Document Recognition and Retrieval VI
Volume 3651

PROCEEDINGS OF SPIE

SPIE—The International Society for Optical Engineering

Document Recognition and Retrieval VI

Daniel P. Lopresti
Jiangying Zhou
Chairs/Editors

27–28 January 1999
San Jose, California

Sponsored by
IS&T—The Society for Imaging Science and Technology
SPIE—The International Society for Optical Engineering

Published by
SPIE—The International Society for Optical Engineering

SPIE P PROCEEDINGS SERIES

Volume 3651

SPIE is an international technical society dedicated to advancing engineering and scientific applications of optical, photonic, imaging, electronic, and optoelectronic technologies.

The papers appearing in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and are published as presented and without change, in the interests of timely dissemination. Their inclusion in this publication does not necessarily constitute endorsement by the editors or by SPIE.

Please use the following format to cite material from this book:
Author(s), "Title of paper," in *Document Recognition and Retrieval VI*, Daniel P. Lopresti, Jiangying Zhou, Editors, Proceedings of SPIE Vol. 3651, page numbers (1999).

ISSN 0277-786X
ISBN 0-8194-3122-2

Published by
SPIE—The International Society for Optical Engineering
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone 360/676-3290 (Pacific Time) • Fax 360/647-1445

Printed in the United States of America.

Conference Committee

Conference Chairs

Daniel P. Lopresti, Lucent Technologies/Bell Laboratories
Jiangying Zhou, Summus, Ltd.

Program Committee

Francine R. Chen, Xerox Palo Alto Research Center
David S. Doermann, University of Maryland/College Park
Michael D. Garris, National Institute of Standards and Technology
Jonathan J. Hull, Ricoh California Research Center
Larry Spitz, Document Recognition Technologies, Inc.
Kazem Taghva, University of Nevada/Las Vegas
Ellen Voorhees, National Institute of Standards and Technology

Session Chairs

1 OCR Systems and Techniques
Jonathan J. Hull, Ricoh California Research Center

2 Handwriting Recognition
Michael D. Garris, National Institute of Standards and Technology

3 Models and Evaluations
Kazem Taghva, University of Nevada/Las Vegas

4 Information Retrieval
Ellen Voorhees, National Institute of Standards and Technology

5 Document Analysis
David S. Doermann, University of Maryland/College Park

Introduction

The sixth annual conference on Document Recognition and Retrieval reflects a number of important changes that we hope will help move this series forward as our field continues to evolve. For the first time, this conference has a formal program committee, consisting of distinguished researchers from government, industry, and academia. It is also the first time that longer, extended abstracts were solicited for review, and that this proceedings is being made available on site. All of these steps, plus additional ones planned for the near future, are intended to strengthen DR&R as a forum for first-class research.

The papers in this volume represent both the high quality of the work being done in document recognition and retrieval today, and the tremendous breadth of interests that fall under this umbrella. These range from traditional problem areas such as character segmentation and recognition, to newer ones including location of text in video, browsing image databases on the WWW, and the issue of "meta-data" and how it might be useful. The papers (and, we expect, the attendees) also reflect the truly international nature of our field.

We are again fortunate to have two invited talks scheduled this year. The first, by Henry Baird of Xerox PARC, is on "Model-directed document image analysis." The second, by Ellen Voorhees of NIST, is a survey titled "Information retrieval: roots and future directions." In addition, we have included a panel session for the first time. The topic is "Meta-Data: What Good Is It?," and the scheduled panelists are Francine Chen (Xerox PARC), Steve Dennis (U.S. Department of Defense), David Doermann (University of Maryland), and Andrew Tomkins (IBM Almaden Research Center).

We hope that everyone finds this year's conference rewarding, and look forward to hearing your opinions on our recent changes as well as possibilities for the future.

Daniel P. Lopresti
Jiangying Zhou

Contents

SESSION 1

OCR Systems and Techniques

Text Enhancement in Digital Video

Huiping Li *[a], Omid Kia[b] and David Doermann[a]

[a]Language and Media Processing Laboratory
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742-3275

[b]Mathematical and Computational Sciences Division
National Institute of Standards and Technology
Gaithersburg, MD 20899

ABSTRACT

One difficulty with using text from digital video for indexing and retrieval is that video images are often in low resolution and poor quality, and as a result, the text can not be recognized adequately by most commercial OCR software. Text image enhancement is necessary to achieve reasonable OCR accuracy. Our enhancement consists of two main procedures, resolution enhancement based on Shannon interpolation and text separation from complex image background. Experiments show our enhancement approach improves OCR accuracy considerably.

Keywords: Text enhancement, Adaptive thresholding, OCR, Shannon Interpolation

1. INTRODUCTION

The increasing availability of online digital imagery and video has rekindled interest in the problems of how to index multimedia information sources automatically and how to browse and manipulate them efficiently. Although content based recognition has not progressed to the point where it is useful for indexing heterogeneous collections, text in digital videos can provide important supplemental metadata such as sports scores, product names, scene locations, speaker names, movie credits, program introductions and special announcements. Complete use of this information requires us to develop algorithms to detect, extract and recognize text robustly. Although in previous work we have developed algorithms to identify text regions in digital video key frames,[4] the stumbling block remains that extracted text can not be recognized by typical commercial OCR systems. Figure 1a shows a text block extracted from a video frame. Even when we manually pick the best threshold to binarize the image (Figure 1b), there is still no output from the OCR software[1], even though the text is clearly readable.

Another difficulty of using OCR is that text in digital video is often embedded in complex scenes. Since most OCR software handles only binary images, the binarization and seperation of text from complex background is a necessary process. These difficulties provide us the motivation to conduct text enhancement to achieve reasonable OCR results using commercial OCR software so the text-based indexing and retrieval is possible.

The remainder of this paper is organized as follows: In Section 2 we address some ralated work. Section 3 describes our text resolution enhancement algorithm in detail and Section 4 addresses the problem of text seperation from scene background. We present experimental results in Section 5 and provide a brief discussion in Section 6.

*Correspondence: Email: huiping@cfar.umd.edu; WWW:http://www.cfar.umd.edu/ huiping
The support of this research by the Department of Defense under contract MDA 9049-6C-1250 is gratefully acknowledged.
[1]We have tested with Caere Omnipage 8.0, Xerox TextBridge Pro98 and Xerox ScanWorx

Part of the IS&T/SPIE Conference on Document Recognition and Retrieval VI • San Jose, California • January 1999
SPIE Vol. 3651 • 0277-786X/99/$10.00

(a) (b)

Figure 1. (a) A text block with a font size of approximately 6-7 pixels, (b) Binarization by manually picking the best threshold. There is no output from commercial OCR software although the text is clearly readable.

2. RELATED WORK

Previous work on text enhancement has focused primarily on binary document images. Hobby and Ho present a method to enhance degraded document images via bitmap clustering and averaging for better display quality and recognition accuracy.[2] OCR accuracy is improved from 6% to 38% for documents with varying quality. Stubberud[8] presented an adaptive technique that restores touching or broken character images to improve performance of OCR system. Liang[5] addressed the problem of document image restoration using morphological filters and achieved nearly 80% OCR accuracy for subtractive noise images and additive noise images.

Some work has also been done on the OCR from scene and *WWW* images. Zhou and Lopresti describes their work on text extraction and recognition from *WWW* images.[11,12] For recognition they do not use commercial OCR software but design two classifiers (surface fitting classifier and n-tuple classifier). The recognition accuracies achieved are 69.7% and 89.3% respectively for two classifiers. Wu and Manmatha describes a text extraction and recognition system and achieves 84% correct OCR rate based only on "*OCRable*" text in images.[10] Lienhart describes a text recognition system in digital video and achieves a recognition result of nearly 80% for artificial text but ignores the scene text.[6] None of these systems, however, perform any enhancement and all rely on the text having substantial resolution.

Our work concentrates on text which has very low resolution but still readable by our human observer.

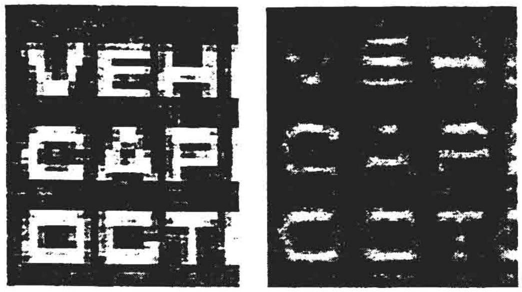
3. IMAGE RESOLUTION ENHANCEMENT USING SHANNON UPSAMPLING

Video frames are typically limited in spatial resolution, but this is overcome by the use of color consistency. This provides a sufficient definition for viewers, yet may not be sufficient for automatic recognition. Furthermore, textual content is almost always rendered with a high-contrast, otherwise, the content provider risks the possibility of the viewer not noticing the text. For document images, 300 dots per inch is commonplace which translates to characters occupying an area as large as 50×50 pixels. This is usually too large for television screens capable of viewing 640×480 pixels or video frames recorded in even lower resolution (In $MPEG-1$, the resolution is 352×240). Instead providers render characters in the neighborhood of 10 pixels using anti-aliasing to reduce digitization effects. To achieve this level of reduction, an anti-aliasing process is performed by low pass filtering the image followed by subsampling. For an image, I, of M rows and N columns and a low pass filter with impulse response of F_{lpf}, the resultant image i subsampled at each Δ pixels or at $1/\Delta$ rate would be represented by

$$i(x,y) = \sum_{i=1}^{M}\sum_{j=1}^{N} I(i,j)F_{lpf}(x-i, y-i), \;\; for \left\{ \begin{array}{l} x = 1,..,N/\Delta \\ y = 1,..,M/\Delta \end{array} \right. \tag{1}$$

A close up of the image shown in Figure 1a is shown in Figure 2a. The process lowers the resolution but increases number of colors (or graylevels) to give impression of contrast and detail.

The reverse process is difficult. The reduction of high frequency components along with cross-over of frequency at subsampling makes it almost impossible to rebuild the original image. The lost information is only recoverable if there is motion or there exists knowledge about the contents of the image. Unfortunately, due to variations in low pass filters, image encoding artifacts, and detectable motion, the task of creating near original images becomes impossible. However, we can increase resolution given information that we have. This is done by extension of the *Nyquist* sampling theorem where a sampled image is a weighted sum of delayed *Sinc* functions.



(a) (b)

Figure 2. (a) Anti-aliased image showing blurred edges and large number of foreground colors. (b) Zoomed image by Shannon's interpolation.

$$i(X,Y) = \sum_{x=1}^{N}\sum_{x=1}^{M} i(x,y) sinc(X-x, Y-y), \ \ for \begin{cases} X \in \Re(1,N) \\ Y \in \Re(1,M) \end{cases} \tag{2}$$

Inter-sample values are then the sum of the Sinc functions at their non-zero crossings. This process is computationaly intensive and while there exists a number of simplifications, time domain processing of this task strains processing resources. The alternative is to pursue a frequency-based approach. The dual of the weighted *Sinc* functions can be performed by Fast Fourier Transforms (FFT) and matrix masking. We increase resolution of our images by copying each pixel to neighboring pixels by the amount of the desired increase in resolution. We then take the two dimensional FFT. The resulting matrix is then multiplied by a mask matrix which zeros the high frequency components. The number of the low frequency that is preserved is the same size as the original image. This in effect is a low pass filter where high frequency components introduced by copying pixels to neighboring pixels are removed. An inverse two dimensional FFT renders the image in the higher resolutional mode. Figure 2b shows the result of the increase in resolution of Figure 2a by a factor of four. This task does not require a lot of processing and can be easily encoded in hardware.

4. TEXT SEPERATION FROM COMPLEX BACKGROUND

4.1. Identification of Inverse Text

After resolution enhancement, we perform text seperation to produce a text image with a clean background, since typical OCR software can handle only binary images with black pixels representing text and white pixels representing background. Before we perform binarization, we need to classify the normal text (Figure 3a) and the inverse text (Figure 3b).

The scheme we use is very simple: First, we calculate a global thresholding **Th**. Considering the background can be defined as the maximum part of the histogram of the image, we can make our decision by comparing the threshold **Th** and the background value **Bg** which corresponds to the maximum part of the histogram. Specifically,

if **Th** > **Bg**, then it is normal text (Figure 3c)

if **Th** < **Bg**, then it is inverse text (Figure 3d)

Although a global threshold is not ideal, it is sufficient for this problem. It is not necessary for us to do any inversion on inverse text at this point. Instead, when we perform the binarization where the text is always written as black pixels as required by OCR software.

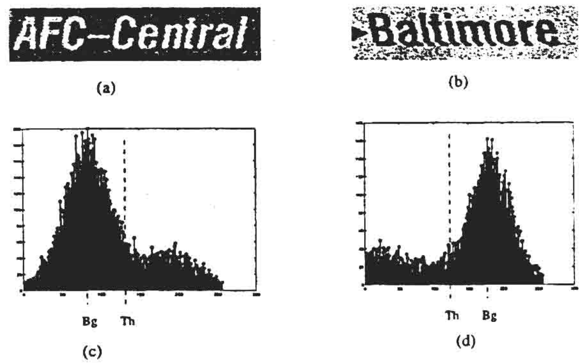


Figure 3. (a) An example of normal text, (b) An example of inverse text, (c) Histogram of (a), (d) Histogram of (b).

4.2. Adaptive Thresholding

Text in digital videos may be located in complex image backgrounds. Due to large variations in text font, size and contrast to the background, selecting an effective universal threshold will be very difficult, if not impossible. Figure 4a shows an example. Since the contrast between text and background is not uniform, we can see the global binarization will mix the text in left side with background. As a result, only the right part of text gets correct OCR result (Figure 4b). Adaptive thresholding techniques are generally required.

Many local adaptive thresholding methods, such as Bernsen's,[3] Mardia's[7] and Niblack's method,[9] create a threshold surface of the same dimension as the original image by calculating an explicit threshold value over a window of a given size for each pixel. Niblack's method varies a threshold over the image, based on the local mean, m, and local standard deviation, s, computed in a small neigborhood of each pixel. A threshold for each pixel is computed from $T = m + k * s$, where k is a user defined parameter. The result of this method is not very good for background areas containing light textures. The reason is that in text areas the deviation s is often large, so k should be kept small ($|k| < 1$). For background areas containing light textures, s is often small, so the threshold T is close to mean m. As shown in Figure 4c, many pixels in the background area will exceed the threshold resulting in bogus OCR results eventhough all the characters in the original image have been correctly recognized. The second problem is that the computational requirements are very high since we need to calculate threshold for each pixel. It takes approximately 3.5 seconds in Ultra Sparc 1 for the binarization of the image in Figure 4a (image size is 348×47).

We incorporate a simple classification scheme with Niblack's algorithm to improve both the binarization result and speed. The idea is based on the observation that text usually has a significant contrast with the background and therefore, has a relatively larger standard deviation than background areas which contains light texture. In order to demonstrate this point, we collected 600 small text blocks from digital video images and calculate the distribution of their standard deviation. As Figure 5 shows, all text blocks' standard deviations are larger than 25. Therefore, if we find the standard deviation in one window is smaller than this threshold, we could simply mark all the pixels in this window as background. We do NOT intend to use this simple method as a classifier to tell text block from nontext block. Our purpose here is to filter out those background areas containing light texture which will deteriorate performance in Niblack's method. As a result, both the binarization and speed is improved considerably. This method (Figure 4d) leads to a correct OCR result, requiring only 0.5 seconds of processing for thresholding.

5. EXPERIMENT RESULTS

5.1. Testing Environment

It should be noted that we do not intend to implement our OCR system since there exists many commercial OCR packages. The OCR software we used in our experiment is Xerox TextBridge Pro98. After binarization of text image, we manully feed the image to TextBridge Pro98 for recognition.

We use *ope*, an OCR evaluation software available by the University of Washington[1] to evaluate our enhancement algorithm. *Ope* performs string matching by measuring the amount of character changes, insertions, and deletions.

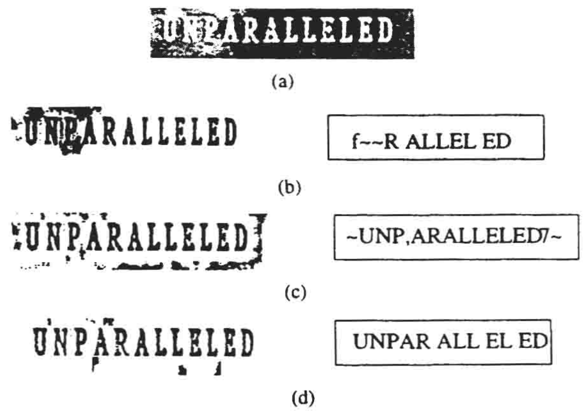


Figure 4. (a) Enhanced gray-scale text blocks, (b) Global Thresholding and OCR result, (c) Niblack's method and OCR result, (d) Our method and OCR result.

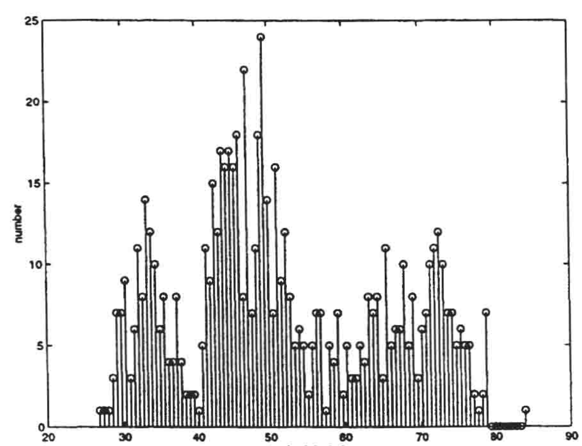


Figure 5. The distribution of the standard deviations of over 600 text regions.

	No Enhancement	Zero order hold Interpolation	Shannon Interpolation
# of Total Blocks	45	45	45
# of Blocks Having OCR Output	13(29%)	36(80%)	45(100%)
# of Total Characters	1452	1452	1452
# of Correctly OCRed Characters	188(13%)	489(34%)	970(66.8%)

Table 1. The Comparision of OCR with no enhancement, zero order hold and Shannon interpolation (interpolation factor is 2).

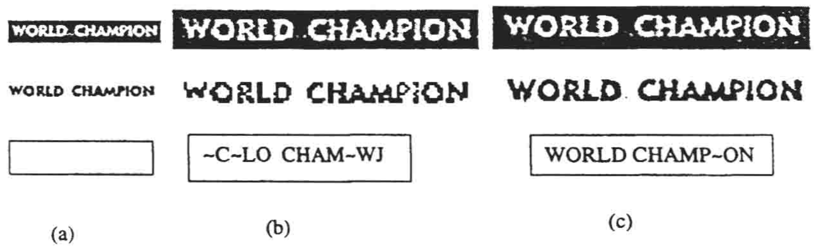


Figure 6. Comparison of OCR result. (a)No enhancement, (b)Zero order interpolation, (c)Shannon interpolation.

5.2. Experiment Result

We have collected 45 text blocks from digital video frames with a standard frame size of 320 × 240. We generated another 45 text blocks by Shannon interpolation to magnify the image in the factor of two. In order to observe how our interpolation scheme can improve OCR result, we generate another group of text blocks by simply copying each pixels in the image to four pixels (Zero order hold interpolation). Both the interpolation schemes increase the resolution by a factor of two. The same binarization scheme mentioned above is applied to all the original text blocks and the interpolated text blocks.

As shown in Table 1, before enhancement, only 13(29%) text blocks have OCR output, but a significant improvement is achieved even with Zero order hold interpolation (36(80%)). All the 45 text blocks have OCR output for Shannon interpolation. The difference here tells us the resolution of most text blocks is beyond the machine recognition capability if no resolution enhancement is performed.

There are a total of 1452 characters in the 45 text blocks. Before enhancement, only 188(13%) are correctly OCRed. This is insufficient for any successful indexing operation. For Zero order hold interpolation, the OCR accuracy rate rises to 34%. We observe 66.8% correct rate for Shannon interpolation (Figure 6). We can see there is no OCR output for original image (Figure 6a). Although Zero order hold interpolation has OCR output, there are considrable number of recognition errors (Figure 6b). Figure 6c is the recognition result for Shannon interpolation.

When we raise the resolution of text by a factor of 2, the average size of text will be 20 × 20 and the OCR accuracy is improved from 13% to 34% for Zero order hold interpolation and 67% for Shannon interpolation. In order to investigate the relation between OCR accuracy and resolution, we interpolate the image by a factor of 4 and

Figure 7. Even with enhancement, there is still no OCR output due to poor image quality.

(a)

Scale 2

VEHICLE ASSEMBLY EUIL~NG
CAPE KENNECY. PLa~DA
OCTOBER 30. IEES

(b)

Scale 4

VEHICLE ASSEMBLY BUiLDING
CAPE KENNEDY, FLORIDA
OCTOBER 30, 1969

(c)

Scale 8

VEHICLE ASSEMBLY BUILDING
CAPE KENNEDY, FLORIDA
OCTOBER 30. 3969

(d)

Figure 8. An example of OCR with Shannon interpolation at different scales. (a) Original image. (b) OCR result with interpolation factor 2, (c) OCR result with interpolation factor 4, (d) OCR result with interpolation factor 8.

Interpolation factor	2	4	8
# of Total Characters	1452	1452	1452
Shannon	970(66.8%)	977(67.3%)	892(61.4%)
Zero order hold	525(36.2%)	598(41.2%)	503(34.7%)

Table 2. The comparision of OCR accuracy with different interpolation factor.

8, respectively for both interpolation methods. The OCR results are shown in Table 2: We can see further improving image resolution does not improve OCR accuracy necessarily. OCR accuracy at factor of 4 is slightly better than that at factor of 2 (41.2% vs. 36.2% for zero order hold and 67.3% vs. 66.8% for Shannon). However, OCR rate at factor of 8 decreases to 34.7% and 61.2% respectively. OCR accuracy at factor of 4 gets the best result since at this scale, the average font size is around 45 - 50 pixels, typical for a document scanned at 300 dpi. Figure 8 is an example of OCR for Shannon interpolation at different factors. Considering the extra calculation cost, Shannon interpolation with a factor of 2 is a good choice in our application.

Although this result is far from perfect, we do not discriminate whether text blocks are OCRable or not. These include some text which is in fact unrecognible (Figure 7). It is very encouraging that we observed over 500% improvement.

6. CONCLUSION AND DISCUSSION

We have addressed the problem of text enhancement in digital videos. Our algorithm consists of resolution enhancement based on Shannon interpolation and text seperation from complex background. The experiments show improvement of the OCR rate from 13% to 67% by our enhancement.

There still remains many interesting concepts worth addressing. For example, use of color instead of our current gray-scale images may provide additional information to the textual content. We can also make use of the fact that the same text elements usually span multiple frames to achieve better enhacement and OCR results. We will focus on these and other related problems in our future work.

REFERENCES

1. S. Chen. OCR performance evaluation software – user's manual. In *The University of Washington Database*.
2. J. Hobby and T.K. Ho. Enhancing degraded document images via bitmap clustering and averaging. In *Proceedings of Fourth International Conference on Document Analysis and Recognition*, pages 394–400, 1997.
3. Bernsen J. Dynamic thresholding of grey-level images. In *Proceedings of ICPR*, pages 1251–1255, 1986.
4. H. Li and D. Doermann. Automatic identification of text in digital video key frames. In *Proceedings of ICPR*, pages 129–132, 1998.

5. J. Liang and R.M. Haralick. Document image restoration using binary morphological filters. In *SPIE Vol. 2660*, 1996.
6. R. Lienhart and F. Stuber. Automatic text recognition in digital videos. In *Proc. of ACM Multimedia 96*, pages 11–20, 1996.
7. K.V. Mardia and T.J. Hainsworth. A spatial thresholding method for image segmentation. In *PAMI(10), No. 6,*, pages 919–927, November 1988.
8. J. Kanai P. Stubberud and V. Kalluri. Adaptive image restoration of text images that contain touching or broken characters. In *ICDAR 95*, pages 778–781, 1995.
9. Niblack W. In *An introduction to image processing*, pages 115–116, Englewood Cliffs, N.J.:Prentice Hall, 1986.
10. V. Wu, R. Manmatha, and E.M. Riseman. Automatic text detection and recognition. pages 707–712. Proceedings of the DARPA Image Understanding Workshop, 5 1997.
11. J. Zhou and D. Lopresti. Ocr for world wide web images. In *Proceedings of SPIE on Document Recognition IV*, pages 58–66, 1997.
12. J. Zhou, D. Lopresti, and T. Tasdizen. Finding text in color images. In *Proceedings of SPIE on Document Recognition V*, pages 130–140, 1998.

Determining the Resolution of Scanned Document Images

Dan S. Bloomberg

Xerox Palo Alto Research Center
Palo Alto, CA 94304

ABSTRACT

Given the existence of digital scanners, printers and fax machines, documents can undergo a history of sequential reproductions. One of the most important determiners of the quality of the resulting image is the set of underlying resolutions at which the images were scanned and binarized. In particular, a low resolution scan produces a noticeable degradation of image quality, and produces a set of printed fonts that cause omnifont OCR systems to operate with a relatively high error rate. This error rate can be reduced if the OCR system is trained on text having fax scanner degradations, but this also requires that the OCR system can determine in advance if such degraded fonts are present.

Methods are found for determining the lowest resolution that a binary document image has been scanned (or printed) in the past. Observing that the signature for a low resolution scan is contained in the boundary contours of the black components, the connected components that constitute the contours are measured, and both the size histogram and its power spectrum are used to determine the underlying low resolution, if any. The primary application is to fax, both standard and fine. Degradation of both font appearance and OCR performance is far more severe for standard than fine fax, so the most important practical problem is to identify documents of binary scanned text that have a standard fax signature. Only a small part of the page image is required to determine the existence of such a signature; consequently, the image is pre-processed to locate a subregion that is well suited for analysis. Results for discriminating between originals, standard fax and fine fax on a small but diverse test suite are given.

Keywords: image resolution, facsimile, fax, pattern spectrum, power spectrum, image analysis

1 Introduction

The purpose of this study is to find accurate and efficient methods for determining if a digital image has been previously scanned at low resolution, measured in pixels/inch (ppi). The most common fax resolutions are *standard*, approximately 200 x 100 ppi, and *fine*, approximately 200 x 200 ppi. Images scanned at standard fax resolution suffer obvious visible degradation. This is also reflected by the generally poor performance of omni-font OCR systems on such images.[3] For example, in the OCR "bakeoff" from the 1996 SDAIR Conference, the character error rate of the two *best* recognizers on standard fax was nearly 6 percent; the others were more than 10 percent. This should be compared with best character error rates of less than 2 percent on 300 ppi binary, and less than 3 percent on 200 ppi binary and fine fax. OCR performance can be improved using special recognizers