PERGAMON

Pattern Recognition 33 (2000) 225–236

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Adaptive document image binarization

J. Sauvola*, M. Pietikäinen

Machine Vision and Media Processing Group, Infotech Oulu, University of Oulu, P.O. BOX 4500, FIN-90401 Oulu, Finland

Received 29 April 1998; accepted 21 January 1999

Abstract

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 final result presentation. The proposed algorithms were tested with images including different types of document components and degradations. The results were compared with a number of known techniques in the literature. The benchmarking results show that the method adapts and performs well in each case qualitatively and quantitatively. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Adaptive binarization; Soft decision; Document segmentation; Document analysis; Document understanding

1. Introduction

Most document analysis algorithms are built on taking advantage of the underlying binarized image data [1]. The use of a bi-level information decreases the computational load and enables the utilization of the simplified analysis methods compared to 256 levels of grey-scale or colour image information. Document image understanding methods require logical and semantic content preservation during thresholding. For example, a letter connectivity must be maintained for optical character recognition and textual compression [2]. This requirement narrows down the use of a global threshold in many cases.

Binarization has been a subject of intense research interest during the last ten years. Most of the developed algorithms rely on statistical methods, not considering the special nature of document images. However, recent developments on document types, for example documents with mixed text and graphics, call for more specialized binarization techniques.

In current techniques, the binarization (threshold selection) is usually performed either globally or locally. Some hybrid methods have also been proposed. The global methods use one calculated threshold value to divide image pixels into object or background classes, whereas the local schemes can use many different adapted values selected according to the local area information. Hybrid methods use both global and local information to decide the pixel label.

The main situations in which single global thresholds are not sufficient are caused by changes in lumination (illumination), scanning errors and resolution, poor quality of the source document and complexity in the document structure (e.g. graphics is mixed with text). When character recognition is performed, the melted sets of pixel clusters (characters) are easily misinterpreted if binarization labelling has not successfully separated the clusters. Other misinterpretations occur easily if meant to be clusters are wrongly divided. Fig. 1 depicts our taxonomy (called MSLG) and general division into thresholding techniques according to level of semantics and locality of processing used. The MSLG can be applied in pairs, for example (ML), (SL), (MG) and (SG).

The most conventional approach is a global threshold, where one threshold value (single threshold) is selected for the entire image according to global/local information. In local thresholding the threshold values

* Corresponding author. Tel.: + 358-40-5890652.
E-mail address: jjs@ee.oulu.fi (J. Sauvola)

0031-3203/99/$20.00 © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(99)00055-2

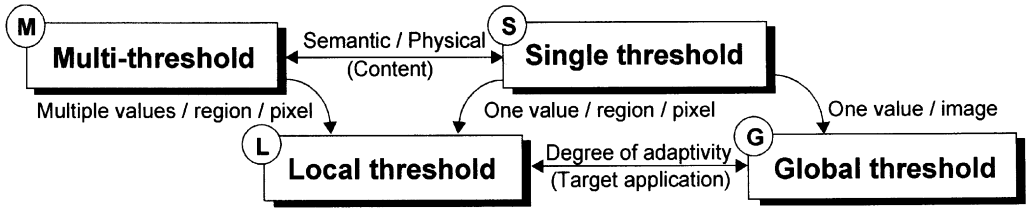

Fig. 1. Taxonomy of thresholding schemes.

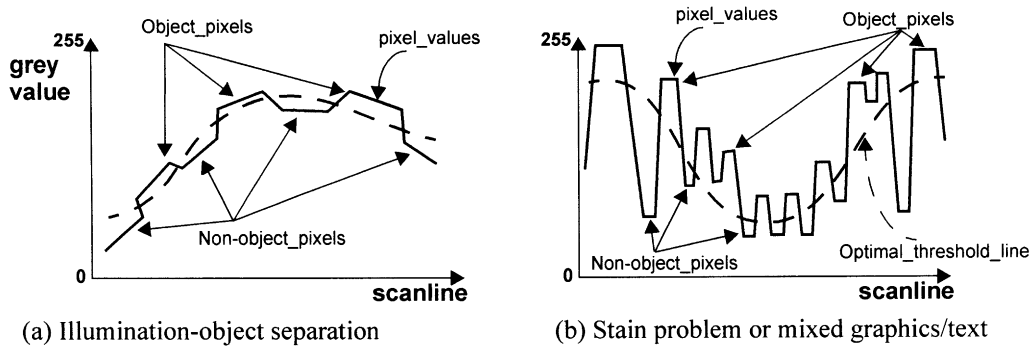

Fig. 2. Examples of document analysis problem types in binarization.

are determined locally, e.g. pixel by pixel, or region by region. Then, a specified region can have 'single threshold' that is changed from region to region according to threshold candidate selection for a given area. Multi-thresholding is a scheme, where image semantics are evaluated. Then, each pixel can have more than one threshold value depending on the connectivity or other semantic dependency related to physical, logical or pictorial contents.

Many binarization techniques that are used in processing tasks are aimed at simplifying and unifying the image data at hand. The simplification is performed to benefit the oncoming processing characteristics, such as computational load, algorithm complexity and real-time requirements in industrial-like environments. One of the key reasons when the binarization step fails to provide the subsequent processing a high-quality data is caused by the different types and degrees of degradation introduced to the source image. The reasons for the degradation may vary from poor source type, the image acquisition process to the environment that causes problems for the image quality directly. Since the degradation is unquestionably one of the main reasons for processing to fail, it is very important to design the binarization technique to detect and filter possible imperfections from becoming the subject for processing and potential cause of errors for post-processing steps. Most degradation types in document images affect both physical and semantic understandability in the document analysis tasks, such as page segmentation, classification and optical character recognition. Therefore, the result after all the desired processing steps can be entirely unacceptable, just because of the poorly performed binarization.

Fig. 2 depicts two types of typical degradation, when dealing with scanned grey-scale document images. In Fig. 2a the threshold 'base line' is changing due to illumination effect or implanted (designed) entity. Then, each object has a different base level that affects the object/non-object separation decision in selecting threshold(s). In Fig. 2b a general type 'stain problem' is presented. In this case, the background and object levels are fluctuating from clear separation to non-clear separation and small level difference between object/non-object. The optimal threshold lines are drawn to both images to depict the base line that a successful binarization algorithm should mimic.

Fig. 3 presents another type of problem, frequently occurring in scanned document images: more than two different levels are visible in textual areas due to transparency of the next page. Then, a binarization algorithm should cope with at least two different threshold candidates: background-transparent text and background-text. The binarized example presents a correct binarization result.

1.1. Survey on document image binarization techniques

The research on binarization techniques originates from the traditional 'scene image' processing needs to

Fig. 3. Example of good binarization on degraded sample image.

optimize the image processing tasks in terms of image data at hand. While the image types have become more complex the algorithms developed have gained wider theoretical grounds. Current trend seems to move forward image domain understanding based binarization and the control of different source image types and qualities. The state-of-the-art techniques are able to adapt to some degree of errors in a defined category, and focus on few image types. In images needing multi-thresholding, the problem seems to be ever harder to solve, since the complexity of image contents, including textual documents has increased rapidly.

Some document directed binarization algorithms have been developed. O'Gorman [3] proposes a global approach calculated from a measure of local connectivity information. The thresholds are found at the intensity levels aiming to preserve the connectivity of regions. Liu et al. [4] propose a method for document image binarization focused on noisy and complex background problems. They use grey-scale and run-length histogram analysis in a method called 'object attribute thresholding'. It identifies a set of global thresholds using global techniques which is used for final threshold selection utilizing local features.

Yang et al.'s [5] thresholding algorithm uses a statistical measurement, called 'largest static state difference'. The method aims to track changes in the statistical signal pattern, dividing the level changes to static or transient according to a grey-level variation. The threshold value is calculated according to static and transient properties separately at each pixel. Stroke connectivity preservation issues in textual images are examined by Chang et al. in Ref. [6]. They propose an algorithm that uses two different components: the background noise elimination using grey-level histogram equalization and enhancement of grey-levels of characters in the neighbourhood using an edge image composition technique. The 'binary partitioning' is made according to a smoothed and equalized histogram information calculated in five different steps.

Pavlidis [7] presents a technique based on the observation that after blurring a bi-level image, the intensity of original pixels is related with the sign of the curvature of the pixels of the blurred image. This property is used to construct the threshold selection of partial histograms in locations where the curvature is significant.

Rosenfeld and Smith [8] presented a global thresholding algorithm to deal with noise problem using an iterative probabilistic model when separating background and object pixels. A relaxation process was used to reduce errors by first classifying pixels probabilistically and adjusting their probabilities using the neighbouring pixels. This process is finally iterated leading to threshold selection, where the probabilities of the background and the object pixels are increased and will be ruled accordingly to non-object and object pixels.

The thresholding algorithm by Perez and Gonzalez [9] was designed to manage situations where imperfect illumination occurs in an image. The bimodal reflectance distribution is utilized to present grey-scale with two components: reflectance r and illumination i, used also in homomorphic filtering. The algorithm is based on the model of Taylor series expansion and uses no a priori knowledge of the image. The illumination is assumed to be relatively smooth, whereas the reflectance component is used to track down changes. The threshold value is chosen from the probabilistic criterion of occurring two-dimensional threshold selection function. This can be calculated in raster-scan fashion.

The illumination problem is emphasized in the thresholding algorithm, called 'edge level thresholding', presented by Parker et al. in Ref. [10]. Their approach uses the principles that objects provide high spatial frequency while illumination consist mainly of low spatial frequencies. The algorithm first identifies objects using Shen–Castan edge detector. The grey-levels are then examined in small windows for finding highest and lowest values that indicate object and background. The average of these values are used to determine the threshold. The selected value is then fitted to all pixels as a surface leading the values above to be judged as a part of an object and a value lower than threshold belongs to background.

Shapiro et al. [11] introduce a global thresholding scheme, where the independency is stressed in the object/background areas ratio, intensity transition slope, object/background shape and noise-insensitivity. The threshold selection is done by choosing a value that maximizes the global non-homogeneity. This is obtained as an integral of weighted local deviations, where the weight function assign higher standard weight deviation in case of background/object transitions than in homogeneous areas.

Pikaz and Averbuch [12] propose an algorithm to perform thresholding for scenes containing distinct

objects. The sequence of graphs is constructed using the size of connected objects in pixels as a classifier. The threshold selection is gained from calculating stable states on the graph. The algorithm can be adapted to select multi-level thresholds by selecting highest stable state candidate in each level.

Henstock and Chelberg [13] propose a statistical model-based threshold selection. The weighted sum of two gamma densities, used for decreasing the computational load instead of normal distributions, are fitted to the sum of edge and non-edge density functions using a five-parameter model. The parameters are estimated using an expectation maximization-style two-step algorithm. The fitted weighted densities separate the edge pixels from non-edge pixels of intensity images.

The enhanced speed entropic threshold selection algorithm is proposed in Ref. [14] by Chen et al. They reduce the image grey-scale levels by quantization and produce a global threshold candidate vector from quantized image. The final threshold selection is estimated only from the reduced image using the candidate vector. The reduction in computational complexity is in the order of magnitude of $O(G^{8/3})$ of the number of grey-scale values, using O-notation. The quality of binarization is sufficient for preliminary image segmentation purposes.

Yanowitz and Bruckstein [15] proposed an image segmentation algorithm based on adaptive binarization, where different image quality problems were taken into consideration. Their algorithm aimed to separate objects in illuminated or degraded conditions. The technique uses variating thresholds, whose values are judged by edge analysis processing combined with grey-level information and construction of interpolated threshold surface. The image is then segmented using the gained threshold surface by identifying the objects by post-validation. The authors indicated that validation can be performed with most of the segmentation methods.

1.2. Our approach

For document image binarization, we propose a new method that first performs a rapid classification of the local contents of a page to background, pictures and text. Two different approaches are then applied to define a threshold for each pixel: a soft decision method (SDM) for background and pictures, and a specialized text binarization method (TBM) for textual and linedrawing areas. The SDM includes noise filtering and signal tracking capabilities, while the TBM is used to separate text components from background in bad conditions, caused by uneven (il)lumination or noise. Finally, the outcome of these algorithms are combined.

Utilizing proper ways to benchmark the algorithm results against ground-truth and other measures is important for guiding the algorithm selection process and directions that future research should take. A well-defined performance evaluation shows which capabilities of the algorithm still need refinement and which capabilities are sufficient for a given situation. The result of benchmarking offers information of the suitability of the technique to certain image domains and quality. However, it is not easy to see the algorithm quality directly from a set of performance values. In this paper we use a goal-directed evaluation process with specially developed document image binarization metrics and measures for comparing the results against a number of well-known and well-performed techniques in the literature [16].

2. Overview of the binarization technique

Our binarization technique is aimed to be used as a first stage in various document analysis, processing and retrieval tasks. Therefore, the special document characteristics, like textual properties, graphics, line-drawings and complex mixtures of their layout-semantics should be included in the requirements. On the other hand, the technique should be simple while taking all the document analysis demands into consideration. Fig. 4 presents the general approach of the binarization processing flow. Since typical document segmentation and labelling for content analysis is out of question in this phase, we use a rapid hybrid switch that dispatches the small, resolution adapted windows to textual (1) and non-textual (2) threshold evaluation techniques. The switch was developed to cover most generic appearances of typical document layout types and can easily be modified for others as well. The threshold evaluation techniques are adapted to textual and non-textual area properties, with the special tolerance and detection to different basic defect types that are usually introduced to images. The outcome of these techniques represent a threshold value

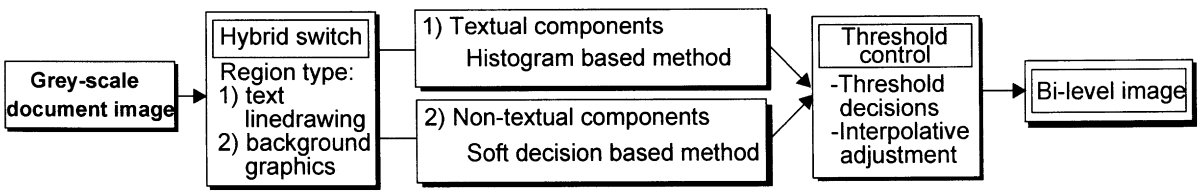

Fig. 4. Overview of the binarization algorithm.

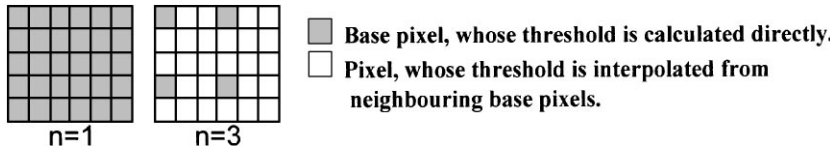

Fig. 5. Interpolation options for binarization computation.

proposed for each pixel, or every *n*th pixel, decided by the user. These values are used to collect the final outcome of the binarization by a threshold control module. The technique also enables the utilization of multi-thresholds region by region of globally, if desired.

3. Adaptive binarization

The document image contains different surface (texture) types that can be divided into uniform, differentiating and transiently changing. The texture contained in pictures and background can usually be classified to uniform or differentiating categories, while the text, line drawings, etc. have more transient properties by nature.

Our approach is to analyse the local document image surface in order to decide on the binarization method needed (Fig. 4). During this decision, a 'hybrid switching' module selects one of two specialized binarization algorithms to be applied to the region. The goal of the binarization algorithms is to produce an optimal threshold value for each pixel. A fast option is to compute first a threshold for every *n*th pixel and then use interpolation for the rest of the pixels (Fig. 5).

The binarization method can also be set to bypass the hybrid switch phase. Then the user can choose which algorithm is selected for thresholding. All other modules function in the same way as in hybrid conditions.

The following subsection describes the region type and switching algorithms. The two different binarization algorithms are then discussed in detail. The final binarization is performed using the proposed threshold values. This process is depicted in the last subsection.

3.1. Region analysis and switching

Threshold computation is preceded by the selection of the proper binarization method based on an analysis of local image properties. First, the document image is tiled to equal sized rectangular windows of 10–20 pixels wide, corresponding to the resolution that linearly varies between >75 and <300 dpi. Two simple features are then computed for each window; these results are used to select the method.

The first feature is simply the average grey value of a window. The second feature, 'transient difference', measures local changes in contrast (Eq. (4)). The difference values are accumulated in each subwindow and then scaled between 0 and 1. Using the limits of 10, 15 and 30% of scaled values, the transient difference property is defined as 'uniform', 'near-uniform', 'differing' or 'transient'. This coarse division is made according to average homogeneity on the surface. According to these labels, a vote is given to corresponding binarization method that is to be used in a window. The labels 'uniform' and 'near-uniform' correspond to background and 'scene' pictures, and give votes to the SDM. The labels 'differing' and 'transient' give their votes to the TBM method.

Selection of a binarization algorithm is then performed as following example rules (1, 2) show:

1. If the average is high and a global histogram peak is in the same quarter of the histogram and transient difference is transient, then use SDM.
2. If the average is medium and a global histogram peak is not in the same quarter of the histogram and transient difference is uniform, then use TBM.

An example result of image partitioning is shown in Fig. 6. The white regions are guided to the SDM algorithm, while the grey regions are binarized with the TBM algorithm.

3.2. Binarization of non-textual components

As in soft control applications, our algorithm first analyses the window surface by calculating descriptive characteristics. Then, the soft control algorithm is applied to every *n*th pixel (Fig. 5). The result is a local threshold based on local region characteristics.

To ensure local adaptivity of threshold selection, we use two different types of locally calculated features:

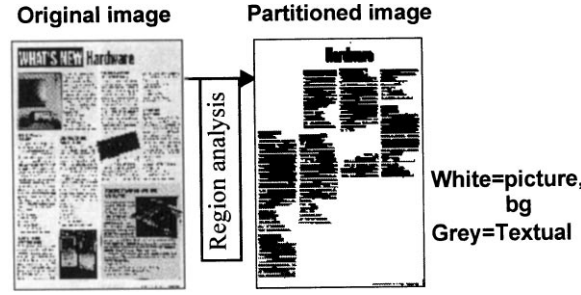

Fig. 6. Example of region partitioning for algorithm (SDM/TBM) selection.

'weighted bound' and 'threshold difference'. The membership function issues, soft decision rules and defuzzification algorithm are presented in the following paragraphs.

3.2.1. Weighted bound calculation

Histogram-based analysis schemes and features are often used in binarization methods. In document analysis the histogram is very useful for detecting and differentiating domains in physical and logical analysis. We use a new approach developed for local detection and weighting of bounds in grey-scale image texture. A new feature called weighted bound (W_b) is introduced and utilized in the soft control algorithm. The W_b is used for characterization of local pixel value profiles by tracking low, medium and high pixels in a small area. In a given surface area of $n \times n$ pixels, where n is a window width gained from the non-overlapping regions analysis tile size (see Section 3.1), three different measures are calculated. The values are collected in a two-dimensional table used to weight and simplify the three envelope curves in soft control membership functions. The measures are minimum, medium and maximum averages given in Eqs. (1)–(3).

Minimum average, $A_{\min}$

$$A_{\min} = \sum_{k=0}^{100/n} \frac{\min_{100/n}(P(i,j))}{100/n}, \tag{1}$$

where $P(i,j)$ is the document image region, and i is the width, and j is the height. n is the static number gained from average window size (see Section 3.1).

Medium average, A_{med}

$$A_{\text{med}} = \sum_{k=0}^{100/n} \frac{\text{med}_{100/n}(P(i,j))}{100/n}. \tag{2}$$

Maximum average, $A_{\max}$

$$A_{\max} = \sum_{k=0}^{100/n} \frac{\max_{100/n}(P(i,j))}{100/n}. \tag{3}$$

These values are stored in an $n \times n \times 3$ table, called a weighted average table (WAT). Using Eqs. (1)–(3), three different histograms are formed where the values are added to their respective bin values (value = bin index). These histograms are then separately partitioned to ten horizontal and three vertical sections, where the number of peaks from histograms are calculated to each section according to sectioning limits.

The horizontal borders are set between bins 0 and 255 with a formula int((256/10)*m), where $m = 1, 2, \ldots, 9$. The number of borders was set to ten. Also a smaller number could be selected, but the penalty is that the original histogram is aliased more. Ten borders equals 25 bins of grey-scale. The two vertical borders are set between 0 and maximum, representing the number of votes calculated for each horizontal bin so that the limits are set to 80% of maximum number of votes and to 40% of the maximum number of votes, respectively. These limits are set according to the tests performed with a large set of images. The higher limit is relatively insensitive to $\pm 10\%$ change. Lowering the lower limit brings more votes to medium peak calculation, thus enhancing the envelope curve in bins where a medium peak appears.

After the peaks are calculated in a 3×10 table, the weighting is performed (Fig. 7). The result is a W_b

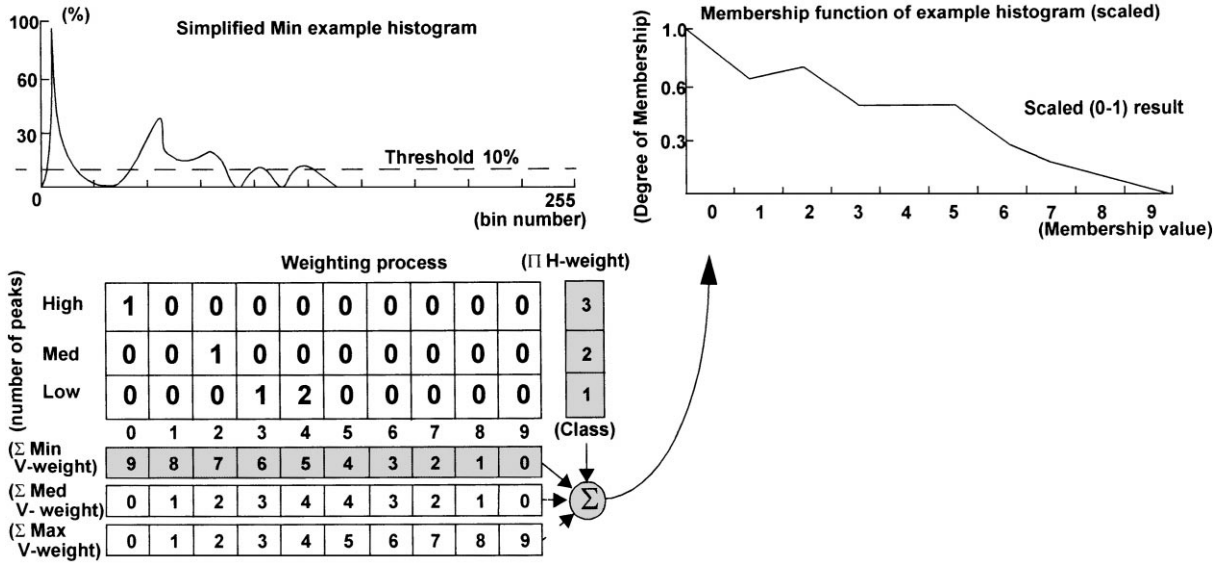

Fig. 7. An example of W_b membership function calculation using $A_{\min}$ histogram.

envelope curve that is used in the soft decision process. The three W_b curves, calculated from $A_{\min}$, A_{med} and $A_{\max}$ are used as membership functions.

3.2.2. Transient difference calculation

The transient difference is aimed at extracting the average amount of variations occurring between the neighbouring pixels (contrast difference) in an $n \times n$ area, i.e. to follow local surface changes. The differences between adjacent pixels are accumulated. The transient difference (TD) of the horizontal and vertical adjacent pixel values is calculated and accumulated. The gained value is then scaled between 0–1 (Eq. (4)). L represents the number of grey-levels in the image.

$$TD = \frac{(\sum_{i=1}^{n}\sum_{j=1}^{n}|2P(i,j) - [P(i-1,j) + P(i,j-1)]|)}{(Ln)^2}. \tag{4}$$

The TD value is used in soft decision making to expose uniform, differential and transient area types when calculating the control value for threshold selection.

3.2.3. Membership function generation

Two different membership functions are used according to the extracted feature values for a given pixel: weighted bound (W_b) and transient difference (TD_m). The first one is calculated dynamically from the image. The transient difference uses predefined membership functions. Fig. 8 depicts these functions using the ideal functions as W_b and the actual membership functions for TD_m.

3.2.4. Soft decision rules and defuzzification

In the soft decision process, we use nine different rules derived from the feature analysis and membership management. For W_b these are (LOW, MIDDLE, HIGH), denoting the local histogram properties. For TD_m we use (UNIFORM, DIFFERING, TRANSIENT), describing the local region property. The rule set is shown in Fig. 9. As in soft control problems, the rules are expressed with clauses, for example:

If W_b **is** $\langle P(i,j)\rangle$ **and** TD_m **is** $\langle TD(i,j)\rangle$
then $T_c(i,j) = \langle 0, 255\rangle$.

The current rule set is designed for pictorial and background-type image regions. Using this set the noise and most illumination defects can be adaptively corrected in the processed areas.

For defuzzification we use Mamdani's method [17]. The result of the defuzzification is a unique threshold value for each pixel n.

3.3. Binarization of textual components

For text binarization we use a modified version of Niblack's algorithm [18]. The idea of Niblack's method is to vary the threshold over the image, based on the local mean, m, and local standard deviation, s, computed in a small neighbourhood of each pixel. A threshold for each pixel is computed from $T = m + k^*s$, where k is a user defined parameter and gets negative values. This method

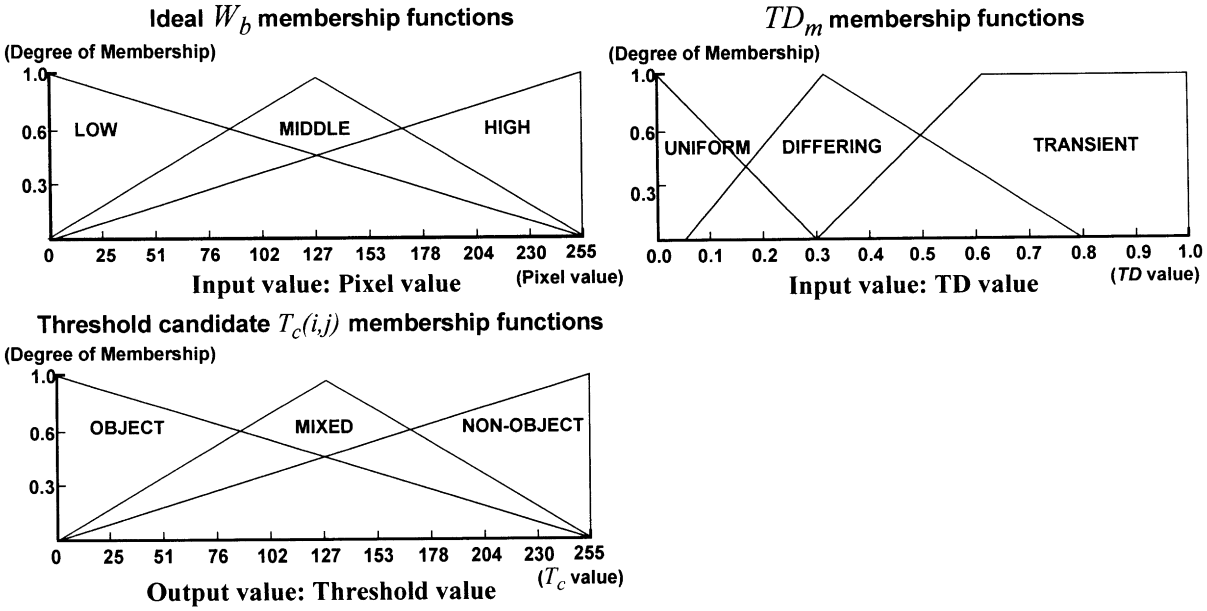

Fig. 8. Input and output membership functions: W_b (ideal), TD_m and T_c.

TD_m \ W_b	LOW	MIDDLE	HIGH
UNIFORM	Object	MIXED	Non-object
DIFFERING	MIXED	MIXED	MIXED
TRANSIENT	Object	MIXED	Non-object

Fig. 9. Example of soft decision rules for threshold candidate $T_c(i, j)$.

does not work well for cases in which the background contains light texture as the grey values of these unwanted details easily exceed threshold values. This results in costly postprocessing as demonstrated in Ref. [19].

In our modification, a threshold is computed with the dynamic range of standard deviation, R. Furthermore, the local mean is utilized to multiply terms R and a fixed value k. This has the effect of amplifying the contribution of standard deviation in an adaptive manner. Consider, for example, a dark text on light dirty-looking background (e.g., stains in a bad copy), Fig. 2. The m-coefficient decreases the threshold value in background areas. This efficiently removes the effect of stains in a thresholded image. In our experiments, we used $R = 128$ with 8-bit gray level images and $k = 0.5$ to obtain good results. The algorithm is not too sensitive to the value of parameter k. Eq. (5) presents the textual binarization formula.

$$T(x, y) = m(x, y) \cdot \left[1 + k \cdot \left(\frac{s(x, y)}{R} - 1\right)\right], \tag{5}$$

where $m(x, y)$ and $s(x, y)$ are as in Niblack's formula. R is the dynamic range of standard deviation, and the parameter k gets positive values. Fig. 10 shows an example threshold line that is adapted to original degraded document image.

3.4. Interpolative threshold selection

After thresholding guided by the surface type, the final thresholds are calculated for background, textual, graphics and line drawing regions. A fast option is to compute first a threshold for every nth pixel and then using interpolation for the rest of the pixels.

The control algorithm has two modes depending on the value of n. If $n = 1$, the threshold values gained from SDM and TBM algorithms are combined directly. If $n > 1$, threshold values for non-base pixels are calculated using the surrounding threshold values.

We have two options to calculate the non-base pixel thresholds: bilinear interpolation and simple averaging. In the interpolation method, the threshold value for a non-base pixel is gained by computing the surrounding base pixels distance to the current one, and using these values as weights, Fig. 11a. This approach gives a more precise, weighted threshold value for each pixel. In the simple averaging method, the average of the surrounding four n pixel threshold candidate values is calculated and used as a final threshold for each non-base pixel between the selected base pixels, Fig. 11b. This approach is used to lower the computational load and is suitable for most images, especially for those with random noise and n larger than five pixels.

4. Experiments

The proposed binarization algorithm was tested with the benchmarking technique and various scenarios

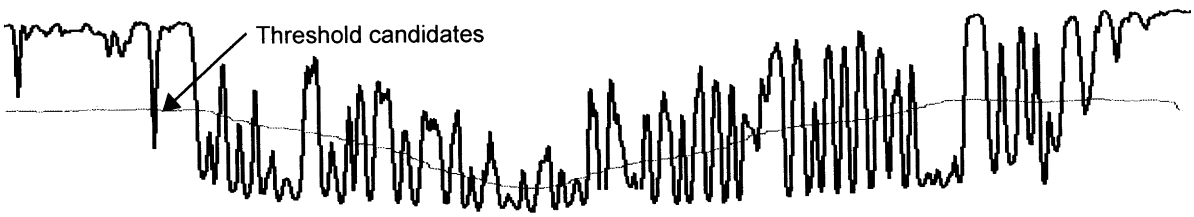

Fig. 10. Example of threshold candidate selection of an example scanline.

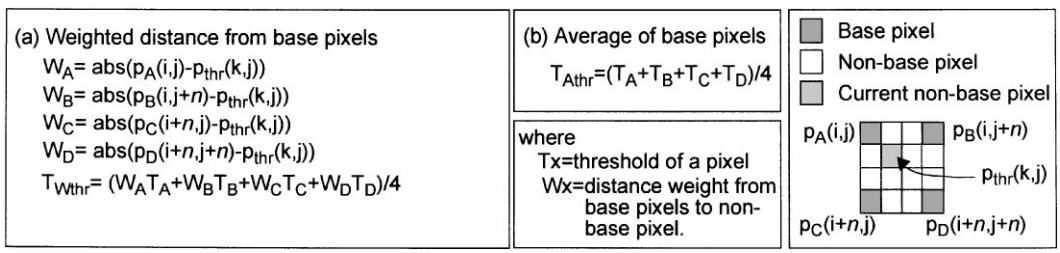

Fig. 11. Two interpolation choices for threshold selection of non-base pixels.

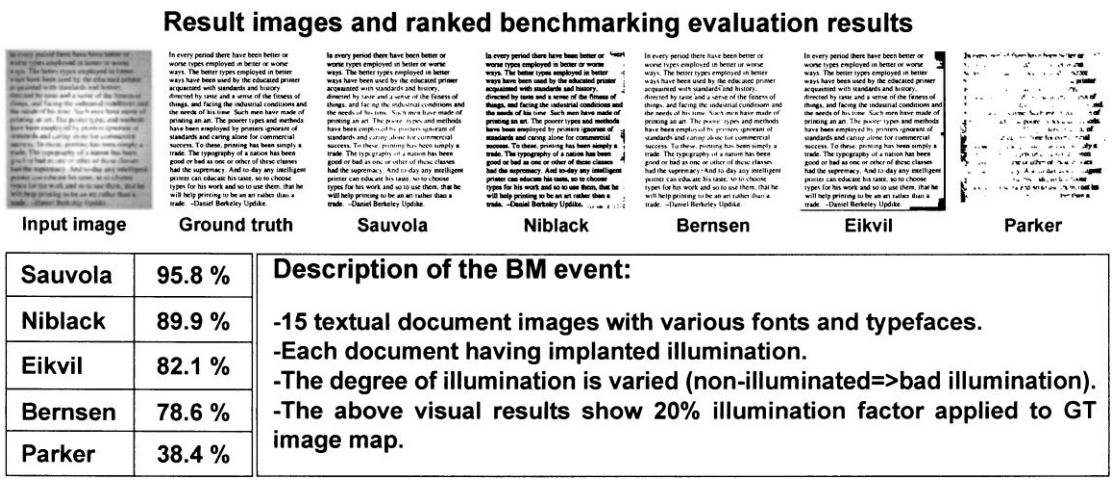

Fig. 12. Visual and numeric results on the comparison algorithms applied to illuminated, textual images.

against several known binarization techniques in the literature [18,20–22]. Using the environment factors (such as different degradations) and available document and test image databases the algorithm results were evaluated and benchmarked against each other, against the ground-truth knowledge by visual and benchmark event(s) evaluation processes. The focus was set on documents with textual content and on multi-content documents, i.e. documents having text, graphics, linedrawings and halftone. The test images were selected from a special database of document image categories, comprising over 1000 categorized document images (e.g. article, letter, memo, fax, journal, scientific, map, advertisement, etc.) [23].

The numerical test and results presented were gained using binarization metrics emphasizing the performance in textual image region binarization. Fig. 12 presents an example benchmarking scene performed to a database of 15 textual document images having illumination. Visual results to a sample input image having 20% of centered illumination defect, an example of a ground-truth image map and the results of the proposed and comparison binarization algorithms. The results show good behaviour of Sauvola's, Niblack's and Eikvil's algorithms, when the limit is set to 80% performance, i.e. the limit where the OCR performance drop is less than 10% using Caere Omnipage OCR package [24]. Bernsen suffered of noise that was introduced to binarized result image, while the Eikvil's threshold ruled some of the darkest areas belong to object pixels. Parker's algorithm adapted poorly to even small changes in lumination, but had sufficient results with relatively 'clean' grey-scale document images.

The visual tests performed for a synthetic test image database were based on ranking according to different objectives set for these types of images. The purpose of the synthetic image database is to allow visual analysis of the nature and behaviour of the benchmarking technique in a different kind of situation, e.g. in edge preservation, the object uniformity preservation, in changing/varying background, etc. This is aimed to aid the suitability selection of different algorithm to differing environmental conditions in terms of adaptability to changes, shape management, object preservation, homogeneousness of region preservation, and so on. An example of the visual results on synthetic images is shown in Fig. 13.

Fig. 13 shows visually the results of our, and comparison, algorithms applied to synthetic grey-scale images having different/differing kind of background(s), object(s), line(s), directions and shapes complying with certain simple test setting rules. As the input grey-scale images were synthetically generated, a set of ground-truth images were generated focusing in different areas of interest in measuring the algorithm performance and behaviour. Therefore, the benchmark results are dependent on the selection of the ground-truth set used, i.e. the target performance group the algorithm behaviour. For example, the ground-truth criteria of object uniformity and edge preservation were tested using ground-truth image in Fig. 13a. The object edge and background/object uniformity was used as a weight criteria, where the Euclidean distance was used as a distance measure between the result and the ground-truth pixel maps. Fig. 13b shows a situation, where the synthetic image has uniformly gliding background from white to black, and thin lines, whose grey-scale value glides on the opposite direction from the background. The test evaluation criterium was set on differentiating lines from background and uniformity of the background. Since the results are highly dependent on the target aims of the binarization,

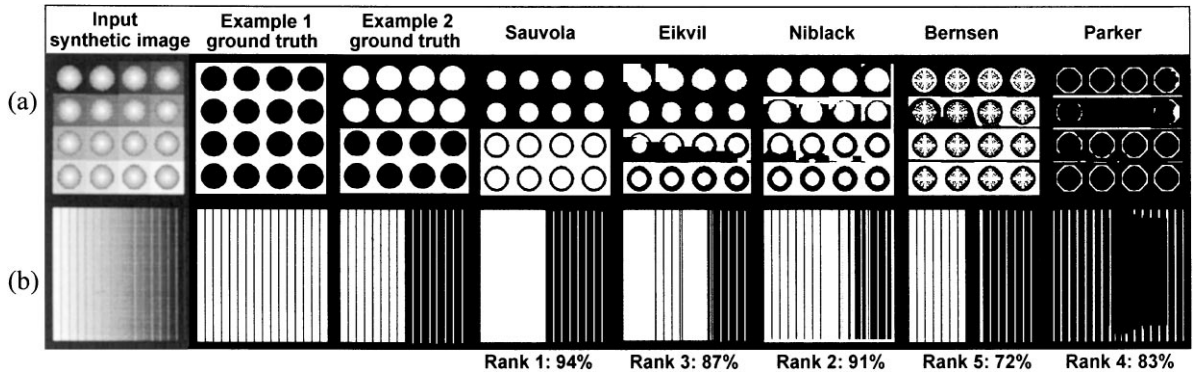

Fig. 13. Results on the comparison algorithms applied to the synthetic graphical images.

Description of the BM event:

-A database of textual documents.
-Each document having different or mixed amount of degradation types.
-The degree of degradation varies: (same document with different degrees of degradation with same degradation type(s)).

BM focus: textual content	Overall weighted performance
Sauvola, (x)	93.2 %
Niblack, (o)	88.1 %
Bernsen, (v)	85.8 %
Eikvil, (+)	84.9 %
Parker, (-)	46.8 %

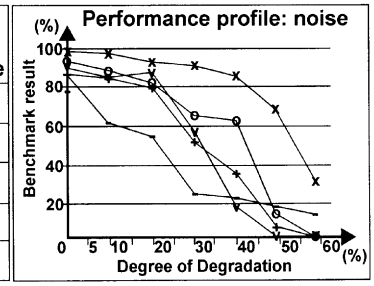

Fig. 14. Overall benchmarked binarization and example profile results on 'text only' document database.

the results are presented also visually. By using the criteria of the uniformity and object shape preservation the proposed algorithm behaves robustly compared to other techniques. Since most of the pixels in synthetic images are judged by the soft control method, the threshold between objects and non-object candidates seems very clear.

Fig. 14 shows benchmarking results performed with the textual image database with small amounts of clean and mixed illumination and noise types. An example performance profile to noise degradation component is shown for all the comparison algorithms. The degree of noise degradation presents the percentage of Gaussian and random noise introduced in the textual image, and the performance using combined pixel map and OCR metrics with equal weight factors. The performance of the proposed and comparison algorithms, excluding Parker's, seems to be sufficient up till 20% noise penetration. The performance profile clearly shows that the performance of the comparison algorithms drops between 20 and 30% penetration, while the proposed algorithm tolerated with severe noise, up to 45% having 80% threshold limit for acceptable value.

Fig. 15 shows the overall results of the proposed and comparison algorithms with various document categories performed to a large database of document images. The test images comprise simple textual documents with and without degradation types and degrees, documents with mixed textual and graphical properties, where the benefits of the hybrid approach of the proposed algorithm can be clearly seen. The methods of Eikvil and Niblack performed best against the proposed algorithm, but they still suffered of poor adaptation to various degradation types and, for example, the font size used in the textual parts was combined with the characters. The Bernsen algorithm shows good results on clean document and did tolerate small amount of one defect type. When the degradation was higher, the algorithm's performance decreased rapidly both in visual and numerical evaluation. Parker's algorithm shows sufficient results with clean document images, but the result quality dropped with even small introduction of document with any defect type.

The algorithm execution times were not measured in this comparison, where only the quality of the result was benchmarked against the metrics in a weighted (textual, graphics, character) process. The computing times for all the evaluated algorithms were tolerable, for example for utilization as a preprocessing step in optical character recognition engines. One question in performing the benchmarking is the arrangement of parametrization. The proposed algorithm had no parameters to set during testing, while Niblack had one, and Bernsen two, Eikvil

Document image resolution 300dpi

Binarization technique	Case: 'clean doc' documents	Case: 'clean doc' small fonts	Case: 'illum_1 doc' documents	Case: 'illum_2 doc' documents	Case: 'noise_2 doc' documents	Case: 'mixed doc' documents	Overall: -percent % -ranked (x)
Sauvola	98.9 % (1)	95.3 % (1)	99.0 % (1)	95.6 % (1)	95.3 % (1)	85.7 % (1)	94.9 % (1)
Niblack	96.8 % (3)	92.5 % (3)	97.6 % (2)	94.8 % (2)	94.3 % (3)	85.6 % (2)	93.7 % (2)
Bernsen	96.3 % (4)	92.1 % (4)	90.6 % (3)	55.4 % (4)	90.4 % (4)	53.1 % (4)	79.7 % (4)
Eikvil	97.3 % (2)	95.1 % (2)	84.2 % (4)	62.6 % (3)	94.7 % (2)	59.4 % (3)	82.2 % (3)
Parker	83.9 % (5)	52.4 % (5)	81.6 % (5)	52.2 % (5)	59.6 % (5)	41.1 % (5)	75.4 % (5)

'clean doc' = cleaned grey-scale image
'illum_1 doc' = light illumination
'illum_2 doc' = heavy illumination
'noise_2 doc' = heavy noise contamination, 40%
'mixed doc' = mixed noise and illumination types

Fig. 15. Overall benchmarked binarization results on textual document database.

used first Otsu's technique with one parameter and their postprocessing with one parameter, Parker's algorithm had four parameters to set. Each algorithm with parameters that needed manual tuning was computed with different parameters, whose result were evaluated and the best was selected to final comparison presented in this paper. When the higher adaptation is required from the algorithm, the number of manually tunable parameters should not exceed two, otherwise the amount of manual work increases too much and cause instability where automated preprocessing is required.

The overall results show good evaluated performance to the proposed, Niblack's and Eikvil's algorithms. The difference if these approaches lies in overall adaptability, the need for manual tunability, target document category domain and environment, where the algorithm is utilized, and finally the threshold performance set for the binarization process. In the latter case the proposed and Niblack's algorithms performance and adaptivity was highest in all test categories in graphical and textual cases.

5. Conclusions

Document image binarization is an important basic task needed in most document analysis systems. The quality of binarization result affects to subsequent processing by offering pre-segmented objects in precise form (object/non-object). In this paper we proposed a new technique to document image binarization, using hybrid approach and taking document region class properties into consideration. Our technique is aimed at generic document types coping also with severe cases of different types of degradation. The result of the quality validation (i.e. benchmarking against other algorithms and ground truth) is an important part of the algorithm development process. The proposed algorithm went over large tests utilizing test image databases having textual, pictorial and synthetically generated document images with ground-truths and degradations. The results show especially good adaptation into different defect types such as illumination, noise and resolution changes. The algorithm showed robust behaviour in most, even severe, situations in degradation and performed well against the comparison techniques.

6. Summary

This paper presents a new algorithm for document image binarization using an adaptive approach to manage different situations in an image. The proposed technique uses rapid image surface analysis for algorithm selection and adaptation according to document contents. The contents is used to select the algorithm type and need for parametrization, if any, and to compute and propose the threshold value for each or every nth pixel (interpolative approach). The document content is used to guide the binarization process: a pictorial content is subjected to a different type of analysis than a textual content. The degradations, such as illumination and noise, are managed within each algorithm structure to effectively filter out the imperfections. The results of the thresholding processes are combined to a binarized image that can either use a fast option, i.e. to compute binarization for every nth pixel and interpolate the threshold value for the in-between pixels, or a pixel by pixel option that computes a threshold value for each pixel separately. The tests were run on a large database of document images having 15 different document types and a number of representative images of each type. Each image was processed with the presence of various amount of different degradation to evaluate the efficiency of the proposed algorithm. The results were compared to those obtained with some of the best-known algorithms in the literature. The proposed algorithm outperformed its competitors clearly and behaved robustly in difficult degradation cases with different document types.

Acknowledgements

The support from the Academy of Finland and Technology Development Centre is gratefully acknowledged. We also thank Dr. Tapio Seppanen and Mr. Sami Nieminen for their contributions.

References

[1] J. Sauvola, M. Pietikäinen, Page segmentation and classification using fast feature extraction and connectivity analysis, International Conference on Document Analysis and Recognition, ICDAR '95, Montreal, Canada, 1995, pp. 1127–1131.
[2] H. Baird, Document image defect models, Proceedings of the IAPR Workshop on Syntactic and Structural Pattern Recognition, 1990, pp. 38–46.
[3] L. O'Gorman, Binarization and multithresholding of document images using connectivity, CVGIP: Graph. Models Image Processing 56 (6) (1994) 496–506.
[4] Y. Liu, R. Fenrich, S.N. Srihari, An object attribute thresholding algorithm for document image binarization, International Conference on Document Analysis and Recognition, ICDAR '93, Japan, 1993, pp. 278–281.
[5] J. Yang, Y. Chen, W. Hsu, Adaptive thresholding algorithm and its hardware implementation, Pattern Recognition Lett. 15 (2) (1994) 141–150.
[6] M. Chang, S. Kang, W. Rho, H. Kim, D, Kim, Improved binarization algorithm for document image by histogram and edge detection, International Conference for Document Analysis and Recognition ICDAR '95, Montreal, Canada, 1995, pp. 636–643.
[7] T. Pavlidis, Threshold selection using second derivatives of the gray scale image, International Conference on Document Analysis and Recognition, ICDAR '93, Japan, 1993, pp. 274–277.
[8] A. Rosenfeld, R.C. Smith, Thresholding using relaxation, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-3 (5) (1981) 598–606.
[9] A. Perez, R.C. Gonzalez, An iterative thresholding algorithm for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-9 (6) (1987) 742–751.
[10] J.R. Parker, C. Jennings, A.G. Salkauskas, Thresholding using an illumination model, ICDAR '93, Japan, 1993, pp. 270–273.
[11] V.A. Shapiro, P.K. Veleva, V.S. Sgurev, An adaptive method for image thresholding, Proceedings of the 11th KPR, 1992, pp. 696–699.
[12] A. Pikaz, A. Averbuch, Digital image thresholding, based on topological stable-state, Pattern Recognition 29 (5) (1996) 829–843.
[13] P.V. Henstock, D.M. Chelberg, Automatic gradient threshold determination for edge detection, IEEE Trans. Image Processing 5 (5) (1996) 784–787.
[14] W. Chen, C. Wen, C. Yang, A fast two-dimensional entropic thresholding algorithm, Pattern Recognition 27 (7) (1994) 885–893.
[15] S.D. Yanowitz, A.M. Bruckstein, A new method for image segmentation, CVGIP 46 (1989) 82–95.
[16] S. Nieminen, J. Sauvola, T. Seppänen, M. Pietikäinen, A benchmarking system for document analysis algorithms, Proc. SPIE 3305 Document Recognition V 3305 (1998) 100–111.
[17] S.T. Welstead, Neural Network and Fuzzy Logic Applications in C/C++, Wiley, New York, 1994, p. 494.
[18] W. Niblack, An Introduction to Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1986, pp. 115–116.
[19] O.D. Trier, A.K. Jain, Goal-directed evaluation of binarization methods, IEEE Trans. Pattern Anal. Mach. Intell. 17 (12) (1995) 1191–1201.
[20] L. Eikvil, T. Taxt, K. Moen, A fast adaptive method for binarization of document images, International Conference on Document Analysis and Recognition, ICDAR '91, France, 1991, pp. 435–443.
[21] J. Bernsen, Dynamic thresholding of grey-level images, Proceedings of the Eighth ICPR, 1986, pp. 1251–1255.
[22] J. Parker, Gray level thresholding on badly illuminated images, IEEE Trans. Pattern Anal. Mach. Intell. 13 (8) (1991) 813–819.
[23] J. Sauvola, S. Haapakoski, H. Kauniskangas, T. Seppänen, M. Pietikäinen, D. Doermann, A distributed management system for testing document image analysis algorithms, 4th ICDAR, Germany, 1997, pp. 989–995.
[24] Caere Ominpage OCR, Users Manual, Caere Corp., 1997.

About the Author—JAAKKO SAUVOLA is a Professor and Director of the Media Team research group in the University of Oulu, Finland, and a member of the affiliated faculty at the LAMP Laboratory, Center for Automation Research, University of Maryland, USA. Dr. Sauvola is also a Research Manager in Nokia Telecommunications, where his responsibilities cover value adding telephony services. Dr. Sauvola is a member of several scientific committees and programs. His research interests include computer-telephony integration, media analysis, mobile multimedia, media telephony and content-based retrieval systems.

About the Author—MATTI PIETIKÄINEN received his Doctor of Technology degree in Electrical Engineering from the University of Oulu, Finland, in 1982. From 1980 to 1981 and from 1984 to 1985 he was a visiting researcher in the Computer Vision Laboratory of the University of Maryland, USA. Currently, he is a Professor of Information Technology, Scientific Director of Infotech Oulu research center, and Director of Machine Vision and Media Processing Group at the University of Oulu. His research interests cover various aspects of image analysis and machine vision, including texture analysis, color machine vision and document analysis. His research has been widely published in journals, books and conferences. He was the editor (with L.F. Pau) of the book "Machine Vision for Advanced Production", published by World Scientific in 1996. Prof. Pietikäinen is one of the founding Fellows of the International Association for Pattern Recognition (IAPR) and a Senior Member of IEEE, and serves as Member of the Governing Board of IAPR. He also serves on program committees of several international conferences.