of structural symmetrical pixels (SSPs) was put forward by Jia [9]. After filtering out the erroneous regions, Valizadeh [10] suggested an adaptive water flow model for binarization in which all pixels covered by water are set to the foreground.

D. Deep Learning based Binarization Methods

Deep learning is an effective alternative technique to overcome the difficulties faced by traditional global and local thresholding methods with degraded historical documents. Initially, experiments conducted by Pastor-Pellicer[11] found that an appropriate CNN architecture will outperform a non-convolutional deep multilayer perceptron counterpart. A convolutional encoder-decoder approach was proposed by Peng [12] and Galego [13] in which the encoder gradually down samples the input image while the decoder generates the binarized output image. Vo [14] and Tensmeyer [15] developed a fully convolutional network that learns features by combining multi-scale outputs. Due to lack of enforcement of spatial smoothness in the predicted output in a fully convolutional network, Kalyan Ram Ayyalasomayajula [16] enhances performance by integrating individual pixel probability with smoothness-based total variation and learning output smoothness and pixel probability. Recently, binarization of degraded documents using GANs has shown promising outcomes. One of the notable approaches is based on a cascaded generator of cGANs [17]. Performance is enhanced by turning the image classification task into a generation task using the cGAN structure and by using the cascaded sub-generator structure to tackle multi scale feature combination.

IV. POPULAR DEGRADED HISTORICAL DOCUMENT BINARIZATION TECHNIQUES

Numerous studies have been done in the past to address document binarization of printed and handwritten documents with various forms of degradations. While the majority of earlier reviews merely summarized the methods in isolation, Table 1 lists the 12 most popular binarization techniques in decreasing order of their performance. The model that performs the best is at the top, while the one that performs poorly is at the bottom.

TABLE I. SUMMARY OF POPULAR BINARIZATION METHODS

Method	Category	Description
Xiong [18]	Deep Learning	Performance is boosted by the presence of blocks called hybrid dilated convolution and spatial pyramid pooling between the encoder and decoder.
Bezmaternykh [19]	Deep Learning	Captures the context and provides pixel-wise classification accuracy on symmetric expanding paths.
Vo [14]	Deep Learning	Combines the Fully Convolutional Network's multiple scale outputs.

Peng [20]	Deep Learning	DRD loss and Multi-scale Convolutional conditional random fields were used to train a fully convolutional network.
Zhao [21]	Deep Learning	Evaluates agreement between predictions and images using a GAN discriminator.
Wolf [4]	Local Thresholding	The Sauvola method is improved with global normalisation.
Sauvola [3]	Local Thresholding	Improved version of Niblack [2].
Bhowmik [22]	Game theory	Use of game theory to extract features for clustering.
Gallego [13]	Unsupervised Tuning	Use of Selectional Auto-Encoder to learn an end-to-end transformation.
Jia [9]	Edge Based	Detection of stroke edges symmetry.
Otsu [1]	Global Thresholding	Maximise global variance and between-class variance of pixel intensity
Niblack [2]	Local Thresholding	Local thresholding with rectangular sliding window and Standard deviation

V. PERFORMANCE AND EVALUATION METRICS

The 12 popular degraded document binarization techniques are assessed on open access datasets from the past document image binarization competitions: DIBCO [23], [24], [25], [26], [27] and H-DIBCO [28], [29], [30], [31], [32]. The DIBCO datasets contain images of both printed and handwritten documents, while the H-DIBCO datasets solely contain handwritten document images with varying degrees of degradation.

We adopt the four-evaluation metrics F-Measure (FM), Pseudo-F-Measure (p-FM), Peak Signal-to-Noise Ratio (PSNR), and Distance Reciprocal Distortion (DRD) used in the degraded document binarization competition. The first two metrics achieve their best and worst values at 1 and 0 respectively. The PSNR quantifies how closely the binarized image resembles the ground truth image; as a result, the greater the PSNR magnitude, the better is the performance. In contrast, the lower value of DRD signifies that the model performs better. The definitions of the evaluation metrics are discussed in the subsections below.

A. F-Measure

F-Measure is calculated using (1) and is the harmonic mean of Precision and Recall where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (1)$$

Here, TP stand for "True Positive," FP for "False Positive," and FN for "False Negative.". The better performing models have high value of F-Measure.

B. Pseudo-F-Measure

Pseudo-F-Measure [33] combines pseudo-Precision (P_{ps}) and pseudo-Recall (R_{ps}) into a single measure (2), using the same equation as F-Measure.

$$p - FM = \frac{2 \times R_{ps} \times P_{ps}}{R_{ps} + P_{ps}} \quad (2)$$

Here, R_{ps} and P_{ps} are computed using distance weights in association with Ground Truth characters.

C. Peak Signal-to-Noise Ratio (PSNR)

PSNR (3) measures the similarity between two images. If the two image are highly similar then the value of PNSR will also be high.

$$PSNR = 10 \log \left(\frac{C^2}{MSE} \right) \quad (3)$$

where, $MSE = \frac{\sum_{x=1}^M \sum_{y=1}^N (L(x,y) - L'(x,y))^2}{MN}$ and the distinction between the foreground and background is defined as C.

D. Distance Reciprocal Distortion (DRD)

DRD is the measure of visual distortion in Binary images which it accurately correlates with human visual perception.

$$DRD = \left(\frac{\sum_k DRD_k}{NUBN} \right) \quad (4)$$

In (4), NUBM stands for the number of 8x8 blocks in the ground truth image that are not uniform, and DRD_k stands for the distortion in the k^{th} flipped pixel.

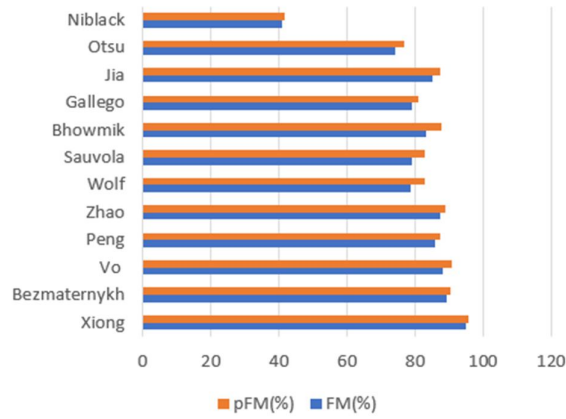


Fig 3. Comparison of binarization methods based on p-FM and FM evaluation metric

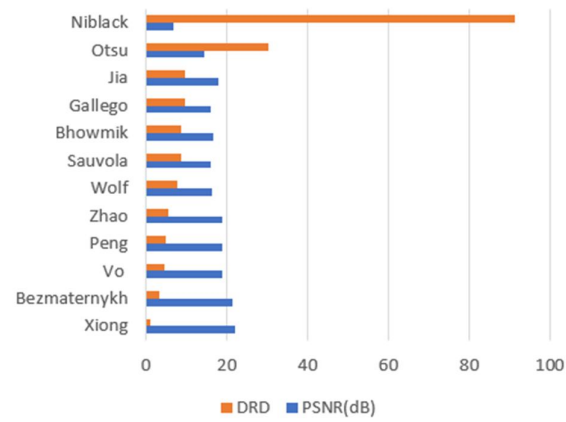


Fig 4. Comparison of binarization methods based on DRD and PSNR evaluation metric

Fig. 3 and Fig. 4. indicate the results of the performance analysis. It is observed that Xiong [18] method achieved the highest performance with F-measure = 95.13, p-FM = 95.8, PSNR = 22.13 and DRD = 1.19 compared to other methods. The lowest value in terms of the evaluation metrics came from the Niblack [2] method where F-measure = 41.12, p-FM = 41.57, PSNR = 6.67 and DRD = 91.23.

VI. CONCLUSION AND FUTURE WORK

This review explores several degradations of handwritten and printed historical documents. The presence of such degradation has led to development of efficient and accurate degraded document binarization methods which ultimately lead to improvement of OCR systems. This paper discusses various methods of binarization based on current research trends. In the current era of digitization, document binarization has a broader scope of application. The primary goal of binarization, which is to separate foreground text from background, is currently being achieved, and future work will focus on improving binarization methods to make them faster and more computationally effective. By combining several machine learning approaches based on the detection of various degradations in the historical document, one may develop a method that automatically determines the best suitable method for binarization. This will significantly increase the accuracy of OCR systems.

REFERENCES

- [1] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, Jan. 1979.
- [2] Wayne Niblack.. An introduction to digital image processing. Strandberg Publishing Company, DNK. 1985.
- [3] J. Sauvola, M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33(2), pp. 225-236 (2000).
- [4] Wolf, C., Jolion, JM. "Extraction and recognition of artificial text in multimedia documents", *Formal Pattern Analysis & Applications* 6, 309-326 (2004).
- [5] Farrahi Moghaddam, Reza & Cheriet, Mohamed, "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization." *Pattern Recognition*. Vol.45. p.p2419-2431.(2012).
- [6] Lu, S., Su, B. & Tan, C.L. Document image binarization using background estimation and stroke edges. *IJDAR* 13, 303-314 (2010).
- [7] Su, Bolan & Tan, Chew Lim & Lu, Shijian. "A Robust Document Image Binarization Technique for Degraded Document Images". *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*. 22(4) (2012).
- [8] J. Canny, "A Computational Approach to Edge Detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986.
- [9] Jia, Fuxi & Shi, Cunzhao & He, Kun & Wang, Chunheng & Xiao, Baihua. Degraded Document Image Binarization using Structural Symmetry of Strokes. *Pattern Recognition*. Vol. 74 (2018).
- [10] Valizadeh, M., Kabir, E. An adaptive water flow model for binarization of degraded document images. *IJDAR* 16, 165-176 (2013).
- [11] Pastor-Pellicer, J. & España-Boquera, S. & Zamora-Martínez, Francisco & Afzal, Muhammad Zeshan & Castro-Bleda, María José. "Insights on the Use of Convolutional Neural Networks for Document Image Binarization." 115-126.(2015).
- [12] X. Peng, H. Cao and P. Natarajan, "Using Convolutional Encoder-Decoder for Document Image Binarization," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, pp. 708-713, (2017).
- [13] Jorge Calvo-Zaragoza, Antonio-Javier Gallego, "A selectional auto-encoder approach for document image binarization," *Pattern Recognition*, vol. 86, pp. 37-47, (2019).
- [14] Vo, Quang Nhat et al. "Binarization of degraded document images based on hierarchical deep supervised network." *Pattern Recognit.* 74: 568-586, (2018).
- [15] Tensmeyer, Chris & Martinez, Tony. Document Image Binarization with Fully Convolutional Neural Networks. 99-104, (2017).
- [16] Kalyan Ram Ayyalasomayajula, Filip Malmberg, and Anders Brun. 2019. PDNet: Semantic segmentation integrated with a primal-dual network for document binarization. *Pattern Recogn. Lett.* 121, C (Apr 2019), 52-60.
- [17] Zhao, Jinyuan & Shi, Cunzhao & Jia, Fuxi & Wang, Yanna & Xiao, Baihua.. Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition*. 96. 106968, (2019).
- [18] W. Xiong, X. Jia, D. Yang, M. Ai, L. Li and S. Wang, "DP-LinkNet: A convolutional network for historical document image binarization," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 5, pp. 1778-1797, (2021).
- [19] Bezmaternykh PV, Ilin DA, Nikolaev DP. U-Net-bin: hacking the document image binarization contest. *Computer Optics* 2019; 43(5): 825-832., (2019).
- [20] X. Peng, C. Wang and H. Cao, "Document Binarization via Multi-resolutional Attention Model with DRD Loss," in 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, pp. 45-50, (2019).

- [21] Zhao, Jinyuan & Shi, Cunzhao & Jia, Fuxi & Wang, Yanna & Xiao, Baihua. . Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition*. Vol. 96 (2019).
- [22] S. Bhowmik, R. Sarkar, B. Das and D. Doermann, "GiB: A Game Theory Inspired Binarization Technique for Degraded Document Images," in *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1443-1455, (2019).
- [23] B. Gatos, K. Ntirogiannis and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009, pp. 1375-1382.
- [24] I. Pratikakis, B. Gatos and K. Ntirogiannis, "ICDAR 2011 Document Image Binarization Contest (DIBCO 2011)," 2011 International Conference on Document Analysis and Recognition, Beijing, China, 2011, pp. 1506-1510.
- [25] I. Pratikakis, B. Gatos and K. Ntirogiannis, "ICDAR 2013 Document Image Binarization Contest (DIBCO 2013)," 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 2013, pp. 1471-1476.
- [26] I. Pratikakis, K. Zagoris, G. Barlas and B. Gatos, "ICDAR2017 Competition on Document Image Binarization (DIBCO 2017)," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 1395-1403.
- [27] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal and I. Marthot-Santaniello, "ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019)," 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 2019, pp. 1547-1556.
- [28] I. Pratikakis, B. Gatos and K. Ntirogiannis, "H-DIBCO 2010 - Handwritten Document Image Binarization Competition," 2010 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, India, 2010, pp. 727-732.
- [29] I. Pratikakis, B. Gatos and K. Ntirogiannis, "ICFHR 2012 Competition on Handwritten Document Image Binarization (H-DIBCO 2012)," 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 2012, pp. 817-822.
- [30] K. Ntirogiannis, B. Gatos and I. Pratikakis, "ICFHR2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014)," 2014 14th International Conference on Frontiers in Handwriting Recognition, Hersonissos, Greece, 2014, pp. 809-813.
- [31] I. Pratikakis, K. Zagoris, G. Barlas and B. Gatos, "ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016)," 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China, 2016, pp. 619-623.
- [32] I. Pratikakis, K. Zagori, P. Kaddas and B. Gatos, "ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018)," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 2018, pp. 489-493.
- [33] K. Ntirogiannis, B. Gatos and I. Pratikakis, "Performance Evaluation Methodology for Historical Document Image Binarization," in *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 595-609, Feb. 2013.