Manual Correction of Misclassified Image Segments

Joseph Czyszczewski*, Michael R. Jensen*, Hong Li*

Keywords: automatic, scanning, analysis, algorithms

Abstract: Since book scanning is often used for low-volume titles, controlling scan cost is critical. Automatic image segmentation promises high quality and low cost scans with a minimal need for user intervention. However, since no segmentation algorithm is perfect, strategies based on business policies are required to manage trade-offs between cost and quality. We describe an adaptive approach that allows the user to control, based on policy, the amount of intervention necessary to verify the segmentation's correctness and to rapidly adjust misclassified segments.

Background

There is a significant market for publishing out-of-print books. These books are usually reprinted in short runs. To satisfy this market, an existing book is commonly scanned, touched up, and printed. Electrophotographic printers and finishers are more economical for these short runs than traditional presses.

In order for these short runs to be profitable, it is critical to minimize pre-press costs. Preparing the book for scanning, performing the scan, and validating the scan quality afterwards is a labor intensive process. Though automated processes facilitate scanning and cleaning up the image, it is often necessary to manually verify that each page is ready to print. Automatic image segmentation provides significant improvement by eliminating the need to manually rescan image zones, yet segmentation is imperfect and intervention is required in some cases. This paper describes a collection of methods that facilitate segmentation editing and minimize the impact of this costly user intervention.

*IBM Printing Systems

A Brief Introduction to Segmentation

It is important to have a general idea of how the segmentation algorithm works to provide a common frame of reference for the remainder of the discussion. Chevion et. al. (2004):

> [Segmentation is dividing] an image of a printed document into "Text," "Images," and "Line-Art." Each segment is then processed according to its specific nature and recombined to form the original document. This procedure is done in order to make a quality print of the document, as close as possible to the original, thus preventing the appearance of unwanted artifacts that would otherwise show up when a scanned document image is printed.

The original image is scanned at 300 DPI 8-bit gray. Extracted text is scaled to 600 DPI and then goes through a threshold process. This produces clear, easily readable binary text. Line art is halftoned through an error diffusion process and is output as 600 DPI binary. 300 DPI, 8-bit halftoned images are descreened to remove the halftone and eliminate moiré effects. The recombined image is optimized for quality and storage size.

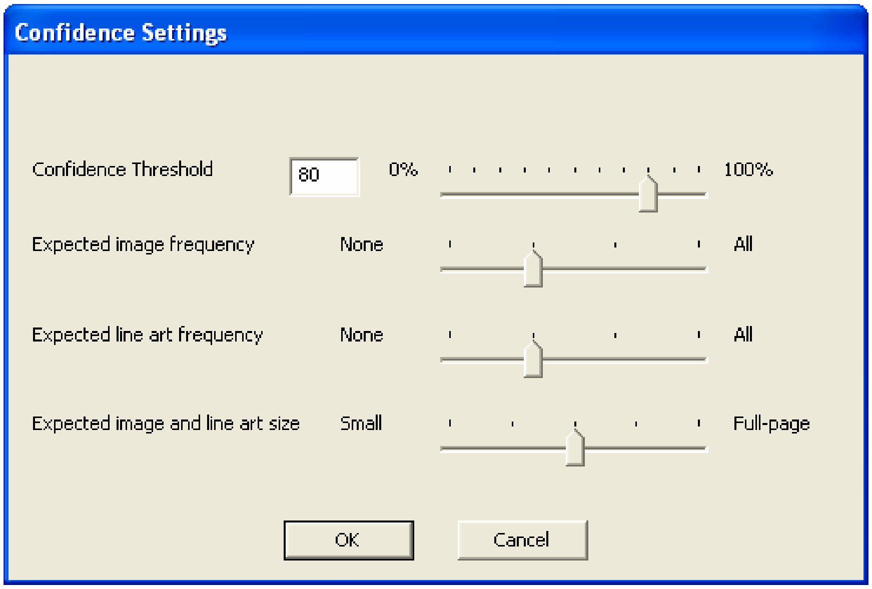
Confidence Value

Each scan job is evaluated in order to ensure quality. The nature of the application is considered. A biology textbook, for example, has a much more rigorous quality standard than does a novel. The business policies thus dictate a balance between the cost of misclassification and the cost of manual verification. This balance is quantified in a value called the *confidence threshold.*

Broadly speaking, a *confidence value* is a variable represented as a percentage that describes how confident we are that a page's segmentation was performed correctly. A number of factors influence this value. First, the system can make use of the operator's fore-knowledge of the characteristics of the book. If, for instance, the operator knows the book is a novel, he can warn the algorithm to expect few halftoned images. Second, the system can look for previously defined suspicious patterns in the segmentation results. Lastly, the program learns from users' interactions the characteristics of segments that the operator had to manually reclassify. Though this poses certain challenges, this adaptive nature contributes to an increasingly accurate confidence value of the page.

Sample Interface for Manual Correction

An operator can flip through a book to get an idea of its nature before scanning. This knowledge is transferred to the program via "sliders." Combined, these values influence the final confidence value of each page.



Confidence Threshold 80

Figure 1 – Confidence value parameters

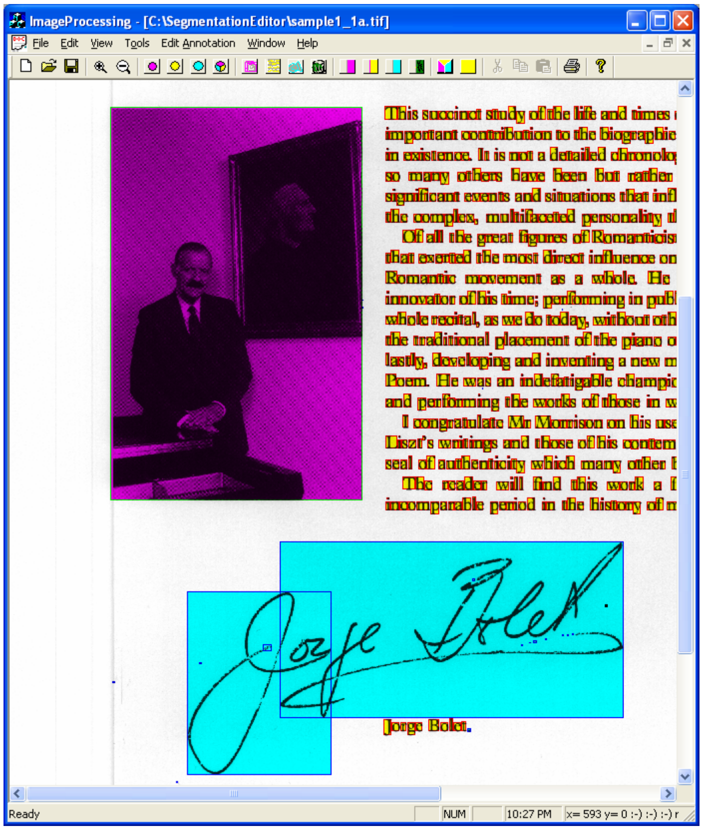
The "Confidence Threshold" slider lets the operator express the minimum quality requirements for the job. After processing, each page whose confidence value is below this threshold is presented to the user for verification and, if needed, correction.

The "Expected Image" slider represents the frequency of halftoned images in the document. Similarly, the "Expected Line Art" slider represents the frequency of line art. These two sliders allow the specification of the following values (from left to right): none, few per page, many per page, and all pages. In this context, "few per page" means one or two instances of images or line art. "Many per page" means more than two instances of image or line art on the page. "All" means each page in the book contains images.

"Expected image and line art size" is measured by page coverage. Set to maximum, the program would expect to find full-page images.

Collectively, the sliders describe the application. Though it is easy to come up with an arbitrary traditional classification of book category (novel, textbook, etc), due to extreme variations within each category it is difficult to map algorithmically meaningful values to the selected category. Having the operator adjust these sliders lets us leverage his knowledge of the application and increase the accuracy of our confidence value algorithm's predictions.

The sample interface facilitates segmentation correction in a variety of ways. First, it displays all segments shaded in different colors. Text is highlighted with yellow, line art with cyan, and halftoned images with magenta. These annotations allow the operator to instantly recognize what parts were categorized into which group. Furthermore, the elements that the program suspects might be misclassified are marked, allowing the elements to be quickly identified by the user. After accepting the operator's changes, the program displays the next page whose confidence value fell below the threshold.


ImageProcessing - [C:\SegmentationEditor\sample1_1a.tif]
File
Edit
View
Tools
Edit Annotation
Window
Help
Ready
NUM
10:27 PM
x= 593 y= 0 :-) :-) :-) r

This succinct study of the life and times
important contribution to the biographic
in existence. It is not a detailed chronolo
so many others have been but rather
significant events and situations that infl
the complex, multifaceted personality
Of all the great figures of Romanticis
that exerted the most direct influence on
Romantic movement as a whole. He
innovator of his time; performing in publ
whole recital, as we do today, without oth
the traditional placement of the piano o
lastly, developing and inventing a new m
Poem. He was an indefatigable champio
and performing the works of those in w
I congratulate Mr Morrison on his use
Liszt's writings and those of his contem
seal of authenticity which many other
The reader will find this work a
incomparable period in the history of m

Jorge Bolet.

Figure 2: Segmentation Editor

Signs of Commonly Misclassified Segments

The program analyzes the results of the segmentation in order to generate a confidence value for each page. The following patterns are often signs of a misclassification:

1. Images or line art that are smaller than expected. If such a segment is found, chances are good that that segment is misclassified. This category maps to the “Expected image and line art size” slider.
2. Unexpected abundance of image or line art. The operator established a bias towards a given frequency of line art or images by adjusting the appropriate slider. If a page has an unexpected number of these segments, it’s likely some were misclassified. This category maps to the “Expected image frequency” and “Expected line art frequency” sliders.
3. Small, isolated, line art or image near the edge of the sheet. Though segmentation expects the image to already be cleaned up (despeckled, deskewed, etc) this type of element is likely an artifact from scanning. This category, and the following categories, describe general characteristics of the application and are not related to a particular slider.
4. Line art or image embedded in lines of text. A paragraph rarely has images interspersed with words on a line.
5. Text embedded in line art or image. Text superimposed on an image should be classified as an image. If it is not, then it was likely misclassified.
6. Adjacent text characters of significantly differing sizes. It is rare to change font size in the middle of a line. This suggests misclassification.

Each of the above categories is assigned an initial weight. These initial values are based upon the corresponding sliders previously positioned by the user. As the program learns from the operator’s corrections, this weight is adjusted. The weight refers to the relative influence this category has on the final confidence value for the page.

Some classification errors are more serious than others. One serious error is classifying text as image. This results in text that is plainly blurred. A less serious error is classifying text as line art. This results in text that is perhaps suboptimal but still acceptable. The default weighting of the above categories reflects this by giving more weight to potential serious errors, such as the one described by category 4.

Furthermore, the evaluation of each page according to these categories generates a score (referred to as a “category score”). This score is a category-specific quantification of the attribute being measured. For example, during the course of calculating a page's confidence value, it is examined for adjacent tiny line art and images (category 1, above). Suppose the analysis found four instances where the line art was unusually small. Because this result causes us to suspect the page, it rates a high score for this category. This score, combined with the category's importance (measured through the category's weight) contributes to the final confidence value of the page.

Another example of the evaluation will be useful at this point. Suppose the operator is scanning a novel, and has adjusted the Expected Line Art and Expected Image sliders accordingly. Specifically, the operator expects no images and very infrequent line art. The weight of the second category (titled "Unexpected abundance of image or line art") is correspondingly increased. A page is encountered with several line art elements on it, thus receiving a high score for category 2. Combining the considerable weight and the score of this category, in conjunction with the weights and scores from the other categories, the system assigns this page a low confidence rating. This page's confidence value falls below the confidence threshold for the book and is presented to the operator for verification and correction.

One danger is that the program will become tuned to a particular type of application. Were this to happen, the program would perform poorly when faced with a scan job to which it is not accustomed. The program's performance would suffer while oscillating between different application types. This oscillation causes unnecessary cost. The problem is avoided by keeping track of what it has learned in conjunction with the application category as defined by the slider settings. Thus what the program experiences while processing one sort of job will be of use when it encounters similar jobs, but will not incorrectly affect it when facing a new situation.

Adaptive Feedback for the Confidence Value

There are only a handful of different segmentation error types. These common misclassifications are delineated above. There are several variables that affect the segmentation's correctness: the algorithm's settings, the properties of the original, and any peculiarities of the scanning hardware. Changing these variables tends to change the frequency, not the type, of the misclassifications. Because the types of misclassification remain fairly constant, they provide a stable framework for the feedback loop.

The formula for a page's confidence value is:

$$V = 100 - \Sigma (W_i \bullet S_i \; / n) \qquad (1)$$

with,

W: the dynamically assigned weight of the misclassification category ranging from 0 to 1
S: the category's score
n: the number of different misclassification categories measured.

A further constraint is that the sum of the *W* over *i* must be exactly 1.0. Also, since a negative confidence value has no meaning, negative values are adjusted to be 0.

By tracking the types of corrections the user makes to the segmentation analysis, the relative weights of the misclassification categories are adaptively adjusted. These adjustments let the user fine-tune the confidence threshold slider to minimize the number of pages that need to be manually examined.

A Workflow Example

The following workflow example illustrates the principles involved in this process. Suppose the first time this system is used, the confidence threshold is set at 100%. This means that users will be asked to review pages with even the slightest suspicion of a segmentation error. As a result, the user will be obligated to review many pages. The operator flips through the book to get a feel for its properties and sets the sliders accordingly. Suppose the pages are generally correctly segmented with the exception of some text incorrectly flagged as line art. One such page (including other minor errors) is evaluated in the following manner according to each misclassification category:

Tiny line art / images: $W_1 = 0.1$, $S_1 = 10$; $W_1 \bullet S_1 / n = 0.1667$
Unexpected abundance of image or line art: $W_2 = 0.2$, $S_2 = 30$; $W_2 \bullet S_2 / n = 1$
Small, isolated image or line art: $W_3 = 0.1$, $S_3 = 30$; $W_3 \bullet S_3 / n = 0.5$
Line art or image embedded in lines of text: $W_4 = 0.4$, $S_4 = 60$; $W_4 \bullet S_4 / n = 4$
Text embedded in line art of image: $W_5 = 0.1$, $S_5= 50$; $W_5 \bullet S_5 / n =0.833$
Adjacent text characters of differing sizes: $W_6 = 0.1$, $S_6 = 40$; $W_6 \bullet S_6 / n =0.667$

Applying the formula, we find this page has a confidence value of 92.8%. As this page falls below the confidence threshold, it is presented to the user for correction. As the user validates the book, he corrects this and other misclassifications. The system notes these corrections and adjusts the weight of the category in question accordingly. When the system presents a suspicious segment to the operator and the operator chooses to not correct it, then that is a signal to the system to not mark so many of that type of error. The weight of that category is reduced. Correspondingly, when the system flags a particular type of error to the user and it is frequently corrected, that signals the system to show more of that sort of error to the user. That error's category is subsequently allotted more weight. For some later book, the user leverages past experience and sets the confidence threshold at 85%. By now the category weights have been adapted to fit this user's segmentation settings and hardware. This lower confidence threshold means that the operator will have fewer pages to manually review (thus decreasing the time and cost for producing this book), while the modified category weights ensure that segmentation problems are still caught.

After this first book has been scanned and validated, the system has gained some experience with the selected book type. Future books of the same type will benefit from the learning of this experience. Future books of a different nature will require similar training for the system.

Conclusion

Applying adaptive feedback to refine the confidence threshold promises to ease the manual burden of post-scan touch-up. Although no known technology eliminates the need for user intervention, our approach mitigates its cost.

Acknowledgements

Dan Chevion, On Amon, Asaf Tazdok, and the IBM Research team in Haifa Israel for their work on the segmentation algorithm upon which our work was based.

Ron Van Buskirk for his contributions to the idea of confidence value.

Literature Cited

Dan Chevion, Ehud Kamin, Gerry Thompson, Asaf Tzadok, Chai Wu, et. al. "Method of Segmenting a Scanned Page into Text, Image, Line-Art and Background", TAGA Proceedings 2004.