Comparing Binarisation Techniques for the Processing of Ancient Manuscripts

Rapeeporn Chamchong, Chun Che Fung, and Kok Wai Wong

School of Information Technology, Murdoch University,
90 South Street, Murdoch, Western Australia 6150
rapeeporn.c@gmail.com, l.fung@murdoch.edu.au,
k.wong@murdoch.edu.au

Abstract. Ancient manuscripts have been preserved by many organizations so as to protect these documents and retrieve traditional knowledge. With the advanced computer technology, digitized media is now commonly used to record these documents. One objective of such work is to develop an efficient image processing system that could be used to retrieve knowledge and information automatically from these ancient manuscripts. Binarization is a preprocessing technique used to extract text and characters from the manuscripts. The output is then used for further processes such as character recognition and knowledge extraction. This paper compares different binarization techniques that could be used for processing of ancient manuscripts. The aim is to improve the binarization techniques with the main objective of developing an automated preprocessing technique for ancient manuscript recognition and knowledge extraction.

Keywords: Binarization, Image segmentation, Ancient Documents.

1 Introduction

Dried palm leaves have been used as one of the most popular written documents for over five hundred years in Thailand. Such materials have been used for recording Buddhist teaching and doctrines, folklores, knowledge and use of herbal medicines, stories of dynasties, traditional arts and architectures, astrology, astronomy, and techniques of traditional massages. Over time, most of these palm leaves, if left unattended, will deteriorate. This could be caused by dampness, fungus, bacteria, insects and bugs. This leads to many projects with the objectives to preserve and protect information from ancient manuscripts. Such projects are initiated and carried out by the Thai libraries, universities and institutes including medical departments and religion organizations. Three examples of these projects are the Digitisation Initiative for Traditional Manuscripts of Northern Thailand Project at Chiang Mai University Library [1, 2], the Palm Leaf Manuscript Preservation Project in Northeastern Region of Thailand at Mahasarakham University [3], and the Thailand Herbal Repository Access Initiative (THRAI) at Kasetsart University [4]. In particular, the THRAI project aims at developing a database for Thai traditional

R. Nakatsu et al. (Eds.): ECS 2010, IFIP AICT 333, pp. 55–64, 2010.

medicine in order to preserve and propagate Thai medical knowledge from the ancient manuscripts.

The number of multimedia databases and the amount of information stored and captured digitally is increasing rapidly with the advances of computer technology. Although the availability of advance imaging tools and effective image processing techniques makes it feasible to process these documents in multimedia formats for future analysis and storage, there is no specific system that is capable to retrieve relevant information efficiently and to extract knowledge from them. It is therefore a key objective of this study to develop an efficient image processing system that could be used to retrieve knowledge and information from these historical manuscripts. However, it is recognized that the process of scanning a digital image of the ancient palm leaves could also presents some difficulties. Most of the original leaves are aged, leading to deterioration of the writing media, with seepage of ink and smearing along cracks, damage to the leaf due to the holes used for binding the manuscript leaves, dirt and other discoloration. These factors lead to poor contrast, smudges, smear, stains, and ghosting noise due to seeping ink from the other side of the manuscripts between the foreground text and the background. Digital image processing techniques are therefore necessary to improve the readability of the manuscripts.

Prior to the stage of knowledge extraction, characters or text on the images have to be recognized. There are three steps which need to be completed prior to the task of character recognition. First, a manuscript is scanned into a RGB image and then it is converted to a gray-scale image. Next, image enhancement is used to enhance the quality of the image. After this stage, binarization is applied and then text and character separation are carried out before character recognition. Binarization is an essential part of the preprocessing step in image processing, converting gray-scale image to binary image, which is then used for further processing such as document image analysis and optical character recognition (OCR). Consequently, both image enhancement and binarization of historical document are crucial to remove unrelated information, noise and background on the documents. If these steps are ineffective, the original characters from the image may be unrecognizable or more noise may be added. Therefore, these techniques are essential to improve the readability of the documents and the overall performance of the process.

Several binarization algorithms have been proposed in the literature [5-14]. However, it is difficult to select the most appropriate algorithm. The comparison of image qualities from those algorithms is not an easy task as there is no objective evaluation process to compare the results. In contrast, some researchers have proposed a quantitative image measurement of binarization. This performance evaluation of binarization algorithms is recognized to significantly dependent on the image content and on the methodologies of binarization. The common approach is to design a set of criteria and scores of criteria. The criteria may be computed by machine or may be decided by visual human. Within this domain, Badekas and Papamarkos [15], applied some binarization techniques with a standard database (Mediateam Oulu Document) [16]. However, these documents are not as complex as palm leaf manuscripts.

In this paper, a comparison of different binarization techniques of manuscripts is reported. In section 2, the binarization techniques are explained. Section 3 describes the framework of appropriate selection of binarization techniques by machine learning. The experimental results are then shown in section 4 and finally, a conclusion and discussion on future research are given in the last section.

2 Binarization Techniques

Binarization is the task of converting a gray-scale image to a binary image by using threshold selection techniques to categorize the pixels of an image into either one of the two classes. Most of studies [5-8, 14, 17] separated the binarization techniques into two main methods that are global thresholding and local adaptive thresholding techniques.

Global Thresholding Techniques. These techniques attempt to find a suitable single threshold value (Thr) from the overall image. The pixels are separated into two classes: foreground (text which is black color) and background (white color). This can be expressed as follows [5]

$$I_b(x,y) = \begin{cases} \text{black} & \text{if} \quad I_f(x,y) \le \text{Thr} \\ \text{white} & \text{if} \quad I_f(x,y) > \text{Thr} \end{cases} \tag{1}$$

where $I_f(x,y)$ is the pixel of the input image from the noise reduction and $I_b(x,y)$ is the pixel of the binarized image.

Otsu's algorithm [11] is a popular global thresholding technique. Moreover, there are many popular thresholding techniques such as Kapur and et al [18], and Kittler and Illingworth [19].

Local Thresholding Techniques [14]. These techniques calculate the threshold values which are determined locally based on pixel by pixel, or region by region. A threshold value (Thr(x,y)) can be derived for each pixel in the image, and the image can be separated into foreground and background as given in expression (2) [5].

$$I_b(x,y) = \begin{cases} \text{black} & \text{if} \quad I_f(x,y) \le \text{Thr}(x,y) \\ \text{white} & \text{if} \quad I_f(x,y) > \text{Thr}(x,y) \end{cases} \tag{2}$$

The conventional local adaptive thresholding techniques have been proposed by Niblack [10] and Sauvola [12].

In this paper, a comparison from nine binarization techniques are reported which are Otsu [11], Kittler and Illingworth [19], Kapur [18], Tsai [20], Huang [21], Yen and et al [22], Niblack [10], Sauvola [12], and Bernsen [14]. The first six techniques are global thresholding techniques and the last three techniques are local adaptive thresholding techniques.

These approaches are then applied to the palm leaves images and also the MediaTeam Oulu Document Database. The objective is to determine whether automated process could be developed in determining the optimal value of threshold for the binarization of the images. A collection of the techniques is given below.

Binarization Tecniques	Criteria
1.Otsu [11]	$\eta(\text{thr}*) = \sigma_B^2(\text{thr}*)/\sigma_T^2$, $\sigma_B^2(\text{thr}*) = \arg\max_{0\le \text{thr} < L-1} \sigma_B^2(\text{thr})$ thr* is optimal threshold,η is separation criteria, $\sigma_B^2(\text{thr})$ is variance between group of histogram at threshold thr and σ_T^2 is variance of histogram
2.Klittler and Illingworth [19]	$T_{opt} = \arg\min\{P(T)\log\sigma_f(T) + [1-P(T)]\log\sigma_b(T) - P(T)\log P(T) - [1-P(T)]\log[1-P(T)]\}$ where $\sigma_f(T)$ and $\sigma_f(T)$ are foreground and background standard deviations.
3.Kapur [18]	$T_{opt} = \arg\max[H_f(T) + H_b(T)]$ where $H_f(T) = -\sum_{l=0}^{T} \frac{p(l)}{P(T)} \log \frac{p(l)}{P(T)}$ and $H_b(T) = -\sum_{l=T+1}^{L} \frac{p(l)}{P(T)} \log \frac{p(l)}{P(T)}$
4.Yen and et al. [21]	$T_{opt} = \arg\max[C_b(T) + C_f(T)]$ where $C_b(T) = -\log\left\{\sum_{l=0}^{T}\left[\frac{p(l)}{P(T)}\right]^2\right\}$ and $C_f(T) = -\log\left\{\sum_{l=T+1}^{L}\left[\frac{p(l)}{1-P(T)}\right]^2\right\}$
5.Huang [21]	$T_{opt} = \arg\min\{-\frac{1}{N^2\log 2}\sum_{l=0}^{L}[\mu_f(l,T)\log(\mu_f(l,T)) + [1-\mu_f(l,T)\log(1-\mu_f(l,T))]p(l)\}$ where $\mu_f[l(i,j),T] = \frac{L}{L+\lvert I(i,j)-m_f(T)\rvert}$
6.Tsai [20]	$T_{opt} = \arg\text{equal}[m_1 = b_1(T), m_2 = b_2(T), m_3 = b_3(T)]$ where $m_k = \sum_{l=0}^{T} p(l)l^k$ and $b_k = P_f m_f^k + P_b m_b^k$
7. Niblack [10]	$T(x,y) = m(x,y)\cdot + k * s(x,y)$ k=-0.2, window size =20x20
8. Sauvola [12]	$T(x,y) = m(x,y)\cdot\left[1 + k\cdot\left(\frac{s(x,y)}{R} - 1\right)\right]$ k=0.5, R=128, window size = 20x20
9.Bernsen [14]	T(x, y) = (Zlow + Zhigh)/2, C(x, y) = (Z_{high} - Z_{low}<) e Window size (r x r) = 15x15 and e = 15

3 Framework of Appropriate Selection of Binarization Techniques

According to human knowledge, human may predict the performance of algorithms on the image and select the best one. However, it is not an easy selection task by computer. If the process of automated algorithm selection can be simulated, this will help people to choose the best one. This study attempt to investigate and propose desire algorithm from image features by using machine learning technique

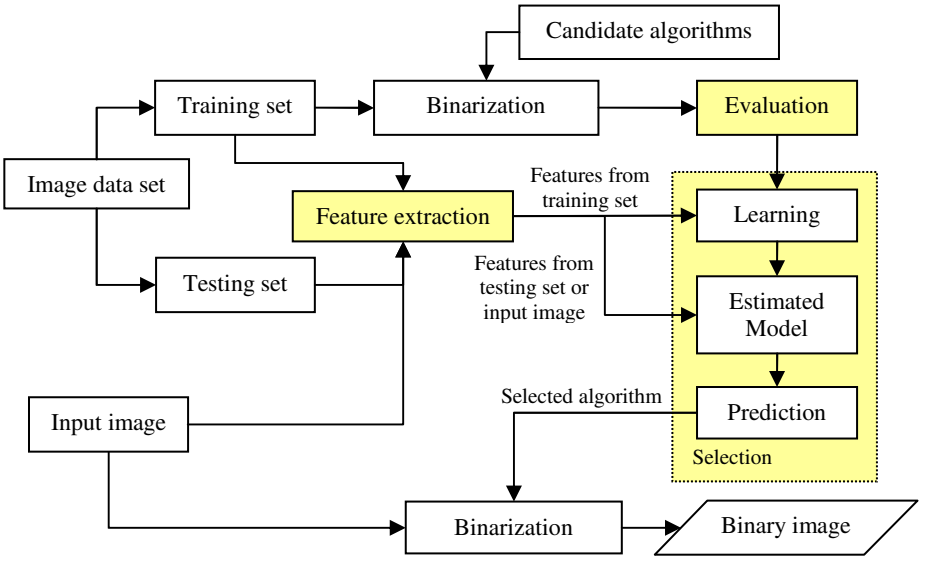


Fig. 1. Framework of appropriate selection of binarization techniques

The propose framework of this system composts of three modules: evaluation, feature extraction, and selection. Firstly, evaluation module is employed with binary image by using k-means algorithm [23]. Secondly, feature extraction module is automated by selecting some characteristics from image. In general, histogram can be used for non-texture image. In addition, average and standard deviation of histogram are used to explain image characteristic. Finally, main module of this system, selection module, is calculated by feeding features from image to learn from training set and then estimated module is generated for predicting the appropriate algorithm. The appropriate selection is done by using Backpropagation [23].

4 Experimental Results

In this experiment, the image data set is based on ancient palm leaf manuscripts. The data set is obtained from the Palm Leaf Manuscript Preservation Project in Northeastern Region, Mahasarakham University [3]. There are 330 palm leaf images which

Table 1. Performance of appropriate selection of binarization algorithms

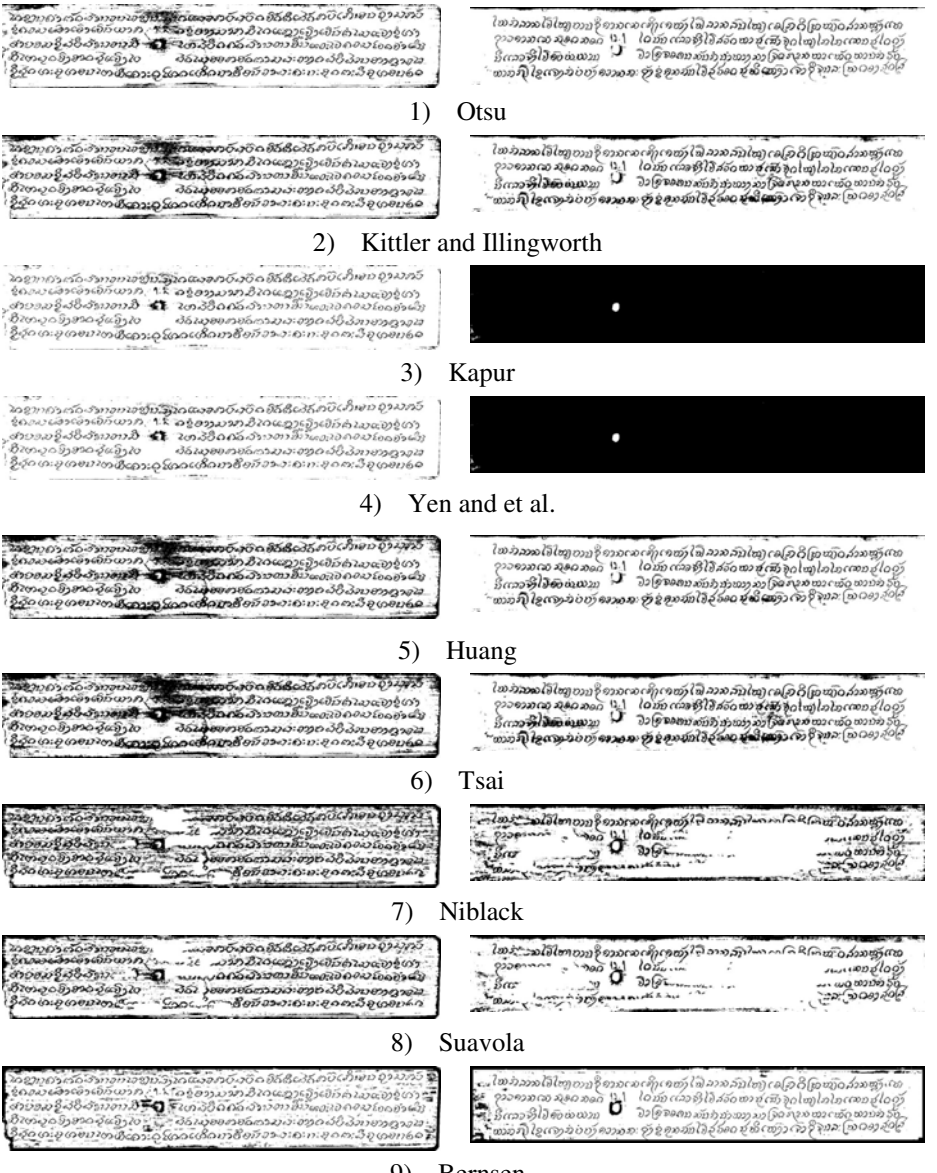
Algorithm	True Positive Rate
1.Otsu	91.05%
2.Klittler and Illingworth	84.00%
3.Kapur	48.72%
4.Yen et al.	8.82%
5.Huang	0%
6.Tsai	71.05%
7.Bernsen	87.81%

a) Palm leaf manuscripts

b) MediaTeam Oulu Document Database

Fig. 2. Samples of original image from the two data sets

a) Example image 1 (L01_9.2) b) Example image 2 (L02_12.1)



1) Otsu

2) Kittler and Illingworth

3) Kapur

4) Yen and et al.

5) Huang

6) Tsai

7) Niblack

8) Suavola

9) Bernsen

Fig. 3. Binary results of nine thresholding techniques on palm leaf manuscripts

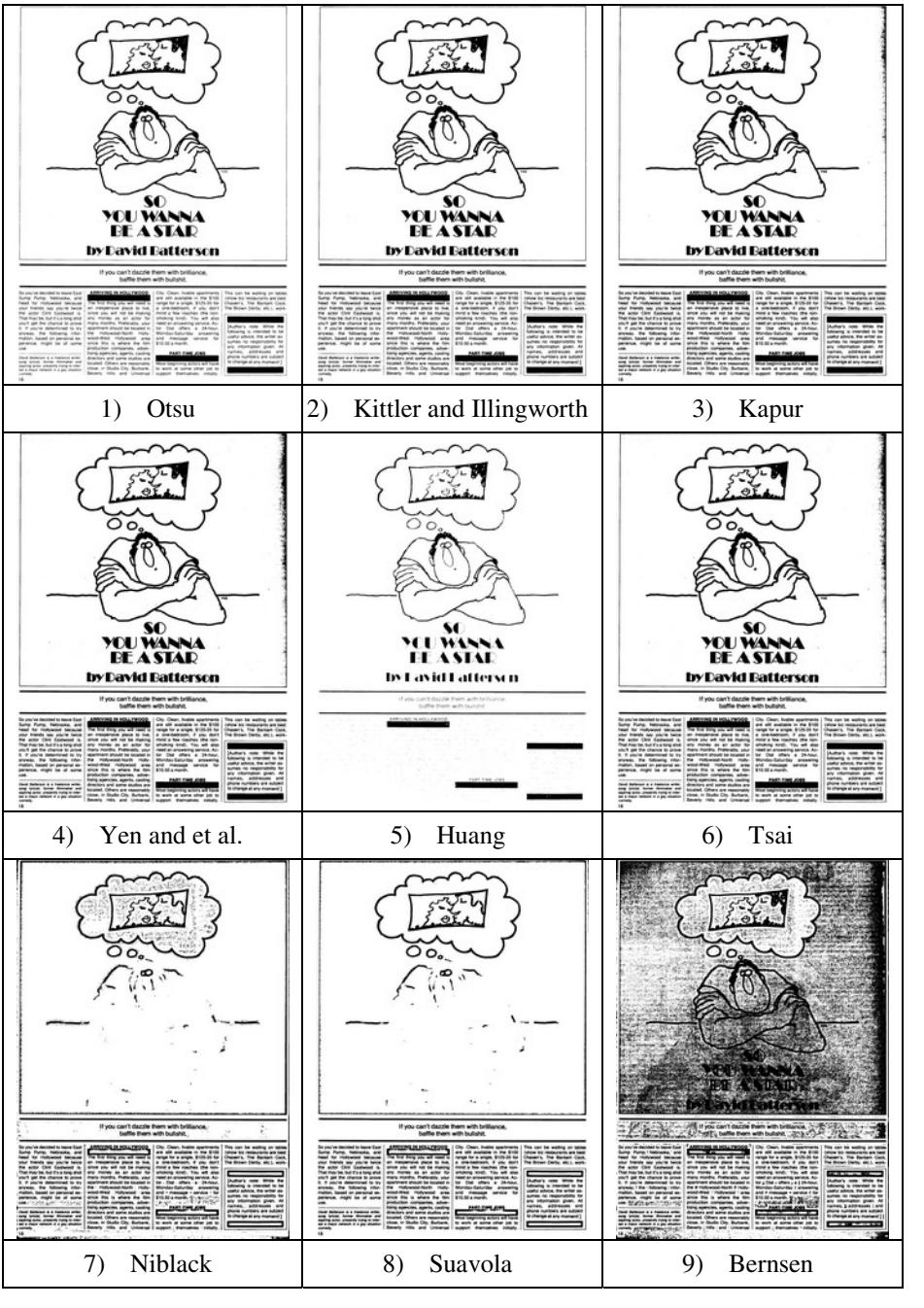


Fig. 4. Binary results of nine thresholding techniques on sample 1 from MediaTeam Oulu Document Database

have been scanned previously. The resolution of the input images is 200x200 dpi in RGB format. In addition, the documents from the MediaTeam Oulu Document Database [16], which is a standard database, is processed to compare binary results with palm leaf manuscripts. There are 158 documents 512 pages and the resolution of the input image is 300x300 dpi in RGB format. Samples of the original images from both data sets are shown in Fig. 2. The input images were converted to gray-scale images and then noise is reduced by Gaussians filtering technique. After that, binarization techniques have been applied. Fig. 3 and 4 show some example binarized results from ancient palm leaf manuscripts and from the MediaTeam Oulu Document Database respectively.

This experiment, 10 fold cross-validation is used for 330 images from palm leaf manuscripts to evaluate the appropriate selection of binarization techniques. According to evaluating by visual human with this data set, there is no best algorithm from Niblack and Suavola, so seven algorithms are selected. The performance of selection system is given in Table 1.

The results from this experiment have shown that Otsu's algorithm, Bernsen's algorithm, and Klittler and Illingworth gave better performance. In contrast, true positive rate of Huang's algorithm and Yean et al. are the lower performance algorithms.

Comparing the results from global thresholding techniques and local adaptive thresholding techniques, it was found that results from local adaptive thresholding techniques have adjusted the output among local areas so that characters will appear to have more stable appearance than the global techniques. The global thresholding techniques may have eliminated noise in some background areas, but the characters in other areas may become unreadable.

5 Conclusion and Future Work

Although many binarization techniques have been successfully applied to the MediaTeam Oulu Document Database [15], such techniques have difficulty in eliminating background noise on palm leaf manuscripts. Most of the reported work has been carried out on standard images that are not as complex and in poor quality as the ancient palm leaf manuscripts. From this study, it can be concluded that there is no single binarization technique that is suitable for all images in each domain. For this reason, how to choose the best binarization technique for the user is the key issue and currently, there is no automatic selection of the optimal binarization technique. At present, machine learning technique is being investigated in order to determine which algorithm is suitable for binarization techniques in different situations. This proposed framework can be applied for automatic selection system.

References

1. Information Heritage, http://www.emc.com/leadership/digital-universe/information-heritage-awards-2008.htm
2. Musiket, Y.: Scanning the Past (2009)

3. Mahasarakham University: Palm Leaf Manuscript Preservation Project in Northeastern Region. Reports for the Financial Year 2004 and 2005 (2005) (in Thai)
4. Thailand Herbal Repository Access Initiative (THRAI), http://thrai.sci.ku.ac.th/
5. Chamchong, R., Fung, C.C.: Comparing background elimination approaches for processing of ancient Thai manuscripts on palm leaves. In: 2009 Int. Conf. Machine Learning and Cybernetics, China (2009)
6. Chen, Y., Leedham, G.: Decompose algorithm for thresholding degraded historical document images. In: IEE Proceeding Visual Image Signal Processing p. 152 (2005)
7. He, J., Do, Q.D.M., Downton, A.C., Kim, J.H.: A comparison of binarization methods for historical archive documents. In: Proc. 8th Int. Conf. Document Analysis and Recognition, vol. 538, pp. 538–542 (2005)
8. Leedham, G., Chen, Y., Takru, K., Joie Hadi Nata, T., Li, M.: Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In: Proc. 7th Int. Conf. Document Analysis and Recognition, pp. 859–864 (2003)
9. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. J. of Electronic Imaging 13, 146–168 (2004)
10. Niblack, W.: An introduction to digital image processing. Prentice-Hall, Englewood Cliffs (1986)
11. Otsu, N.: A threshold selection method from gray-level histogram. IEEE Trans. Systems Man Cybernet 9, 62–66 (1979)
12. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recognition 33, 225–236 (2000)
13. Sezgin, M., Sankur, B.: Selection of thresholding methods for nondestructive testing applications. In: Proc. 2001 Int. Conf. Image Processing, vol. 3, pp. 763, 764–767 (2001)
14. Trier, O.D., Jain, A.K.: Goal-directed evaluation of binarization methods. IEEE Trans. Pattern Analysis and Machine Intelligence 17, 1191–1201 (1995)
15. Badekas, E., Papamarkos, N.: Document binarisation using Kohonen SOM. IET Image Process. 1, 67–84 (2007)
16. Suavola, J., Kauniskangas, H.: MediaTeam Document Database II, a CD-ROM collection of document images. University of Oulu, Finland (1999)
17. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice-Hall, New Jersey (2002)
18. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A new method for gray-level picture thresholding using the entropy of the histogram. Graph. Models Image Process. 29, 273–285 (1985)
19. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recognition 19, 41–47 (1986)
20. Tsai, W.H.: Moment-preserving thresholding: A new approch. Graph. Models Image Process. 19, 377–379 (1985)
21. Huang, L.-K., Wang, M.-J.J.: Image thresholding by minimizing the measures of fuzziness. Pattern Recognition 28, 41–51 (1995)
22. Jui-Cheng, Y., Fu-Juay, C., Shyang, C.: A new criterion for automatic multilevel thresholding. IEEE Transactions on Image Processing 4, 370–378 (1995)
23. Heijden, F.v.d., Duin, R.P.W., Ridder, D.d., Tax, D.M.J.: Classification, parameter estimation and state estimation: an engineering approach using MATLAB. John Wiley & Sons, Ltd., West Sussex (2004)