Extraction of Handwritten Text from Carbon Copy Medical Form Images

Robert Milewski and Venu Govindaraju

University at Buffalo,
Center of Excellence for Document Analysis and Recognition,
520 Lee Entrance, UB Commons Suite 202,
Amherst NY 14228
{milewski, govind}@cedar.buffalo.edu

Abstract. This paper presents a methodology for separating handwritten foreground pixels, from background pixels, in carbon copied medical forms. Comparisons between prior and proposed techniques are illustrated. This study involves the analysis of the New York State (NYS) Department of Health (DoH) Pre-Hospital Care Report (PCR) [1] which is a standard form used in New York by all Basic and Advanced Life Support pre-hospital healthcare professionals to document patient status in the emergency environment. The forms suffer from extreme carbon mesh noise, varying handwriting pressure sensitivity issues, and smudging which are further complicated by the writing environment. Extraction of handwriting from these medical forms is a vital step in automating emergency medical health surveillance systems.

1 Introduction

This research evaluates several algorithms which extract handwriting from medical form images (see Figure 1) to eventually provide the best handwriting recognition performance. The research copy of the NYS PCR [1] is a yellow-gray carbon mesh where both the handwriting and the mesh around the handwriting have approximately the same intensity. While the density and connectedness of handwriting is heavier then the mesh residue surrounding the handwriting, pressure sensitivity issues can affect the differentiation between the handwriting stroke and the background. The absence of sufficient pen pressure while writing leads to the loss of character information in the carbon copy. This causes character strokes to break after binarization which leads to recognition failures. Prior binarization algorithms have been reported to handle noisy and complicated surfaces [6][10][12]. However, the broken/unnatural handwriting due to ambulance movement and emergency environments, and carbon smearing from unintentional pressure to the form add further complexity to the binarization task. A lexicon driven word recognizer (LDWR) [13] is used for evaluation of the binarization methods. Analysis of the LDWR, as well as a full view of an actual NYS PCR image, can be found in [5].

Section 2 presents an examination of the carbon mesh paper image. Section 3 discusses the results of prior work on the medical forms. Section 4 proposes a new

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 106 – 116, 2006.

strategy for the handwriting foreground extraction. Section 5 compares all algorithms using the LDWR [13]. Section 6 summarizes the findings of this work.

2 Carbon Paper Image

Figure 1 shows an example of the "Objective Assessment" region of the NYS PCR form. It provides an overview of the complex nature of the handwriting on the carbon paper. Figure 2 shows a 400% zoom of one word from Figure 1. It shows the carbon paper mesh integrated with the carbon handwriting stroke. The displayed word *ABD*, in Figure 2, is a common abbreviation for *abdomen*. Since both the background paper and the handwriting are affected by the carbon paper, both the foreground and some parts of the background have identical intensities. This causes many binarization algorithms, which rely on thresholding small areas of the document to fail. The details of these failures are discussed in the following sections. This paper presents an algorithm for binarizing the handwriting on carbon paper while preserving the handwriting stroke connectivity better then the prior algorithms. The inconsistent carbon paper, which shows varying grayscale intensities (see Figures 1 and 2), is referred to as *carbon mesh*.

Pressure sensitivity issues, as a result of light strokes in penmanship, affects the extent to which character connectivity is maintained after binarization. In order for the carbon copy to receive a reasonable representation of the top copy original, the healthcare professional needs to press down firmly with the writing instrument. Since the emergency environment is not conducive to good penmanship, the binarization and cleanup algorithms need to compensate.

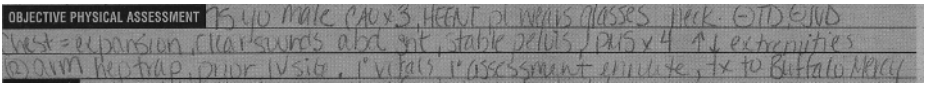

Fig. 1. NYS PCR Object Physical Assessment Example

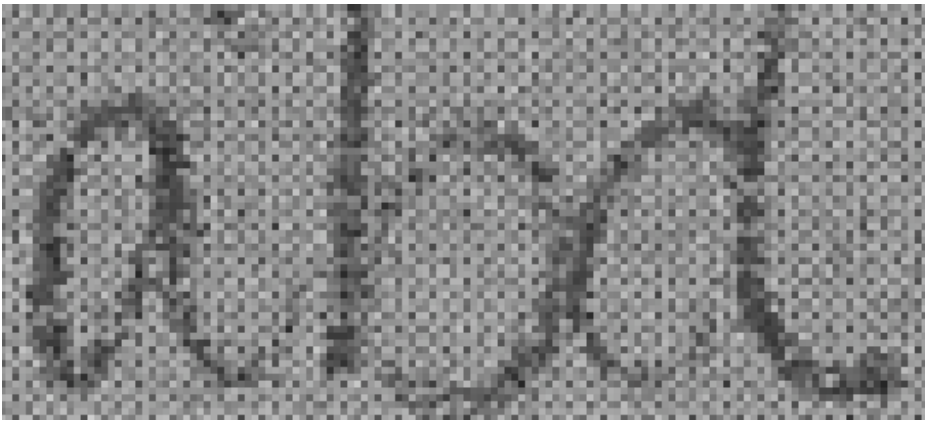

Fig. 2. Grayscale 256 Carbon Mesh Handwriting Example (400% Zoom)

The carbon paper forms also contain guide lines which often interfere with the character strokes. These lines can be detected by those pixels with a grayscale value less then 40; this is consistent across all forms. To reduce stroke fragmentation, it is sufficient to retain the pixels near the line thus keeping most character ascenders and descenders reasonably connected. This form drop out step is performed before binarization.

3 Prior Work

In this section, methods previously described in the literature are compared with the algorithm presented in this paper.

Gaussian, median and mean filtering/smoothing are often used as a base step or an integrated step for noise removal and image enhancement [2][11][18][21]. Mean filter (Figure 3b), shows the least damage to strokes. Median filter (Figure 3c) shows severe character damage. Gaussian Filter (Figure 3d) shows the characters being washed into the background.

Global thresholding algorithms determine a single threshold and apply it to the entire image. In the PCR application, the high pressure sensitive areas are binarized well, whereas medium to low pressure areas run the risk of being classified as background.

Many authors have cited and shown superiority over the Otsu [8][24] algorithm. Since a global threshold is computed, the background paper in many areas of the document merge with the foreground pixels. The Wu/Manmatha [12] method expects at least two histogram intensity peaks. The PCR handwriting strokes have intensities that are the same as the background image with equivalent frequency. This causes large portions of the handwriting to be lost to the background, rendering this technique ineffective.

The Niblack binarization [22] algorithm is an adaptive technique which has been used and compared in many applications such as image and video text detection and

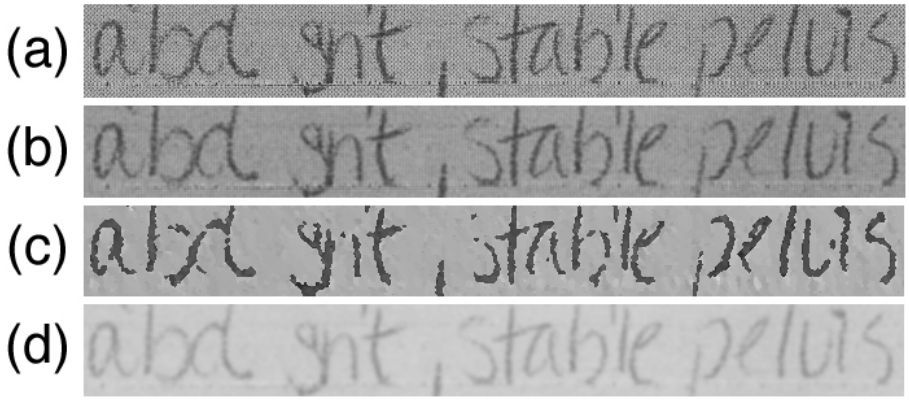

Fig. 3. Smoothing Operations (a) original image + form drop out (b) mean filter (c) median filter (d) Gaussian filter

extraction [7], low quality camera images [9], low quality grayscale utility maps (such as cable and hydro maps with various intensity and noise issues [18]), and low quality historical documents [3]. This algorithm resulted in severe noise, jagged edges and broken character segments. While post-processing improves the algorithm performance substantially, the broken character strokes resulted in lower performance.

Sauvola binarization [14] is a modification of the Niblack algorithm [22] which attempts to suppress noisy areas. Gatos [3] introduced an algorithm which outperformed both Niblack [22] and Sauvola [14]. The Gatos method illustrated that the Sauvola method removed more noise than the Niblack on low quality historical documents. This indicated the potential for the Sauvola method to handle the noise found on the PCR forms. Experimental results in the following sections found the Sauvola method performed better than the Niblack (depending on lexicon size), especially before post-processing.

Logical binarization uses heuristics for evaluating whether a pixel belongs to the foreground or background. It is also common to integrate other adaptive binarization strategies with such heuristics. The Kamel/Zhao algorithm [19] is a logical algorithm which finds stroke locations and then later removes the noise in the non-stroke areas using an interpolation and thresholding step. Various stroke width combinations from 1-10 pixels were tried. However, this algorithm would often classify the stroke as the background thereby making it ineffective.

The Yang/Yan [10] algorithm is a variant of the Kamel/Zhao [19] algorithm. The modifications are to handle low quality images affected by varying intensity, illumination, and artifacts, such as smearing. However, the run analysis step in this algorithm is computed using only black pixels. Neither the foreground or background of the carbon copy medical forms have black pixels; nor are the foreground pixels the same throughout. While both the background and foreground carbon have the same intensities on a specific form, this is not universal. Therefore, the stroke width computation, which is dependent on the run length computation of black or any other specific pixel intensity range, cannot be determined with the carbon paper forms.

In addition to the binarization algorithms, various post-processing strategies are used. The despeckel algorithm is a 3x3 mask which removes a foreground pixel that has no D8 neighbors [11]. The blob removal algorithm is a 9x9 mask which removes small pixel regions that have no neighbors [11]. The Niblack [22] + Yanowitz and Bruckstein method [21] was found to be the best combination strategy here [18]. The Shi and Govindaraju is an image enhancement strategy which has been used on postal mailpieces [17].

4 Proposed Algorithm

4.1 Methodology

Prior algorithms have relied on techniques such as histogram analysis, edge detection, and local measurements. However, these techniques are less effective (see section 5). This motivated the proposed algorithm to use a larger central NxN mask, which determined the intensity of one region, and compare it with the intensities of multiple dynamically moving smaller PxP regions (see Figures 4 and 5).

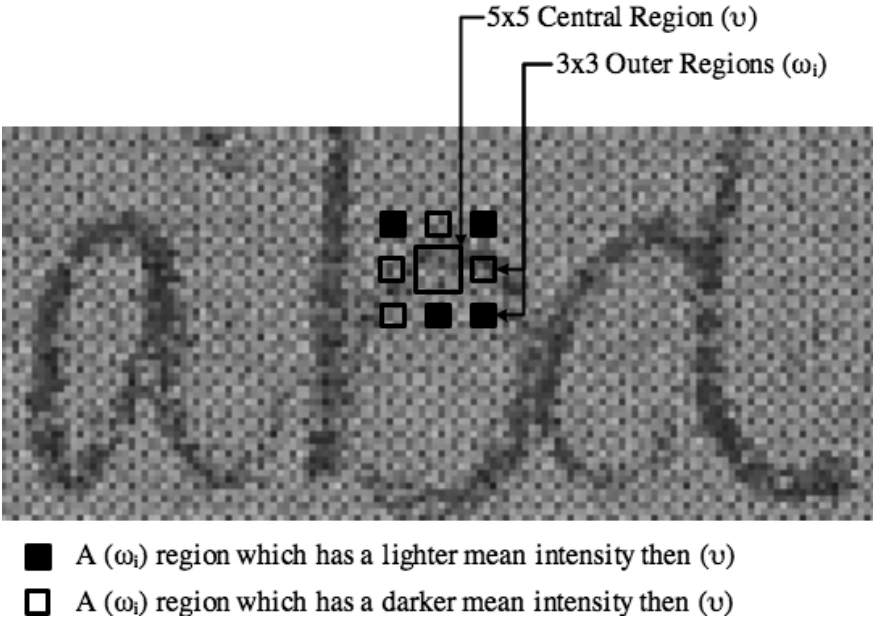

Fig. 4. Initial Mask Placement Example (N=5 and P=3)

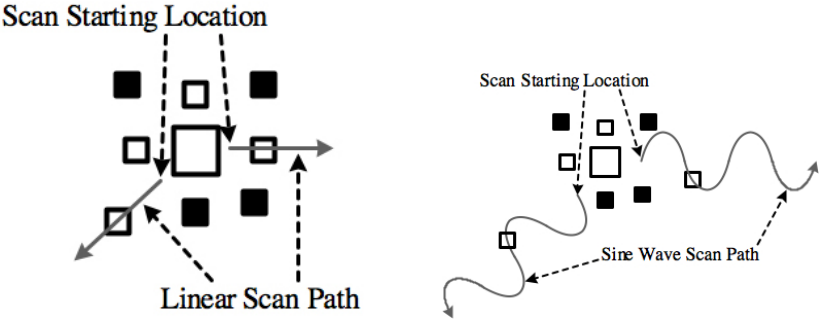

Fig. 5. Scanning Approaches: Linear and Sine Wave

One hypothesis in managing the varying intensities of the carbon mesh and its similarity with the stroke, was to use a wave trajectory for the D8 positioned masks, as opposed to a linear trajectory (see Figure 5). The experiments show that the use of a wave is beneficial for four reasons: (i) the notion of evading a stroke, (ii) finding a background region as close as possible to the central mask (note that the further out from the center mask, the more likely that the carbon mesh of the background can change), (iii) that the best background region to compare to a handwriting stroke may or may not be the edge of the stroke, and (iv) areas surrounding a stroke, in the same trajectory, can be observed. With the inclusion of a stopping condition, our approach does not behave like confined square mask windows, which are relative to a central

position, as other global and adaptive approaches pursue. In this context, the wave trajectory for scanning can be thought of as searching for lighter pockets in the intensity fluctuation of the carbon mesh (see Figures 4 and 5).

A sine wave trajectory offers the benefit of beginning at the origin and allowing a continuous trajectory regardless of distance. It allows the control of frequency and amplitude which are necessary to adjust for stroke width. Sinusoidal waves have been used in other contexts for the modeling of human motor function for on-line handwriting recognition feature extraction and segmentation [15], shape normalization of Chinese characters [4], and signal canceling of pathological tremors while writing [16]. Based on these studies, and the knowledge of the English character set, it was possible to scan out from a character stroke, at a certain frequency, allowing a handwriting stroke to be maneuvered, as opposed to traced, in the search for background regions. The sine trajectory can be thought of as a path which continually crosses the handwritten strokes. This allows the background paper on both sides of the stroke, in all directions, and with a dynamic distance, to be evaluated. Intuitively, more space can be searched and both sides of the stroke can be evaluated in the same computational step at variable distances. It is also presumed that in a moving ambulance, carbon smearing is more likely since the writer will press their hand harder to maintain balance in the vehicle. While strokes in the English language contain both curves and straight lines, at the pixel level, they can be considered piecewise linear movements such that a linear scan will trace the stroke and reduce the likelihood of finding the background. Furthermore, holistic features, such as the area in the letter "D", are typically small, and missing the carbon paper inside of such character holes may result in missed background analysis. This motivated the use of a higher sine wave frequency so that the trajectory would pass through the center of holistic features as frequently as possible. Additionally, since the thickness of characters fluctuate, it is difficult to precisely calculate the true stroke width.

4.2 Algorithm

An input grayscale image 0 (black) to 255 (white) is the input and a binarized image is the output. At a given position on the image, there are 9 masks. A single mask is denoted as Ξ. The mean intensity of a single mask is denoted by $M(\Xi)$. The central mask which slides across the image is denoted by (υ) and has a size NxN, such that $N \geq 3$ and is odd (e.g. 3x3, 5x5, and 7x7). The size of (υ) is based on the estimated stroke width constant denoted by ϕ. The value of ϕ has been estimated to be 5 pixels, therefore (υ) is of size 5x5. At each (υ) position over the image, 8 masks are initially stationed in each D8 position (Figure 4) and are denoted by ω_i where $1 \leq i \leq 8$. The mask size of (ω_i) is PxP such that $3 \leq P \leq \lceil N/2 \rceil$. Note that $P \leq \lceil N/2 \rceil$ allowing a small mask the opportunity of preserving small holistic features, when moving on the sine curve, while also making sure that the mask will not overlap (υ). Each (ω_i) is initially stationed as close to (υ) as possible so as to avoid the mask overlapping between (ω_i) and (υ). Each (ω_i) moves in its respective D8 direction, either linearly (see Figure 5) or via a sinusoidal wave (see Figure 5). The $M(\omega_i)$is computed at each position along the trajectory and stops at a position after one cycle and when the current average is less then the previous average on the sine trajectory. A list of mean values, for each position on that trajectory, are denoted by $M(\omega_i)_q$ where q is a coordinate on the sine curve. The minimum mean value for one trajectory is represented by the equation

$M(\omega_i)_{min} = \min(M(\omega_i)_{\forall_q})$). Next, a comparison of all the D8 $M(\omega_i)_{min}$ positions are made against $M(\upsilon)$. If there are at least 3-4 (empirically determined) out of 8 of the $M(\omega_i)_{min}$ values which satisfy the equation $|M(\upsilon) - M(\omega_i)_{min}| \geq \kappa$, such that κ is a small constant (we use $\kappa = 10$), then the center pixel of (υ) is classified as a foreground pixel. The value κ defines a tolerance with respect to the intensity fluctuation of the carbon paper and is denoted the *carbon intensity similarity rule*. It is assumed that the new image has been initialized to white background pixels, therefore, it is only necessary to mark the foreground pixels when they are found. A dynamic programming step is used to store each $M(\omega_i)$, corresponding to the appropriate region on the image, to improve the runtime performance.

The sinusoidal path is defined by equation (1).

$$y = 2\phi \sin(½ x) \tag{1}$$

The coordinate (x, y), on a sinusoidal trajectory, is relative to its starting location (origin). A nearest neighbor approach is sufficient for conversion of real coordinates to integer coordinates. Each ω_i is computed on the sine curve trajectory (see Figures 4 and 5). Note that using ϕ amplitude value in equation (1), without the coefficient, will result in a distance of 2ϕ between the highest and lowest y-axis points (ϕ is the stroke width). It is further beneficial to use 2ϕ as the amplitude, yielding a distance of 4ϕ, to account for the possibility of 2 touching strokes (e.g. two touching letters). This places a reasonable guarantee that the curve will efficiently exit a stroke while searching for the background. The constant ½ is used in equation (1) so that the sine frequency does not trace the handwritten stroke.

5 Experimental Results

All tests were performed on 30 PCR's comprising of 1,440 word images and various size lexicons (identified in the LEX column of the tables 1 and 2) using the LDWR handwriting word recognition engine [13]. The linear strategy was outperformed by the Otsu [24] method and a despeckel. However, our technique outperformed all algorithm combinations.

The table columns are abbreviated as follows: (W)u/Manmatha, (K)amel/Zhao, (N)iblack, (S)auvola, (O)tsu, and (SW) for our sine wave approach.

Table 1 Shows the performance of the binarization algorithms without post-processing. Our algorithm has 9-21% improvement, with various lexicon sizes, over Otsu [24].

Table 1. Binarization Performance (Figure 7)

LEX	W	K	N	S	O	SW
100	3.1%	4.6%	< 1%	19.4%	35.3%	56.5%
1K	< 1%	1.5%	0.0%	10.1%	17.1%	26.9%
4K	0.0%	0.0%	0.0%	4.4%	11.2%	20.3%

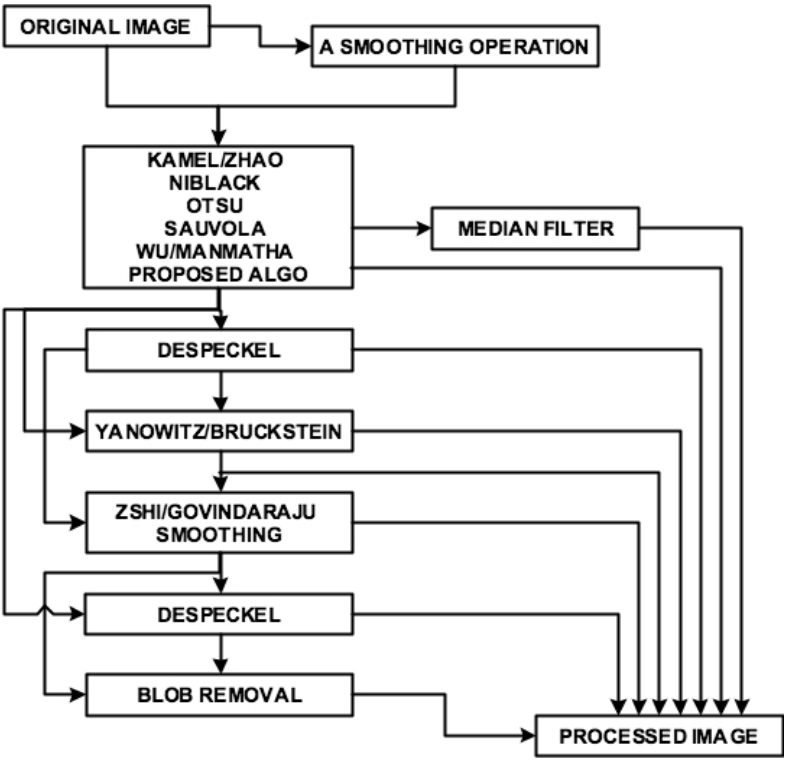

Fig. 6. Image Processing Combinations

Table 2. Combination Performance (Figure 8)

LEX	W	K	N	S	O	SW
100	3.1%	4.6%	46.9%	48.3%	52.6%	59.1%
1K	< 1%	1.5%	26.0%	26.4%	36.1%	40.1%
4K	0.0%	0.0%	17.8%	12.7%	21.2%	25.3%

Figure 7 shows the output of the aforementioned binarization strategies with no post-processing support. The handwriting phrase, from NYS PCR medical form, "abd snt, stable pelvis" means *abdominal soft-not-tender, stable pelvis*. Figure 7e contains heavier noise that is not removed by the LDWR algorithm's pre-processing step, whereas Figures 7f and 7g have noise which is less severe and therefore more easily removed.

Table 2 Shows performance of the binarization algorithms with their best respective post-processing combination from Figure 6. The proposed algorithm + despeckel + blob removal offers 4-7% improvement, with various lexicon sizes, over Otsu [24] + Despeckel.

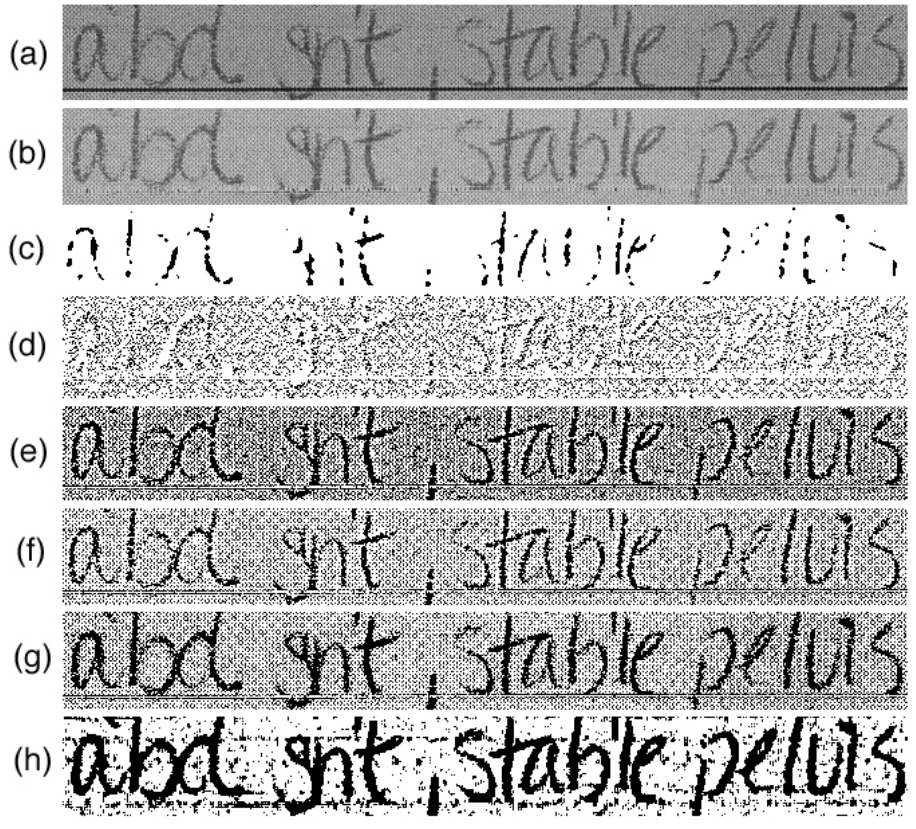

Fig. 7. Comparison of Binarization Algorithms Only: (a) original image (b) original image with form drop out (c) Wu/Manmatha Binarization (d) Kamel/Zhao Binarization (e) Niblack Binarization (f) Sauvola Binarization (g) Otsu Binarization (h) Sine Wave Binarization

Figure 8 shows the output of the aforementioned binarization strategies with each algorithms best post-processing combination.

The percentages in all result tables neglect accept/reject rates due to the complex nature of medical handwriting. The percentages reflect the output of the recognizer, regardless of confidence. Recognizer accept and reject analysis will be included in the handwriting recognition algorithms of future work. Words in the form region were manually segmented by a human. Future research will include the analysis of automated segmentation as well. Stopwords were omitted from both the recognizer lexicon and word images from the form. In addition, the LDWR algorithm uses pre-processing strategies for its own noise removal and smoothing before executing its recognition algorithm [13][20][23]. Therefore, a noisy image submitted to the LDWR algorithm will be internally pre-processed by the handwriting recognizer.

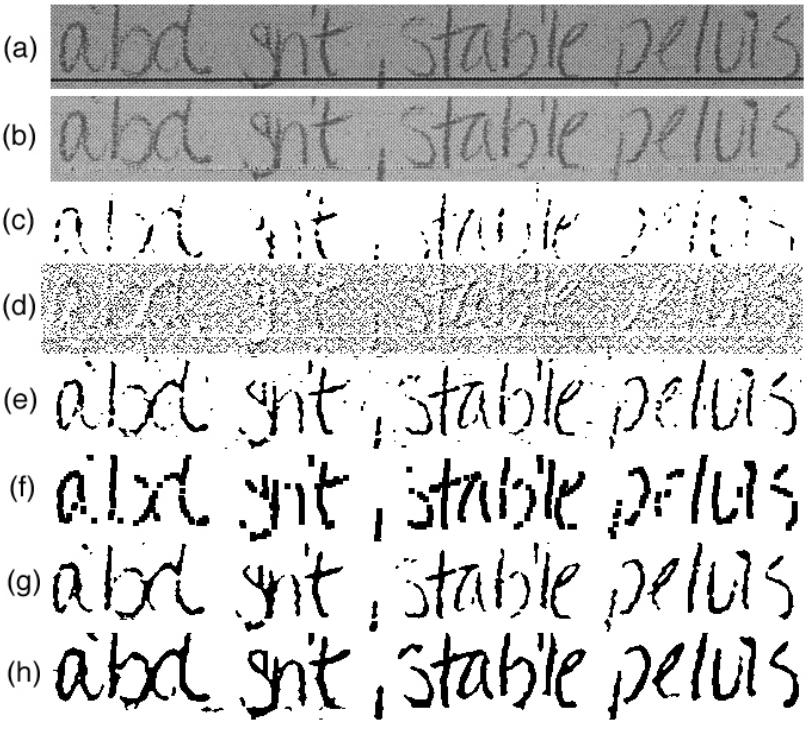

Fig. 8. Comparison of Binarization Algorithms with their Best Post-Processing Strategy: (a) original image (b) original image with form drop out (c) Wu/Manmatha Binarization + unassisted (d) Kamel/Zhao Binarization + unassisted (e) Niblack Binarization + Despeckel (f) Sauvola Binarization + Despeckel + Shi/Govindaraju Region Smoothing (g) Otsu Binarization + Despeckel (h) Sine Wave Binarization + Despeckel + Blob Removal

6 Conclusions

In this paper we describe a binarization algorithm for handling carbon paper medical documents. Improvements of approximately 9-21% (using various lexicon sizes) are obtained over prior binarization algorithms. Approximately 4-7% improvement is obtained using post-processing.

References

1. Western Regional Emergency Medical Services. Bureau of Emergency Medical Services. New York State (NYS) Department of Health (DoH). Prehospital Care Report v4.
2. Hatami, Safar., Hosseini, R., Kamarei, M., Ahmadi, H. Wavelet Based Fingerprint Image Enhancement. IEEE International Symposium on Circuits and Systems (ISCAS). C2005.
3. Gatos, B., Pratikakis, I., Perantonis, S.J. An Adaptive Binarization Technique for Low Quality Historical Documents. 6th International Conference on Document Analysis Systems (DAS). C2004.

4. Liu, C.L., Marukawa, K. Global Shape Normalization for Handwritten Chinese Character Recognition: A New Method. International Workshop on Frontiers of Handwriting Recognition. C2004.
5. Milewski, R and Govindaraju, V. Handwriting Analysis of Pre-Hospital Care Reports. IEEE Proceedings. Seventeenth IEEE Symposium on Computer-Based Medical Systems (CBMS). C2004.
6. Leedham, G., Varma, S., Patankar, A., Govindaraju, V. Separating Text and Background in Degraded Document Images – A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding. Proceedings Eighth International Workshop on Frontiers of Handwriting Recognition. C2002.
7. Wolf, C., Jolion, J.M., Chassaing, F. Text Localization, Enhancement and Binarization in Multimedia Documents. 16th International Conference on Pattern Recognition (ICPR) C2002.
8. Liao, P.S., Chen, T.S., Chung, P.C. A Fast Algorithm for Multilevel Thresholding. Journal of Information Science and Engineering. C2001.
9. Seeger, M., Dance, C. Binarising Camera Images for OCR. Xerox Research Center Europe. 6th International Conference on Document Analysis and Recognition C2001.
10. Yang, Y., Yan, H. An Adaptive Logical Method for Binarization of Degraded Document Images. The Journal of the Pattern Recognition Society. C2000.
11. Sonka, M., Hlavac, V., Boyle, R. Image Processing, Analysis, and Machine Vision; 2nd Edition. PWS Publishing. C1999.
12. Wu, V. and Manmatha, R. Document Image Clean-Up and Binarization. Proc. SPIE Symposium on Electronic Imaging. C1998.
13. Kim, G., and Govindaraju, V.: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. IEEE Trans. PAMI 19(4): 366-379. C1997.
14. Sauvola, J. Seppanen, T. Haapakoski, Pietikainen, M. Adaptive Document Binarization. In International Conference on Document Analysis and Recognition, Volume 1. C1997.
15. Beigi, H. Processing, Modeling and Parameter Estimation of the Dynamic On-Line Handwriting Signal. Proceedings World Congress on Automation. C1996.
16. Hsu, D.S., Huang, W.M., Thakor, N.V. StylPen: On-line Adaptive Canceling of Pathological Tremor for Computing Pen Handwriting. IEEE Transactions on Biomedical Engineering. C1998.
17. Shi, Z. and Govindaraju, V. Character Image Enhancement by Selective Region-growing. Pattern Recognition Letters, 17. C1996.
18. Trier, O.D. and Taxt, T. Evaluation of Binarization Methods for Document Images. , IEEE Trans. PAMI, 17 (3). C1995.
19. Kamel, M., Zhao, Extraction of Binary Character/Graphics Images from Grayscale Document Images. CVGIP: Graphics Models Image Processing; 55 (3). C1993.
20. Schurmann, J., et al. Document Analysis-From Pixels to Contents. Processing IEEE Vol. 80 No.7. C1992.
21. Yanowitz, S.D. and Bruckstein, A.M. A New Method for Image Segmentation. Computer Vision Graphics and Image Processing Vol. 46 No.1. C1989.
22. Niblack, W. An Introduction to Digital Image Processing. Englewood Cliffs, N.J. Prentice Hall. C1986.
23. Brown, M.K., Ganapathy, S. Preprocessing Techniques for Cursive Word Recognition. Pattern Recognition, Vol. 16 No. 5. C1983.
24. Otsu, N. A Threshold Selection Method from Gray-Level Histogram. IEEE Transactions on System Man Cybernetics, Vol. SMC-9, No. 1. C1979.