Document Analysis System for Automating Workflows

Steven J. Simske¹ and Jordi Arnabat²

¹ Hewlett-Packard Labs, Mailstop 85, 3404 E. Harmony Road,
Fort Collins, CO 90528, USA
Steven.Simske@hp.com

² Hewlett-Packard Espanola, Av. Graells, 501
98190 Sant Cugat del Valles, Spain
Jordi.Arnabat@hp.com

Abstract. When a user places a document in a capture device—copier, multi-functional printer [MFP], or scanner—the user expects good output to be produced regardless of the document type. There are a variety of means to achieve improved output, in which the settings on the copying device are tuned to the content characteristics of the document. These settings can be automated across the range of scanned context extremes from photo (blurring, no snapping) to fully-text (sharpening, aggressive snapping) documents. This procedure is “document auto typing”, and relies on a fast and accurate assessment of the content of the captured image. We herein describe the development of seven distinct systems for document analysis, and through the comparison of these systems arrive at an efficient and accurate document analysis system for automating the copying settings. We discuss the applicability of this method to other automated workflows in document capture.

1 Problem Statement

We have developed a repertoire of seven different document analysis systems to analyze the content scanned. Each of the algorithms described here can be performed at low (30-100 ppi) resolution. We define them and provide a description of their relative advantages and disadvantages here (see also Table 1):

1. High-Level Document Classification: Photo, Text and Mixed. This is a trinary classification: we determine if the document has text or photo, or else mixed (text AND photo). This algorithm has excellent performance, since it ends its search for text/photo as soon as it verifies the presence of text/photo. This classification scheme implicitly imposes a "region manager" on the classification with the following characteristics: uniform background regions and line drawings are considered “text” regions, since they do not require the same type of specialized image processing (blurring, no snapping) as photos, and are not deleteriously affected when treated as text (sharpening, aggressive snapping). The main disadvantage of this approach appears in multi-page document copying, in which image quality (IQ) variability of the text on different pages of a multi-page document may occur due to page-to-page classification variability.

2. Low-Level Document Classification: Percent Text, Photo. In addition to assigning a document to {text/photo/mixed}, we determine the relative amount

(percentage) of text and photo present. This option can be accommodated with a simple “slider” motif: the minimum/maximum for the slider is 0/100 percent text. However, using this, text will vary in its sharpness considerably with a mixed document set, and the output is more sensitive to classification errors.

3. Neighborhood Analysis—High and Low Frequency Sites. Here, each pixel is classified by the amount of variance in its neighborhood. This is roughly equivalent to edge detection. Only the resulting edges are sharpened. All non-edge pixels are treated as if they are photos. This algorithm is quite fast, and provides uniform appearance across multiple-page documents. However, it can result in artifacts in some regions that lower image quality (this is dependent on print quality, text font, etc.). It will, however, generally result in similar-appearing text regions throughout a multiple-page document.

4. Neighborhood Analysis—Filtered High-Frequency Sites. Simple zoning analysis is performed to prevent sharpening of large, low-edge containing (e.g. photo) regions. This further improves the correlation of sharpening with text and line drawings.

5. Region Analysis and Region-Specific Pipelines. A zoning analysis engine (e.g. [1]) is used to segment and classify specific region types, including but not limited to text, line drawings, tables, photos and graphics. This provides a uniform appearance of like region types across corpora of any size. There is, generally, a decline in performance by roughly an order of magnitude in comparison to the previous methods. Moreover, when there are zoning mistakes, the effects on the output can be quite noticeable—for example, when photos and text are mistaken.

6. By Region with (OCR) Feedback. In this option, region analysis as described in option 5 above is combined with optical character recognition (OCR) to improve the segmentation and classification. That is, OCR results are used to determine the probabilities that zoning was correct in the first pass, and thus improve the zoning output in the second pass. While this greatly improves text classification, it has high memory, performance and other resource costs. However, it does provide searchable text as output, with obvious implications in searching, clustering, summarizing, archiving and topic-tracking.

Table 1. Relative comparison of the 7 options above. All seven of these systems have been built—or incrementally improved—and tested during the past year. These data are for a Windows-based laptop computer with Pentium III processor, 256 MB of RAM, and 1.13 GHz CPU clock.

Option	Time (/page)	Error (pct)	Appearance Flaws
1. High-level	0.3 sec	2-25%	Moderate
2. Pct	0.5 sec	<20%	Moderate
3. Edge-based	0.2 sec	0%	Severe
4. Filtered-edge	0.7 sec	0-2%	Severe
5. Region-based	1.5 sec	2-30%	Moderate
6. Regions+OCR	7.5 sec	0.5-10%	Moderate
7. Template	1.6 sec	0.5-5%	Moderate

7. By Region with Template/Corpus Feedback. When the output closely matches a template, the zoning results are “snapped” to the template, providing increased accuracy without a substantial impact on performance. However, this method generally fails in practice, since most documents to be copied have no template in place.

2 Design and Implementation

We chose Option 1 (with an estimate for Pct, Option 2, which can be used with a slider) for use on a high-speed, walk-up copier. This improves the output image quality for text-only or photo documents over the default “mixed” mode otherwise applied to all documents independent of their content. For this to succeed, most text and photos must be typed correctly, with virtually no mixed documents mistyped. Performance is also an issue (<0.5 sec on the test system). The steps involved are outlined here:

1. Threshold. We used dual-Gaussian fitting (Modified from [2]). The sub-threshold black pixels form a Threshold Map.

2. Find Solids. We then find solid areas of non-uniform hue or intensity, which represent photos. We first scrub any black shadows along the edge of the scanned, then adaptively determine the amount of run-length smearing to generate sufficient regions to estimate skew correctly but not so many that we sacrifice performance. This estimate is based on the number of black runs rather than using full region formation. After smearing the map, we form regions from the black pixels on the map. Solid (potential photo) regions are, for example, greater than 1 square inch in area and contain 70% or more black pixels, among other possible factors. In general, solid regions have areas with long black runs while “nonsolid” (potential text) regions do not have a high percentage of long black pixel runs. These solid regions are then scrubbed from the Threshold Map, but their boundary information is stored for later analysis.

3. Nonsolid Skew Estimate. The left over pixels are then tested for skew. We use a robust, fast modified Hough transform method.

4. Estimate Text Area. Projection profiles are then formed both horizontally and vertically at the angle of skew to estimate the amount of text on the document. The projection profile statistics are: (1) The number of pixels in the run (length of run in pixels), (2) The number of black pixels in the run (number of “1’s” in the thresholded map for the particular run) and (3) The percent of black pixels = (2) / (1). Text can be easily gleaned from projection profiles. EFIGS text, for example, will generally have ascenders and descenders around an x-height, with typically 15-50% black pixels in the x-height area for the length of text in the projection (Asian and other languages will tend to have, like EFIGS languages, other characteristics, such as $\leq 3X$ variability in black pixels in all text runs). Because salient text will generally be spaced in repetitive lines, the statistics for text presence go up with more text lines. After this step, if text is present we know the final classification will now be “text” or “mixed”. If text is not present, we will need to search for text using projection profiles over solid regions, as discussed in the next section.

5. Estimate Photo and Drawing Areas. The solid regions are next investigated. First an Edge Map is created. The edges overlying every solid region are then smeared to

create a New Threshold Map that can be analyzed for text presence, as in (4) above. The amount of potential text over these regions is added to the text area in step (4) and the decision on text is made.

6. Determine Typing. The solid regions are then analyzed—using the histograms of the non-edge pixels—to see if they are uniform backgrounds. If any of the solid regions are not uniform backgrounds, then the final typing is Mixed (text already present) or Photo (no text present).

3 Results

The document auto-typing above is then used to control the copier settings for tone-scale, color table, sharpening and snapping. The software was developed testing on an existing 50-document corpus and then tested on a separate 205-document corpus (generated and maintained by the HP copier image quality team in Boise, ID, USA). For this test set, only 1.6% of documents were made worse than the default (Mixed) setting. Thus, the system provides a precision of 98.4%. Additionally, 87.8% of those scanned documents requiring an altered setting were so identified (this corresponds to recall in machine learning vernacular). Accuracy, which is determined from $2pr/(p+r)$ where p is precision and r is recall, is thus 92.8%. In everyday usage, the value for accuracy depends on the relative mixture of document types scanned, and will approach 100% for simple text document copying. The remaining 12.2% (those which did not have correct "recall") of the documents tested were typed as default (importantly, there were no extreme Text/Photo or Photo/Text misclassifications—this means that none of the documents was mistyped to be worse than a default setting on the copier).

The algorithm was tested on a Windows-based laptop computer with Pentium III processor, 256 MB of RAM, and 1.13 GHz CPU clock (comparable to resources available on a walk-up copier). The mean processing time for this "test system" was 0.28 sec/document, with a standard deviation of 0.07 sec and a range of 0.17-0.76 sec.

4 Competitive Approaches

Currently shipping copiers allow settings to be made manually on the copier control panel. This includes multifunctional printers and document scanners, which focus on business documents, and usually default to text or near-text mixed modes (i.e. they use sharpening and snapping). All-in-one devices also have multiple modes, but many times default to a more photo-centric processing mode. However, low-end devices such as desktop scanners and all-in-ones have shipped a region-based zoning analysis approach [3] to improving output quality. This analysis, however, is prone to more mistakes since so many (segmentation, classification, bit-depth, etc.) decisions are made on each document page. The per-page failure rate is thus greater than 5%, and the time for analysis is typically an order of magnitude greater than that of the method presented here. Allowing users to correct the analysis, such as is commonplace for a preview-scan workflow motif [4], is not generally acceptable for copier auto-typing. Comparatively, our method is faster (0.3 sec on the "test system") and more accurate,

and is suitable for implementation in high throughput devices such as 50 ppm machines without deleteriously affecting performance or processing accuracy.

5 Discussion

Seven methods of improving the automatic image quality of a scanned document are outlined herein. The approach used for a given system will be chosen based on performance requirements, resources available, and the accuracy expected by the users. For example, our choice of Option 1 was natural given it provided high performance (mean processing time of < 0.3 sec), relatively high accuracy ($> 90\%$) and no severe errors (no "text only" documents in the sample corpus were misclassified as "photo only" documents, and vice versa) occurred. Additionally, only two different "quality" settings for text are observed: high quality for "text only" documents and moderate quality for "mixed" documents.

The automatic classification of documents is of interest for a number of other applications, in addition to the need for copier output quality as described here. One such application is the automatic archiving of documents that are scanned and copied. This may be required for auditing, payment, security, etc., reasons. When the documents are assigned to "text", "photo" and "mixed" classes, there can be a considerable savings over a default archiving scheme. For example, "text" pages are typically archived at 400 ppi, 1-bit/pixel before compression, while "photo" pages are typically archived at 200 ppi, 24-bit/pixel before compression. Archiving a "mixed" page requires preserving both the higher resolution and the greater bit depth settings of the "text" and "photo" classes, respectively. Before compression, then, "mixed" pages require 24 times as much memory as "text" pages, and 4 times as much memory as "photo" pages. Compression further exacerbates the difference in archiving. Other applications of the automatic document analysis system used herein include preliminary template matching and preliminary document classification for a document indexing system.

Acknowledgments

The authors thank helpful colleagues—especially Nehal Dantwala, Peter Bauer and Óscar Martínez—for their comments and suggestions in this work.

References

1. Wahl, F.M., Wong, K.Y. and Casey, R.G.: Block segmentation and text extraction in mixed/image documents. *Computer Vision Graphics and Image Processing*, Vol. 2, pp.375-390, 1982.
2. Kittler, J. and Illingworth, J.: Minimum error thresholding. *Pattern Recognition*, Vol 19, no. 1, pp. 41-47, 1986.
3. Lee, J.P., Simske, S.J. and Dawe, J.T.: Segmenting a document into regions associated with a data type, and assigning pipelines to process such regions. U.S. Patent 6,880,122, Apr. 12, 2005.
4. Simske, S.J. and Arnabat, J.: User-directed analysis of scanned images. *Proc. DocEng 2003*, Grenoble, pp.212-221, 2003.